Université Nice Sophia Antipolis

Data Science Seminars

Information Retrieval from the Web

Date: 12 – June - 2017

GONZÁLEZ HUESCA Juan Manuel

Introduction

Information retrieval is an important part of the web science because it helps to find information from unstructured data sources in order to fulfill a knowledge need. There are many techniques and approaches depending on the situation and needs, but it's important to acknowledge its selection will have an impact on the performance and accuracy of the information.

Goal

The objective of the laboratory is to apply the techniques of information retrieval in 10 articles about "BREXIT" to create the term-document incidence matrix, the inverted index matrix and perform some queries.

Development

The first activity was to search and download 10 BREXIT articles and store them in txt format. Once in the local drive, I was able to clean the text and create the term-document incidence matrix as taught in the lecture.

X	Articles/Article1.txt	Articles/Article2.txt	Articles/Article3.txt	Articles/Article4.txt	Articles/Article5.txt	Articles/Article6.txt	Articles/Article7.txt	Articles/Article8.txt	Articles/Article9.txt	Articles/Article10.txt
closeness	0	1	0	0	0	0	0	0	0	0
capture	0	0	1	0	0	0	0	0	0	0
according	1	1	1	1	0	0	0	1	0	0
fact	0	1	1	0	0	0	0	1	0	0
autumn	0	1	0	0	0	0	0	0	0	0
fly	0	0	1	0	0	0	0	0	0	0
£32750	0	0	1	0	0	0	0	0	0	0
unclear	1	0	0	1	0	0	0	1	0	0
reinforced	0	0	0	0	0	0	1	0	0	0
proposals	1	1	0	0	0	0	0	0	0	0
nonexistent	0	0	1	0	0	0	0	0	0	0
farreaching	0	0	1	0	0	0	0	0	0	0
notably	0	1	0	0	0	0	1	0	0	0
bernie	0	0	0	0	0	0	1	0	0	0
commitments	0	0	0	0	0	0	0	1	0	0
coal	0	0	0	0	0	1	0	0	0	0
distributed	0	0	1	0	0	0	0	0	0	0

Figure 1. Term-document incidence matrix

This matrix allows us to discover all the unique words in all the 10 files together, and which one of these words appears on which article. From this matrix, it's possible also to do the inverted index matrix to indicate in a more efficient and clear way what are the documents containing each one of the unique words.

closeness	{2}
capture	{3}
according	{8, 1, 2, 3, 4}
fact	{8, 2, 3}
autumn	{2}
fly	{3}
£32750	{3}
unclear	{8, 1, 4}
reinforced	{7}
proposals	{1, 2}
nonexistent	{3}
farreaching	{3}
notably	{2, 7}
bernie	{7}
commitments	{8}
coal	{6}
distributed	{3}

Figure 2. Inverted index matrix

At this point, as also mentioned in the lecture, it was obvious the term-document incidence matrix is not scalable and the inverted index matrix comes to fill this gap, when we have a huge collection, because we are just saving the index of the file containing the word.

At this point we have enough data to perform the queries with the operators “AND”, “NOT” and “OR” which are mainly based in the merge algorithm.

BREXIT AND negative

```
[4]
```

UKIP AND NOT(bill OR market)

```
[9, 10]
```

(withdrawal OR EU) AND NOT (Scotland OR consequences)

```
[1, 3, 5]
```

Figure 3. Execution and report of queries

Evaluation

The results obtained during the laboratory were impressive, the data preprocessing was very useful to decrease the number of unique words to 4675 and to execute with a high degree of performance all the computation needed to create the matrixes.

Conclusion

Information retrieval is a very important and challenging area of the web science, and it’s good that we, as data science student, are aware of the opportunities areas and that we start to work on some solutions with the approaches explained in the lecture. The exercise provided a good way to consolidate the knowledge acquired during the theoretical section of the seminar, and in an opportunity to discover and get deeper into the development of “real” algorithms which in this case become a first approach into solving this problem.