Universidade do Minho
Escola de Engenharia

Introdução ao Processamento de Linguagem Natural
Trabalho Prático 3

Programa Split-words

Relatório de Desenvolvimento

José André Martins Pereira
(a82880@alunos.uminho.pt)

Ricardo André Gomes Petronilho
(a81744@alunos.uminho.pt)

3 de Fevereiro de 2020

0.1 Contextualização e objetivos

Na unidade curricular de Introdução ao Processamento de Linguagem Natural, foram-nos propostas várias opções para o trabalho prático três.

Desta forma, a equipa escolheu a opção dois, que consiste no desenvolvimento de um programa, que recebe um texto, onde as palavras se encontram juntas ("coladas"). O objetivo do programa desenvolvido, consiste em determinar as possíveis palavras desse texto, colocando espaços no mesmo.

0.2 Solução para o problema

A solução encontrada pela equipa, para satisfazer os requisitos dos docentes para a opção dois, consiste na construção de *N-Gramas* (N variável), sendo este um treino/conhecimento, feito através de textos, para ser utilizado no programa **Split-words**.

A execução do programa necessita sempre deste treino. No entanto, para facilitar o processo, bem como reduzir o tempo de execução, guarda-se o conhecimento em ficheiros, para reutilização.

O tipo de dados utilizado para guardar o treino, consiste num dicionário, onde a chave é a combinação de palavras, e o valor o número de ocorrências da mesma, nos textos utilizados para treino.

Desta forma, o programa, percorre o dicionário, isto é, as chaves (combinações de palavras) do mesmo, e procura-as no texto recebido, que contém as palavras juntas, substituindo as mesmas (colocando espaços), quando ocorre *match*. Importante referir, como as palavras do texto encontram-se juntas, as chaves do dicionário também são colocadas dessa forma, isto é, juntas, no entanto, numa cópia, mantendo-se o estado original no dicionário. Quando ocorre *match*, visto que se guardou o estado original no dicionário, basta substituir a chave no texto.

No entanto, não se consegue garantir grande precisão, pois, pode existir confusão com determinadas combinações de palavras, como: *"com o", "como"*. Do mesmo modo, a qualidade do treino torna-se muito importante para obtenção de bons resultados, tanto em quantidade, como no tipo de linguagem. A solução possível para resolução do problema referido acima, seria a avaliação das palavras próximas, isto é, a identificação do tipo de palavras (verbos, nomes, determinantes, etc ...), antes e depois da palavra em questão. No entanto, a equipa não teve tempo para implementar esta possível solução.

Como funcionalidade adicional, ou melhoria de performance, decidiu-se garantir a possibilidade de se aplicar o programa a vários ficheiros. No entanto, tal como referido, foram feitas melhorias de performance a este nível, pois, a execução de vários ficheiros realiza-se em paralelo. Na verdade, são criadas tantas *Threads*, quanto o número de ficheiros, reduzindo-se aproximadamente o tempo em $\mathbf{y}$ (y = número de ficheiros) vezes menor.

0.3 Observações

A realização do teste prático do programa deve ser efetuada, lendo previamente o ficheiro **README**, onde se especifica todas a funcionalidades disponíveis, os argumentos necessários, entre outras notas importantes.

0.4 Conclusões

Em suma, o programa final, não obtém resultados excelentes, tal como referido acima, devido à necessidade de uma melhor avaliação das palavras em questão, no entanto, conclui-se, mesmo assim, que se obtém

bons resultados.

A realização deste trabalho, permitiu, a aplicação dos conhecimentos sobre *N-Gramas*, e uma melhor percepção da utilidade dos mesmos, para diversos problemas. Do mesmo modo, o trabalho contribuiu para o melhor conhecimento da ferramenta *Python* e poder da mesma.

Outros conhecimentos lecionados na unidade curricular, foram aplicados, tais como: expressões regulares, tipos de dados utilizados (dicionário), ordenações, obtenção de argumentos da linha de comandos.

No entanto, neste trabalho prático, decidiu-se explorar um pouco o paradigma de orientado a objetos, criando-se classes necessárias e respetivas instâncias.