Universidade do Minho

Mestrado Integrado em Engenharia Informática

Introdução ao Processamento de Linguagem Natural

Trabalho Individual

António Gonçalves (A85516)

25 de março de 2021

Conteúdo

1 Tema

O tema que eu escolhi para este Trabalho Individual foi *Named entity-based Text Anonymization*. Para isto será lido um ficheiro de texto, identificando as várias entidades presentes utilizando a biblioteca *spaCy*, sendo depois substituídas pelas respetivas categorias (neste caso, *labels*), seguidas de um identificador, por exemplo, substituir um nome de uma Pessoa por *PERSON1*.

2 Resolução

Para começar, testei alguns textos portugueses utilizando os 3 tamanhos de exemplos que a biblioteca *spaCy* tem acesso e percebi que não se encontra muito desenvolvida. A identificação das entidades era pouco correta, portanto decidi realizar o resto do trabalho em Inglês.

A utilização de exemplos menores(como o *en_core_web_sm*) é bastante mais rápido a identificar as entidades, mas mesmo assim usar *en_core_web_lg* faz com que a análise seja mais correta.

Para realizar este trabalho usei principalmente a utilidade do *spaCy* que identifica e classifica as várias entidades presentes no texto. Após tratar o texto usando *en_core_web_lg*, percorri as várias entidades encontradas, e conforme a sua *label*, ia substituindo pela própria *label*, seguida de um número para as identificar.

De seguido apresento as várias *label* que o *spaCy* permite identificar:

```
PERSON:      People, including fictional.
NORP:        Nationalities or religious or political groups.
FAC:         Buildings, airports, highways, bridges, etc.
ORG:         Companies, agencies, institutions, etc.
GPE:         Countries, cities, states.
LOC:         Non-GPE locations, mountain ranges, bodies of water.
PRODUCT:     Objects, vehicles, foods, etc. (Not services.)
EVENT:       Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART: Titles of books, songs, etc.
LAW:         Named documents made into laws.
LANGUAGE:    Any named language.
DATE:        Absolute or relative dates or periods.
TIME:        Times smaller than a day.
PERCENT:     Percentage, including "%".
MONEY:       Monetary values, including unit.
QUANTITY:    Measurements, as of weight or distance.
ORDINAL:     "first", "second", etc.
CARDINAL:    Numerals that do not fall under another type.
```

Figura 1: Diferentes *labels* disponíveis

Permiti a utilização de várias *flags* para tornar o programa mais flexível e prático. Assim:

- -h Imprime todas as *tags* que o *spaCy* consegue identificar, bem como uma pequena descrição;
- -t Permite escolher o ficheiro de texto a ser lido;

- -i Número de vezes que cada *label* aparece no texto;
- -l Seguido de uma ou mais *label*. Apresenta a lista de entidades dessa *label*, bem como o número de vezes que essa entidade se repete ao longo do texto;
- -p Seguida de uma ou mais *labels*. Substitui todas as entidades das *labels* apresentadas;
- -d Adição à anterior, apresenta um dicionário com os valores associados a cada entidade.

Vou analisar agora a parte principal do trabalho, a substituição e anonimização de entidades, mais detalhadamente.

Uma vez que permito que seja substituída mais do que uma *label* de cada vez, irei, sobre cada uma, percorrer todas as suas entidades, criando assim uma lista com todas as entidades de cada *label*. Finalmente, percorro essas mesmas listas, substituindo o nome da entidade pela *label* em si e o índice da lista em que este se encontra. Por exemplo, imaginemos que a entidade "Harry Potter"seria a primeira com a *label* "PERSON". Após correr o programa neste texto, passaremos a encontrar "PERSON1"em todos os sítios onde a entidade se repete.

Os seguintes prints são obtidos analisando o primeiro capítulo do primeiro livro da Saga "Harry Potter".

```
keeper17@17Keeper:~/Desktop/IPLN/Trabalhos/ti$ ./tpi -i -t "hpc1.txt"
{'PERSON': 193, 'ORG': 6, 'WORK_OF_ART': 11, 'CARDINAL': 33, 'TIME': 16, 'DATE': 21, 'NORP': 8, 'FA
C': 9, 'ORDINAL': 4, 'GPE': 7, 'QUANTITY': 1, 'LOC': 1, 'PRODUCT': 1, 'LANGUAGE': 1}
```

Figura 2: Número de entidades de cada *label*

```
keeper17@17Keeper:~/Desktop/IPLN/Trabalhos/ti$ ./tpi -l PERSON -t "hpc1.txt"
{'JESSICA': 1, 'DI': 1, 'Nicolas Flamel': 1, 'Norbert': 1, 'Dursley': 46, 'Dursleys': 7, 'Dudley':
9, 'Potter': 4, 'Grunnings': 1, 'baker': 1, 'Harry': 16, 'Harvey': 1, 'Harold': 1, 'Muggles': 1, 'N
ext Door': 1, 'Jim McGuffin': 1, 'Jim': 1, 'Ted': 1, 'Howard': 1, 'Albus Dumbledore': 2, 'Dumbledor
e': 35, 'McGonagall': 26, 'Dedalus Diggle': 1, 'Voldemort': 6, 'Pomfrey': 1, 'Lily': 3, 'James Pott
er': 1, 'James': 2, 'Harry Potter': 4, 'Hagrid': 14, 'Sirius Black': 1, 'whiskery': 1}
Total de PERSON encontrados: 32
```

Figura 3: Número de vezes que foi encontrada cada *PERSON*

```
None of them noticed a large, tawny owl flutter past the window.
At half past eight, Mr. PERSON5 picked up his briefcase, pecked Mrs.
PERSON5 on the cheek, and tried to kiss PERSON7 good-bye but missed, because
PERSON7 was now having a tantrum and throwing his cereal at the walls.
```

Figura 4: Exemplo de substituição de PERSONs

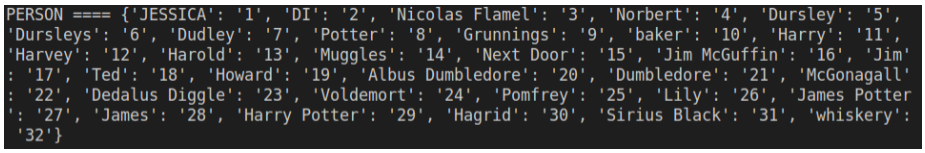

Figura 5: Números associados a cada PERSON

3 Conclusão

Penso que consegui alcançar o principal objetivo do trabalho, que seria implementar anonimização de entidades. Para além disto permiti aos utilizadores decidir quais as entidades a alterar, e implementei algumas funcionalidades adicionais que facilitam a interação com o utilizador.

Para além disto, consegui melhorar o meu conhecimento sobre *python* e explorei mais a biblioteca *spaCy*.

4 Bibliografia

- http://www.getfreestories.weebly.com/uploads/7/9/0/2/79020522/harry _potter _and_the_sorcerers_-_j.k._rowling.pdf
- https://towardsdatascience.com/explorations-in-named-entity- recognition-and-was-eleanor-roosevelt-right-671271117218
- https://spacy.io/