FACULDADE ENGENHARIA

Departamento de Informática

Plataforma Online para Extração Eficiente de Padrões Textuais

Proposta de Projeto

Orientador: Sebastião Pais(sebastiao@di.ubi.pt)

Objetivos

Information Retrieval (IR) é uma área atualmente em expansão dada a quantidade crescente de documentos, em formato eletrónico, disponível na Web e em outros tipos de redes ou sistemas similares. Esta imensidão de documentação necessita ser tratada e classificada para que possa ser acedida duma forma fácil e rápida.
Os motores de pesquisa de documentos na Web (e.g Google, Bing), são a face mais visível da aplicação da investigação e desenvolvimento efetuados em IR. No entanto, podem ser considerados como sistemas ainda muito primitivos se atendermos ao modo como os documentos estão classificados, basicamente recorrendo a palavras-chave, ao formato admitido para os pedidos de informação , expressões lógicas simples do tipo "palavra and palavra or palavra"e à, normalmente, baixa precisão e cobertura das respostas, obrigando o utilizador a perder muito tempo até obter a resposta desejada. São sistemas que ainda estão mais virados para os dados (i.e. documentos, palavras) que para a informação lá contida (i.e. conceitos, contextos). Nos próximos anos, é essencial que o foco passe a estar, cada vez mais, na informação e menos nos dados. Para tal, são precisas ferramentas cada vez mais poderosas, eficazes e eficientes, para a organização, classificação e pesquisa de informação em documentos.
É no contexto desta nova realidade que os sistemas de extração automatizada de termos, expressões relevantes ou *MWUs*, ganham importância. Estes sistemas não são mais que aplicações que permitem a identificação automática de sequências, contíguas ou não contíguas, de unidades lexicográficas (e.g. palavras, sinais de pontuação) que constituam uma *MWU*, ou seja, que estejam associadas a um conceito perfeitamente identificável. Estes padrões textuais são, por exemplo, substantivos compostos, expressões idiomáticas, verbos compostos, locuções preposicionais, ou locuções adverbiais. Genericamente, são expressões que ocorrem mais vezes do que o simples acaso faria prever.
A extração de *MWUs* é importante não só para a classificação e indexação de documentos mas para muitas outras áreas como seja a tradução automática, ou o alinhamento de textos paralelos.
Existem, basicamente três abordagens para a extração de *MWUs* dum documento ou dum conjunto de documentos agrupados num corpus:

1. Utilização de técnicas baseadas em métodos linguísticos;
2. Utilização de métodos puramente estatísticos onde o reconhecimento de *MWUs* é um processo totalmente independente da língua base aos documentos;
3. Um misto das duas anteriores onde são estabelecidos limiares a partir dos quais certos padrões textuais são assumidos como relevantes.

Exemplos da primeira e da terceira vias são a utilização de padrões ou modelos linguísticos e ou de etiquetas morfosintáticas para a extração de termos. São métodos dependentes da língua ou então dependentes de heurísticas baseadas na utilização de determinados padrões ou sequências de tipos de palavras ou de etiquetas. Obrigam à existência de bases de dados, atualizadas, de padrões linguísticos e ao desenvolvimento de medidas e dos respectivos limiares de aceitação.
A segunda via tem a vantagem de ser completamente independente de qualquer conhecimento prévio sobre a língua e sobre a estrutura do corpus a analisar. Existem sistemas propõem um método estatístico para a identificação de MWUs baseado no cálculo eficiente do número de ocorrências de cada sufixo do texto em análise.
Assim, este projeto tem como objetivo conceptualizar, construir e disponibilizar à comunidade cientifica uma plataforma online para extração eficiente de padrões textuais.

Tarefas a Realizar e Cronologia

T1 Investigação Preliminar e Especificação de Requisitos Iniciais;

T2 Investigação, Conceptualização e Desenvolvimento Experimental de métodos não supervisionadas e independentes da língua, para extração de padrões textuais;

T3 Conceptualização e Desenvolvimento de uma Plataforma Web para disponibilização de serviços;

T4 Integração, Testes e Avaliação;

T5 Escrita do relatório de projeto.

Requisitos Técnicos / Académicos

Interesse pela área de Inteligência Artificial (sub-área de Processamento da Linguagem Natural), programação Web.

Elementos de Avaliação a Entregar

Para além do relatório, o(a) aluno(a) deverá entregar todos os *scripts* e código fonte desenvolvido.

Resultados Esperados

* Uma plataforma web;
* O levantamento do estado da arte e trabalhos relacionados;
* Relatório de projeto.

Contactos

Sebastião Pais (sebastiao@di.ubi.pt)