



Universidade do Minho

Mestrado Integrado em Engenharia Informática

Introdução ao Processamento de Linguagem Natural

Trabalho Prático 2

António Gonçalves (A85516)
Ricardo Costa (A85851)

25 de março de 2021

Conteúdo

1	Tema	3
2	Resolução	4
2.1	Leitura	4
2.2	Tratamento	5
2.3	Exportar para JSON	6
2.4	Funcionalidades Adicionais	6
3	Conclusão	8
4	Bibliografia	9

1 Tema

O tema que nós escolhemos foi a análise e alinhamento de *TED Talks*, utilizando uma API específica. No nosso caso, utilizamos a API *ted2srt*, que foi fornecida pelo professor. Como principais objetivos para o trabalho temos:

- Alinhamento de *TED Talks* frase a frase e parágrafo a parágrafo, para um par de línguas;
- Descarregar este mesmo alinhamento para um formato XML ou JSON;
- Permitir descarregar várias *TED Talks* para um par de línguas;
- Criar uma listagem de *TED Talks* a partir de linguagens selecionadas;

2 Resolução

2.1 Leitura

Iremos começar por analisar a API utilizada. No URL "https://ted2srt.org/api/talks" encontramos as informações referentes às várias *TED Talks* presentes na primeira página de "https://ted2srt.org". Esta página contém todas as transcrições mais recentes, tal como se encontram no site oficial das *TED Talks*. Uma vez que esta informação está num formato *json*, a sua leitura e tratamento foi algo simples de fazer. Assim, retiramos alguns dados que são importantes para a leitura do texto das *TED Talks* em si:

- **ID**: A partir do *id* da *Talk* iremos aceder ao seu texto, algo que será explicado mais à frente;
- **Languages**: Lista de todas as línguas em que a *TED Talk* está disponível, bem como os *languageCode* a elas associados;
- **Description**: Descrição geral da *TED Talk*;

Cada "página" do URL inicial tem um total de 20 *TED Talks* (da 0 até à 20, ordenadas da mais recente à menos recente). Assim, caso se queira ler a informação de mais, teremos de utilizar, por exemplo, "https://ted2srt.org/api/talks?offset=40", que nos dá acesso às *TED Talks* entre 21 e 40.

A transcrição de cada *TED Talk* pode ser consultada utilizando:

"https://ted2srt.org/api/talks/ID/transcripts/txt?lang=COD"

Onde ID será o id de cada *TED Talk* (representado por um número de 5 dígitos) e COD será o código associado à língua em que esta está escrita (por exemplo "pt" para Português).

Algo que não é consultável utilizando esta API são as *tags* associadas às *TED Talks* e os *time stamps* que se encontram ao longo das transcrições no site oficial. Posto isto, não conseguimos utilizar esta informação para por exemplo identificar cada parágrafo pelo *time stamp*, ou utilizar a *tag* de cada texto para depois fornecer todas as *TED Talks* com o mesmo tema.

2.2 Tratamento

Após identificar como iríamos retirar as transcrições das várias palestras, para várias línguas, começamos então o seu tratamento para fazer o alinhamento.

Como primeira abordagem utilizamos um *split* por todos os "`\W`". Assim conseguimos dividir o mesmo texto em duas línguas diferentes em parágrafos, sendo o alinhamento dos mesmos algo bastante simples. O "`\W`" serve para identificar tudo o que não seja um carácter ou número antes do "`\n`", já que todos os parágrafos devem acabar com algum tipo de sinal de pontuação. De seguida apresenta-se um exemplo com 2 parágrafos de um texto através do alinhamento por parágrafo:

```
***** Id do texto: 21996, línguas: pt/en *****

Mesmo depois de escrever 11 livros e de ter ganho vários prémios de prestígio, Maya Angelou não conseguia evitar
a dúvida persistente de que não merecia esses enclómos. Albert Einstein também sentia uma coisa semelhante. Des
crevia-se a si mesmo como um "escroque involuntário" cujo trabalho não merecia tanta atenção como a que recebia
----
Even after writing eleven books and winning several prestigious awards, Maya Angelou couldn't escape the naggi
ng doubt that she hadn't really earned her accomplishments. Albert Einstein experienced something similar: he
described himself as an "involuntary swindler" whose work didn't deserve as much attention as it had received
=====
São raros os feitos ao nível de Angelous ou de Einstein mas este sentimento de fraude é extremamente comum. Porq
ue é que tantas pessoas se debatem com sentimentos de que não merecem os seus êxitos ou de que as suas ideias e
talentos não merecem a atenção dos outros
----
Accomplishments at the level of Angelou's or Einstein's are rare, but their feeling of fraudulence is extremel
y common. Why can't so many of us shake feelings that we haven't earned our accomplishments, or that our ideas
and skills aren't worthy of others' attention
```

Figura 1: Alinhamento por parágrafos apresentado no terminal

De salientar que nem todas as traduções têm a mesma estrutura. Alguns textos em certas línguas podem estar bem separados por parágrafos, mas outros estarem em texto corrido. Isto acontece devido ao próprio site oficial das *TED Talks* ter uma estrutura incoerente entre os diferentes línguas. Assim, apesar de conseguirmos alcançar um alinhamento ao parágrafo correto, a própria estrutura que não segue nenhum conjunto de regras entre transcrições não permite que haja divisão dos mesmos para todos os casos.

De forma a fazer o alinhamento frase a frase, foi feito na mesma um *split* através de uma expressão regular em que tentámos englobar a maioria dos casos possíveis. Tal como no caso dos parágrafos, aqui também surgiram ainda mais problemas relativamente à estrutura incoerente entre textos (por exemplo: pontos finais a serem uma vírgula na tradução correspondente). Isto resulta em certos desalinhamentos entre os textos (original e respetiva tradução) de uma dada *TED Talk* para alguns casos.

De qualquer forma, a expressão regular final utilizada para dar *split* entre as frases foi a seguinte:

```
split(r'[\!?\?:|]|--|(?<!\.)\. [\n]|\.{3}[ ]?[\nA-Z]', t2)
```

- `[\!?\?:|]`: para encontrar os sinais "!", "?", ":" e "—"

- --: para encontrar --"
- (?<!\.)\. [\n] : para encontrar um ponto final que não seja precedido de outro ponto final (por causa do caso das reticências) e seguido de um espaço em branco ou "\n"
- \.{3}[]?[\nA-Z] : para o caso das reticências seguidas de um "\n" ou palavra começada por maiúscula.

De seguida apresenta-se um exemplo deste alinhamento no terminal:

```
***** Id do texto: 21996, línguas: pt/en *****

Mesmo depois de escrever 11 livros e de ter ganho vários prémios de prestígio, Maya Angelou não conseguia evitar
a dúvida persistente de que não merecia esses encómios
----
Even after writing eleven books and winning several prestigious awards, Maya Angelou couldn't escape the nagging
doubt that she hadn't really earned her accomplishments
=====
Albert Einstein também sentia uma coisa semelhante
----
Albert Einstein experienced something similar
```

Figura 2: Alinhamento por frases apresentado no terminal

2.3 Exportar para JSON

De forma a melhor apresentar os resultados para ser perceptível o alinhamento, foi exportado o resultado final para um ficheiro em formato JSON. Para isto, a informação foi guardada num dicionário em *python*, que permite uma conversão mais fácil para JSON através da biblioteca *json* existente em *python*. O formato final foi o seguinte:

```
▼ 0:
  ▼ pt: "Mesmo depois de escrever 11 livros e de ter ganho vários prémios de prestígio, Maya Angelou não conseguia evitar a dúvida persistente de que não merecia esses encómios"
  ▼ en: "Even after writing eleven books and winning several prestigious awards, Maya Angelou couldn't escape the nagging doubt that she hadn't really earned her accomplishments"
▼ 1:
  pt: "Albert Einstein também sentia uma coisa semelhante"
  en: "Albert Einstein experienced something similar"
```

Figura 3: Formato do alinhamento em JSON para frases

Como podemos ver, temos uma chave para cada par de frases (original e respetiva tradução) e para cada chave temos 2 outras chaves que identificam a língua cujo valor respetivo (que é a frase na dada língua) se encontra.

2.4 Funcionalidades Adicionais

De forma a facilitar a procura de *TED Talks* com as características desejadas criamos um segundo programa. Assim, fornecendo o código de duas línguas, teremos acesso a todos os IDs das *TED Talks* que têm transcrições para as mesmas.

De forma a ser o máximo interactivo e pratico possível criamos várias *flags* que o utilizador posso dizer aquilo que procura:

- **-l** Esta *flag* terá de vir seguida por dois códigos de linguas separados por `"/"`. Permite encontrar as palestras com as línguas pretendidas;
- **-t** Seguida de um número inteiro múltiplo de 20. Representa o total de palestras que serão lidas. Caso não seja utilizada, é usado o tamanho mínimo, que é de 20;
- **-i** Devolve a lista dos IDs das palestras que têm tradução para as línguas pretendidas;

De seguida temos alguns exemplos de utilizações das flags apresentadas:

```
ID: 68976
Inspired by the rising movement against racism in the US, WNBA champion Renee Montgomery made an unexpected decision: she opted out of her dream job. As she say
s in this stirring talk, she wanted to "make it felt," and that meant turning her attention from the court to the community. But you don't have to be a basketba
ll star to make it felt; anyone can turn important moments into meaningful momentum. How will you?
-----
en English
====
Total de TEDs: 40
```

Figura 4: Execução com a *flag*: `-t 40`

Na figura apresentada temos uma representação de uma palestra, tendo a sua descrição, o seu *id* e as línguas em que se encontra traduzida. Algumas *TED Talks* não têm nenhum código de língua, uma vez que existem algumas em que não há transcrições, apenas vídeo. Com esta *flag* temos então acesso às 40 *TED Talks* mais recentes.

```
70235
69933
69647
69333
68987
```

Figura 5: Execução com as *flags*: `-i -l en/pt -t 40`

Com estas *flags* conseguimos uma lista de todos os *IDs* (-i) de palestras que tenham a língua Portuguesa e Inglesa (-l en/pt), nas 40 mais recentes (-t 40). A ideia principal seria, fazendo um redireccionamento desta execução para um ficheiro, este seria lido pela outra parte do trabalho, fazendo o alinhamento de todas as *TED Talks* com estes *IDs*. Assim, foi adicionada uma nova opção no programa principal, através de uma nova *flag* `-a` que permite ler os *IDs* do ficheiro criado no outro programa e então, alinhar vários textos de uma vez, exportando cada texto nas duas línguas seleccionadas para um ficheiro novo dentro de uma pasta nova que também é criada denominada de "data".

3 Conclusão

Tendo em conta que o objetivo principal deste trabalho seria o alinhamento das palestras podemos dizer que a nossa solução poderia estar mais completa. Contudo, a própria estrutura das transcrições não permite que haja um forma "correta" de fazer a divisão de igual forma para todas as linguagens e todos os textos. Ainda assim, conseguimos obter resultados corretos para algumas *TED Talks*.

Para além disso, implementamos alguns extras que pensamos ser bastante úteis para a consulta das palestras, quer antes ou depois do alinhamento.

Conseguimos também desenvolver as nossas capacidades no que diz respeito à utilização do formato *json*, assim como um estudo mais aprofundado sobre os *requests*.

4 Bibliografia

- <https://www.ted.com/talks>
- <https://ted2srt.org/>
- https://www.w3schools.com/js/js_json_intro.asp