Mathematics and Big Data

Text Mining - Part 2

Sundus Zafar

Department of Mathematics
Autonomous University of Barcelona

1. Text Mining
   - Vector Space Model
   - Text Classification
   - Corpus Representation
   - Text Visualization
   - Text Visualization with Python
   - Document Similarity
   - Document classification Techniques

Vector Space Model

Document Representation

Definition

Let C be the corpus consisting of n terms $t_i, i = 1, ..., n$, then document d in corpus C would be represented with the vector $d = \{w_1, w_2, ..., w_n\}$, where w_i are weights associated with terms t_i.

- Each $d_i \in C$ is represented as a multidimensional vector.
- Each $t_i \in C$ represents one dimension of the vector space.
- Distance among the documents d_i defines relationship in them in this multidimensional space.
- The document that are close to one another are assumed to have similar meaning in this multidimensional space.

Term Frequency & Inverse Term Frequency

This highlights each word's relevance in the entire document. By evaluating Term frequency and Inverse term frequency we find that:

- How useful a word is to a document.
- How useful a word is to a sentence.
- Helps to ignore irrelevant words.
- Helps to understand common theme in all the documents.

TF - IDF

For a given term w_i and document d_i, let n_{ij} be the number of occurences of w_j in a document d_i, $|d_i|$ is the number of words in document d_i, n is the number of documents and n_j is the number of documents that contain w_j then term frequency can be found as

$$TF_{ij} = \frac{n_{ij}}{|d_i|}, \tag{1}$$

and Inverse term frequency can be found as

$$IDF_j = log\frac{n_j}{n}, \tag{2}$$

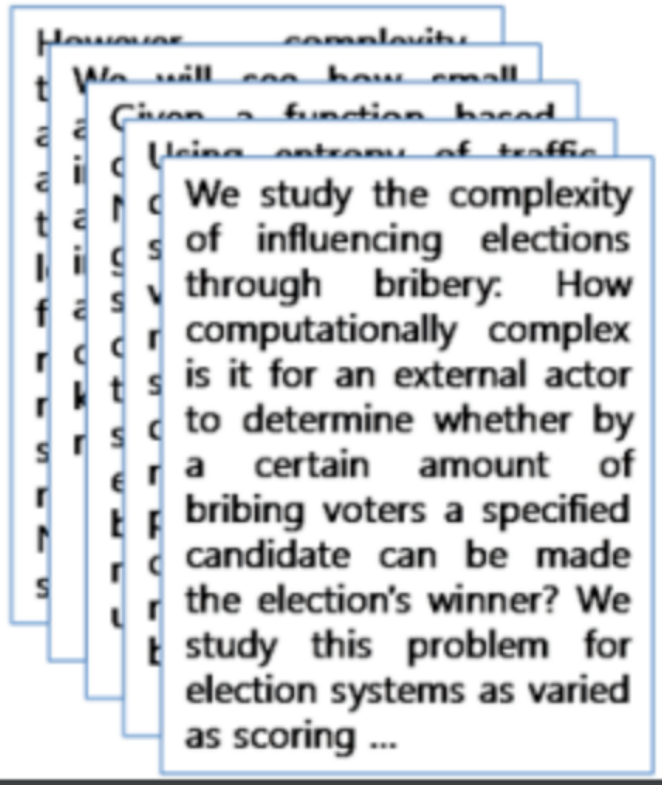
$$x_{ij} = TF_{ij} \cdot IDF_j \tag{3}$$

Term Document Matrix

In vector space model Corpus C is represented in the form of Term Document Matrix, where TDM is an $m \times n$ matrix with the following features:

- Terms $t_i, i = 1, ..., m$ from corpus C are represented as rows.
- Documents $d_j, j = 1, .., n$ from corpus C are represented as columns.
- Cell $i\,j$ stores the weight $w_{i\,j}$ of the term t_i in the document d_j.

Term Document Matrix



Vector-space representation

	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Term-document matrix

Term Document Matrix

In the following a similar numerical representation is given with switch between rows and columns. This form is often called as document term matrix.

Example

Documents	Terms	Camera	Memory	Data	Results	Analysis
	Document 1	0	2	0	1	1
	Document 2	1	0	1	2	0
	Document 3	1	1	0	2	0

TDM with Python

Term Document Matrix TDM can be found by using the following Python libraries:

```
>>> import textmining
>>> import pandas as pd
>>> import sklearn.feature_extraction.text as skt
>>> from skt import CountVectorizer
>>> tdm = textmining.TermDocumentMatrix()
>>> tdm.add_doc(___add document here___)
>>> ____Add code here to get TDM___
>>> docs = [___add documents here___]
>>> vec = CountVectorizer()
>>> X = vec.fit_transform(docs)
>>> ___Add code here to get TDM___
>>> ___Compare the results of two libraries___.
```

WordCloud

A Word Cloud is a visual representation of words that give greater prominence to the most frequent words in the text e.g.

WordCloud with Python

WordCloud can be found by using the following Python libraries:

```
>>> import wordcloud
>>> import matplotlib
>>> import matplotlib.pyplot as plt

>>> doc = [___add your documents here___]
>>> wc = WordCloud().generate(doc)
>>> plt.imshow(wc, interpolation = 'bilinear')
>>> plt.axis("off")
>>> plt.show()
```

Document Similarity

How to measure if two documents are similar?

"Two documents are similar if their vectors are similar."

Document Similarity

It is found by calculating distance between documents.The following distance calculating measures are useful:

1. Distance based matching

Definition

Let $X \in R^n$ and $Y \in R^n$ be two document vectors then the euclidean distance between these two document vectors $X = [x_1, x_2, ..., x_n]$ and $Y = [y_1, y_2, ..., y_n]$ is calculated as

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{4}$$

Document Similarity

2. Angle based matching : Cosine similarity
Documents in the same direction are closely related.

Definition

Let $X \in R^n$ and $Y \in R^n$ be two document vectors then the cosine distance between two vectors $X = [x_1, x_2, ..., x_n]$ and $Y = [y_1, y_2, ..., y_n]$ is defined as

$$d(X,Y) = cos(X,Y) = \frac{X^T Y}{||X|| \cdot ||Y||}$$

$$=> cos(X,Y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2}}$$

Document classification Techniques

- Decision trees
- K - nearest neighbours
- Naive Bayes classifier
- Support Vector Machines
- Neural networks
- Document clustering