



Universidade do Minho

Mestrado Integrado em Engenharia Informática

Introdução ao Processamento de Linguagem Natural

Trabalho Prático 1

António Gonçalves (A85516)
Ricardo Costa (A85851)

25 de março de 2021

Conteúdo

1	Tema	3
2	Resolução	4
2.1	Menu	4
2.2	Inserir nome do ficheiro	4
2.3	Contagem de ocorrências	5
2.4	<i>Blacklist</i> de palavras	5
2.5	Número de capítulos	5
2.6	Assuntos do texto (frequências relativas logarítmicas)	6
2.7	Encontrar uma palavra e o seu contexto	7
2.8	Tratamento de exceções	7
3	Conclusão	8
4	Bibliografia	9

1 Tema

O tema escolhido pelo nosso grupo foi o tema 1.a. (identificar os principais temas de um texto através do uso das frequências relativas logarítmicas de cada palavra).

Além de conseguirmos identificar os principais assuntos de um texto, decidimos também desenvolver outras funcionalidades no nosso programa a partir daquilo que também tínhamos vindo a trabalhar nas aulas, já que faz tudo parte do mesmo tema de analisar texto e conseguirmos obter informações úteis sobre o mesmo.

Então, tendo por base aquilo que fizemos nas aulas, além de mostrarmos os principais assuntos de um dado texto, desenvolvemos também as seguintes funcionalidades:

- Permitir ao utilizador inserir o nome do ficheiro que contém o texto que pretende analisar (para permitir analisar qualquer texto que o utilizador pretenda)
- Contagem de ocorrências de cada palavra no texto (valor esse que vamos precisar para calcular a frequência relativa logarítmica)
- Definir uma lista de palavras que não queremos que sejam consideradas para a contagem do número de ocorrências e no cálculo das frequências relativas logarítmicas
- Determinar o número de capítulos do dado texto que estamos a analisar
- Encontrar todas as ocorrências de uma dada palavra e o seu contexto (ou seja, algumas palavras que vêm antes e depois da palavra pedida)

2 Resolução

A resolução deste trabalho pode ser dividida nas seguintes fases:

2.1 Menu

Como decidimos adicionar várias funcionalidades e para permitir que o utilizador pudesse escolher livremente o que quer fazer a seguir, achamos por bem implementar um menu clássico no terminal. Para isso usámos um ciclo *while* com a condição a *True* que apenas sai se o utilizador escolher a opção "q: Sair", acabando com a execução do programa.

```
----- Analisador de texto -----  
1 Introduzir o nome do texto que pretende analisar.  
2 Contagem de ocorrências de cada palavra no texto.  
3 Definir blacklist (palavras a não ser consideradas).  
4 Número de capítulos do texto.  
5 Frequências relativas logarítmicas para inferir os principais temas do texto.  
6 Encontrar uma certa Palavra.  
  
q Sair.  
Escolha uma opção:
```

Figura 1: Menu da aplicação

2.2 Inserir nome do ficheiro

No nosso programa permitimos que o utilizador insira o ficheiro que pretende analisar. Depois do utilizador introduzir o nome do ficheiro, abrimo-lo de forma segura (tal como aprendemos nas aulas), lemos tudo para uma variável e imprimimos uma mensagem de sucesso. Como o utilizador pode introduzir mal o nome do ficheiro, ou seja, não existir, usámos uma cláusula *try* e *except* para apanhar a exceção de *FileNotFoundError*.

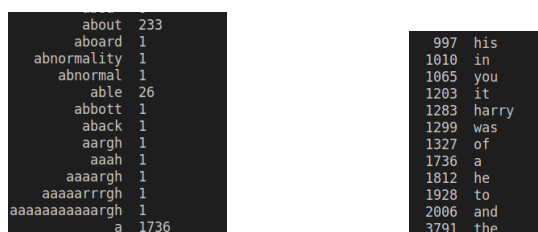
Além disso, tudo isto se encontra dentro de um *while* até que a seguinte condição se verifique: valor da variável em que guardamos o texto deixar de ser nula.

Adicionalmente, outras opções que necessitam do texto para serem corridas informam que o ficheiro ainda não foi lido, caso isto seja verdade, de forma a impedir contagem do texto enquanto este está vazio, por exemplo.

2.3 Contagem de ocorrências

A segunda funcionalidade permite contar o número de ocorrências de cada uma das palavras no texto. Para isso guardamos numa lista tudo o que é palavra e depois percorremos essa lista e vamos adicionando num dicionário (em que a chave é a palavra e o valor o número de ocorrências dessa palavra) as ocorrências de cada palavra.

Por fim, é dada a opção ao utilizador de imprimir esta lista por ordem numérica de ocorrências, inserindo "num" no menu, ou imprimir a lista por ordem alfabética, inserindo "alf" no menu.



```
about 233
aboard 1
abnormality 1
abnormal 1
able 26
abbott 1
aback 1
aargh 1
aaah 1
aaaargh 1
aaaaarrgh 1
aaaaaaaaaargh 1
a 1736
```

```
997 his
1010 in
1065 you
1203 it
1283 harry
1299 was
1327 of
1736 a
1812 he
1928 to
2006 and
3791 the
```

Figura 2: Output da contagem por ordem alfabética e por ordem numérica

Estes *prints* foram feitos lendo o livro do Harry Potter que utilizamos nas aulas. O mesmo se aplica aos restantes *prints* do relatório.

2.4 *Blacklist* de palavras

Relativamente a definir as palavras que não sejam consideradas para os efeitos da contagem de ocorrências e da frequência relativa logarítmica foi necessário apenas permitir ao utilizador introduzir as palavras que quer separadas por uma vírgula. Estas palavras são depois guardadas numa lista.

2.5 Número de capítulos

Para conseguirmos contar o número de capítulos de um dado texto foi apenas necessário guardar numa lista todas as ocorrências das palavras "Chapter" seguidas de uma outra palavra ou número referente ao número do capítulo. Permitimos também que todas as outras letras sem ser a primeira apareçam como maiúsculas ou minúsculas. Para termos a certeza que se trata de facto de um novo capítulo e a palavra não está escrita por acaso no texto em si, obrigamos a que apareça no início da linha.

Depois foi só preciso contar o número de elementos da lista para saber o número de capítulos presentes no texto.

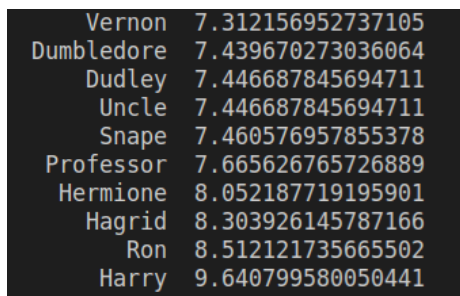
2.6 Assuntos do texto (frequências relativas logarítmicas)

Para podermos determinar os assuntos de um dado texto é necessário calcular a frequência relativa logarítmica de cada palavra no texto em questão e depois comparar o seu valor com o da frequência relativa logarítmica da mesma palavra num dado *corpus*, fazendo a diferença entre os dois valores.

Um *corpus* é uma compilação gigantesca de textos numa dada língua que tem como objetivo servir de dados para estudos realizados sobre os demais temas relacionados com a língua (tal como a sua evolução, etc). No nosso caso, e para não aumentar em demasia o tempo de execução desta funcionalidade, escolhemos apenas uma amostra de um dado *corpus* da língua inglesa através de um *site* referido na bibliografia.

Assim, calculamos a frequência relativa (número de ocorrências / número total de palavras do texto) de cada palavra no texto em questão, multiplicamos por 1 milhão para os resultados serem de leitura mais fácil e fazemos o logaritmo desse valor. De seguida, é realizado o mesmo processo para cada uma das palavras no texto mas na amostra do *corpus*. Por fim, é feita a diferença entre os dois valores e ordenamos por ordem crescente os valores obtidos para cada palavra.

As palavras com os valores mais altos são aquelas palavras que aparecem muito mais no texto em questão do que na compilação de texto aleatória, o que nos permite induzir os temas/assuntos principais do texto.



Vernon	7.312156952737105
Dumbledore	7.439670273036064
Dudley	7.446687845694711
Uncle	7.446687845694711
Snape	7.460576957855378
Professor	7.665626765726889
Hermione	8.052187719195901
Hagrid	8.303926145787166
Ron	8.512121735665502
Harry	9.640799580050441

Figura 3: Algumas das principais palavras do livro

2.7 Encontrar uma palavra e o seu contexto

A última funcionalidade permite encontrar uma dada palavra introduzida pelo utilizador em todas as vezes que ela aparece no texto, mais algumas palavras que venham atrás e à frente dessa palavra.

Para isso é apenas necessário ter numa lista todas as ocorrências da palavra mais o contexto através do uso do *findall* e depois percorrer essa lista e imprimir cada elemento.

```
...e back of his neck. harry knew he shouldn't h...  
...said the very thing harry had been thinking h...  
...ey could see famous harry potter now, he thou...  
... on the newspaper!" harry moved gladly into t...  
...ink cocktail dress. harry washed his hands an...  
...to the living room, harry caught a glimpse of...  
... boy – one sound –" harry crossed to his bedr...  
...998 by warner bros. harry potter characters, ...  
...© warner bros. ent. harry potter publishing r...
```

Figura 4: Exemplo de output para a palavra "Harry"

2.8 Tratamento de exceções

Inicialmente, e porque estamos a tratar de textos em inglês, decidimos considerar as contrações existentes na língua inglesa como, por exemplo, *don't* e *I'm* como uma única palavra. Mas depois ao usarmos a amostra do *corpus* reparamos que essas expressões apareciam com um espaço antes do apóstrofe (ou seja, *do n't* e *I 'm* e então tínhamos duas alternativas: alterar as contrações do texto ou da amostra do *corpus*.

Como o texto escolhido para analisar vai ser sempre substancialmente menor do que a amostra, optamos por alterar as contrações do texto para o formato da amostra quando carregamos o texto para memória (funcionalidade 1 do menu). Assim, criamos um dicionário *excecoes* com as contrações mais frequentes e respetiva alteração e depois de carregar o texto para memória procedemos à substituição através da operação *sub* para cada elemento do dicionário *excecoes*.

3 Conclusão

Com este trabalho conseguimos aprofundar e aplicar algumas das estratégias utilizadas nas aulas para analisarmos texto e compilar todas num único programa com várias funcionalidades.

Apesar de ser uma versão bastante simplificada e termos usado uma amostra bastante reduzida do que seria o tamanho normal de um *corpus* (para facilitar em termos de tempo de execução) para realizar estes tipos de estudos, conseguimos na mesma obter resultados relativamente viáveis sobre quais são os principais assuntos de um texto, que era a tarefa principal deste trabalho.

Relativamente ao que poderíamos melhorar neste trabalho seria:

- permitir que se pudessem analisar textos em português, utilizando uma amostra de um *corpus* da língua portuguesa,
- melhorar a viabilidade do cálculo dos assuntos de um texto utilizando uma amostra maior e que não influenciasse negativamente o tempo de execução
- e outras funcionalidades novas que se poderia acrescentar como por exemplo detetar a língua presente no texto.

4 Bibliografia

- <https://www.corpusdata.org/formats.asp>
- <https://sherlock-holm.es/stories/pdf/a4/1-sided/houn.pdf>
- http://www.getfreestories.weebly.com/uploads/7/9/0/2/79020522/harry_potter_and_the_sorcerers_-_j.k._rowling.pdf