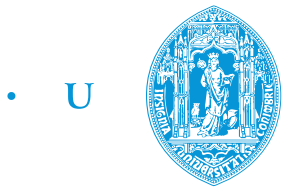

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

KUSCO@TICE.Mobilidade: Enriquecimento Semântico de Lugares e Eventos

Tese de Mestrado em Engenharia Informática

Relatório Final

Jorge Alexandre Gomes de Oliveira Santos

jasant@student.dei.uc.pt

Sob Orientação de

Francisco Câmara Pereira e Ana Oliveira Alves

Orientador DEI: Pedro Abreu

11 de Julho de 2012

Resumo

Dada a recente globalização dos dispositivos móveis baseados em localização e a respectiva difusão das redes sociais nestes, a quantidade de informação descritiva sobre lugares e eventos presente na *web* tem vindo a crescer exponencialmente. Este tipo de informação permite-nos ter outra perspectiva do espaço, caracterizando-o pelas unidades atómicas que o compõem: os pontos de interesse. No entanto, nem sempre é possível encontrar este tipo de informação devidamente organizada: os conteúdos encontram-se dispersos por uma vasta gama de serviços de forma desequilibrada, o que tende a dificultar o acesso a esta informação.

Neste trabalho, pretende-se desenvolver uma metodologia que visa enriquecer semanticamente lugares e eventos. A solução caracteriza-se pela integração de informação proveniente de diferentes fontes de dados (que podem ir desde directórios comerciais a redes sociais) e de fontes de conhecimento como a *Wikipedia*, através da utilização de técnicas de processamento de linguagem natural e extracção de informação.

Os resultados atingidos no final deste trabalho apresentam-se sobre a forma de um módulo de enriquecimento semântico, permitindo a recolha de dados em múltiplas fontes de forma dinâmica e cuja integração permite caracterizar lugares de forma específica. Este módulo agrega ainda a etiquetagem automática de locais de forma a caracterizá-los, podendo estes corresponder a pontos de interesse ou a uma área geográfica. No futuro, este módulo será integrado numa plataforma transversal do *TICE.Mobilidade*, um projecto de mobilidade que está a ser desenvolvido para o território Português.

Palavras-chave: extracção automática de informação, enriquecimento semântico, etiquetagem automática, processamento de linguagem natural

Agradecimentos

Gostaria de agradecer a algumas pessoas, sem as quais este trabalho não teria sido possível. Aos orientadores desta tese: Professor Francisco Câmara Pereira, por sempre ter acreditado no meu trabalho e por me ter proporcionado um leque de oportunidades que sem dúvida me fizeram crescer como a pessoa que hoje em dia sou. Professora Ana Oliveira Alves, que embora estando num momento decisivo da sua carreira encontrou sempre tempo para me guiar ao longo deste trabalho. Os meus desejos de boa sorte na conclusão desta etapa que todos nós sabemos que vai correr pelo melhor. Professor Pedro Henriques Abreu, que sempre teve uma voz activa no meu trabalho, alertando-me sempre para os erros cometidos não só para zelar pelo bom rumo do trabalho mas também pelo meu crescimento enquanto pessoa.

Aos meus pais, por sempre terem apoiado as minhas decisões de carreira compreendendo que o pouco tempo que passo com eles destina-se à construção do meu futuro.

Ao meu irmão, por todo o companheirismo ao longo do percurso académico e por toda a ajuda prestada ao longo dos últimos anos.

Ao pessoal do AmILab, por todo o apoio que me deram ao longo do último ano e pelo óptimo ambiente de trabalho proporcionado.

Por fim, gostaria de agradecer a todos os meus amigos, colegas e outras pessoas que fizeram da minha vida académica nestes últimos anos um período que ficará para sempre marcado em mim.

Conteúdo

Resumo **iii**

Lista de Tabelas **vii**

Lista de Figuras **ix**

Lista de Acrónimos **xi**

1 Introdução **1**

1.1 Contexto e Motivação 1

1.2 Objectivos 4

1.3 Estrutura do documento 5

2 Estado da Arte **6**

2.1 Extracção de Informação 6

2.1.1 Ferramentas 7

2.1.2 Análise de Ferramentas de Extracção de Informação 9

2.2 Fontes de Enriquecimento Semântico 13

2.2.1 Fontes Estudadas 14

2.2.2 Termos de Utilização das Fontes 18

2.3 Métricas de Similaridade entre *Strings* 20

2.4 Etiquetagem Automática 22

2.4.1 OpenCalais 22

2.4.2 Zemanta 22

2.4.3 AlchemyAPI 24

2.4.4 Yahoo Content Analysis API 24

2.4.5 Análise Comparativa 25

2.5 Enriquecimento Semântico 26
2.5.1 Kusco 26
2.5.2 Topica 28
2.5.3 Abordagens baseadas em actividades 29
2.5.4 Análise Comparativa 30

3 Abordagem **31**
3.1 Especificação Funcional 31
3.1.1 Análise de Requisitos 31
3.1.2 Arquitectura do PPS 2 - SEMA 33
3.1.3 Arquitectura do Módulo de Enriquecimento Semântico 35
3.1.4 Especificação dos Dados a extrair das Fontes de Enriquecimento . . 35
3.2 Metodologia 38
3.2.1 Pesquisa em fontes de informação 38
3.2.2 Integração e detecção de duplicados 40
3.2.3 Extracção de Termos 41
3.2.4 Validação e contextualização de termos 42
3.2.5 Cálculo de relevância 46

4 Implementação **48**
4.1 Módulos Implementados 48
4.1.1 Extracção de Dados nas fontes 48
4.1.2 Integração de Recursos 49
4.1.3 Extracção de Informação 50
4.1.4 Validação e contextualização de Termos 51
4.1.5 Cálculo de Relevância 52
4.1.6 Camada de Acesso a Dados 52
4.1.7 API de Acesso 53
4.2 Integração na plataforma TICE.Mobilidade 54

5 Experiências e Resultados **55**
5.1 Extracção de Recursos 55
5.2 Integração de Recursos 56
5.3 Etiquetagem Automática 60
5.3.1 Coerência de Termos 60

5.3.2 Ordem de Relevância 63

6 Conclusões **66**

6.1 Trabalho Futuro 67

Bibliografia **72**

Apêndices

A Estudo das Fontes de Enriquecimento Semântico **74**

B Modelos de Dados **76**

C Especificação da API REST **79**

D Instalação do Módulo de Enriquecimento Semântico **84**

Lista de Tabelas

2.1 Comparação entre as ferramentas de Extracção de Informação 10
2.2 Sub-conjuntos de Teste 11
2.3 Resultados para a tarefa de Etiquetagem Morfológica 12
2.4 Identificação de Sintagmas Nominais para o Inglês 12
2.5 Identificação de Sintagmas Nominais para o Português 12
2.6 Reconhecimento de Entidades Mencionadas para o Inglês 12
2.7 Reconhecimento de Entidades Mencionadas para o Português 12
2.8 Características utilizadas para classificar as fontes de enriquecimento 15
2.9 Exemplo dos Resultados do OpenCalais 23
2.10 Comparação entre as diferentes ferramentas de Etiquetagem Automática . . 25

3.1 Requisitos Funcionais 32
3.2 Requisitos Não-Funcionais 33
3.3 Dados disponíveis nas diversas fontes de enriquecimento 37

5.1 Valores resultantes da Extracção de Recursos 56
5.2 Exemplo simplificado de um grupo de duplicados 57
5.3 Resultados da validação do processo de integração de recursos 58
5.4 Exemplo de verdadeiros e falsos positivos 59
5.5 Exemplo simplificado da folha de cálculo para validação da Coerência de Termos 63
5.6 Resultados da validação relativa à experiência da coerência de termos . . . 63
5.7 Teste do Qui-Quadrado relativo à experiência da Coerência de Termos . . . 63
5.8 Resultados da validação relativa à experiência de ordem de relevância . . . 64
5.9 Teste do Qui-Quadrado relativo à experiência da ordem de relevância . . . 64

A.1 Quadro comparativo das Fontes de Enriquecimento Semântico estudadas . . 75

Lista de Figuras

1.1 Arquitectura TICE.Mobilidade 3

2.1 *Workflow* do OpenCalais 23
2.2 Modelo de Enriquecimento Semântico de Lugares do Kusco 27
2.3 Arquitectura do Topica 28

3.1 Arquitectura do PPS2 - SEMA 34
3.2 Arquitectura do Módulo de Enriquecimento Semântico 36
3.3 Modelo de Enriquecimento Semântico 39
3.4 Número de palavras por título na versão Portuguesa da *Wikipedia* 43
3.5 Número de palavras por título na versão Inglesa da *Wikipedia* 44

4.1 Visualização dos dados do Módulo de Enriquecimento Semântico 54

5.1 Frequência de termos extraídos por POI 61
5.2 Frequência de palavras por Termo 62

B.1 Modelo de dados 77

Lista de Acrónimos

ANOVA	*Analysis of Variance* 58
API	*Application Programming Interface* 8
ASA	*Atenção Selectiva Artificial* 2
CRF	*Conditional Random Field* 8
DAO	*Data Access Object* 53
ES	*Enriquecimento Semântico* 2
IS	*Interoperabilidade Semântica* 2
HTML	*Hypertext Markup Language* 6
HTTP	*Hypertext Transfer Protocol* 35
JSON	*JavaScript Object Notation* 14
KUSCO	*Knowledge Unsupervised Search for populating Concepts on Ontologies* . 26
NLP	*Natural Language Processing* 6
POI	*Ponto de Interesse* 1
PPS	*Processo, Produto ou Serviço* 1
RDF	*Resource Description Framework* 29
REST	*Representational State Transfer* 33
SEMA	*Selecção de informação baseada em mecanismos de atenção selectiva, enriquecimento semântico e interoperabilidade semântica* 2
SPARQL	*SPARQL Protocol and RDF Query Language* 29
SOAP	*Simple Object Access Protocol* 7
TF – IDF	*Term Frequency x Inverse Document Frequency* 29
URL	*Uniform Resource Locator* 24
XML	*eXtensible Markup Language* 14

Capítulo 1

Introdução

1.1 Contexto e Motivação

Através do crescimento contínuo da cultura *web* na sociedade actual, verifica-se o crescimento *online* de entidades representativas de recursos e acontecimentos com que nos deparamos no dia-a-dia e uma evolução na maneira como estes são representados. Tanto os POIs (*pontos de interesse*) como os eventos são um bom exemplo desta evolução, dado que actualmente a quantidade destas representações de lugares e acontecimentos reais *online* cresce de dia para dia, tanto por parte de entidades comerciais ou culturais que queiram divulgar os seus serviços como por parte de indivíduos que queiram assinalar a sua presença em determinado local, um movimento que recentemente foi impulsionado pelas redes sociais mais abrangentes e pela massificação dos dispositivos móveis baseados em localização.

A mobilidade nos grandes centros populacionais tem sido objecto de estudo nos últimos anos [1], tentando inferir como as populações se movimentam mediante as suas actividades e os acontecimentos que caracterizam o espaço. A classificação de um espaço tendo em conta os serviços e o tipo de acontecimentos que o compõem [2] é um passo importante para este tipo de abordagem, sendo para isto necessária uma representação enriquecida destes.

Tendo em conta as abordagens já desenvolvidas nesta área e os resultados obtidos [3], o presente trabalho pretende aplicar estes conceitos em território português através do projecto **TICE.Mobilidade – Sistema de Mobilidade Centrado no Utilizador**. Este projecto visa disponibilizar uma plataforma digital de serviços de mobilidade centrados no utilizador, combinando mobilidade, optimização energética e gestão de espaços urbanos. O projecto encontra-se organizado em unidades estruturais, os PPS's(*Processos, Produtos*

ou Serviços), sendo subdivididos em duas categorias: PPS's transversais e de serviços (ou verticais). Estes últimos fornecem uma gama de serviços de apoio à mobilidade para o utilizador final, enquanto que os PPS's transversais asseguram a recolha de dados facilitando o acesso a estes através de mecanismos de interoperabilidade para com os restantes PPS's.

O presente trabalho insere-se no PPS 2 - SEMA (*Selecção de informação baseada em mecanismos de atenção selectiva, enriquecimento semântico e interoperabilidade semântica*), um dos dois PPS's transversais da plataforma TICE.

Como podemos observar na figura 1.1, existem actualmente 10 PPS's, sendo os dois primeiros considerados transversais e os seguintes verticais. O PPS 1 - *One.Stop.Transport* consiste numa plataforma de aquisição, tratamento e análise de dados, que podem ser relativos a trânsito, meteorologia, pontos de interesse e afins. Estes dados são provenientes de diversos fornecedores de informação, como operadores de mobilidade, serviços de *WebGIS* e outros agregadores de conteúdos.

Os serviços fornecidos pelos PPS's 3 a 10 compreendem soluções que passam pela gestão de veículos autónomos, planeamento de rotas de mobilidade, *bike-sharing*, eficiência energética, entre outros. Os dados adquiridos pelo PPS 1 são aqui utilizados, mas para isso necessitam de ser sujeitos a algumas transformações. É aqui que entra o PPS 2 - SEMA, uma plataforma que transforma a informação recebida através do PPS 1 em informação semântica que constitui conhecimento importante para os PPS's verticais.

Estando o presente trabalho enquadrado neste último PPS, efectuamos uma descrição mais detalhada do seu funcionamento. Este divide-se em três módulos distintos:

- IS (*Interoperabilidade Semântica*)
- ASA (*Atenção Selectiva Artificial*)
- ES (*Enriquecimento Semântico*)

O primeiro, como o nome indica, fornece uma camada de interoperabilidade semântica sobre os dados provenientes do PPS 1, anotando-os com metadados semânticos de modo a que os PPS's superiores saibam interpretá-los correctamente.

O serviço de ASA é dotado pela capacidade de gerir a atenção selectiva do utilizador, através de mecanismos que filtrem a informação fornecida pelo PPS 1 utilizando métricas de utilidade, atenção selectiva e diversidade. Assim, é garantida a relevância da informação apresentada tendo em conta o perfil de cada utilizador e o contexto em que este se encontra. Assim, os restantes PPS's podem tirar partido de alertas específicos sobre eventos (por

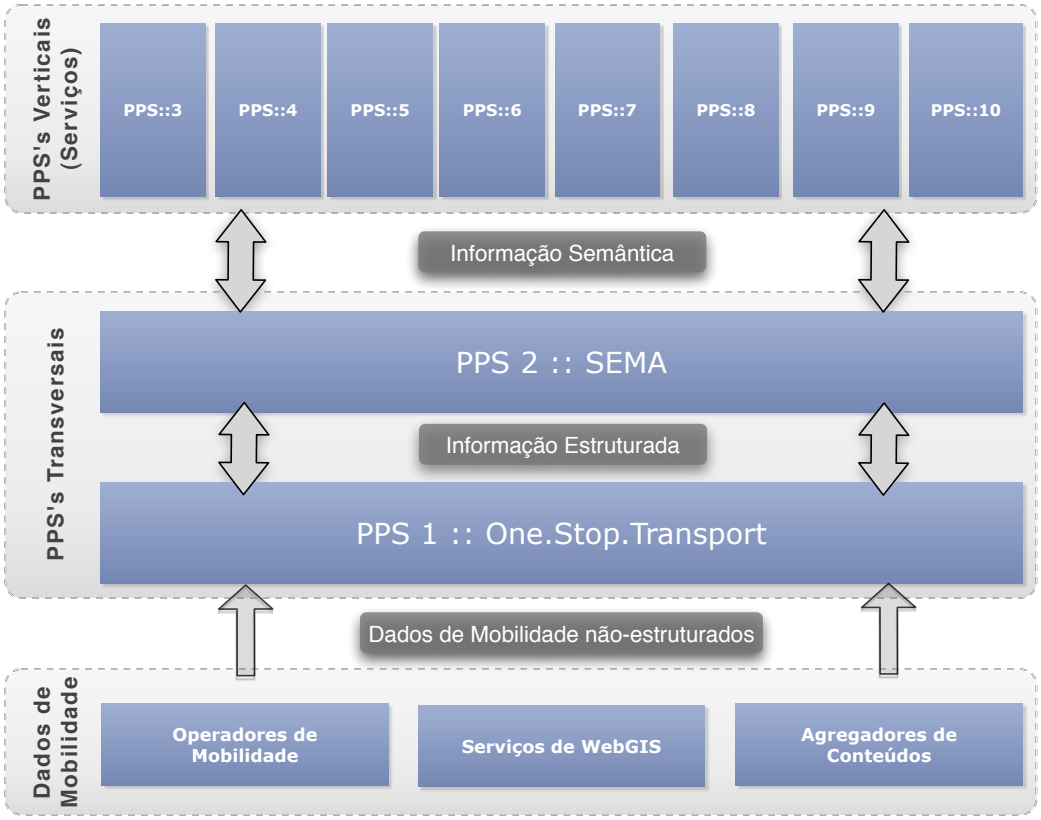

Figura 1.1: Arquitectura TICE.Mobilidade

exemplo, um evento que gera congestionamento da mobilidade na rota em que o utilizador se encontra).

Por fim, o serviço de ES é onde se centra este trabalho, tendo como principal funcionalidade o enriquecimento semântico de recursos. Numa perspectiva geral, este processo consiste numa indexação dos dados enriquecidos do espaço, o que permite fazer uma análise da mobilidade através da classificação e segmentação do espaço em função dos recursos que lhe estão associados.

O processo de enriquecimento semântico é feito com recurso a fontes de enriquecimento semântico e de contextualização, com o intuito de obter dados relevantes que aprimoram a informação geoespacial.

1.2 Objectivos

O principal objectivo deste estágio centra-se no desenvolvimento de um módulo de enriquecimento semântico de lugares (entre outros tipos de recursos, como eventos e notícias) para a língua portuguesa, devendo ainda ser capaz de classificar áreas geográficas de acordo com os serviços oferecidos ao seu redor.

Este objectivo complementa em parte o principal objectivo do SEMA, que passa pelo estabelecimento de uma ponte entre o PPS 1 - *One.Stop.Transport* (que faz toda a aquisição de dados) e os outros PPSs.

Pretende-se que através do enriquecimento de dados provenientes do PPS 1, seja possível obter dados complementares que criem valor para utilização *a posteriori* destes.

Deste modo, pretende-se que o módulo de enriquecimento semântico seja dotado de funcionalidades como:

- **Extração e Integração de POIs e Eventos**;
- **Etiquetagem Semântica Automática**;
- **Enriquecimento Semântico de Recursos já existentes**;

A extracção e integração de POIs e eventos consiste numa recolha exaustiva de dados representativos destes, com recurso a diversas fontes de informação *online*. Para manter a coerência dos dados, esta recolha deve garantir que não existam recursos duplicados, utilizando para isso métricas de comparação adequadas para este tipo de dados. Os dados resultantes deste processo serão armazenados numa ontologia, permitindo assim a partilha de informação estruturada para outros serviços ou módulos da plataforma TICE.Mobilidade.

A funcionalidade de Etiquetagem Semântica Automática passa pela criação de um serviço que permite classificar um local ou recurso através de uma lista de conceitos (*tags*). Para cada recurso serão associadas *tags* extraídas de fontes de informação e contextualização (Ex: *Wikipedia*), sendo estas ordenadas por ordem de relevância. Esta ordem de relevância é calculada através de métricas estatísticas que têm em conta o factor de diferenciação que cada *tag* representa no conjunto de documentos disponível. Por exemplo, para o POI "Exploratório de Coimbra"[1], seriam obtidas *tags* como "Ciência Viva", "Museu", "Aprendizagem".

[1] Centro interactivo de ciência viva, localizado em Coimbra

O Enriquecimento Semântico de Recursos já existentes tem como objectivo obter novas informações relevantes para um determinado recurso. Recorrendo a fontes de enriquecimento (que podem ser as próprias fontes de POIs) e a fontes de contextualização, tenta-se enriquecer um recurso já existente na base de conhecimento. Um bom exemplo deste objectivo pode ser um restaurante sobre o qual actualmente apenas conhecemos o nome e a localização geográfica. Após o processo de enriquecimento semântico, podem-se obter dados mais relevantes, como as especialidades da casa, horário de funcionamento, entre outros.

Estas funcionalidades deverão ser disponibilizadas através do Módulo de Enriquecimento Semântico tanto para os outros módulos do SEMA como para os outros PPS's, de modo a que todos possam beneficiar deste serviço de uma forma simplificada.

1.3 Estrutura do documento

Este primeiro capítulo tem como função familiarizar o leitor com o contexto deste trabalho e a problemática que aborda, passando também pela definição dos objectivos que se pretendem atingir.

No Estado da Arte (capítulo 2) é feita uma análise detalhada sobre os conceitos teóricos e tecnologias envolvidas neste trabalho. Estes elementos constituem um suporte essencial para o desenvolvimento do sistema segundo a metodologia descrita em 1.2.

Com a abordagem (capítulo 3), são apresentadas as decisões tomadas na implementação da metodologia. São também aqui definidas métricas necessárias para refinamento de resultados.

A implementação da abordagem proposta é descrita em detalhe no capítulo 4, especificando as estratégias utilizadas no desenvolvimento do módulo e as tecnologias envolvidas.

No capítulo 5 são apresentadas as diferentes experiências realizadas e os respectivos resultados, com o intuito de validar a abordagem proposta no capítulo 3.

Por fim, o capítulo 6 apresenta as conclusões obtidas ao longo do desenvolvimento deste trabalho, sendo discutidas algumas ideias que podem levar a melhorias do sistema proposto e da abordagem implementada para trabalho futuro.

Capítulo 2

Estado da Arte

Através deste capítulo são clarificados os conceitos teóricos e tecnologias envolvidos neste trabalho, de modo a familiarizar o leitor com as metodologias desenvolvidas e respectiva contextualização.

Numa abordagem *bottom-up*, partimos dos avanços que têm sido feitos no campo do NLP(*Natural Language Processing*), passando por áreas como a extracção de informação, métricas de similaridade entre *strings*, etiquetagem automática e enriquecimento semântico.

2.1 Extracção de Informação

A Extracção de Informação é considerada uma sub-tarefa de Recuperação de Informação (*Information Retrieval*), que tem como principal objectivo a extracção automática de informação estruturada de fontes de informação não-estruturadas [4]. Para atingir este objectivo, são utilizadas diversas ferramentas de NLP.

Destacam-se algumas sub-tarefas da Extracção de Informação utilizadas neste trabalho:

- **Remoção de Ruído**, que consiste na filtragem do conteúdo de modo a eliminar termos que não trazem qualquer valor para a tarefa a executar. Por exemplo, ao extrair conteúdos de páginas *web*, é normal que surjam *tags* HTML(*HyperText Markup Language*), termos técnicos como "*http*"ou "*ftp*", palavras como "*email*", "contacto"ou outras palavras simples que em nada acrescentam valor aos dados que estamos a analisar. Existem diversas ferramentas de NLP que facilitam esta tarefa: no caso do *screen-scraping*, omitem as *tags* HTML ao processar o conteúdo de uma página *web*. Quanto ao restante ruído, normalmente são usadas listas de palavras comuns no

léxico em questão, denominadas *stopword lists*, de modo a que se possam ignorar as ocorrências destes termos.

- **Extracção de Termos**, que consiste na extracção automática de termos relevantes de um conjunto de documentos. As abordagens mais comuns para esta tarefa utilizam técnicas como Etiquetagem Morfológica e identificação de Sintagmas Nominais [5].

  A etiquetagem morfológica é utilizada para classificar cada palavra presente num texto de acordo com a sua classe gramatical. As palavras são classificadas em categorias, de acordo com o papel que estas desempenham no contexto [6].

  Já a identificação de sintagmas nominais consiste na divisão de um texto em conjuntos de palavras sintacticamente correlacionados [7]. Por exemplo, a frase "A cidade possui uma vasta oferta de restaurantes"seria transformada em "A cidade"e "uma vasta oferta de restaurantes", sendo que o primeiro conjunto é um sintagma nominal que constitui o sujeito da frase e o segundo um sintagma nominal que possui a função de predicado.

- **Reconhecimento de Entidades Mencionadas**, que consiste na classificação dos substantivos próprios existentes num texto em categorias pré-definidas, que podem ser referências a pessoas, locais, organizações, expressões temporais/numéricas ou outras entidades relevantes[8].

Dado que o âmbito deste trabalho não foca directamente o desenvolvimento de ferramentas para execução destas tarefas, iremos centrar-nos na análise das ferramentas já existentes para executá-las, de modo a seleccionar aquelas que revelem funcionalidades interessantes e nos proporcionem melhores resultados em termos de precisão estatística.

2.1.1 Ferramentas

- **F-EXT-WS**[1] [9] - É uma ferramenta disponibilizada através de um *webservice* SOAP (*Simple Object Access Protocol*) que permite executar algumas das tarefas mencionadas em 2.1, nomeadamente o reconhecimento de entidades mencionadas, identificação de sintagmas nominais e etiquetagem morfológica para a língua portuguesa, estando algumas destas tarefas também disponíveis para a língua inglesa.

- **Freeling**[2] [10] - É uma ferramenta *open-source* que permite a execução de diversas

[1] Mais informações em: `http://www.learn.inf.puc-rio.br/fextws/`

[2] Mais informações em: `http://nlp.lsi.upc.edu/freeling/`

tarefas de NLP (cobrindo todas as mencionadas em 2.1) para um vasto conjunto de línguas (incluindo o português), sendo esta bastante extensível e personalizável. A ferramenta é estruturada como uma biblioteca em C++, mas oferece diversas API's (*Application Programming Interface*) de integração com outras linguagens de programação, como *Java*, *Python* e *Perl*.

- **Apache OpenNLP**[3] - É um conjunto de ferramentas baseado em *Machine Learning* para NLP. Oferece funcionalidades como etiquetagem morfológica, reconhecimento de entidades mencionadas, identificação de sintagmas nominais entre outros. Fornece modelos para executar estas funcionalidades em diversos idiomas, incluindo o Português. Para além destas sub-tarefas de extracção de informação, permite-nos ainda treinar os modelos existentes com dados específicos, oferecendo ainda ferramentas que medem o desempenho destes modelos, que se revelam bastante úteis na sua avaliação.

- **Stanford CoreNLP**[4] - Fornece um conjunto de ferramentas de NLP, englobando ferramentas que permitem a execução de algumas das tarefas mencionadas em 2.1, entre outras. Por pré-definição, esta ferramenta está disponível para a língua inglesa. No entanto, é possível treinar alguns dos módulos que a compõem para outras línguas, mediante o fornecimento de textos de treino nesses idiomas.

- **CMS Chunker** - Uma ferramenta de análise sintáctica a partir da qual podemos extrair de um texto previamente etiquetado elementos como sintagmas nominais. Esta ferramenta foi desenvolvida como parte integrante de um projecto[11] do CISUC[5]. Para identificação dos sintagmas, esta ferramenta utiliza um conjunto de regras extraídas do recurso *Bosque*[12], disponibilizado pela Linguateca[6].

- **PT-NER** - Trata-se de um reconhecedor de entidades mencionadas para o português, desenvolvido por um estudante de doutoramento da Universidade de Coimbra, Filipe Rodrigues, para o AmILab[7] do CISUC. Para execução da tarefa, utiliza CRF's (*Conditional Random Fields*) [13] recorrendo ainda a algumas funcionalidades do *OpenNLP*, nomeadamente a etiquetagem morfológica. Este foi treinado através do corpus do

[3] Mais informações em: `http://incubator.apache.org/opennlp/`
[4] Mais informações em: `http://nlp.stanford.edu/software/corenlp.shtml`
[5] Mais informações em: `https://www.cisuc.uc.pt/`
[6] Mais informações em: `http://linguateca.pt/`
[7] *Ambient Intelligence Laboratory*

segundo *HAREM*[14], uma avaliação conjunta de sistemas de reconhecimento de entidades mencionadas em colecções de documentos em português, criado pela Linguateca.

2.1.2 Análise de Ferramentas de Extracção de Informação

As ferramentas apresentadas diferem bastante entre si, independentemente das funcionalidades que oferecem. O facto de nem todas suportarem a mesma linguagem ou de se cingirem apenas a uma pequena parte das sub-tarefas de extracção de informação leva-nos a ponderar uma forma de as poder avaliar conjuntamente.

Deste modo, foi tomada a decisão de separar estas ferramentas em subconjuntos distintos, sendo o principal elemento discriminador a língua que cada uma suporta. A tabela 2.1 evidencia as principais diferenças entre as ferramentas, o que nos permite proceder à criação destes subconjuntos.

Classificadas as ferramentas, foi possível criar subconjuntos de teste, como podemos observar na tabela 2.2.

A criação destes sub-conjuntos de teste foi feita após uma análise inicial das ferramentas, onde foram detectadas falhas ou mais-valias que as conduziram ao sub-conjunto de teste onde se encontram. Destacamos aqui algumas das decisões: O *Freeling*, devido ao facto de ter uma API muito pouco documentada não entrou em alguns grupos de teste. Outro factor que o excluiu de alguns testes, nomeadamente a etiquetagem morfológica para o português foi o facto de não obedecer aos *standards* de etiquetagem desta tarefa, catalogando cada palavra com *tokens* do léxico inglês. Outros testes com esta ferramenta foram descartados devido ao facto de a documentação ser muito escassa. O *Apache OpenNLP*, que por pré-definição não suporta o reconhecimento de entidades mencionadas para português, é dotado de um módulo para treinar o reconhecedor com diferentes *corpus* linguísticos, uma funcionalidade que nos pareceu interessante. Este facto levou a que esta ferramenta fosse incluída nos testes para reconhecimento de entidades mencionadas para o português.

A medida de avaliação mais comum para as tarefas de reconhecimento de entidades mencionadas e detecção de sintagmas nominais consiste na verificação dos resultados obtidos face a resultados que já estejam confirmados como correctos (*ground truth*). Esta verificação é feita recorrendo à medida F1[4], também conhecida como modelo de *precision/recall*. Esta medida baseia-se na precisão — *precision* (número de resultados correctos dividido pelo número de resultados) e na abrangência — *recall* (número de resultados correctos dividido pelo número de resultados correctos do *ground truth*) dos testes efectuados. O valor F1

Tabela 2.1: Comparação entre as ferramentas de Extracção de Informação

Ferramenta	Tarefas Suportadas	Idiomas Suportados
F-EXT-WS	Etiquetagem Morfológica	Português, Inglês
	Identificação de Sintagmas Nominais	Inglês
	Reconhecimento de Entidades Mencionadas	Português
Freeling	Etiquetagem Morfológica	Inglês, Espanhol, Catalão, Galego, Italiano, Português
	Identificação de Sintagmas Nominais	Inglês, Espanhol, Catalão, Galego, Italiano, Português
	Reconhecimento de Entidades Mencionadas	Inglês, Espanhol, Catalão, Galego, Italiano, Português
Apache OpenNLP	Etiquetagem Morfológica	Inglês, Português, Sueco, Holandês, Alemão, Dinamarquês, Espanhol, Tailandês
	Identificação de Sintagmas Nominais	Inglês
	Reconhecimento de Entidades Mencionadas	Inglês, Espanhol, Holandês,
Stanford CoreNLP	Etiquetagem Morfológica	Inglês, Francês, Alemão, Árabe
	Reconhecimento de Entidades Mencionadas	Inglês
CMS Chunker	Identificação de Sintagmas Nominais	Português
PT-NER	Reconhecimento de Entidades Mencionadas	Português

varia entre 0 e 1 (podendo também ser expresso em percentagem), sendo definido por:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2.1}$$

Através desta medida, é-nos possível medir o desempenho das tarefas executadas pelas ferramentas em questão. No caso da etiquetagem morfológica, o seu desempenho é medido apenas pela precisão da etiquetagem que faz.

Tabela 2.2: Sub-conjuntos de Teste

Tarefa	**Português**	**Inglês**
Reconhecimento de Entidades Mencionadas	F-EXT-WS, Apache OpenNLP, PT-NER	Apache OpenNLP, Stanford CoreNLP
Identificação de Sintagmas Nominais	Apache OpenNLP, CMS Chunker	F-EXT-WS, Apache OpenNLP
Etiquetagem Morfológica	F-EXT-WS, Apache OpenNLP	F-EXT-WS, Freeling, Apache OpenNLP, Stanford CoreNLP

Para treinar o reconhecedor de entidades mencionadas do *Apache OpenNLP* foi pensado inicialmente utilizar o *corpus* do segundo *Harem*, mas para este reconhecedor funcionar devidamente é necessário treiná-lo com um *corpus* que contenha mais de 15.000 frases, um valor bastante superior às dimensões do segundo *Harem*. Foi então utilizado o *corpus* Amazônia[15], também fornecido pela Linguateca, sendo este mais robusto mas não tão preciso quanto o segundo *Harem*, dado que não foi integralmente revisto por linguistas. Quanto à língua inglesa, não foi possível obter o *corpus* do *CONLL*[8] *2003* de modo a utilizá-lo como *ground truth*. Dada a ausência destes dados, limitamo-nos a colocar aqui os resultados encontrados nas suas páginas oficiais[9] para este *corpus*.

Quanto à etiquetagem morfológica e detecção de sintagmas nominais, para o Português utilizamos o *corpus* Bosque, dado que este é revisto por linguistas. Já para a língua inglesa utilizamos o corpus do *CONLL 2000*.

As tabelas 2.3, 2.4, 2.5, 2.6 e 2.7 contêm os resultados dos testes efectuados, sendo uma pequena análise dos resultados obtidos descrita de seguida.

Durante o decorrer deste trabalho, o *webservice* que mantém o F-EXT-WS entrou em manutenção, um factor que pôs em causa a sua viabilidade de utilização. Como consequência, foi tomada a decisão de utilizar exclusivamente ferramentas sobre as quais tenhamos controlo total, de modo a não comprometer as tarefas fulcrais da abordagem a implementar.

Relativamente à tarefa de etiquetagem morfológica, o *Stanford CoreNLP* revelou um melhor desempenho para a língua inglesa, enquanto que para a língua portuguesa o *Apache OpenNLP* mostrou-se superior ao *F-EXT-WS*. Tendo em conta os problemas que tivémos

[8] *Conference on Natural Language Learning*, onde anualmente são avaliadas ferramentas de NLP.

[9] Dados presentes em `https://cwiki.apache.org/OPENNLP/` e `http://nlp.stanford.edu/`

Tabela 2.3: Resultados para a tarefa de Etiquetagem Morfológica

Ferramentas	***Precision***	
	Português	Inglês
F-EXT-WS	0.9031	0.9395
Freeling	—	0.8628
Apache OpenNLP	0.9260	0.8923
Stanford CoreNLP	—	0.9658

Tabela 2.4: Identificação de Sintagmas Nominais para o Inglês

Ferramenta	***Precision***	***Recall***	***F1***
F-EXT-WS	0.9210	0.9136	0.9173
Apache OpenNLP	0.9258	0.9222	0.9240

Tabela 2.5: Identificação de Sintagmas Nominais para o Português

Ferramenta	***Precision***	***Recall***	***F1***
Apache OpenNLP	0.9369	0.9306	0.9337
CMS Chunker	0.5407	0.5187	0.5295

Tabela 2.6: Reconhecimento de Entidades Mencionadas para o Inglês

Ferramenta	***Precision***	***Recall***	***F1***
Apache OpenNLP	0.8539	0.6529	0.7400
Stanford CoreNLP	0.9328	0.9271	0.9299

Tabela 2.7: Reconhecimento de Entidades Mencionadas para o Português

Ferramenta	***Precision***	***Recall***	***F1***
F-EXT-WS	0.0459	0.0157	0.0234
Apache OpenNLP	0.4624	0.1357	0.2098
PT-NER	0.6495	0.5832	0.6145

com esta última ferramenta, este resultado revela-se bastante favorável. Na identificação de sintagmas nominais para o inglês, o *Apache OpenNLP* revelou novamente um desempenho

superior. Já para a língua portuguesa, esta ferramenta também apresentou um desempenho bastante satisfatório, com um valor F1 na ordem dos 93% contra um valor na ordem dos 53% do *CMS Chunker*.

Finalmente, para a tarefa de reconhecimento de entidades mencionadas o *Stanford CoreNLP* apresentou um melhor desempenho para a língua inglesa. No caso da língua portuguesa, os valores obtidos pela maioria das ferramentas foram extremamente baixos; no caso do *Apache OpenNLP* isto pode-se dever ao facto de ter sido treinado pelo *corpus* Amazônia, sendo este muito extenso e pouco preciso. Mesmo assim apresentou um desempenho superior ao *F-EXT-WS*, cujos resultados ficaram muito aquém do esperado. O *PT-NER* foi a ferramenta que apresentou melhor desempenho neste sub-conjunto.

Após esta análise, a decisão sobre quais as ferramentas utilizar torna-se mais simples, sendo esta aqui apresentada:

- Etiquetagem Morfológica
  - Português - *Apache OpenNLP*
  - Inglês - *Stanford CoreNLP*
- Identificação de Sintagmas Nominais
  - Português - *Apache OpenNLP*
  - Inglês - *Apache OpenNLP*
- Reconhecimento de Entidades Mencionadas
  - Português - *PT-NER*
  - Inglês - *Stanford CoreNLP*

2.2 Fontes de Enriquecimento Semântico

As fontes de Enriquecimento Semântico de onde serão constantemente recolhidas informações sobre lugares, eventos ou que servem para efectuar a contextualização de termos, são sem dúvida um suporte essencial para este trabalho.

Tendo em vista a sua extrema importância, procedeu-se a um estudo detalhado das diversas fontes de enriquecimento semântico, incidindo sobre um conjunto específico de atributos capazes de classificar cada fonte através de diversas perspectivas. Por exemplo,

caso uma fonte seja especializada num certo tipo de categoria de POIs ou eventos (museus, restaurantes) podemos classificá-la como específica para um dado domínio, enquanto que outras mais genéricas poderão abranger diversas categorias de entidades.

A tabela 2.8 permite sintetizar que características poderão ser utilizadas para classificar as fontes de enriquecimento semântico, apresentando para cada uma destas os valores possíveis.

Existe uma grande variedade de fontes de enriquecimento, estruturadas de diferentes formas quanto à organização da informação *online*, que disponibilizam dados sobre diversos tipos de entidades na *web*.

Numa perspectiva mais generalista, estas podem ser consideradas como: estruturadas — com dados devidamente organizados em formato *standard* (ex.: XML (*Extensible Markup Language*), JSON (*Javascript Object Notation*), normalmente disponibilizados sob a forma de uma API); semi-estruturada — com um misto de texto e campos presentes na mesma página *web*; não-estruturada — com uma grande proporção de texto livre [4].

2.2.1 Fontes Estudadas

A escolha das fontes a serem estudadas foi baseada não só na popularidade de cada uma, mas também na qualidade e quantidade de dados oferecidos. O risco de recolher informações na *web* sob a consequência de estas não serem correctas foi tido em conta: algumas das fontes (nomeadamente as redes sociais) possuem estatísticas que podem ser utilizadas para criar métricas de popularidade e confiabilidade. No entanto, estas estatísticas muitas vezes não são suficientemente precisas de modo a filtrar todo o ruído que pode surgir na informação extraída. Com base nisto, torna-se necessário atribuir um grau de confiabilidade a cada fonte, dando prioridade a informações oficiais, normalmente fornecidas por directórios com moderadação activa de conteúdos.

Foi também necessário algum trabalho de pesquisa relativamente aos termos e condições de utilização de cada uma, de modo a inferir se o uso destas fontes seria legalmente correcto, visto que diversos serviços apresentam políticas bastante restritas no que toca ao uso dos dados que fornecem.

Enumeramos assim as fontes que foram objecto de estudo, fazendo uma sucinta descrição de cada uma:

- **SAPO**[10] - É um directório comercial que pode ser acedido através de uma API

[10] Mais informações em: `http://www.sapo.pt/`

Tabela 2.8: Características utilizadas para classificar as fontes de enriquecimento

Atributo	**Valores Possíveis**	**Descrição**
Natureza	Colaborativa Comercial/Institucional Mista	Classifica como é gerida a fonte de enriquecimento: através de participantes registados ou através da própria empresa/instituição detentora da fonte.
Estrutura	Estruturada Semi-Estruturada Livre	O modo como os dados estão organizados na fonte de enriquecimento, através de campos e/ou tabelas. Nestes campos podem estar valores pré-definidos (estruturados) ou descrições em texto (semi-estruturados). Caso os dados das fontes de enriquecimento estejam disponibilizados apenas através de texto sem qualquer estrutura comum entre as várias entidades presentes nesta fonte, esta é considerada como texto livre.
Acesso	Livre Registado Pago	Termos de utilização e licença a ter em caso de armazenamento de dados.
Informação	Geográfica Nominal Categorias Atributos	Âmbito da informação disponibilizada, que pode ser através de dados georeferenciados (coordenadas GPS ou morada), nome da entidade, categoria ou outros valores que sejam apresentados para cada entidade.
Domínio	Específico Genérico	Fonte especializada num conjunto específico (Ex: apenas relacionado com a alimentação: cafés, bares, etc.) ou de um vasto conjunto de categorias.
Geografia	Restrita a uma zona Nacional Internacional	Cobertura geográfica das entidades presentes na fonte de enriquecimento.
Tipos de Entidades	POIs Eventos	Quais as entidades representadas na fonte de enriquecimento.
Organização Categorias	Taxonomia Lista Conjunção de Categorias	Como estão organizadas as categorias da fonte de enriquecimento para cada entidade. Se estão estruturadas de forma hierárquica através de uma taxonomia, ou apenas através de uma lista de categorias. Além disso, mais do que uma categoria poderá ser usada para classificar cada entidade.
Meio de Extracção	API Dump da BD Web Scraping	Meio pelo qual é possível consultar a informação presente na fonte.

fornecida pelo portal SAPO. Apesar do conjunto de serviços relacionados com Sistemas de Informação Geográfica (SIG) já não se encontrar listado nos serviços de domínio público, este continua a estar disponível mediante o pagamento de uma taxa mensal. Os dados possíveis podem ser acedidos através de uma API com diversos serviços que possibilitam a obtenção de informação sobre POIs ou eventos, consoante os parâmetros de pesquisa que se decidam usar.

- **Factual**[11] - Uma plataforma aberta que disponibiliza dados georeferenciados em larga escala. Estes dados vão desde pontos de interesse a geometrias que descrevem uma determinada área (como uma freguesia, um concelho ou um distrito), apresentando uma vasta cobertura a nível global: 58 milhões de entidades distribuídas por 50 países, contendo cerca de 250.000 pontos de interesse só em território Português. Estes dados podem ser acedidos mediante uma API que suporta pedidos georeferenciados, de modo a obter todos os pontos de interesse numa determinada área.

- **Lifecooler**[12] - Segundo é afirmado no próprio, é o portal de Turismo e Lazer mais visitado em Portugal, contendo informações bastante descritivas sobre pontos de interesse e eventos recorrentes que ocorrem em determinada região, englobando diversos artigos sobre cada entidade. Os dados aqui presentes encontram-se bastante estruturados, sendo possível obter informações sobre pontos de interesse com base no distrito, município ou freguesia onde se encontram, estando estes devidamente categorizados se for de maior interesse uma pesquisa baseada em categorias.

- **Gowalla**[13] - É uma rede social baseada em localização, onde os utilizadores assinalam a sua presença num local fazendo *check-in* neste, através de uma lista de pontos de interesse localizados nas proximidades de onde o utilizador se encontra. Estes pontos de interesse são criados pela comunidade de utilizadores, podendo muitas das vezes surgir locais que não constituam pontos de interesse de valor significante (por exemplo: "Casa do João"). No entanto, a API fornece-nos dados suficientes para descartar estes locais. A plataforma vai recompensando os utilizadores por terem feito *check-in* em locais de grande aderência, com itens que podem ser trocados com outros utilizadores, levando assim um pouco à criação de um espírito de *gamification*. A cada ponto de

[11] Mais informações em: `http://www.factual.com/`
[12] Mais informações em: `http://www.lifecooler.com/`
[13] Mais informações em: `http://blog.gowalla.com/`

interesse podem ser anexados comentários, fotos ou *highlights*. Este últimos consistem no encaixe do POI numa categoria mais específica (ex.: "Coffee Wonderland"). Durante este estudo a plataforma foi comprada pela rede social *Facebook*, anunciando assim o fim do seu serviço nos próximos meses, oferecendo aos utilizadores a possibilidade de exportarem os seus dados do serviço. Dado isto, esta plataforma deixa de nos ser útil, visto que não haverá possibilidade de obter dados a longo prazo.

- **Foursquare**[14] - À semelhança do *Gowalla*, o *Foursquare* também consiste numa rede social *location-based*. Aqui, os locais (*venues*) encontram-se categorizados sob uma hierarquia mais robusta, recorrendo a categorias e subcategorias. É também permitido aos utilizadores inserir conteúdos em cada local, como fotos, *links* e comentários (aqui apelidados de *tips*). Através da sua API, podemos obter a listagem dos pontos de interesse mais próximos consoante vários atributos, como os locais onde mais utilizadores fizeram *check-in* nos últimos tempos e os locais mais populares/recomendados.

- **Google Places**[15] - É um serviço que fornece informação sobre locais como estabelecimentos comerciais, localizações geográficas e pontos de interesse específicos. Os POIs fornecidos por este serviço são os mesmos que encontramos ao pesquisar no *Google Maps*. A sua API permite-nos obter uma lista de pontos de interesse dado um par de coordenadas, um raio de acção (em metros), uma *query* de texto (opcional) e uma lista de categorias pré-definidas, o que nos dá alguma especificidade no caso de querermos obter um ponto de interesse em particular.

- **Facebook**[16] - É a rede social mais popular da actualidade. Todos os dias, grandes quantidades de informação são geradas pelos seus utilizadores, informação esta que pode simbolizar a presença de pessoas num determinado local (*Facebook Places*), número de pessoas que irão estar presente num evento (*Facebook Events*) ou até mesmo informações sobre um determinado local, como promoções, horário, morada e afins (*Facebook Pages*). A API permite-nos aceder a todos estes serviços de modo contínuo (se nos limitarmos a informação de domínio público). A informação mais íntima dos utilizadores (testemunhos e informações pessoais) não está acessível para qualquer utilizador da API.

[14] Mais informações em: `http://foursquare.com/`
[15] Mais informações em: `https://developers.google.com/maps/documentation/places/`
[16] Mais informações em: `http://www.facebook.com/`

- **PAPEL**[17] - O Palavras Associadas Porto Editora - Linguateca (PAPEL)[16] é um recurso léxico criado a partir do Dicionário da Língua Portuguesa (DLP) da Porto Editora e desenvolvido no pólo de Coimbra da Linguateca. Neste recurso estão disponibilizadas relações semânticas entre as palavras existentes no dicionário, como por exemplo sinónimos. Ao contrário de outros recursos lexicais para o português de que temos conhecimento, o PAPEL é público, gratuito e utilizável por todos os actores de processamento da língua que o quiserem usar, encontrando-se aberto para subsequente melhoria pela comunidade. Através dele é-nos possível contextualizar termos simples, isto é, se dado termo simples existir no recurso, então existe uma, ou mais entrada(s) para esse termo no dicionário, o que o torna num conceito com significado. É dito termo simples porque poderão existir termos compostos como "Administração Interna"ou "actos jurídicos"e o recurso PAPEL apenas contém termos simples.

- **Wikipedia**[18] - É uma fonte genérica de enriquecimento semântico que permite a contextualização de informação genérica (tal como a descrição e características gerais de um restaurante) como específica para uma dada entidade. Os artigos da Wikipedia estão organizados em categorias (ex.: Restaurantes, Museus, etc.) que permitem obter novas entidades que ainda não estejam presentes na base de conhecimento. Além disso, não sendo propriamente um recurso lexical, a Wikipedia permite reconhecer termos compostos que não estejam presente no PAPEL. Desta forma é possível identificar um maior conjunto de palavras-chaves (simples e compostas) associadas a cada entidade a ser enriquecida semanticamente. As vantagens na utilização da Wikipedia para o processamento de linguagem natural têm sido alvo de discussão de diversos autores[17], dando especial relevância à abrangência do recurso, ao facto de ser disponibilizado abertamente e evidenciando que este é actualizado e revisto continuamente mediante a colaboração dos seus utilizadores.

No anexo A encontra-se uma tabela (A.1) comparativa das fontes de enriquecimento estudadas, segundo as características de classificação apresentadas na tabela 2.8.

2.2.2 Termos de Utilização das Fontes

Ao extrair dados de fontes *online*, é conveniente saber até que ponto este processo será legal. Dado isto, foram efectuados contactos com cada uma das fontes estudadas, de modo

[17]Mais informações em: `http://www.linguateca.pt/PAPEL/`
[18]Mais informações em: `http://pt.wikipedia.org/`

a inferir a viabilidade de utilização destas. Descrevemos assim sucintamente os progressos feitos até agora nesta área:

- **Foursquare**: A equipa do *Fourquare* respondeu prontamente às nossas questões, mas infelizmente o resultado não foi de todo o esperado. Foi-nos transmitido explicitamente que poderíamos utilizar o seu serviço como base de dados de pontos de interesse, mas a utilização dos recursos textuais (como comentários de utilizadores, críticas) não poderia ser feita. No entanto, a sua API continua a ser-nos útil para obter métricas de popularidade sobre locais e para comparar os dados mais simples dos pontos de interesse, como a morada ou *websites* associados.

- **Gowalla**: Embora não tenhamos obtido qualquer resposta da equipa do *Gowalla*, decidimos abandonar o estudo desta fonte devido ao que já foi referido anteriormente: a compra do serviço por parte do *Facebook*. Dado que a plataforma vai ser encerrada, não nos oferece perspectivas futuras para utilização dos seus dados.

- **Factual** - Sendo esta uma plataforma aberta de dados, estes são livres para ser usados por qualquer pessoa que tenha uma conta registada. A única condicionante aqui será obedecer aos limites de utilização da API disponibilizada.

- **Lifecooler** - Neste ponto apenas é mencionado que a apropriação indevida de criação intelectual não é aprovada pelo serviço. Dado que se pretende registar onde foi originado cada um dos dados extraídos, o devido crédito é atribuído aos autores que criam conteúdo e moderam este directório.

- **Google Places**: Os termos de utilizaçao do *Google Places* inicialmente só requerem que seja apresentada uma imagem em conjunto com o serviço, alusiva à utilização deste (Ex: *Powered by Google Places*). No entanto, já foram feitos contactos para clarificar todo o uso que queremos fazer da plataforma, cuja resposta ainda não foi obtida.

- **Facebook**: Nos termos de utilização do *Facebook* não foi encontrada nenhuma directriz que afectasse o nosso trabalho. Apenas é limitada a venda de dados obtidos através da aplicação. Graças às práticas de privacidade do serviço, quando um utilizador publica informação em domínio público, a informação pode ser utilizada por todos. Dado que o nosso interesse passa maioritariamente por informações públicas

(páginas de entidades, pontos de interesse e eventos), não existirão problemas com a utilização deste serviço.

- **SAPO**: Embora o serviço mais actual do SAPO esteja disponível mediante pagamento, estão a decorrer negociações entre o consórcio TICE.Mobilidade e o SAPO, de modo a ser possível a disponibilização deste serviço de forma gratuita para os elementos do consórcio.

2.3 Métricas de Similaridade entre *Strings*

Ao lidar com dados provenientes de várias fontes de informação, é imperativo ter em atenção a possibilidade de surgirem elementos duplicados. Para evitar este tipo de situações torna-se necessária a comparação minuciosa dos conteúdos textuais (como os nomes), podendo assim integrar dados que apresentem um elevado grau de semelhança.

Através de um conjunto de métricas que avaliam as semelhanças ou diferenças entre duas *strings*, é possível identificar semelhanças em nomes de entidades. Esta temática tem sido abordada em diferentes áreas como estatística, bases de dados e inteligência artificial, existindo diversos estudos que comparam as abordagens existentes [18, 19, 20].

Tendo como base os estudos referenciados, é notória a popularidade da distância de *Levenshtein*[21] (ou *edit distance*), uma métrica simples que se define pelo menor número de operações de edição (inserção, substituição ou remoção) necessárias para transformar uma *string* noutra. Por exemplo, distância de *Levenshtein* para as *strings* "teste"e "testar"é 2, dado que se verificam duas operações de edição para transformar a primeira *string* na segunda. Esta distância pode ser convertida para uma métrica de similaridade (entre 0.0 e 1.0) através da fórmula 2.2.

$$dist_{lv}(s1, s2) = 1.0 - \frac{dist_{lv}(s1, s2)}{max(|s1|, |s2|)} \tag{2.2}$$

sendo *s1* e *s2* as duas *strings* a comparar e $dist_{lv}(s1, s2)$ a função que representa a distância de *Levenshtein*, retornando o valor 0 se as *strings* forem iguais ou um número positivo que corresponde ao número de edições caso estas sejam diferentes. Este valor é simétrico, sendo sempre garantido que $0 \leq dist_{lv}(s1, s2) \leq max(|s1|, |s2|)$ e $abs(|s1| - |s2|) \leq dist_{lv}(s1, s2)$. Esta última condição permite a filtragem rápida de pares de *strings* cuja diferença de tamanho seja grande. A complexidade algorítmica desta métrica é de $O(|s1| \times |s2|)$.

A métrica de *Jaro*[22] é algo semelhante a esta abordagem, contabilizando as inserções, remoções e transposições de caracteres. O algoritmo calcula o número c de caracteres comuns em duas *strings*, *s1* e *s2* (apenas considerando transposições compreendidas em distâncias menores que metade do comprimento da maior *string*) e o número de transposições t. Esta métrica de similaridade, representada como $sim_{jaro}(s1, s2)$, é calculada através da fórmula 2.3.

$$sim_{jaro}(s1, s2) = \frac{1}{3}\left(\frac{c}{|s1|} + \frac{c}{|s2|} + \frac{c-t}{c}\right) \tag{2.3}$$

A complexidade algorítmica desta métrica é de $O(|s1| + |s2|)$, revelando-se assim mais económica a nível computacional.

Uma variante da métrica de Jaro, proposta por *Winkler*[23] apresenta melhorias face às anteriores, assumindo que normalmente os erros na comparação entre *strings* acontecem no início destas. Utiliza o comprimento P do maior prefixo comum entre *s1* e *s2*, podendo este ter o valor máximo de 4, sendo $P' = max(P, 4)$. A métrica de similaridade de *Jaro-Winkler*, cujo valor é representado por $sim_{Jaro-Winkler}(s1, s2)$ é calculada através da equação 2.4.

$$sim_{Jaro-Winkler}(s1, s2) = sim_{jaro}(s1, s2) + \frac{P'}{10} \cdot (1.0 - sim_{jaro}(s1, s2)) \tag{2.4}$$

Estudos prévios[18] comprovam que as duas últimas métricas (*Jaro* e *Jaro-Winkler*) apresentam melhores resultados para *strings* de tamanho reduzido, o que se enquadra na tarefa de calcular similaridade entre nomes de recursos.

Actualmente, existem ferramentas onde este tipo de métricas já se encontra totalmente implementado e testado: o *SecondString*[19] é uma biblioteca disponível para a linguagem *Java* que engloba as métricas que aqui foram mencionadas. Este irá ser usado não só para determinar a similaridade entre nomes de recursos mas também para comparar nomes de entidades no processo de contextualização de recursos.

[19]Mais informações em: `http://secondstring.sourceforge.net/`

2.4 Etiquetagem Automática

Dado que um dos objectivos deste trabalho foca a etiquetagem automática de recursos, analisámos algumas ferramentas com funcionalidades semelhantes ao serviço a desenvolver:

2.4.1 OpenCalais

O OpenCalais[20] [24] é um serviço disponibilizado por intermédio de uma API, que recebendo texto não-estruturado cataloga automaticamente entidades (pessoas, empresas, lugares, áreas geográficas), relações e eventos. Embora não exista muita informação sobre como este processo é feito, devido ao facto de ser uma ferramenta proprietária sabe-se que são utilizados técnicas de NLP e *Machine Learning*. Actualmente, suporta apenas três linguagens: Inglês, Espanhol e Francês. Na figura 2.1 podemos observar detalhadamente o tipo de dados que pode ser extraído de texto não-estruturado. Para demonstrar o funcionamento do *OpenCalais*, utilizamos a seguinte texto[21] (em inglês):

José Manuel Durão Barroso (born 23 March 1956) is a Portuguese politician. He is President of the European Commission, since 23 November 2004. He served as Prime Minister of Portugal from 6 April 2002 to 17 July 2004.

Obtemos a classificação presente na tabela 2.9, que torna perceptível como o serviço cataloga entidades, eventos e relações. No que toca às entidades, estas são devolvidas juntamente com o seu valor de relevância de modo a destacar a importância de cada entidade. De modo a preservar a estrutura devolvida pelo *OpenCalais*, os resultados não foram traduzidos.

O acesso à API é gratuito, mas no entanto existe um limite de utilização de 50.000 *queries* por dia. Para necessidades acima destes limites, existe também uma versão paga, o *CalaisProfessional*, que fornece um serviço equivalente mas com limites cinco vezes superiores aos da versão gratuita.

2.4.2 Zemanta

O Zemanta[22] [25] é um serviço que inicialmente era usado para enriquecer artigos em *blogs* com diversos conteúdos, imagens, ligações e *tags* através de uma análise semântica destes.

[20] Mais informações em: `http://www.opencalais.com/`

[21] Proveniente do artigo da versão inglesa da Wikipedia presente em `http://goo.gl/G1gTF`

[22] Mais informações em: `http://www.zemanta.com/`

Tabela 2.9: Exemplo dos Resultados do OpenCalais

Topics	Politics	
Social Tags	Politics, Portuguese, People, ...	
Entities	**Country**	Portugal (0.2)
	Organization	European Commission (0.33)
	Person	José Manuel Durão Barroso (0.81)
	Position	Politician (0.4)
		President (0.33)
		Prime Minister (0.23)
Events & Facts	**Generic Relations**	José Manuel Durão Barroso, President of the European Commission
	Person Carreer	**Past:** José Manuel Durão Barroso, Prime Minister, Portugal
		Current: José Manuel Durão Barroso, President, European Commission

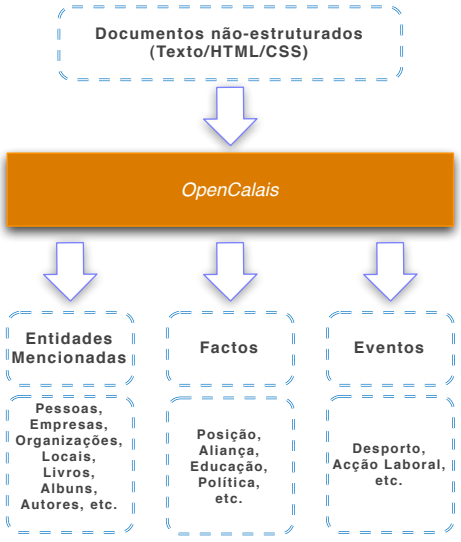

Figura 2.1: *Workflow* do OpenCalais
(Adaptado de [24])

Actualmente disponibiliza uma API que recebendo texto estruturado, devolve conteúdos relacionados com este, podendo estes ser entidades mencionadas, artigos de outros utilizadores do serviço, categorias, imagens e *tags*. Para isto, trabalha com conteúdos pré-indexados de diversas fontes de conteúdos, como a *Wikipedia*, *Twitter*, *Flickr*, entre outros. Os dados recebidos são processados com recurso a ferramentas de NLP e algoritmos semânticos, comparando este conteúdo com os dados pré-indexados.

À semelhança da API do *OpenCalais*, esta também é gratuita para uma utilização básica (10.000 *queries* por dia), sendo possível o pagamento de uma taxa por utilizações mais intensivas.

2.4.3 AlchemyAPI

AlchemyAPI[23] é uma ferramenta que engloba um vasto leque de funcionalidades sobre o conteúdo fornecido, como *sentiment analysis*, extracção de *tags*, detecção de linguagem, extracção de entidades, categorização de texto, *scraping* de páginas *web*, entre outros. A maioria destas funcionalidades funciona com recurso a métodos estatísticos de NLP e algoritmos de *machine learning*. Estas funcionalidades estão disponíveis por intermédio de uma API, mediante o envio de conteúdos textuais ou apenas um *URL* (*Uniform Resource Locator*). O acesso a esta API é gratuito para uma utilização básica, até 30.000 *queries* diárias. Existem diversos planos pagos dependendo da utilização que se quer dar a este serviço.

2.4.4 Yahoo Content Analysis API

A *Yahoo Content Analysis*[24] é uma ferramenta capaz de extrair entidades, conceitos, categorias e relações em texto não-estruturado. Para além de associar meta-dados às *tags* extraídas, ainda as ordena pelo nível de relevância que apresentam no conteúdo fornecido, podendo ainda interligá-las com páginas da *Wikipedia*, se estiverem disponíveis. Este serviço encontra-se disponível através de uma API, tendo a sua utilização gratuita limitada a 10.000 *queries* diárias.

[23] Mais informações em: `http://www.alchemyapi.com/`
[24] Mais informações em: `http://developer.yahoo.com/contentanalysis/`

2.4.5 Análise Comparativa

De modo a avaliar as ferramentas descritas nesta secção, foram escolhidas métricas descritivas simples que esclarecem o que cada destas oferece. Estas métricas encontram-se esquematizadas relativamente a cada fonte na tabela 2.10. A métrica "Conteúdos fornecidos"explicita o que cada uma das ferramentas oferece ao utilizador, para além do processo de etiquetagem, nos "Limites diários da API"está descrito o limite máximo de utilização gratuita de cada ferramenta e nos "Idiomas suportados"é descrito o conjunto de línguas (ou língua) que a ferramenta consegue processar.

Tabela 2.10: Comparação entre as diferentes ferramentas de Etiquetagem Automática

Ferramenta	**Conteúdos fornecidos**	**Limites Diários da API**	**Idiomas Suportados**
OpenCalais	Categorias, entidades, *tags sociais*, factos e eventos	50.000 pedidos (Versão Gratuita)	Inglês, Francês, Espanhol
Zemanta	*Tags*, *Links*, artigos relacionados, imagens	10.000 pedidos (Versão Gratuita)	Inglês (No entanto algumas funcionalidades são suportadas noutras línguas)
AlchemyAPI	Idioma, *tags*, entidades, categorias	30.000 pedidos (Versão Gratuita)	Inglês, Francês, Alemão, Italiano, Português, Russo, Espanhol, Sueco
Yahoo! Content Analysis	Categorias, *tags*, entidades, relações, imagens	10.000 pedidos (Versão Gratuita)	Inglês, Chinês

Dado que os objectivos deste trabalho envolvem processamento de linguagem natural para o português, a ferramenta *AlchemyAPI* foi a que nos despertou mais curiosidade, tendo em conta o elevado número de funcionalidades que apresenta e a vasta gama de idiomas suportados. No entanto, ignorando o facto de não suportarem a língua portuguesa, o resto

das ferramentas estudadas também mostram bons resultados, sendo de sublinhar a rapidez com que nos apresentam resultados relacionados com os textos ou *links* fornecidos, o que demonstra que a pré-indexação de conteúdos (utilizada na maioria destes serviços) é uma mais-valia neste tipo de tarefa.

O factor de diferenciação que este trabalho apresenta relativamente às ferramentas acima enunciadas passa pelo facto de poder etiquetar uma maior gama de recursos e não apenas texto e *links*. Foi decidido não recorrer a estes serviços de modo a criar um serviço independente, havendo assim uma melhor tolerância a falhas visto que não estamos dependentes de serviços externos.

2.5 Enriquecimento Semântico

Nesta secção, apresentamos algumas ferramentas já existentes na área do enriquecimento semântico de lugares. O facto de ser uma área bastante específica faz com que o conjunto de abordagens aqui apresentadas seja bastante reduzido.

2.5.1 Kusco

O KUSCO(*Knowledge Unsupervised Search for populating Concepts on Ontologies*)[3] é uma ferramenta desenvolvida por Ana Alves, uma estudante de Doutoramento da Universidade de Coimbra e co-orientadora deste estágio, que permite associar um conjunto de conceitos extraídos de textos em inglês presentes na *web* a lugares ou eventos, de modo a poder enriquecer estes com informação semântica.

Dado que os dois primeiros módulos representados são considerados os mais relevantes para o processo de enriquecimento semântico, descrevemos estes em maior detalhe. O fluxo do processo pode ser descrito da seguinte forma:

- É fornecida ao sistema uma fonte de POIs (um conjunto de documentos, um directório comercial ou uma API) que é mapeada no modelo de dados conceptual de modo a popular automaticamente uma base de dados de POIs. Esta extracção intensiva, apelidada de *POI Mining* é feita através de *screen scraping* ou por intermédio de uma API, caso esteja disponível.

- O módulo de "Extracção de Informação numa perspectiva"efectua uma pesquisa de documentos sobre cada POI nas fontes de informação disponíveis. De modo a obter

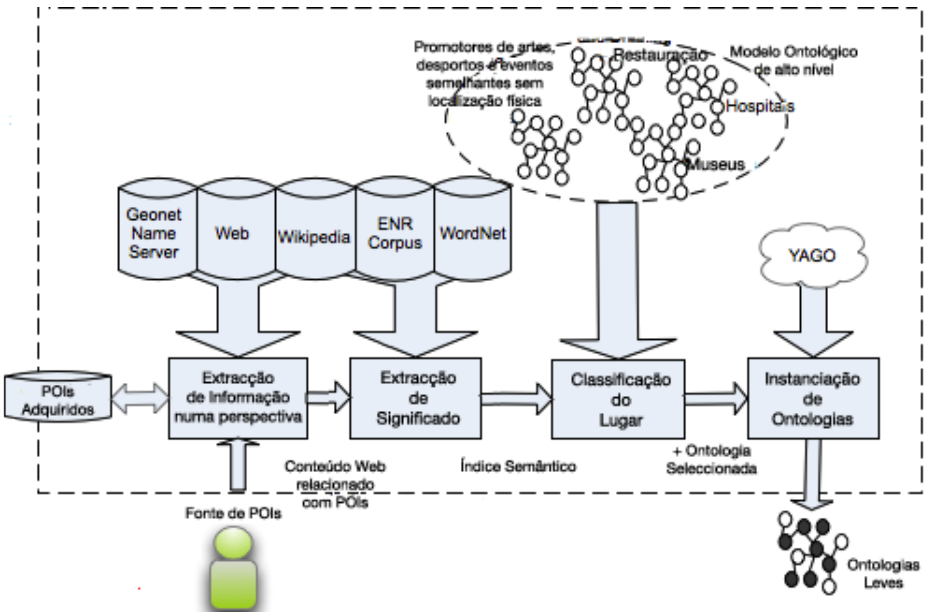

Figura 2.2: Modelo de Enriquecimento Semântico de Lugares do Kusco (Adaptado de [3])

estes documentos, recorre-se a uma pesquisa *web* e à *Wikipedia*, aplicando duas abordagens diferentes a cada uma destas pesquisas, intituladas "perspectivas": utilizando a *web* através de uma API de pesquisa, existe uma perspectiva específica direccionada ao *website* do POI, e uma perspectiva mais alargada abrangendo documentos de diversos *sites*; utilizando a *Wikipedia* e a sua API, existe também uma perspectiva mais específica, *Yellow Wiki*, que vai à procura da página de um POI, enquanto que na perspectiva *Red Wiki*, mais genérica, são utilizadas as páginas das categorias do POI. Estas categorias são geralmente fornecidas pelo directório de onde o POI é extraído.

- O módulo "Extracção de Significado"extrai conceitos relevantes dos documentos obtidos na fase anterior, contendo estes conceitos um significado associado, sendo a desambiguação de cada termo feita neste módulo. Deste módulo resulta o índice semântico, através do qual cada termo extraído é pesado recorrendo a métricas de relevância estatísticas.

2.5.2 Topica

Topica[25] [26] é uma aplicação *web* que através de dados publicados em redes sociais permite modelar características dinâmicas de POIs, de modo a fornecer ao utilizador informações sobre uma determinada área, baseando-se nas entidades nela existentes. Através da representação básica de um POI, é possível obter mais informação deste, utilizando para este meio as interacções que diversos utilizadores fazem com este ao longo da sua actividade nas redes sociais. A abordagem divide-se em duas etapas: indexação de dados na ontologia (onde é feito o processo de enriquecimento semântico) e representação de POIs.

Figura 2.3: Arquitectura do Topica
(Adaptado de [26])

A indexação na ontologia é feita da seguinte forma: para uma determinada área (um conjunto de coordenadas delimitando uma região), são extraídos os locais do *Facebook* (*Facebook Places*). Utilizando as propriedades fornecidas pela API da rede social (nome, morada, descrição) é feita uma interligação com a página do POI (tipicamente uma página institucional, também conhecida como *fanpage*) de modo a extrair os comentários aqui presentes. Estes comentários são enriquecidos com recurso a diversos serviços, como o *OpenCalais*, *Zemanta* e *DBPedia Spotlight*. Através destes serviços, são extraídas entidades, *tags* e páginas relacionadas. Estes dados serão filtrados de modo a gerar uma lista de recursos a enviar para a *DBPedia*, que para cada recurso irá assinalar categorias que permitem enriquecer cada recurso. Por exemplo, se um recurso chamado "Sushi"esteja presente na lista, estas categorias poderiam incluir: "Cozinha Japonesa"ou até mesmo categorias relacionadas, como

[25]Mais informações em: `http://nebula.dcs.shef.ac.uk/sparks/topica/`

"Cozinha Oriental". As categorias obtidas através dos comentários são submetidas a um cálculo de relevância, sendo o seu peso obtido pelo cálculo do TF-IDF(*Term Frequency* x *Inverse Document Frequency*)[26]. Após este cálculo de relevância, os POIs são devidamente estruturados numa ontologia, através do formato RDF(*Resource Description Framework*), permitindo o acesso a estes através de *queries* SPARQL(*SPARQL Protocol and RDF Query Language*).

A etapa de representação dos POIs é feita com auxílio das APIs do *Open Street Maps* e do *Google Maps*. O utilizador navega pelo mapa visualizando os POIs de forma agregada, consoante o nível de *zoom* actual. Ao interagir com um POI é apresentada a descrição de alto nível deste (nome, morada, *tags* e conteúdos de mensagens presentes em redes sociais, como críticas do local). Esta visualização permite ainda uma filtragem de POIs baseada em conceitos presentes numa lista que varia consoante o local onde o utilizador se encontra no mapa, de modo a ser possível o foco do utilizador naquilo que realmente lhe interessa.

2.5.3 Abordagens baseadas em actividades

Recentemente, alguns autores apresentaram abordagens relacionadas com o que é feito neste trabalho, mais especificamente a utilização de fontes de dados de cariz colaborativo para a classificação de um espaço de acordo com os serviços que o compõem.

Dearman e Truong[27] utilizam os conteúdos do *Yelp*[27] (uma rede social que combina um directório de POIs com críticas fornecidas pela sua comunidade) de modo a identificar o tipo de actividades efectuadas num determinado local. Aplicando técnicas de extracção de informação (como divisão do texto em blocos e etiquetagem morfológica) aos conteúdos textuais inseridos pelos utilizadores, os autores conseguem extrair pares de palavras constituídos por um verbo e um nome de modo a representar actividades possíveis que podem ocorrer num determinado local, como "*eat pizza*"ou "*order latte*". Foi também feito por estes autores um trabalho de validação de modo a provar que é possível obter conhecimento importante a partir de dados gerados por uma comunidade.

A.N. Alazzawi et al.[28] propõem uma metodologia que permite a identificação de conceitos representativos de serviços oferecidos num determinado local e as actividades que podem ocorrer neste. Através da detecção de padrões linguísticos relacionados com locais

[26]Enquanto o TF mede a frequência de uma palavra no documento actual, o IDF identifica a frequência da palavra num universo de documentos, tendo palavras comuns um valor de IDF baixo. Esta métrica é descrita detalhadamente na secção 3.2

[27]Mais informações em `http://www.yelp.com/`

em textos etiquetados morfologicamente, os autores aplicam estes padrões a recursos *online* de modo a extrair conceitos relacionados com serviços e actividades. Desta maneira, é possível contextualizar locais dando relevância às actividades e serviços que ali têm lugar (Ex.: "*Church*"obtém conceitos associados como "*worship*", "*ChristianWorship*"ou "*HoldChristianServices*").

2.5.4 Análise Comparativa

As abordagens aqui apresentadas apresentam metodologias algo semelhantes, mas enquanto que no Kusco podemos utilizar diversas fontes de dados no processo de enriquecimento semântico, o *Topica* é algo limitado neste aspecto. O Kusco lida com as tarefas de extracção de informação de uma forma controlada, utilizando para isto ferramentas que não dependem de recursos de terceiros. Já o *Topica* recorre a serviços como o *OpenCalais* e o *Zemanta*, estando assim dependente da disponibilidade destes, o que a certa altura pode ser negado tendo em conta os limites de utilização impostos por estes serviços.

Durante este trabalho não nos foi possível testar intensivamente o *Topica*, devido ao facto de o código não ser aberto. Na verdade, a única demonstração disponível desta aplicação era bastante restrita, contendo apenas dados numa janela temporal bastante reduzida e para uma pequena área geográfica. Desta maneira concluímos que o Kusco é uma abordagem mais madura, tendo já bastante trabalho de investigação envolvido na área do enriquecimento semântico.

No que toca às abordagens baseadas em actividades, estas diferem da abordagem que se pretende seguir na medida em que estas se focam nas acções à volta dos conceitos, enquanto que neste caso é dado maior ênfase aos conceitos, visto que as actividades podem ser induzidas através da utilização de fontes de conhecimento externas, como o PAPEL ou mesmo a *Wikipedia*. Ex.: Uma *pizza* é um prato que pode ser encomendado, servido e provado (pela abordagem de *Dearman e Turong*), enquanto que pela nossa abordagem o que nos interessa é a *pizza* simplesmente.

A abordagem levada a cabo neste trabalho assenta sobre o trabalho desenvolvido no Kusco, adaptando a metodologia deste à língua portuguesa (dado que o KUSCO apenas se encontra disponível para a língua inglesa) sob a forma de uma plataforma mais robusta e dinâmica, suportando a integração de novas fontes de informação e garantindo a interoperabilidade deste sistema com serviços externos.

Capítulo 3

Abordagem

3.1 Especificação Funcional

Esta secção pretende familiarizar o leitor com os requisitos do módulo de enriquecimento semântico, através de uma breve descrição destes e da apresentação da arquitectura proposta para a sua implementação.

3.1.1 Análise de Requisitos

As primeiras reuniões entre os vários elementos do SEMA consistiram no levantamento de casos de uso para a plataforma, através da criação e discussão de *user stories*. Os casos de uso distinguem-se pela sua modularidade, estando diferentes casos de uso à responsabilidade de módulos específicos do SEMA. O módulo de enriquecimento semântico está responsável por dois dos casos de uso em questão:

- **Pedir Etiquetagem de recursos**
- **Pedir Extracção de novo conhecimento sobre um recurso**

O primeiro caso de uso permite ao actor **pedir etiquetagem de recursos** (podendo estes ser POIs, eventos ou notícias), processo no qual o módulo obtém um conjunto de *tags* relevantes associadas ao recurso em questão. Por exemplo, para um POI denominado "Exploratório"em Coimbra será possível associar conceitos como "ciência civa", "museu", "associação"e "aprendizagem".

Nesta fase são recolhidos conceitos-chave dos elementos textuais existentes sobre um determinado recurso através de técnicas de extracção de informação, podendo assim obter

conceitos (*tags*) associados ao recurso em questão. Após obter estes conceitos, é calculada a respectiva coerência entre as diferentes fontes de enriquecimento e a relevância, de modo a poder devolver uma lista de conceitos ordenada por ordem de relevância (à semelhança de uma *tag cloud*).

O segundo caso de uso permite ao actor **pedir a extracção de novo conhecimento para um recurso** já existente na base de conhecimento mas que não se encontre devidamente categorizado/descrito (ex.: o menu no caso de um restaurante ou existência de zona de fumadores num bar). Este recurso será enriquecido semanticamente, sendo possível obter novos dados a cada ocorrência deste caso de uso (caso estejam disponíveis nas fontes de enriquecimento registadas), através de técnicas de similaridade espacial, lexical ou temporal. Estes dados podem ser características-base do recurso, como no caso de um POI o horário de funcionamento e serviços fornecidos ou até testemunhos ou críticas sobre este. No final deste processo, o recurso é actualizado na base de conhecimento e devolvido.

Na primeira iteração do SEMA, foi definido que inicialmente o módulo de enriquecimento semântico deveria focar-se primariamente no enriquecimento de POIs, dado o interesse específico dos restantes parceiros neste tipo de recursos. Este trabalho apresenta assim maior foco no enriquecimento de POIs, sendo as experiências efectuadas com recurso essencialmente a este tipo de dados. No entanto, a definição da arquitectura e as metodologias propostas têm em vista o suporte a outro tipo de recursos, tais como eventos ou até mesmo recursos não georeferenciados (Ex.: notícias).

Com base no levantamento de casos de uso foram delineados os requisitos funcionais e não-funcionais do módulo de enriquecimento semântico, estando estes representados nas tabelas 3.1 e 3.2.

Tabela 3.1: Requisitos Funcionais

Requisitos Funcionais	
Número	**Descrição**
1	Pedir etiquetagem de recursos
2	Pedir extracção de novo conhecimento sobre um recurso

Tabela 3.2: Requisitos Não-Funcionais

Requisitos Não-Funcionais		
Número	**Categoria**	**Descrição**
1	Modularidade	Receber os recursos de forma genérica, podendo ser apresentados em listas mistas (POIs, eventos e afins), normalmente correspondendo a uma rota.
2	Desempenho	A aplicação deve ter em conta o tempo médio de execução de cada pedido que é feito, de modo a avaliar se é exequível responder ao pedido em tempo real ou despoletar uma tarefa em *background* e devolver uma resposta provisória. Deste modo, o desempenho de outros serviços não será afectado pelo tempo de espera necessário para tarefas com grande complexidade.
3	Coerência de Dados	Os dados obtidos nas fontes devem ser sujeitos a uma validação específica, de modo a evitar falsos positivos ou redundância de dados.
4	Suporte Multilinguístico	A plataforma deve ser capaz de lidar com conteúdos em Português e Inglês, dado que alguns dos PPS's verticais podem necessitar de conteúdos em múltiplas línguas (Ex.: para aplicações na área do turismo, de modo a que utilizadores internacionais possam usufruir do serviço fornecido).

3.1.2 Arquitectura do PPS 2 - SEMA

Como mencionado em 1.1, o SEMA é constituído por diversos módulos, estando estes acessíveis entre si e para outros PPS's através de *webservices* REST(*Representational State Transfer*).

Em traços gerais, o módulo de ES permite enriquecer semanticamente e etiquetar automaticamente os recursos presentes numa base de conhecimento, o módulo de ASA permite seleccionar estes recursos com base em conceitos de relevância, surpresa e diversidade e finalmente o módulo de IS é o responsável por todas as interacções com a base de conhecimento. Esta consiste num servidor de Ontologias com base em triplos que contém a Ontologia onde são mapeados todos os dados do SEMA. Deste modo a gestão de dados é feita de forma transparente para os outros módulos, visto que as operações de leitura, inserção, remoção,

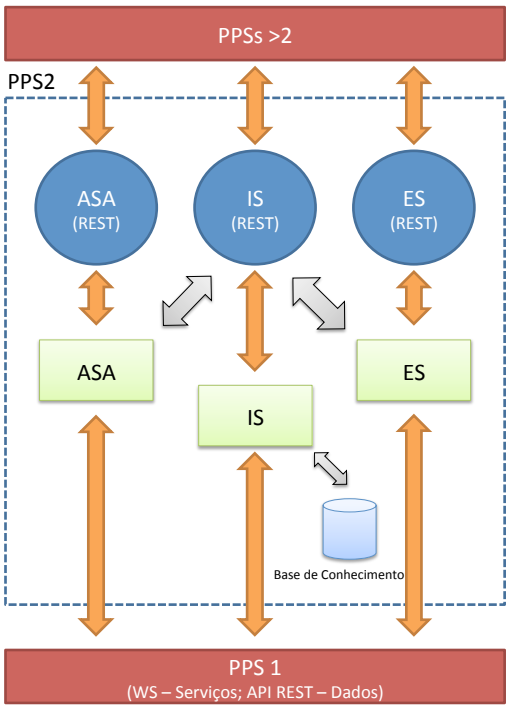

Figura 3.1: Arquitectura do PPS2 - SEMA

actualização e pesquisa semântica serão asseguradas pelo *webservice* de interoperabilidade semântica. Enquanto este serviço não estiver disponível, o armazenamento de dados provisório será feito com recurso a uma base de dados PostGIS[1], através de uma camada de acesso a dados, tornando a futura migração mais simples.

Como podemos observar na figura 3.1, o *web service* do módulo de ES pode ser acedido através do PPS 1 ou de qualquer outro PPS. A nível interno, o módulo comunica directamente com o serviço de interoperabilidade semântica, de modo a poder gerir dados presentes na base de conhecimento.

Em termos de *workflow*, as acções do módulo de enriquecimento semântico podem ser despoletadas por qualquer outro módulo/PPS que necessite de operações de enriquecimento ou etiquetagem de um recurso presente na base de conhecimento. Assim, o recurso será actualizado com nova informação e posteriormente armazenado/actualizado nesta, recorrendo à interface do serviço de IS.

Através desta arquitectura, os serviços fornecidos pelos módulos do PPS2 podem receber recursos do PPS1 (como dados de mobilidade, entre outros), ficando estes armazenados de

[1]Mais informações em: `http://postgis.refractions.net/`

acordo com a ontologia especificada na base de conhecimento. Deste modo é garantida a interoperabilidade de informação com os PPS's superiores, que assim podem utilizar a informação semântica gerada pelo PPS2 de forma a assegurar os serviços fornecidos ao utilizador final.

3.1.3 Arquitectura do Módulo de Enriquecimento Semântico

O módulo de enriquecimento semântico deve ser capaz de receber dados georeferenciados como POIs, eventos associados a POIs, conjuntos de POIs e eventos ou até recursos não-georeferenciados, como notícias. Com estes dados, acedemos às fontes de enriquecimento semântico previamente estudadas (2.2) de forma a recolher mais informação sobre os dados fornecidos. A comunicação com o *web service* é feita através de uma interface RESTful, sob o formato JSON. Esta interface poderá ser acedida mediante a execução de pedidos nos diversos verbos do protocolo HTTP (*Hypertext Transfer Protocol*), consoante o tipo de operação a efectuar.
Na figura 3.2 está representada a arquitectura do módulo de enriquecimento semântico, sendo esta dividida em três camadas distintas: Camada de interligação com os outros módulos, onde está disponível a API que será utilizada por terceiros; Camada lógica, que engloba a lógica do processo de enriquecimento semântico, como os extractores de informação para cada uma das fontes, a contextualização e integração de recursos e ainda as ferramentas de extracção de informação que permitem, entre outros, o reconhecimento de entidades mencionadas. Por fim, a camada de acesso a dados, onde são feitas todas as operações relacionadas com a base de conhecimento utilizada.

3.1.4 Especificação dos Dados a extrair das Fontes de Enriquecimento

Para conseguir uma especificação coerente das estruturas a implementar no serviço, procedeu-se a uma análise dos dados que podemos extrair de cada uma das fontes de enriquecimento previamente estudadas. Nesta subsecção iremos descrever os dados mais importantes que cada fonte de enriquecimento nos pode fornecer, representados na tabela 3.3. A partir desta podemos descrever os dados provenientes de cada uma das fontes, visto que os dados descritos podem nem sempre estar presentes ou estarem sob uma formatação mais específica.

Algumas das fontes de enriquecimento são redes sociais, pelo que é importante ter em conta que grande parte do conteúdo é gerado pelos utilizadores, sendo possível que os

Figura 3.2: Arquitectura do Módulo de Enriquecimento Semântico

dados referidos não estejam sempre presentes em alguns recursos (Ex: Um utilizador criou um ponto de interesse, mas dado que não sabia a morada exacta, deixou este campo por preencher). Tendo em conta que pode existir alguma parcialidade nestes conteúdos, torna-se necessária uma comparação dos dados extraídos nas redes sociais com dados provenientes de fontes mais confiáveis, como é o caso dos directórios comerciais ou páginas institucionais.

Tendo como base esta listagem de dados passíveis de obter com recurso às fontes de enriquecimento, descrevemos alguns dos menos evidentes de modo a explicar como se podem revelar úteis no processo de enriquecimento semântico.

- **Categorias**: É extremamente importante saber como categorizar um recurso. Foram analisadas as taxonomias das diferentes fontes de modo a verificar como se distribuem as categorias nas diferentes fontes de enriquecimento, tendo em conta o nível de profundidade e especificidade de cada categoria. Após uma análise das diversas taxonomias, a do *Foursquare* foi a que revelou maior detalhe. Apresenta uma profundidade máxima de 4 níveis (categorias e subcategorias) e os nomes das categorias podem ser obtidos em diversas línguas através da sua API.

- **Comentários/Sugestões**: São dados textuais que acompanham diversos recursos, constituindo na sua maioria críticas a um recurso que podem ser utilizadas para obter dados importantes sobre este. No entanto, é necessária atenção redobrada ao lidar com estes conteúdos: muitas vezes podem não passar de *spam*.

Tabela 3.3: Dados disponíveis nas diversas fontes de enriquecimento

Atributos	***Foursquare***	***Gowalla***	***Lifecooler***	***Factual***	***Google Places***	***Facebook***	***SAPO***
Nome	v	v	v	v	v	v	v
Coordenadas	v	v	v	v	v	v	v
Morada	v	v	v	v	v	v	v
Website	v	v	v	v	v	v	v
Telefone	v	v	v	v	v	v	v
Horários	-	-	v	-	-	v	-
Categorias	v	v	v	v	v	v	v
Email	v	v	v	v	v	v	v
Preços	-	-	v	-	-	-	-
Acessos	-	-	v	-	-	v	-
Conteúdos Multimédia	v	v	v	-	-	v	-
Comentários / Sugestões	v	v	-	-	-	v	-
Tags	v	v	-	-	-	-	-
Locais Relacionados	v	v	-	-	-	-	-
Estatísticas	v	v	-	-	v	v	-
Referências para outras fontes	-	v	-	-	-	v	v

- **Estatísticas**: Com este atributo é-nos possível inferir a popularidade de um dado recurso ou até mesmo verificar se as informações apresentadas são confiáveis. Pode surgir sob várias formas: no caso das redes sociais baseadas em localização, cada POI apresenta um *check-in count* e um *user count* — a conjunção destes dois valores discriminam quantos utilizadores diferentes já assinalaram a sua presença no POI; no caso do *Facebook*, é-nos possível obter o número de *check-in's* que já foram feitos num determinado POI, quantos *likes* (métrica utilizada pelo *Facebook* que indica quantos utilizadores seguem as actualizações de um recurso) e quantos utilizadores têm mencionado este POI na sua actividade recente.

- **Locais Relacionados**: Esta característica é apresentada por algumas das fontes, podendo ir desde uma lista de recursos relacionados até uma espécie de roteiro turístico, onde são listados recursos sobre uma temática específica (Ex: "Roteiro das Francesi-

nhas do Porto").

- **Referências para outras fontes**: Existem fontes que referenciam outras, de modo a evitar replicação de informação. Esta característica revela-se extremamente importante do ponto de vista de integração de recursos provenientes de diferentes fontes.

Esta especificação compreendeu também a criação de extractores de dados para cada uma das fontes, um conjunto de bibliotecas que através da utilização de *screen-scraping* ou API's nos ajudam a obter estes dados de cada uma das fontes quando tal é permitido pelos seus termos de utilização (de acordo com os contactos estabelecidos e descritos em 2.2.2).

Tendo em conta estas características, foi criado o modelo de dados do serviço, presente no anexo B.

3.2 Metodologia

Na figura 3.3 está representado o modelo de enriquecimento semântico. Este baseia-se na metodologia utilizada pelo Kusco (2.5.1), destacando-se pela inclusão de novas fontes de dados e um diverso leque de ferramentas que são utilizadas conforme as características dos dados a tratar (língua e tipo de recurso, principalmente). Podemos dividir o processo em cinco fases distintas, correspondendo as duas primeiras fases (Pesquisa em fontes de informação, integração e detecção de duplicados) ao primeiro caso de uso, "Pedir extracção de novo conhecimento", enquanto as que se seguem abordam essencialmente o caso de uso "Pedir etiquetagem de recursos". Ambos os casos de uso podem ser despoletados mediante um pedido à API disponibilizada, contendo uma referência ao recurso sobre o qual se pretende efectuar a operação desejada.

3.2.1 Pesquisa em fontes de informação

Nesta primeira fase acedemos às fontes de informação disponíveis em busca de novos conteúdos sobre um determinado POI. Primariamente é efectuada uma pesquisa específica baseada nos dados geográficos e textuais presentes no POI e caso não sejam encontrados resultados é feita uma pesquisa genérica (considerando elementos como o nome e a categoria) à *Wikipedia*. Esta pesquisa é feita através dos extractores de informação criados para cada fonte, já mencionados em 3.1.3. Cada um dos extractores de informação utilizados conta com um valor de confiabilidade (descrito em pontos percentuais entre 0 e 1), de modo a

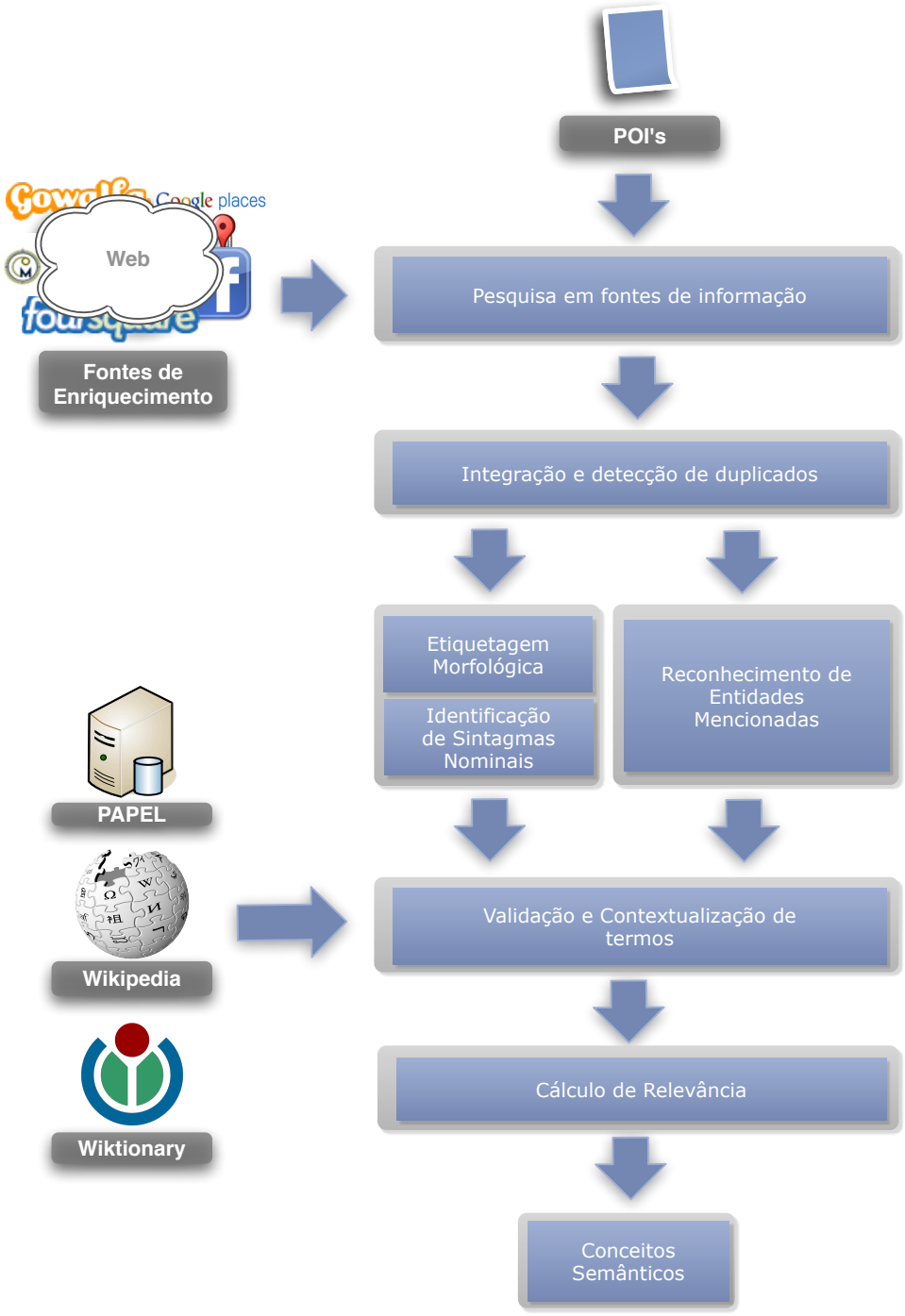

Figura 3.3: Modelo de Enriquecimento Semântico

classificar o rigor dos dados que extrai. Este valor depende do tipo de fonte associada ao extractor, sendo necessário ter em conta se esta é devidamente moderada e qual a veracidade dos seus conteúdos. Embora este valor seja inicialmente fixo, varia consoante os dados atribuídos são substituídos ou inseridos num POI. Dado que a qualidade dos conteúdos presentes numa fonte pode variar com o tempo, o cariz dinâmico deste valor adequa-se à tarefa. Esta operação terá como resultado uma lista de representações de POIs, que podem ou não corresponder ao POI sobre a qual a operação foi efectuada.

3.2.2 Integração e detecção de duplicados

Os dados resultantes da fase anterior são analisados de modo a verificar se correspondem ao POI original. Dado que algumas das fontes de onde extraímos este tipo de recursos se baseiam em contribuições dos utilizadores (tanto no caso das inserções como das actualizações), corremos o risco de existirem variações entre o nome, *website* associado e falta de precisão na informação geográfica ou temporal entre os recursos que temos presentes e os das fontes, o que torna o processo de comparação mais complexo.

De modo a lidar com estas inconsistências, foram definidos diversos factores, focando essencialmente a proximidade geográfica entre recursos e a similaridade entre nomes e *websites*[2]. Assim, consideramos que dois POIs são iguais caso se enquadrem nos seguintes grupos:

- A distância entre ambos os POIs é menor do que 100 metros, a similaridade entre nomes é maior que 90% e nenhum dos POIs contém *websites* associados;
- A distância entre ambos os POIs é menor do que 100 metros, a similaridade entre nomes é maior que 70% e o maior valor de similaridade entre os *websites* associados[3] é maior que 50%.

Estes valores de limiar resultaram de um processo iterativo, começando por valores menores e analisando os POIs semelhantes que eram obtidos no processo de modo a verificar se os resultados eram aceitáveis. Dado que o primeiro grupo apenas se baseia em dois atributos, foi necessário utilizar um valor alto para o limiar de similaridade entre nomes, de modo a evitar falsos positivos.

[2] A similaridade entre recursos textuais é calculada com recurso à ferramenta mencionada em 2.3.

[3] Como um POI pode ter múltiplos *websites*, considera-se o conjunto com maior similaridade.

Caso algum dos POIs analisados encaixe neste grupo, as informações deste serão utilizadas para acrescentar conteúdo ao POI original, tendo esta operação em conta o grau de confiabilidade da fonte que originou o POI candidato à integração.

Durante este processo de integração, torna-se necessário efectuar algum tratamento aos conteúdos fornecidos pelos extractores. Os blocos de texto de grandes dimensões presentes em atributos como as críticas ou descrições muitas vezes contêm ruído que pode consistir em *tags* HTML dentro dos textos e termos comuns de páginas *web* que não trazem qualquer mais-valia para o processo. Antes de ser feita qualquer persistência deste tipo de dados, verifica-se se existe ruído deste género e caso exista, é removido dos conteúdos a armazenar.

3.2.3 Extracção de Termos

Os conteúdos textuais de maior dimensão presentes no recurso a enriquecer (como descrições ou críticas) são submetidos ao processo de extracção de termos, de modo a extrair palavras-chave destes textos.

Este processo consiste na execução em paralelo de duas das sub-tarefas de extracção de informação já mencionadas em 2.1.1: identificação de sintagmas nominais e reconhecimento de entidades mencionadas. Para executar a primeira é necessário dividir o texto fornecido em blocos unitários (como palavras e sinais de pontuação), etiquetando estes blocos morfologicamente através da utilização de ferramentas de etiquetagem morfológica, sendo assim possível fornecer estes conteúdos ao identificador de sintagmas nominais. Paralelamente a este processo, termos que representam entidades como pessoas, locais e organizações são extraídos do mesmo texto através do reconhecimento de entidades mencionadas. As entidades identificadas como locais são aqui descartadas: dado que já existe informação geográfica sobre o recurso em questão, as entidades desta categoria tornam-se irrelevantes para a abordagem em questão.

A execução em paralelo destas duas tarefas produz duas listas: uma de sintagmas nominais e outra de entidades mencionadas, podendo estas corresponder a termos simples ou compostos (formados por mais do que uma palavra). Estas duas listas são unificadas numa só, e no caso de um termo ou parte dele figurar nas duas listas dá-se maior relevância ao que está presente na lista de entidades mencionadas.

3.2.4 Validação e contextualização de termos

De modo a verificar se os termos resultantes da fase anterior são válidos, é necessário criar um conjunto de regras que define se um termo é válido. Para tal, começamos por definir o que é para nós um termo: Um conjunto de palavras que descreve claramente características importantes de um local: um tipo de restaurante, um prato típico ou até pequenas descrições objectivas como "boa música".

Para determinar se um termo é válido, há que começar pela sua forma unitária: a palavra. Apresentamos assim as heurísticas que nos permitem inferir se uma palavra é ou não válida:

- Não deve constar na lista de *stopwords* do idioma do texto[4];
- Deve conter pelo menos uma vogal;
- Não pode conter uma letra maiúscula após uma letra minúscula (Ex.: cAsa);
- Não deve conter números, sinais de pontuação (à excepção do hífen em palavras justapostas) ou outros símbolos como parênteses e barras;
- Deve conter pelo menos três letras.

Estas heurísticas são aplicadas essencialmente a termos simples (compostos por uma única palavra), sendo estes termos descartados caso não as cumpram. Para os termos compostos a situação é diferente, visto que existem situações em que palavras constituintes do termo não cumprem estes pressupostos (Ex.: o "de"que consitui o termo "Casa de Fados").

Neste caso, são descartados termos que contenham informação coincidente com o nome do POI ou informações relativas à sua morada (rua, freguesia, concelho, distrito, país), termos cuja primeira ou a última palavra não sejam consideradas válidas (pelas heurísticas definidas acima) e termos que contenham mais do que três palavras. Esta última heurística tem por base a análise de todos os títulos das versões portuguesa e inglesa da *Wikipedia* (contendo a primeira mais de um milhão e a segunda mais de nove milhões de títulos), para contabilizar a frequência do número de palavras por título, inferindo assim o número máximo de palavras por termo. Como podemos observar pelos histogramas presentes nas figuras 3.4 e 3.5, os títulos com mais de três palavras apresentam uma frequência bastante inferior às

[4]Presentes em: `http://snowball.tartarus.org/algorithms/`

restantes. Na língua Portuguesa, por exemplo, os termos com mais de três palavras têm um alto nível de especificidade, como o nome de uma pessoa ou de um local (Ex.: "Diana, Princesa de Gales", "Aeroporto Francisco Sá Carneiro", "Bom Jesus de Braga"), algo que não traz valor para descrever um local.

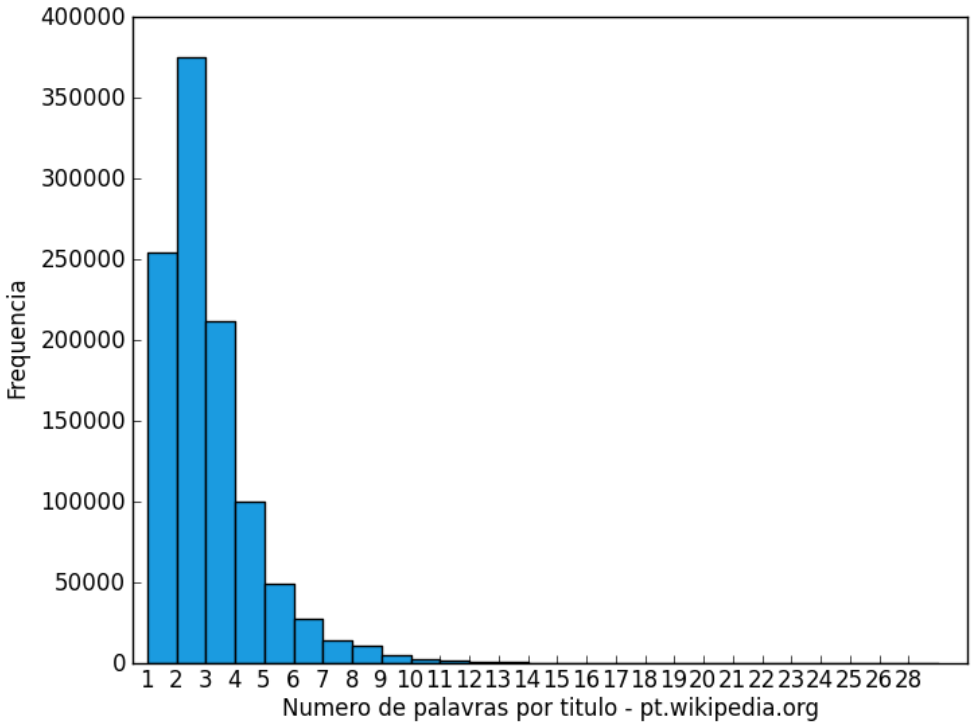

Figura 3.4: Número de palavras por título na versão Portuguesa da *Wikipedia*

Após a filtragem dos termos inválidos da lista produzida em 3.2.3, é necessário contextualizar os termos restantes, inferindo se estes existem no léxico do idioma do texto de onde foram extraídos.

Para este processo de contextualização utilizam-se os seguintes recursos: o PAPEL (para a língua Portuguesa, sendo este constituído essencialmente por termos simples), a *Wikipedia* (para as duas línguas, utilizando as respectivas versões e essencialmente para termos compostos) e ainda o *Wiktionary*[5], um projecto colaborativo semelhante à *Wikipedia* cuja missão passa por produzir um dicionário poliglota livre, fornecendo significados e etimolo-

[5]Mais informações em `http://www.wiktionary.org/`

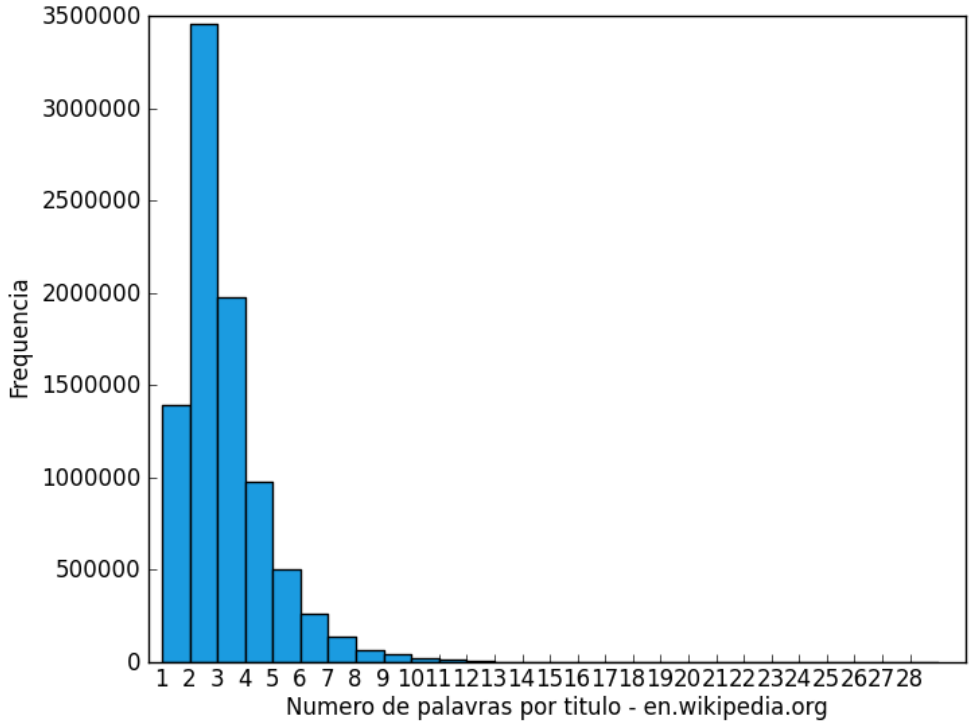

Figura 3.5: Número de palavras por título na versão Inglesa da *Wikipedia*

gias essencialmente para termos simples.

O algoritmo 1 descreve detalhadamente como é feito o processo de contextualização de um termo. É aqui feita uma pequena descrição das funções externas que este envolve, de modo a contextualizar o leitor com o pseudo-código apresentado: a função *pesquisaPAPEL*, tem um cariz binário, verificando se o termo está contido no dicionário e devolvendo uma resposta conforme o resultado desta pesquisa. O mesmo acontece na função *stopword*, embora nesta situação se o termo aqui estiver contido não nos interessa (é uma *stopword*). As funções *pesquisaWikipedia* e *pesquisaWiktionary* são bastante semelhantes: a primeira acede à *Wikipedia* fazendo uma pesquisa baseada no nome do termo de entrada devolvendo uma lista de resultados obtidos, efectuando a segunda o mesmo processo mas por sua vez acedendo ao *Wiktionary*.

A função *geraCombinações* devolve uma lista de todas as combinações possíveis do termo removendo palavras à esquerda e à direita, resultando esta operação numa lista de termos menores ordenados por ordem descendente de tamanho. Por fim, a função *similari-*

dade devolve a similaridade entre duas *strings*, utilizando a mesma métrica de similaridade utilizada em 3.2.2, cujo valor varia entre 0 e 1.

Algoritmo 1: Processo de contextualização de termos

```
Recebe : Termo a contextualizar
Retorna: Termo contextualizado ou null, caso o termo não tenha significado
combinações ⟵ geraCombinações(termo);
for combinação ∈ combinações do
    if !stopWord(combinação) then
        // Verifica se é um termo composto ou simples
        if |combinação| > 1 then
            resultados ⟵ pesquisaWikipedia(combinação);
            for resultado ∈ resultados do
                if similaridade(combinação, resultado) > 0.9 then
                    return combinação;
                end
            end
        else
            if pesquisaPAPEL(combinação) then
                return combinação;
            else
                resultados ⟵ pesquisaWiktionary(combinação);
                for resultado ∈ resultados do
                    if similaridade(combinação, resultado) > 0.9 then
                        return combinação;
                    end
                end
            end
        end
    end
end
return null;
```

A ordem de acesso às fontes de conhecimento aqui descrita não é arbitrária: A *Wikipedia* é apenas acedida no caso de o termo ser composto. Quando o termo é simples, recorre-se primeiro à fonte de conhecimento de acesso mais rápido, o PAPEL, visto que está disponível localmente. Apenas na situação em que o termo não seja aqui encontrado é que se recorre ao *Wiktionary*. Uma solução para acelerar o processo de contextualização dos termos compostos poderia passar pela utilização da *Wikipedia* localmente, mas aqui iria ser posta em causa uma das suas vantagens: a constante actualização e moderação dos seus conteúdos.

Este processo é aplicado a todos os termos da lista, e no caso de o valor de retorno da operação para um determinado termo ser nulo (valor *null*), este termo é removido da lista, considerando-se que não tem significado. Assim é garantido que todos os termos presentes nesta lista tenham algum significado associado, aumentando assim as hipóteses de que estes tragam algum valor na descrição do ponto de interesse ao qual estão associados.

3.2.5 Cálculo de relevância

Neste ponto, existe uma lista de conceitos associados ao POI considerados válidos, que pretendemos ordenar por ordem de relevância, de modo a evidenciar os termos mais significativos para classificar o POI. Para calcular a relevância de cada termo, recorremos ao TF-IDF[4], uma métrica estatística que atribui um peso $tf\text{-}idf_{t,d}$ a um termo t presente num documento d, definida por:

$$tf\text{-}idf_{t,d} = tf_{t,d} \times idf_t \tag{3.1}$$

onde $tf_{t,d}$ corresponde à frequência com que o termo t ocorre no documento d (*Term Frequency*) e idf_t corresponde à frequência inversa do documento, que quantifica a relevância geral do termo num conjunto de documentos (*Inverse Document Frequency*). Este valor pode ser dado por:

$$idf_t = \log \frac{|D|}{|\{d : t \in d\}|} \tag{3.2}$$

sendo $|D|$ a cardinalidade de D, ou seja, o número total de documentos analisados e $|\{d : t \in d\}|$ o número de documentos onde o termo t aparece (ou seja, $tf_{t,d} \neq 0$). Caso o documento não apareça no universo de documentos pode ser gerada uma indeterminação, devido à divisão por zero. Neste caso, normalmente ajusta-se a fórmula para $1+|\{d : t \in d\}|$. Nesta abordagem tal indeterminação não ocorre, dado que só efectuamos este cálculo sobre os termos que vamos encontrando nos documentos disponíveis. Independentemente disto,

a nossa base de conhecimento deve ser populada com alguma informação inicial, de modo a garantir melhores resultados.

Assim sendo, o valor de relevância através do TF-IDF será mais alto quando um termo ocorre muitas vezes num número pequeno de documentos, baixo quando um termo ocorre poucas vezes num documento ou está presente em diversos documentos. Poderá ainda ter um valor mais baixo quando o termo ocorre praticamente em todos os documentos.

Neste caso específico, consideramos que cada recurso extraído de uma fonte de enriquecimento é um documento, tendo em conta os atributos que o compõem (descrições, críticas), sendo o universo de documentos representado pelo conjunto geral de recursos já armazenados na base de conhecimento.

Capítulo 4

Implementação

Neste capítulo é descrito em detalhe como foi feita a implementação da metodologia proposta, seguindo para este efeito os pontos-chave desta. São descritos os módulos implementados (segundo a arquitectura proposta em 3.1.3), evidenciando o papel de cada um no funcionamento da plataforma. Por fim, é descrito o processo de integração com a plataforma TICE.Mobilidade, familiarizando o leitor com a instalação do módulo desenvolvido.

4.1 Módulos Implementados

4.1.1 Extracção de Dados nas fontes

O módulo responsável pela extracção de dados nas fontes de enriquecimento caracteriza-se essencialmente pela sua vertente dinâmica, permitindo a adição de novas fontes de dados de forma pouco intrusiva. O acesso às fontes de enriquecimento é feito através de *scripts* individuais que devem seguir uma *interface pattern*, regulamentando o tipo de dados que os extractores podem receber e devolver.

Os parâmetros de entrada de cada extractor cingem-se a um par de coordenadas (latitude e longitude), um raio de acção (para extrair dados numa área circular) e ainda uma *query* textual (sobre a qual se faz a pesquisa). Um extractor pode utilizar apenas a informação geográfica caso a fonte não suporte pesquisa textual ou caso a fonte não contenha informações georeferenciadas, utilizar apenas a informação textual. Idealmente, as duas deverão ser utilizadas para assegurar uma pesquisa eficaz.

Cada extractor deve então aceder à respectiva fonte de modo a extrair todos os recursos obtidos através da pesquisa pelos parâmetros de entrada, devolvendo estes resultados sob

a forma de uma lista de recursos no formato JSON. Não englobando acessos à base de conhecimento, estes extractores podem ser desenvolvidos e testados sem ter que interagir directamente com a plataforma.

No que toca à avaliação de cada fonte de enriquecimento em termos de confiabilidade, este parâmetro é deixado ao critério de quem produz o extractor, devendo ser tido em conta se a fonte é uma rede social, um directório comercial e se os seus conteúdos são devidamente moderados, de modo a agilizar o futuro processo de integração.

Estando os extractores integrados na plataforma, e sendo requerida por terceiros a extracção de novos dados para um determinado recurso, estes são executados numa *pool* de *threads* enviando os resultados destas *threads* para o módulo de integração aquando do término da sua execução. Para evitar bloqueios, cada extractor tem um tempo máximo de execução, que caso seja excedido o seu resultado é descartado.

De modo a identificar o idioma em que se encontram os conteúdos textuais de maiores dimensões, foi utilizada uma biblioteca[1] que permite extrair o idioma em que se encontra representado um bloco de texto mediante uma análise probabilística. Segundo o autor[2], esta biblioteca apresenta uma precisão acima de 99% para 49 linguagens distintas, sendo o modelo utilizado treinado através do *corpus* da *Wikipedia* para cada uma destas línguas e validado em fontes de notícias para cada um dos idiomas. Esta biblioteca é utilizada essencialmente nos textos de grandes dimensões aos quais aplicamos técnicas de extracção de informação, como é o caso das descrições aqui extraídas.

4.1.2 Integração de Recursos

A informação proveniente do módulo de extracção de dados é analisada de modo a detectar a similaridade entre o recurso que se pretende enriquecer e os resultados provenientes das fontes, fazendo uso das métricas de integração descritas em 3.2. O cálculo da similaridade entre *strings* é feito com recurso à biblioteca *SecondString*, mais concretamente utilizando uma métrica híbrida combinando *Jaro-Winkler* (destacada em 2.3) com o *TF-IDF*. Segundo os autores do *SecondString*, esta tem-se revelado a métrica mais eficiente[18] da biblioteca.

Ao calcular a similaridade entre *websites*, há que ter em conta que os URLs apresentam na maior parte das situações diversas sequências de caracteres em comum (como "*http://*"ou "*www*"), que embora façam parte da especificação deste formato, neste processo não nos

[1]Mais informações em: `http://code.google.com/p/language-detection/`
[2]Mais informações em: `http://goo.gl/pOqAE`

trazem qualquer mais-valia. Tendo este factor em conta, antes de dois *websites* serem comparados omitimos estas sequências de caracteres de modo a obter resultados mais precisos. A inclusão destas sequências influenciaria de forma negativa o cálculo de similaridade, dado que o URL de um *website* pode ser representado utilizando estas duas sequências ou só uma, dependendo da configuração do domínio. Para além disto, em *websites* estas sequências podem constituir a maior parte do URL (Ex.: `http://www.iol.pt`).

Aquando da detecção de duplicados, é necessário fundir dois recursos distintos. Este método tem em conta o nível de confiabilidade de cada fonte de modo a inferir quais os dados que pode ou não sobrepor. Para isto, são percorridos todos os campos do recurso original para aferir quais os atributos em falta de forma a preencher estes. Caso o novo recurso seja proveniente de uma fonte com maior nível de confiabilidade, os campos simples (como a morada ou o código-postal) já existentes são comparados com estes e caso não se revelem mais significativos que os novos dados[3] são substituídos. Este tipo de interacção modifica o valor de confiabilidade da fonte, subindo-o num ponto percentual caso contenha dados mais significativos e vice-versa.

O processo de limpeza de ruído nos dados provenientes dos extractores é feito com auxílio de um *parser* de HTML, o *JSoup*[4]. Este fornece-nos uma API que permite extrair e manipular de forma eficaz conteúdos presentes em documentos HTML, tal como a limpeza de *tags* HTML em grandes blocos de texto.

4.1.3 Extracção de Informação

A implementação do módulo de Extracção de Informação englobou essencialmente a integração das ferramentas estudadas em 2.1. Esta integração consistiu na criação de uma camada que normaliza as dados de entrada e formata os resultados obtidos pelas ferramentas, ou seja, faz o tratamento dos dados de entrada em texto não-estruturado consoante os parâmetros de entrada de cada ferramenta (Ex.: Algumas ferramentas fazem todo o tratamento de informação textual por si, enquanto que outras necessitam de algum pré-processamento como divisão dos componentes do texto em blocos unitários).

O conjunto de ferramentas aqui utilizado não corresponde de forma exacta aos resultados obtidos em 2.1.2. Para as ferramentas que lidam com a língua Portuguesa, os resultados analisados aquando da implementação deste módulo não se mostraram nada satisfatórios,

[3] Ex.: uma morada que não corresponda à localização especificada ou um *website* que já não se encontre disponível.

[4] Mais informações em: `http://jsoup.org/`

tendo sido testadas as ferramentas que supostamente teriam pior desempenho com base nos testes efectuados.

Esta situação verificou-se principalmente na identificação de sintagmas nominais, onde o *Apache OpenNLP* tinha obtido resultados claramente superiores ao *CMSChunker* mas ao lidar dinamicamente com dados provenientes das fontes estudadas não produziu os melhores resultados. Esta situação deveu-se ao facto de o identificador de sintagmas nominais do *Apache OpenNLP* ter apenas sido treinado com o *corpus* do Bosque e não se terem especificado regras para este, algo que foi feito no *CMSChunker*.

No que toca ao reconhecimento de entidades mencionadas, o *PT-NER* não foi utilizado devido ao facto de ainda não se encontrar totalmente pronto para ser utilizado em produção, necessitando ainda de algum trabalho para disponibilizar uma API que possa ser utilizada externamente. Esta decisão teve em conta o tempo que seria necessário para efectuar tais modificações na ferramenta, e dado que a janela temporal conferida ao processo de integração das ferramentas de extracção de informação era reduzida, optou-se por utilizar o reconhecedor de entidades mencionadas do *OpenNLP*, tendo este apresentado resultados aceitáveis neste processo.

4.1.4 Validação e contextualização de Termos

As heurísticas que nos permitem validar palavras presentes nos termos foram implementadas com base em expressões regulares, sendo verificado se a palavra era aceite pelo padrão definido na expressão regular de modo a considerar se esta era ou não válida. Já a contagem de palavras por termo é feita com recurso a bibliotecas já usadas em 3.2.3 para dividir texto em blocos unitários de palavras.

No que toca à contextualização de termos, as interacções com a *Wikipedia* e o *Wiktionary* são feitas de modo semelhante: Ambos os serviços são baseados na *Media Wiki*[5], que disponibiliza uma API genérica[6] permitindo operações de pesquisa e extracção de conteúdos. Este módulo oferece uma interface de acesso a estes recursos de forma transparente, através das APIs disponibilizadas.

Quanto ao PAPEL, o acesso a este recurso na sua forma original (ficheiro de texto com aproximadamente 200.000 linhas) é incomportável, por isso indexaram-se os conteúdos deste em estruturas nativas sendo estas guardadas num ficheiro interno, providenciando um

[5]Mais informações em: `http://www.mediawiki.org/wiki/MediaWiki`
[6]Mais informações em: `http://www.mediawiki.org/wiki/API:Main_page`

acesso mais rápido a este recurso.

4.1.5 Cálculo de Relevância

Embora o cálculo de relevância já tenha sido abordado em 3.2.5, é importante explicitar como este foi implementado.

O cálculo da frequência de um termo relativamente ao conjunto de informações associado ao POI pode ser calculado de duas formas distintas: Se o termo for simples, basta dividir o número de ocorrências deste no conjunto de informações pelo número total de palavras. No entanto, para os termos compostos esta operação não será a mais adequada. Utiliza-se nesta situação a frequência composta do termo[29], ou seja, somam-se as frequências de todas as palavras que constituem o termo, desprezando apenas as que se encontram na lista de *stopwords*, cuja contagem iria influenciar o valor aqui calculado.

O valor de relevância (*Term Frequency x Inverse Document Frequency*) não se encontra apenas associado ao termo, dado que este não depende apenas da relevância no conjunto total de documentos (*Inverse Document Frequency*), mas também da frequência com que o termo ocorre nas informações associadas a um determinado POI (*Term Frequency*). Tendo em conta que um termo pode estar associado a múltiplos POIs, é importante garantir a separação destes valores. Para isto, o valor do *Term Frequency* fica armazenado na ligação entre o recurso e o termo (entidades *Resource* e *Tag*, como pode ser observado no apêndice B) enquanto que o valor do *Inverse Document Frequency* fica armazenado numa entidade distinta (TF_IDF), sendo este valor actualizado aquando da inclusão de novos dados na base de conhecimento, variando a relevância de cada termo consoante as interacções com os dados persistidos.

4.1.6 Camada de Acesso a Dados

Estando a base de conhecimento acessível através do módulo IS, todas as interacções exteriores ao módulo são baseadas em pedidos REST. No entanto, dado que o módulo IS está a ser desenvolvido em paralelo com o ES, nem todas as operações necessárias ao nosso módulo se encontravam disponíveis aquando do seu desenvolvimento.

Para superar esta restrição, no ambiente de desenvolvimento recorreu-se a uma base de dados espacial PostGIS. De modo a agilizar o futuro processo que engloba a migração da base de conhecimento, todas as operações de acesso foram especificadas através da utilização

de uma *software pattern Java*, conhecida como DAO(*Data Access Object*)[7]. Esta permite uma maior abstracção da camada de dados, através da criação de diferentes objectos que tratam do processo de inserção, modificação, actualização e eliminação de dados mediante a base de conhecimento utilizada (como uma API externa ou base de dados).

Deste modo foi possível desenvolver em paralelo os mecanismos de acesso a dados tanto para o processo de desenvolvimento como para o futuro processo de integração, sem necessidade de haver grandes modificações a nível de código.

4.1.7 API de Acesso

A API de acesso ao módulo de enriquecimento semântico é a principal via de comunicação entre este e os serviços exteriores, encontrando-se documentada em detalhe no apêndice C. Todos os dados são apresentados no formato JSON, um consenso adoptado entre todas as equipas responsáveis pelos módulos do PPS2.

Dado que algumas operações do módulo levam algum tempo a ser executadas (devido maioritariamente à ligação a API's externas e outros serviços), foram tidos em conta alguns detalhes de modo a minimizar o tempo de espera na origem do pedido. Pedidos cujo objectivo seja apenas despoletar uma determinada acção no módulo (como a recolha de novo conhecimento ou extracção de novas "tags") são executadas de forma assíncrona, sendo apenas devolvida uma resposta de confirmação sobre o sucesso ou insucesso do pedido, ficando a tarefa a ser executada em *background*.

Tendo em conta que nesta fase não existiam serviços complementares que permitissem a visualização dos dados produzidos pelo módulo de enriquecimento semântico, foi criada uma pequena aplicação que utiliza diversos métodos da API fornecida para representar o conhecimento aqui presente. Esta baseia-se essencialmente em pedidos REST através de *Javascript* que permitem representar os recursos enriquecidos com recurso à API do *Google Maps*. Esta aplicação pode ser consultada em `http://ubiquo.dei.uc.pt:8080/SE/`, estando também representada na figura 4.1. Aqui, os POIs presentes no mapa estão agrupados em *clusters* geográficos que ao serem clicados é apresentada uma lista dos POIs aqui contidos. Ao clicar num determinado POI é apresentada a informação disponível sobre este, e caso existam, os termos que o descrevem ordenados por relevância.

[7]Mais informações em: `http://java.sun.com/blueprints/corej2eepatterns/Patterns/DataAccessObject.html`

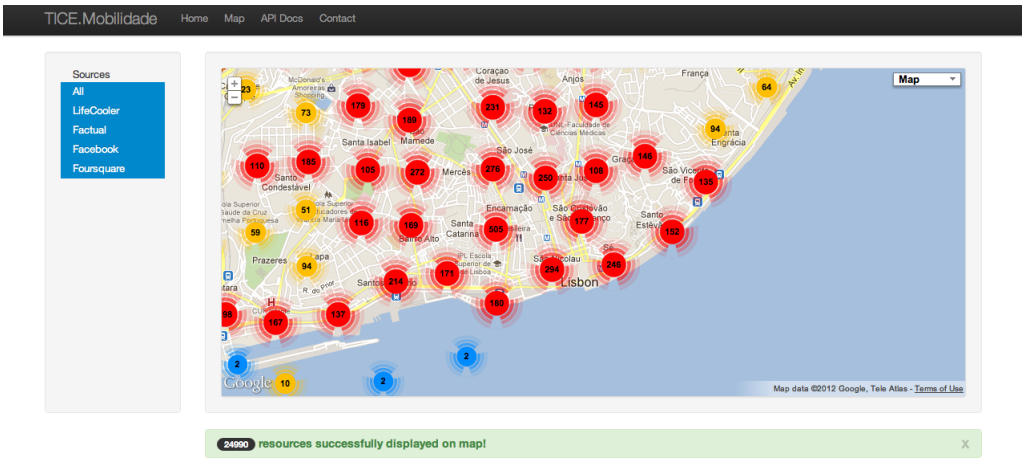

Figura 4.1: Visualização dos dados do Módulo de Enriquecimento Semântico

4.2 Integração na plataforma TICE.Mobilidade

O processo de integração do módulo de enriquecimento semântico na plataforma TICE.Mobilidade (ou em outro qualquer serviço que pretenda interagir com a API disponibilizada) foi idealizado de forma a poder ser executado por qualquer outro membro do consórcio sem estar familiarizado com o código-fonte.

O facto de o módulo não conter um mecanismo de autenticação individual foi fruto da decisão geral dos membros do PPS2: Este encontra-se a ser desenvolvido pelo PPS1, de modo a homogeneizar os acessos a todos os serviços contidos na plataforma TICE.Mobilidade.

Os detalhes relativos à instalação deste módulo podem ser consultados no apêndice D.

Capítulo 5

Experiências e Resultados

Este capítulo foca-se nas experiências realizadas ao longo da implementação descrita no capítulo anterior, apresentando também os resultados de cada uma de modo a obter uma validação do trabalho realizado.

5.1 Extracção de Recursos

Dado que a nossa abordagem se foca essencialmente no território Português, as experiências aqui descritas englobam a área metropolitana da Grande Lisboa, de modo a obter uma diversificada gama de recursos.

Para recolher recursos contidos nesta área, obtiveram-se os limites geográficos da área da Grande Lisboa através da API de *geocoding* do *Google*[1], que mediante a pesquisa pelo nome de uma cidade retorna, entre outros dados, os limites geográficos desta. Estes limites são representados por dois pares de coordenadas, correspondendo um ao ponto mais a sudoeste da área e o outro ao ponto mais a nordeste. Com este conjunto de coordenadas é possível criar uma geometria capaz de englobar toda a área visada, que percorremos sobre a forma de uma grelha iterando sobre pares de coordenadas que são fornecidos aos vários extractores para recolher recursos contidos num raio de 500 metros destas coordenadas.

Para o processo de extracção foram escolhidas quatro das fontes de enriquecimento estudadas em 2.2: duas redes sociais (*Facebook* e *Foursquare*) e dois directórios comerciais (*Factual* e *Lifecooler*), diversificando desta maneira os conteúdos utilizados nas experiências.

A tabela 5.1 apresenta valores resultantes deste processo de extracção, de modo a quan-

[1]Mais informações em: `https://developers.google.com/maps/documentation/geocoding/`

Tabela 5.1: Valores resultantes da Extracção de Recursos

Fonte	Facebook	Foursquare	Factual	Lifecooler	Total
Número de Recursos	4139	6168	11754	4782	26843
Número de Descrições	4476	0	0	3465	7941
Número de Categorias	116	308	236	106	766

tificar o volume de dados aqui utilizados. Como se pode observar, a diferença de valores entre o número de pontos de interesse extraídos e o número de alguns campos que figuram nestes (como as descrições) é algo díspar. Isto deve-se ao facto de que apenas algumas fontes nos fornecem recursos textuais, como é o caso do *Facebook* e do *Lifecooler*. Embora as fontes *Factual* e *Foursquare* não nos forneçam grandes dados textuais, podem ser importantes no processo de integração de recursos ou até mesmo na extracção de novo conhecimento. O valor médio de caracteres por descrição é de 529.66 caracteres, revelando-se assim bastante positivo para o tipo de tarefas que se pretende efectuar sobre estes campos, nomeadamente extracção de informação e processamento de linguagem natural.

5.2 Integração de Recursos

De forma a medir o desempenho do processo de integração de recursos, iterou-se sobre todo o conjunto de dados extraído em 5.1 verificando se cada recurso se encontrava duplicado seguindo a abordagem proposta em 3.2. De modo a reduzir a complexidade do processo, ao iterar sobre cada elemento analisamos todos os POIs presentes num raio de 100 metros, visto que a distância é o primeiro elemento discriminador no processo de integração. Finda esta análise, obteve-se um conjunto de POIs considerados duplicados do item actual, denominado grupo de duplicados. No fim de todo o processo, obteve-se uma lista de grupos de duplicados. A lista obtida ao executar este processo sobre o conjunto de dados extraídos em 5.1 continha 2216 conjuntos de duplicados, perfazendo um total de 4979 POIs. Ou seja, o tamanho do conjunto original foi reduzido a 2216, agrupando assim as características dos POIs considerados duplicados. A tabela 5.2 exemplifica como é constituído um grupos de duplicados, sendo o POI original identificado a negrito seguido pelos candidatos a duplicados.

Dada a inexistência de *ground truth* para validar este processo de forma eficaz (em ter-

Tabela 5.2: Exemplo simplificado de um grupo de duplicados

Nome	Morada	*Websites*	Distância (m)	Categorias
Restaurante La Rúcula	**Rua Rossio Olivais**	**[larucula.com.pt]**	—	[***Food & Beverage, Restaurants***]
Restaurante La Rúcula	Rua Rossio Olivais	—	61.53	[*Local business*]
La Rúcula	Rossio dos Olivais	—	77.21	[*Italian Restaurant*]

mos de precisão e cobertura), recorreu-se a um conjunto de voluntários[2] com a finalidade de validar a lista de duplicados manualmente. De modo a agilizar o processo, a lista resultante do processo descrito acima foi dividida em oito folhas de cálculo diferentes, cada uma constituída por cerca de 300 grupos de duplicados. Para facilitar a tarefa de validação, cingimos a representação de cada POI aos elementos mais discriminadores, como o nome, morada, *websites*, distância e categorias associadas. Este conjunto de folhas de cálculo foi distribuído pelos voluntários, sendo pedido a estes para assinalarem apenas os duplicados correctamente identificados. Deste modo, consideraram-se como verdadeiros positivos os duplicados assinalados pelos voluntários e como falsos positivos os duplicados não assinalados. Através destes valores é-nos possível avaliar a precisão do processo, tal como foi feito em 2.1.2.

Os resultados do processo de validação podem ser observados na tabela 5.3. Dado que a validação dos resultados é dividida em vários subconjuntos, podemos avaliar a precisão da abordagem de duas maneiras distintas: Fazendo a média dos resultados de precisão de cada conjunto (método conhecido como *macro-average precision*) ou através da precisão calculada com base no somatório dos verdadeiros e falsos positivos de todos os conjuntos (*micro-average precision*)[4]. A primeira é normalmente utilizada em situações em que desejamos saber como a abordagem se comporta em diferentes conjuntos de teste, enquanto que a segunda é idealmente utilizada para inferir a precisão geral da abordagem em conjuntos de teste com diferentes tamanhos.

Neste caso foram utilizadas as duas metodologias, não só com o intuito de quantificar

[2]Este conjunto de voluntários caracteriza-se por um grupo de 8 estudantes do ensino superior de áreas distintas (Ciências, Humanidades e Artes).

Tabela 5.3: Resultados da validação do processo de integração de recursos

Conjuntos	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Total**
Total de POIs	701	702	685	693	688	687	671	152	4979
Grupos de Duplicados	299	293	307	304	309	307	322	75	2216
Verdadeiros Positivos	395	390	368	383	352	326	286	44	2544
Falsos Positivos	7	19	10	6	27	54	54	33	210
Precisão	0.98	0.95	0.97	0.98	0.93	0.86	0.82	0.57	—

a precisão da abordagem mas também para encontrar padrões nos falsos positivos detectados. Assim, os resultados revelaram-se bastante satisfatórios, com uma *micro-average precision* de 0.92, sendo a *macro-average precision* quantificada com um valor médio de 0.88 e apresentando um desvio padrão de 0.14.

Com base nos valores de precisão divergentes em cada conjunto de validação, foram analisados os conjuntos com maior taxa de falsos positivos, verificando-se um padrão nestes conjuntos: A maioria dos falsos positivos surge em situações onde os POIs analisados estão contidos dentro de outros. Um exemplo comum desta situação pode ser uma loja contida num centro comercial ou um restaurante contido num hotel, como podemos observar na tabela 5.4. Normalmente estes POIs apresentam grande similaridade em atributos como o nome, contendo até vários atributos em comum, como a morada e os *websites*, algo que dificulta a diferenciação entre recursos. Esta falha pode ser ultrapassada com a utilização de um sistema de regras baseado em expressões regulares aquando do cálculo da similaridade entre os nomes dos recursos. Para tal teríamos que discriminar todas as combinações possíveis de nomenclaturas deste género, representando uma tarefa algo complexa e ainda assim susceptível a falhas.

Dadas as diferenças de precisão nos diferentes conjuntos de teste, efectuou-se uma validação estatística com o intuito de testar a hipótese de os conjuntos serem iguais entre si. Havendo oito amostras para comparar, decidimos recorrer a um teste não-paramétrico para amostras independentes, a *ANOVA (*Analysis of Variance*) de Kruskal-Wallis*[30]. Este teste permite-nos testar se duas ou mais amostras provém de populações diferentes com a mesma distribuição, adequando-se assim aos nossos resultados.

Através da utilização do *SPSS*[3], efectuou-se este teste com uma probabilidade de erro $\alpha = 0.05$ tendo sido obtido um P-Valor $p = 0.429 > \alpha = 0.05$. Este resultado aceita assim

[3]Mais informações em: `http://www-01.ibm.com/software/analytics/spss/`

Tabela 5.4: Exemplo de verdadeiros e falsos positivos

Nome	Morada	*Website*	Distância (m)	Categorias
IBM Portugal	**Rua do Mar da China, Lote 1.07.2.3**	—	—	[***Office***]
Companhia IBM Portuguesa	Rua Mar da China, Lt. 1.07.2.3	`www.ibm.com/pt`	14.31	[*Business & Professional Services, Equipment, Supplies & Services, Office*]
Altis Belem Hotel & Spa, Lisboa	**Doca do Bom Sucesso**	—	—	[***Local business***]
Restaurante Feitoria - Altis Belém Hotel & SPA	Doca do Bom Sucesso	—	40.84	[*Local business*]
Altis Belém Hotel & SPA	Doca do Bom Sucesso	—	69.04	[*Hotel*]

a hipótese de que as oito populações seguem a mesma distribuição relativamente à precisão apresentada.

5.3 Etiquetagem Automática

Para validar a funcionalidade de etiquetagem automática, torna-se necessário avaliar os resultados em duas vertentes distintas: a coerência dos termos associados a cada recurso (inferindo se as *tags* extraídas se adequam ao recurso em questão) e se estes se encontram devidamente ordenados em termos de relevância, tendo esta por base os valores obtidos na pesagem dos termos (3.2.5).

Assim sendo, o processo de validação engloba duas experiências distintas: A primeira avalia a coerência dos termos, inferindo se os termos resultantes da etiquetagem automática realmente se coadunam ao recurso a que estão associados (Ex.: O termo "serviços de beleza"é coerente face ao POI a que se encontra associado, "Skinclinic - Clínica de Estética e Terapia"). Já a segunda experiência tem como objectivo avaliar o cálculo da relevância através da qual a lista de termos associados a um recurso é ordenada.

O conjunto de dados aqui utilizado corresponde ao número de recursos para os quais o módulo de extracção de informação conseguiu extrair termos, perfazendo um total de 3279 POIs com 9539 termos associados, num universo de 4939 termos distintos. Uma visão geral da dimensão dos termos extraídos pode ser observada nas figuras 5.1 e 5.2, onde estão representadas a frequência de termos extraídos por POI e a frequência de palavras por termo extraído, respectivamente, onde podemos verificar que a maior parte dos termos extraídos neste processo são simples (com apenas uma palavra), estando em média aproximadamente três termos associados a cada POI.

Tal como na experiência anterior, a validação dos resultados para estas duas experiências é feita recorrendo a um conjunto de voluntários[4], pelo qual são distribuídas folhas de cálculo com os resultados obtidos com a finalidade de os validar manualmente.

5.3.1 Coerência de Termos

De modo a avaliar se os termos extraídos se adequam ou não ao recurso a que estão associados, é pedido aos voluntários que classifiquem cada conjunto de termos face a um

[4]Neste caso os voluntários foram pessoas diferentes das que validaram a experiência anterior, sendo estes 10 estudantes do ensino superior de diferentes àreas.

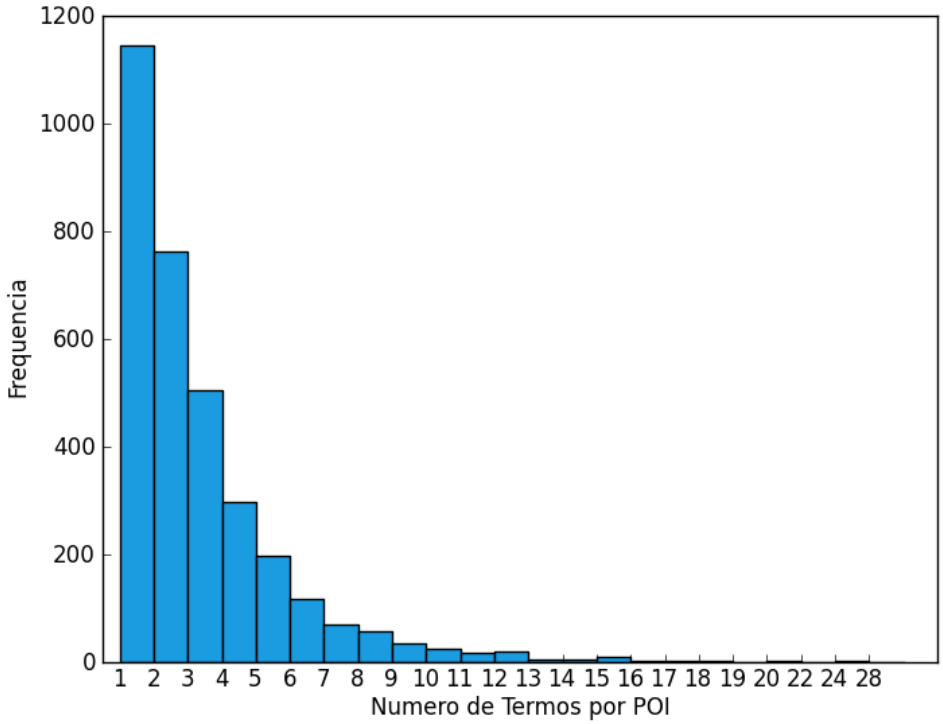

Figura 5.1: Frequência de termos extraídos por POI

determinado recurso como sendo:

1. Pouco Relevante
2. Algo Relevante
3. Muito Relevante

O formato das folhas de cálculo distribuídas é exemplificado na tabela 5.5, onde estão representados exemplos das três situações possíveis de validação. Enquanto que o último conjunto de termos descreve o POI com rigor, os dois primeiros já deixam algo a desejar, não tendo o primeiro conjunto de termos algum cariz descritivo sobre o POI.

A contagem destas respostas permite-nos avaliar a qualidade dos termos extraídos, sendo que os resultados obtidos podem ser observados na tabela 5.6. A cada resposta está associado o respectivo número de ocorrências desta e a percentagem que representa no conjunto total de POIs com termos associados.

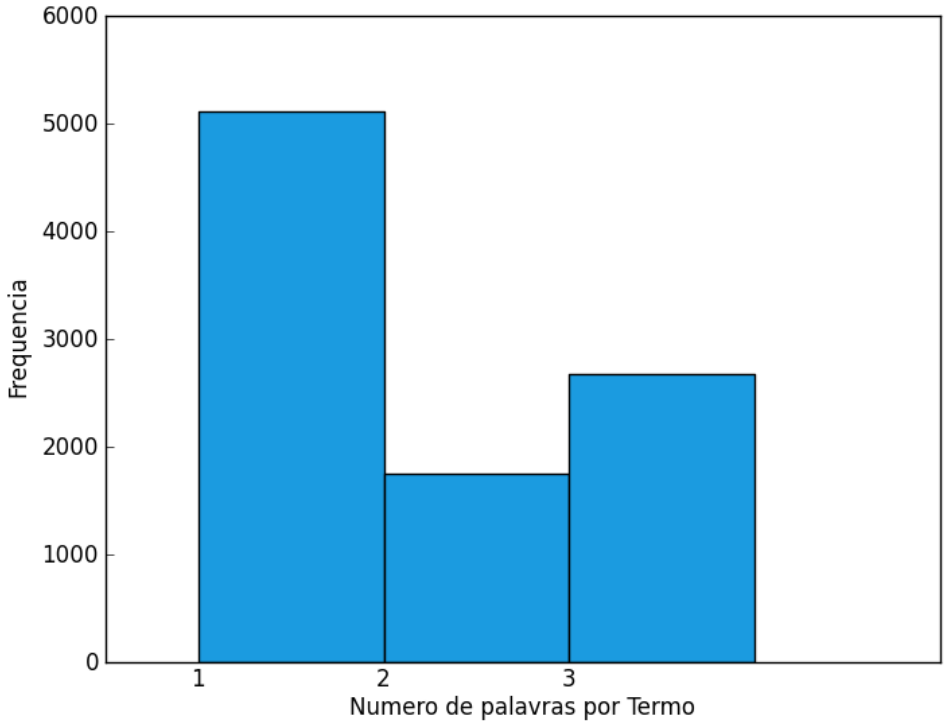

Figura 5.2: Frequência de palavras por Termo

Os resultados obtidos apontam para uma predominância de termos relevantes para cada recurso, rondando a soma das percentagens de ocorrência das respostas "algo relevante"e "muito relevante"os 73%, mostrando-se este valor claramente superior à percentagem de conjuntos de termos não-relevantes, que ficou a rondar os 27%.

O processo de validação utilizado (em diferentes conjuntos de teste) permitiu-nos mais uma vez averiguar padrões nos termos mais incoerentes, que normalmente consistem em palavras que embora não constem na lista de *stopwords* utilizada também não acrescentam valor descritivo a qualquer recurso. Palavras como nomes próprios (muitas vezes associados a figuras públicas que frequentam um determinado local ou até mesmo proprietários) e palavras genéricas para serviços como "loja", "estabelecimento"ou "casa".

De modo a verificar se as respostas por parte dos voluntários seguem uma distribuição uniforme, foi efectuada a seguinte validação estatística: Caso a distribuição da amostra seja uniforme, em termos teóricos a percentagem de cada categoria a ser atribuída será igual para todo o conjunto, ou seja, $\frac{1}{3}$. Procedeu-se ao teste do Qui-Quadrado[31], de modo a

Tabela 5.5: Exemplo simplificado da folha de cálculo para validação da Coerência de Termos

Resposta	**Nome**	**Categorias**	**Termos**
2	Cervejaria Portugália	Noite e Restaurantes, Restaurantes	bacalhau, mesa
1	Camisaria Moderna	Turismo de Compras, Outras lojas	slogan, amigo, cliente
3	AEFCSH - UNL	Organization	associação de estudantes, faculdade de ciências

Tabela 5.6: Resultados da validação relativa à experiência da coerência de termos

Resposta	**Número de Ocorrências**	**Percentagem**
Pouco Relevante (1)	875	26.69%
Algo Relevante (2)	577	17.57%
Muito Relevante (3)	1827	55.73%
Total	3279	100%

aferir a qualidade de ajuste, sendo os seus resultados apresentados na tabela 5.7.

Como podemos observar, os valores de N observados não correspondem aos valores esperados, verificando-se assim a não-uniformidade das respostas resultantes do processo de validação, sendo o valor do Qui-Quadrado $\chi^2 = 779.9^2$ com dois graus de liberdade.

5.3.2 Ordem de Relevância

Nesta experiência, a tarefa de validação passa pela etiquetagem dos resultados de forma binária, respondendo à questão: "O conjunto de palavras referentes ao recurso em questão

Tabela 5.7: Teste do Qui-Quadrado relativo à experiência da Coerência de Termos

	N Observado	**N Esperado**	**Resíduos**
Pouco Relevante	875	1093.0	-218.0
Algo Relevante	577	1093.0	-516.0
Muito Relevante	1827	1093.0	734.0
Total	3279		

Tabela 5.8: Resultados da validação relativa à experiência de ordem de relevância

Resposta	**Número de Ocorrências**	**Percentagem**
Sim (1)	1279	72.71%
Não (2)	480	27.2882319499716%
Total	1759	100%

Tabela 5.9: Teste do Qui-Quadrado relativo à experiência da ordem de relevância

	N Observado	**N Esperado**	**Resíduos**
Sim	1279	879.5	399.5
Não	480	879.5	-399.5
Total	1759		

encontra-se correctamente ordenado por relevância?". O carácter binário da questão permite avaliar o desempenho desta funcionalidade de forma rígida simplificando também a tarefa de validação para os voluntários.

Já existindo resultados da validação relativa à coerência de termos, utilizamos neste ponto POIs onde os termos associados foram classificados como "algo relevante"ou "muito relevante", de modo a reduzir o tamanho do conjunto de dados de validação focando dados consistentes. Outro ponto de corte foram os POIs cujo tamanho do conjunto de termos é unitário, dado que não faria sentido validar a ordem de relevância numa lista de um só elemento. Desta forma, o tamanho do conjunto de validação foi reduzido a 1759 POIs. De forma a agilizar o processo de validação, cada voluntário validou o mesmo conjunto de dados da tarefa anterior (dentro do possível), mas devidamente filtrado. Estando o voluntário familiarizado com os resultados a analisar, é possível obter coerência nos resultados de validação da exactidão da ordem de relevância atribuída a cada termo.

Os resultados desta validação podem ser observados na tabela 5.8, onde se verifica que em aproximadamente 73% dos casos de teste fornecidos aos voluntários a abordagem foi bem-sucedida.

Como no processo de validação anterior, recorremos ao teste do Qui-Quadrado com o intuito de verificar se as respostas seguem uma distribuição dos resultados, estando os resultados deste representados na tabela 5.9.

Verifica-se que os valores de N observados não correspondem aos valores esperados,

verificando-se assim a não-uniformidade das respostas resultantes do processo de validação, sendo o valor do Qui-Quadrado $\chi^2 = 362.93^2$ com um grau de liberdade unitário.

Capítulo 6

Conclusões

O trabalho aqui apresentado consistiu na implementação de uma metodologia de enriquecimento semântico de lugares e eventos para a língua Portuguesa, tendo como base o projecto KUSCO[3]. Foi desenvolvido um módulo capaz de adquirir dinamicamente conhecimento sobre este tipo de recursos acedendo a fontes de dados *online* e etiquetando-os com termos cujo significado consiga descrever um determinado local de forma eficaz, com recurso a técnicas de extracção de informação e processamento de linguagem natural.

Ao longo da primeira fase do trabalho foram discutidos e analisados os requisitos deste trabalho, em conjunto com os do PPS 2 - SEMA. Foram analisadas ferramentas necessárias à implementação da metodologia e abordagens semelhantes, de modo a sondar o que poderia ser utilizado neste projecto e por onde se podia inovar. Dado que a recolha de dados *online* é um dos pilares deste trabalho, foi feito um estudo intensivo sobre as fontes de informação a utilizar, tendo em conta o tipo de recursos possível de extrair e a sua cobertura em território português.

A segunda fase deste trabalho englobou essencialmente o desenvolvimento da metodologia proposta, a sua experimentação e a consequente validação de resultados. Os resultados demonstraram que apesar dos problemas comuns ao utilizar técnicas de extracção de informação e processamento de linguagem natural em texto não-estruturado, é possível obter informações bastante descritivas sobre locais a partir de fontes de conhecimento presentes na *web*, tendo apresentado a tarefa de integração de recursos uma precisão que ronda os 92% e as tarefas envolvidas no processo de extracção de termos significativos apresentam uma taxa de sucesso na ordem dos 70%.

Os principais contributos deste projecto passam pela aplicação da metodologia do

KUSCO para a língua Portuguesa assente numa plataforma robusta de cariz modular em termos de fontes de conhecimento e conteúdos, de modo a poder ser utilizada no âmbito do projecto TICE.Mobilidade. Todas as funcionalidades que figuravam nos requisitos do projecto foram implementadas, abrindo agora o caminho para uma melhoria contínua nos resultados aqui apresentados.

Deste trabalho, resultou um artigo que descreve a metodologia implementada e consequente validação, "Semantic Enrichment of Places for the Portuguese language", submetido e aceite no *Infórum 2012*[1] - Simpósio de Informática. Este será apresentado no simpósio, a decorrer em Setembro do presente ano.

6.1 Trabalho Futuro

A análise de resultados presente neste relatório permitiu-nos identificar eventuais melhorias que podem ser feitas na abordagem proposta. Uma delas passa pela imposição de regras mais rígidas na tarefa de integração de recursos, na qual foram encontradas falhas, mais concretamente em situações onde um ponto de interesse está contido em outro de maior dimensão.

No que toca à extracção de termos, os resultados podem ser melhorados, passando este processo pelo ajuste a algumas das ferramentas de extracção de informação usadas. A validação e contextualização de termos poderá fazer mais uso do processo de etiquetagem morfológica, analisando os padrões dos termos que não foram considerados coerentes de modo a filtrá-los no futuro.

O cálculo de relevância pode ser melhorado com recurso ao pré-calculo dos valores do *Inverse Document Frequency*, utilizando para isto *corpus* mais abrangentes do que os dados presentes na base de conhecimento do módulo de Enriquecimento Semântico.

A inclusão de recursos como eventos ou dados não georeferenciados no processo de enriquecimento semântico está também agendada, passando esta inclusão pelo ajuste de alguns pontos na metodologia apresentada, como é o caso da detecção de duplicados (que no caso dos eventos deverá conter uma componente temporal e no caso dos recursos não georeferenciados a omissão da componente geográfica).

Deverá também ser desenvolvido uma metodologia que permita a utilizadores da plataforma TICE.Mobilidade com menos experiência criarem extractores de dados para novas

[1] Mais informações em: `http://inforum.org.pt/INForum2012/`

fontes. No caso das fontes onde é necessário *screen-scraping*, por exemplo, pode passar pela utilização de serviços que criam APIs de forma intuitiva, como é o caso do *APIfy*[2].

Por fim, e de modo a obter uma integração completa com o projecto TICE.Mobilidade, é apenas necessário que as dependências requeridas por este módulo estejam implementadas pelos módulos complementares, permitindo assim o bom funcionamento de toda a estrutura do PPS 2 - SEMA.

[2]Mais informações em: `http://apify.heroku.com/`

Bibliografia

[1] A. Vaccari, L. Liu, A. Biderman, C. Ratti, F. Pereira, J. Oliveirinha, and A. Gerber, "A holistic framework for the study of urban traces and the profiling of urban processes and dynamics," *2009 12th International IEEE Conference on Intelligent Transportation Systems*, pp. 1–6, 2009.

[2] A. Alves, F. Rodrigues, and F. Pereira, "Tagging space from information extraction and popularity of points of interest," in *Ambient Intelligence* (D. Keyson, M. Maher, N. Streitz, A. Cheok, J. Augusto, R. Wichert, G. Englebienne, H. Aghajan, and B. Kröse, eds.), vol. 7040 of *Lecture Notes in Computer Science*, pp. 115–125, Springer Berlin / Heidelberg, 2011.

[3] A. Alves, F. Pereira, F. Rodrigues, and J. Oliveirinha, "Place in perspective: Extracting online information about points of interest," in *Ambient Intelligence* (B. de Ruyter, R. Wichert, D. Keyson, P. Markopoulos, N. Streitz, M. Divitini, N. Georgantas, and A. Mana Gomez, eds.), vol. 6439 of *Lecture Notes in Computer Science*, pp. 61–72, Springer Berlin / Heidelberg, 2010.

[4] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, July 2008.

[5] R. Milidiú, C. dos Santos, and J. Duarte, "Portuguese corpus-based learning using etl," *Journal of the Brazilian Computer Society*, vol. 14, pp. 17–27, 2008. 10.1007/BF03192569.

[6] C. Nogueira dos Santos, R. Milidiú, and R. Rentería, "Portuguese part-of-speech tagging using entropy guided transformation learning," in *Computational Processing of the Portuguese Language* (A. Teixeira, V. de Lima, L. de Oliveira, and P. Quaresma,

eds.), vol. 5190 of *Lecture Notes in Computer Science*, pp. 143–152, Springer Berlin / Heidelberg, 2008.

[7] R. L. Milidiú, C. N. Santos, and J. C. Duarte, "Phrase chunking using entropy guided transformation," in *in Proc. of ACL-08: HLT*, pp. 647–655, 2008.

[8] R. Zanoli, E. Pianta, and C. Giuliano, "Named entity recognition through redundancy driven classifiers," *In Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence, 12th December 2009, Reggio Emilia, Italy*, vol. 9, 2009.

[9] E. N. Motta, E. R. Fernandes, and R. L. Milidiú, "A web service for natural language processing," *World Wide Web Internet And Web Information Systems*, pp. 139–139, 2007.

[10] L. Padró, M. Collado, S. Reese, M. Lloberes, and I. Castellón, "Freeling 2.1: Five years of open-source language processing tools," *LREC*, pp. 931–936, 2010.

[11] R. Rodrigues, H. Gonçalo Oliveira, and P. Gomes, "Uma abordagem ao Págico baseada no processamento e análise de sintagmas dos tópicos," *Linguamática*, vol. 4, pp. 31–39, Abril 2012.

[12] C. Freitas, P. Rocha, and E. Bick, "Floresta Sintá(c)tica: Bigger, Thicker and Easier," in *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)* (A. Teixeira, V. L. S. de Lima, L. C. de Oliveira, and P. Quaresma, eds.), vol. Vol. 5190, pp. 216–219, Springer Verlag, Setembro 2008.

[13] J. Lafferty, A. McCallum, and F. Pereira, *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*, pp. 282–289. Citeseer, 2001.

[14] C. Freitas, C. Mota, D. Santos, H. G. Oliveira, and P. Carvalho, "Second harem: Advancing the state of the art of named entity recognition in portuguese," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pp. 3630–3637, May 2010.

[15] C. Freitas and D. Santos, "Blogs, Amazônia e a Floresta Sintá(c)tica: um corpus de um novo gênero?," in *Atas do ELC2010*, 2012.

[16] H. Gonçalo Oliveira, D. Santos, and P. Gomes, "Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação," *Linguamática*, vol. 2, pp. 77–93, Maio 2010. Nova versão, revista e aumentada, da publicação Gonçalo Oliveira et al (2009), no STIL 2009.

[17] L. Ferreira, A. Teixeira, and J. P. da Silva Cunha, *REMMA - Reconhecimento de entidades mencionadas do MedAlert*, pp. 213–229. Linguateca, Dezembro 2008.

[18] W. Cohen, P. Ravikumar, and S. Fienberg, "A comparison of string metrics for matching names and records," in *ACM International Conference on Knowledge Discovery and Data Mining (KDD) 09, Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, pp. 73–78, Citeseer, 2003.

[19] P. Christen, "A comparison of personal name matching: Techniques and practical issues," *Sixth IEEE International Conference on Data Mining Workshops ICDMW06*, no. September, pp. 290–294, 2006.

[20] W. E. Winkler, W. E. Winkler, and N. P, "Overview of record linkage and current research directions," tech. rep., Bureau of the Census, 2006.

[21] V. Levenshtein, "Binary codes capable of correcting deletions and insertions and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[22] M. A. Jaro, "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 414–420, 1989.

[23] W. E. Winkler, "String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage," *Proceedings of the Section on Survey Research Methods American Statistical Association*, pp. 1184–1187, 1990.

[24] M.-G. Butuc, "Semantically enriching content using OpenCalais," *Interface*, vol. 9, no. 2, pp. 77–80, 2009.

[25] B. Spetic, "Zemanta opening its semantic API and announcing commercial package," *Group*, 2008.

[26] E. Cano, G. Burel, A.-S. Dadzie, and F. Ciravegna, "Topica: A tool for visualising emerging semantics of POIs based on social awareness streams," in *10th International Semantic Web Conference (ISWC2011) (Demo Track)*, 2011.

[27] D. Dearman and K. N. Truong, "Identifying the activities supported by locations with community-authored content," in *Proc. of the 12th ACM int. conference on Ubiquitous computing*, Ubicomp '10, (New York, NY, USA), pp. 23–32, ACM, 2010.

[28] A. N. Alazzawi, A. I. Abdelmoty, and C. B. Jones, "What can I do there? Towards the automatic discovery of place-related services and activities," *Int. Journal of Geographical Information Science*, vol. 26, no. 2, pp. 345 – 364, 2012.

[29] M. Russell, *Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites*. Head First Series, O'Reilly Media, 2011.

[30] J. Maroco, *Análise estatística com utilização do SPSS*. Sílabo, 3ª ed., 2007.

[31] A. Field, *Discovering Statistics Using SPSS*. SAGE Publications, 2005.

Apêndices

Apêndice A

Estudo das Fontes de Enriquecimento Semântico

Tabela A.1: Quadro comparativo das Fontes de Enriquecimento Semântico estudadas

Classificação	SAPO	PAPEL	Wikipedia	Lifecooler	Factual	Gowalla	Foursquare	Google Places	Facebook
Natureza	Mista	Institucional	Colaborativa	Institucional	Institucional	Colaborativa	Colaborativa	Colaborativa	Colaborativa
Estrutura	Estruturada	Estruturada	Semi-Estruturada	Semi-Estruturada	Estruturada	Estruturada	Estruturada	Estruturada	Estruturada
Acesso	Registado / Pago	Livre	Livre	Livre	Registado	Registado	Registado	Registado	Registado
Informação	Geográfica, Nominal, Categorias e Atributos	Nominal	Geográfica, Nominal, Categorias e Atributos	Geográfica, Nominal, Categorias e Atributos	Geográfica, Nominal, Categories e Atributos	Geográfica, Nominal, Categorias e Atributos	Geográfica, Nominal, Categorias e Atributos	Geográfica, Nominal, Categorias e Atributos	Geográfica, Nominal, Categorias e Atributos
Domínio	Genérica	Genérica	Genérica	Genérica	Genérica	Genérica	Genérica	Genérica	Genérica
Geografia	Nacional	-	Internacional	Nacional	Internacional	Internacional	Internacional	Internacional	Internacional
Tipo de Entidades	POIs e Eventos	-	POIs e Eventos	POIs e Eventos	POIs	POIs	POIs	POIs	POIs e Eventos
Organização das Categorias	Lista	-	Taxonomia	Lista e Conjunção de Categorias	Lista e Conjunção de Categorias	Taxonomia	Taxonomia	Lista	Taxonomia
Meio de Extracção	API	Dump	Dump ou API	Web Scraping	API	API	API	API	API

Apêndice B

Modelos de Dados

Tendo em conta a especificação dos dados a extrair das fontes de enriquecimento semântico (3.1.4), foi criado o modelo de dados do serviço, presente na figura B.1.

A entidade *Resources* representa a estrutura de um recurso recebido/devolvido pelo módulo de enriquecimento semântico. Este contém os campos que podem ser obtidos através do processo de enriquecimento semântico assim como os valores possíveis. Deste modo, o processo pode ser gradual, actualizando o recurso à medida que obtemos dados relevantes. O recurso pode ou não ser georeferenciado, podendo conter dados como localização espacial (*geometry*), morada, código postal, cidade, país e informações sobre como chegar ao local (transportes públicos, estradas). Esta entidade contém ainda uma relação "many-to-many"para si própria, visto que um recurso pode estar relacionado com vários outros recursos (Ex.: Um POI inserido na mesma rota turística que outro conjunto de POIs). Existe ainda outra relação semelhante a esta, mas neste caso cinge-se a situações como um recurso fazer parte de outro recurso (Ex.: Uma peça de arte pertencente a um museu).
A maioria das restantes entidades podem ser classificadas como parte integrante do recurso, estando assim separadas de modo a permitir uma relação do tipo "one-to-many", onde cada recurso pode ter vários elementos das outras entidades (Ex.: Um recurso pode conter diversos *websites* alusivos a este, diversas críticas de utilizadores e até mesmo várias fontes de enriquecimento, caso os dados tenham sido obtidos através de diversas fontes). Insere-se neste conjunto a entidade *Reviews* (que consiste em críticas feitas por indivíduos a um certo recurso), *Websites* (sítio web oficial ou não-oficial do recurso), *Descriptions* (descrições sobre o recurso), *EnrichmentSources* (fontes de enriquecimento usadas para enriquecer o recurso em questão).

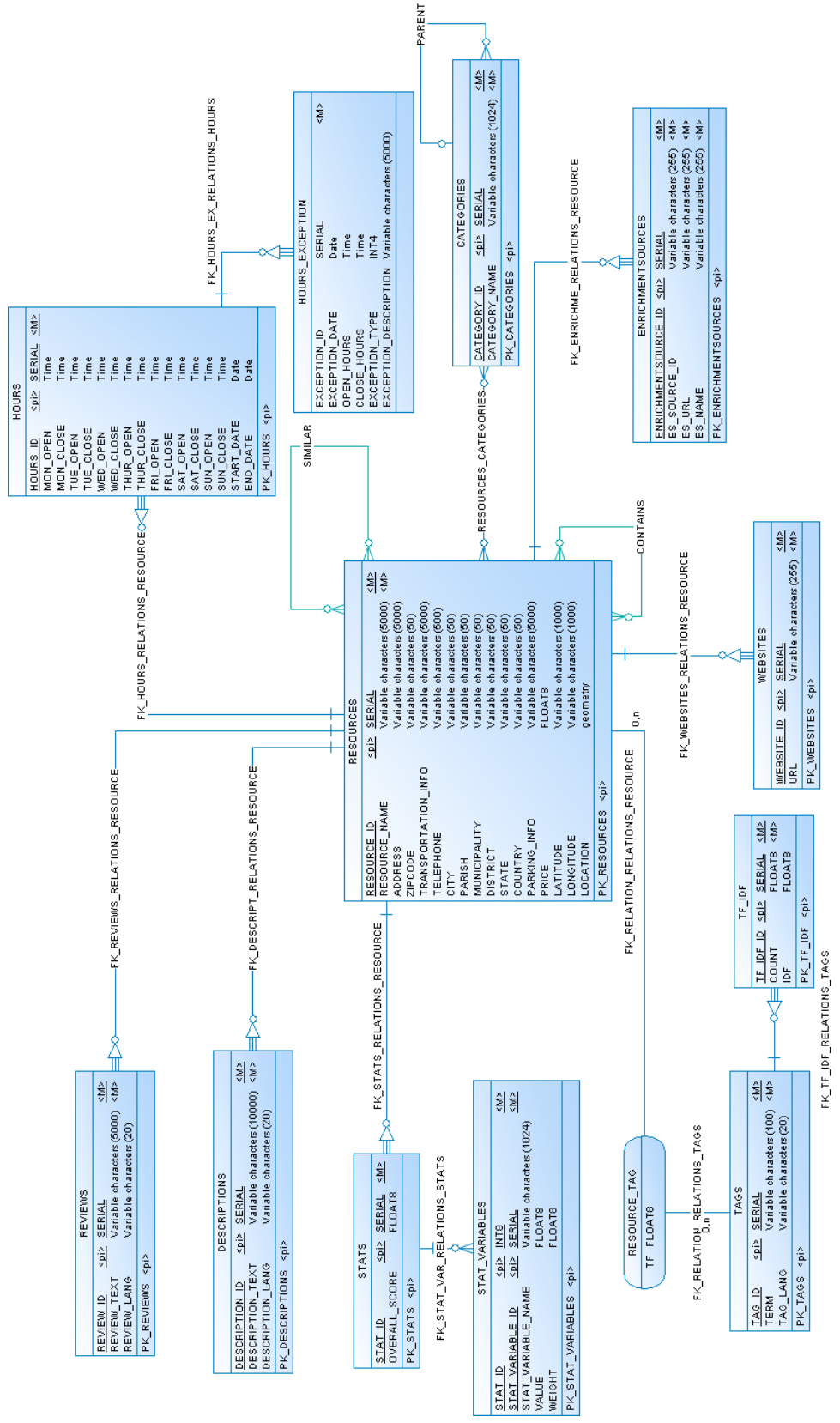

Figura B.1: Modelo de dados

A entidade *Tags* (termos associados ao recurso) encontra-se associada a um recurso através de uma relação "many-to-many", visto que um determinado conceito pode servir para expressar mais do que um recurso e vice-versa (um recurso pode ser expresso por vários termos). Esta entidade encontra-se ainda associada ao peso de cada termo, calculado através do TF-IDF.

A entidade *Hours* foi criada com inspiração no modelo de dados que a API do *Facebook* fornece para representar os horários de um determinado POI. Consegue ser específica ao ponto de discriminar diferentes horários para cada dia da semana, o que em algumas situações pode vir a ser necessário. A esta entidade alia-se ainda a entidade *Hours_Exception*, que nos permite especificar uma excepção nos horários de um POI (o dia de encerramento semanal, ou outra data incomum). Cada recurso pode ter uma destas entidades associada.

A entidade *Stats* reúne dados estatísticos provenientes das diversas fontes de enriquecimento, que nos permitem inferir a popularidade de um dado recurso. De modo a assegurar uma forma genérica de adição de variáveis para obter estes dados, criou-se uma entidade auxiliar, *stats_variables*, que nos permite inserir dinamicamente outras métricas de popularidade.

De modo a agregar todas as variáveis obtidas sobre a popularidade de um determinado recurso, criou-se um campo chamado *overall_score*, que será o resultado de um cálculo ponderado sobre a popularidade do recurso.

Apêndice C

Especificação da API REST

Nesta secção são descritas as chamadas disponíveis à API do módulo de Enriquecimento Semântico, assim como os parâmetros necessários em cada uma. O formato de resposta é sempre JSON, sendo necessário especificar o cabeçalho "Content-type=application/json"aquando do envio de dados neste formato (Ex.: Num pedido POST com dados em JSON no corpo).

Actualmente, a API está disponível no seguinte endereço:

`http://ubiquo.dei.uc.pt:8080/SE/`

As operações disponíveis através da API REST e respectivos parâmetros são descritos na lista abaixo:

- **GET /SE/resources**
  Devolve uma lista de recursos contidos numa determinada área, podendo esta ser representada por uma caixa delimitadora[1] ou por uma área circular.
  Parâmetros:
  Os parâmetros deste *endpoint* podem ser fornecidos de duas maneiras distintas:
  - **Área circular**
    - *latitude*: Latitude do ponto central de onde se pretendem recolher recursos.
    - *longitude*: Longitude do ponto central de onde se pretendem recolher recursos.
    - *radius*: Raio de acção (em metros).

[1] Mais informações em: `http://wiki.openstreetmap.org/wiki/Bounding_Box`

  - **Caixa delimitadora**
    - *minLatitude*: Latitude do ponto sudoeste.
    - *minLongitude*: Longitude do ponto sudoeste.
    - *maxLatitude*: Latitude do ponto nordeste.
    - *maxLongitude*: Longitude do ponto nordeste.

  Exemplo:
  `http://ubiquo.dei.uc.pt:8080/SE/resources?latitude=38.70734&longitude=-9.14309&radius=100`
  ou
  `http://ubiquo.dei.uc.pt:8080/SE/resources/nearby?minLatitude=38.704302228778594&minLongitude=-9.152668174743667&maxLatitude=38.7176971437664&maxLongitude=-9.115331825256362`

- **GET /SE/resources/id**
  Devolve um recurso dado o seu identificador (atribuído pela base de conhecimento).
  Parâmetros:
  - *id*: Identificador do recurso.

  Exemplo:
  `http://ubiquo.dei.uc.pt:8080/SE/resources/18500`

- **GET /SE/resources/id/enrich**
  Despoleta o processo de enriquecimento semântico. Este pedido pode ser feito de forma síncrona ou assíncrona, dependendo das necessidades da origem do pedido.
  Parâmetros:
  - *id*: Identificador do recurso.

  Exemplo:
  `http://ubiquo.dei.uc.pt:8080/SE/resources/18500/enrich`

- **POST /SE/resources**
  Insere um recurso na base de conhecimento, mediante uma representação deste no formato JSON no corpo do pedido.

Cabeçalho: *Content-type=application/json*

Corpo do pedido:

Exemplo de Código C.1: Conteúdo do ficheiro resource.json

```
{
  "address":"Sacavem",
  "name":"Estacao Ferroviaria de Sacavem",
  "location":"0101000020957F0000D52E41DADAAF2F411A5C40B8EC8C4EC1",
  "country":"Portugal",
  "stats":{
    "userCount":{
      "name":"userCount",
      "value":38.0,
      "weight":0.0
    },
    "checkinsCount":{
      "name":"checkinsCount",
      "value":107.0,
      "weight":0.0
    }
  },
  "categories":[{
    "name":"Train Station"
    }],
  "sources":[{
    "name":"Foursquare",
    "link":"https://foursquare.com/v/4cc5c16b3d7fa1cdd6a0b35f",
    "idOnSource":"4cc5c16b3d7fa1cdd6a0b35f"
  }],
  "city":"Loures",
  "latitude":"38.795742077583235",
  "longitude":"-9.099510435878573"
}
```

Exemplo:

Executar a seguinte instrucção na linha de comandos[2]:

Exemplo de Código C.2: Pedido POST para inserção de recursos

[2]A aplicação cURL necessita de estar instalada, mais informações em `http://curl.haxx.se/`

```
curl -v -H "Accept: application/json" -H "Content-type: application/json
    " -X POST -d resource.json  http://ubiquo.dei.uc.pt:8080/SE/
    resources
```

- **GET /SE/resources/source/name**
  Devolve uma lista de recursos já presentes na base de conhecimento extraídos de uma fonte de dados dado o nome desta.
  Parâmetros:
  - *name*: Nome da fonte de dados.

  Exemplo:
  `http://ubiquo.dei.uc.pt:8080/SE/resources/source/LifeCooler`

- **GET /SE/tags**
  Devolve uma lista de conceitos obtidos através dos recursos de uma determinada área, ordenados por relevância.
  Parâmetros:
  - *latitude*: Latitude do ponto central de onde se pretendem recolher recursos.
  - *longitude*: Longitude do ponto central de onde se pretendem recolher recursos.
  - *radius*: Raio de acção (em metros).

  Exemplo:
  `http://ubiquo.dei.uc.pt:8080/SE/tags?latitude=38.70734&longitude=-9.14309&radius=300`

- **GET /SE/resources/id/tags**
  Devolve uma lista de conceitos obtidos através dos conteúdos textuais do recurso de identificador id, ordenados por relevância.
  Parâmetros:
  - *id*: Identificador do recurso.

  Exemplo:
  `http://ubiquo.dei.uc.pt:8080/SE/resources/18500/tags`

- **POST /SE/resources/id/tags**

  Despoleta o processo de extracção de termos sobre os conteúdos textuais do recurso de identificador id, fazendo também o cálculo de relevância dos mesmos

  Parâmetros:

  – *id*: Identificador do recurso.

  Exemplo:

  Executar a seguinte instrucção na linha de comandos:

  Exemplo de Código C.3: Pedido POST para despoletar o processo de extracção de termos

```
curl -X POST http://ubiquo.dei.uc.pt:8080/SE/resources/18500/tags
```

Apêndice D

Instalação do Módulo de Enriquecimento Semântico

O módulo é distribuído sobre o formato *WAR*[1], num único ficheiro onde todos os elementos necessários para a utilização do serviço estão compreendidos.

Antes de instalar o módulo é necessário criar um ficheiro de configuração onde constam as informações dinâmicas necessárias de modo a que seja possível comunicar com os outros módulos do PPS2, mais especificamente o módulo de Interoperabilidade Semântica. Este trata-se de um ficheiro *.properties*[2], e dado que actualmente o módulo de IS não suporta todas as operações necessárias ao bom funcionamento do nosso serviço, é necessário, para além do endereço deste incluir algumas propriedades relacionadas com a base de dados espacial PostGIS. A estrutura do ficheiro de configuração pode ser vista no exemplo de código D.1.

Exemplo de Código D.1: Exemplo do ficheiro de configuração

```
# Properties to define the required Semantic Enrichment connections
# Spatial Database Connection
db.driverClassName = org.postgresql.Driver
db.url = localhost
db.name = SE
db.username = postgres
db.password = secretPassword
# IS Module URL
```

[1]Mais informações em: http://en.wikipedia.org/wiki/WAR_file_format_(Sun)
[2]Mais informações em: http://en.wikipedia.org/wiki/.properties

```
is.url = localhost:8080/onto
```

Estes campos devem ser preenchidos conforme as configurações da máquina onde se pretende instalar o módulo de enriquecimento semântico, desde as credenciais de acesso à base de dados espacial até ao endereço onde está disponível o módulo IS. O ficheiro deve ser colocado numa pasta intitulada "*.se*"contida na pasta pessoas do utilizador responsável pelo módulo (Em ambientes *UNIX* "*~/.se/*"ou em ambientes *Windows* "C:/Users/user/.se/"[3]) sob o nome "*semanticenrichment.properties*".

Após criado este ficheiro de configuração, basta instalar o ficheiro *.war* num *container* adequado, como o *Tomcat*[4] ou o *Jetty*[5]. Este irá proceder à descompressão do ficheiro ficando o serviço pronto a utilizar no endereço mencionado pelo *container*.

[3]Sendo "user"o nome do utilizador.

[4]Mais informações em: `http://tomcat.apache.org/`

[5]Mais informações em: `http://jetty.codehaus.org/`