IT550: Information Retrieval

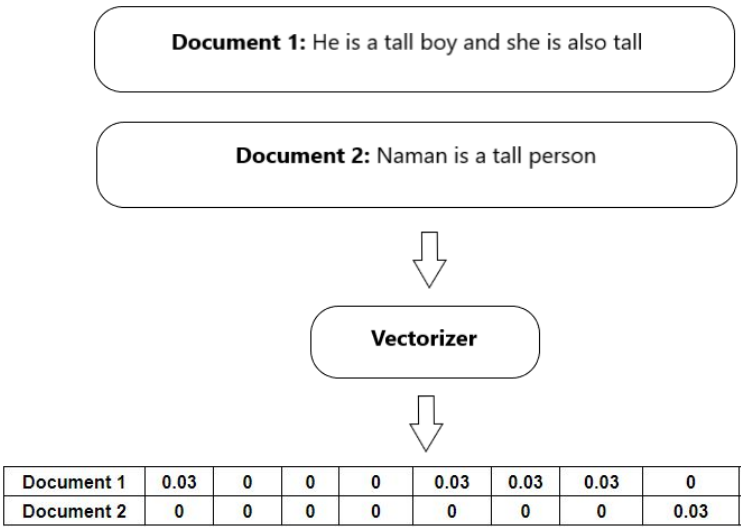
Assignment 2: Document Representation (Part 1)

Instructor: Prasenjit Majumder

Learning Outcome: After this assignment you will learn how to represent document using tf-idf. And also you will learn how to parse XML documents and use it for your further assignments.

1 Problem description

Text Representation is a method that converts text into a numbers form called a vector. Vectorizing text is very useful. For example, the sentence "This man is so tall" given a human. It is easy for a human to understand the sentence as they know the semantics of the words and the sentence. But how will the computer understand this sentence? The computer can understand any data only in the form of numerical value. So, for this reason, we represent the text by projecting it into vector space. Thus the text is represented by a vector.Text representation mainly divided into two types discrete representation and distributed representation. This assignment focuses on discrete representation.



Document 1	0.03	0	0	0	0.03	0.03	0.03	0	0	0.03
Document 2	0	0	0	0	0	0	0	0.03	0	0

2 TF-IDF

Term Frequency-Inverse Document Frequency (tf-idf) technique is divide into two-part.

- Term Frequency (tf): The number of times a term(t) appears in a document(d). It is given by the equation 1

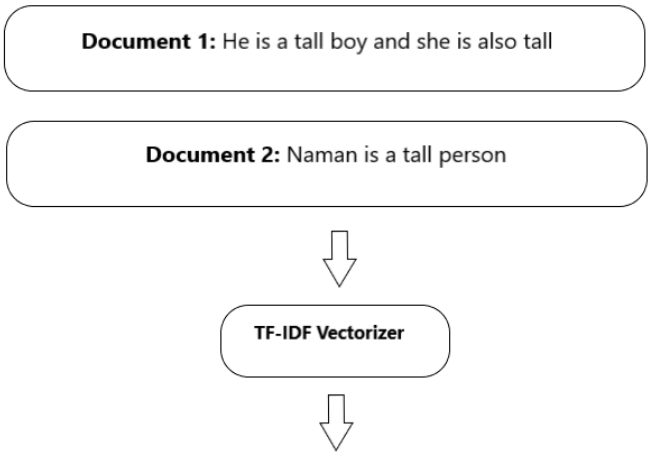
$$tf_{t,d} = \frac{\text{Number of times t appears in d}}{\text{Total number of terms in d}} \tag{1}$$

- Inverse Document Frequency (idf): In this we first calculate document frequency (df_t) which is frequency of a term(t) over a collection of documents(N). We then calculate the inverse document frequency as shown in equation 2

$$idf_t = \log \frac{N}{df_t} \tag{2}$$

By multiplying equation 1 and 2 we get the tf-idf weight of each term given by equation 3

$$tf - idf_{t,d} = tf_{t,d} * idf_t \tag{3}$$



	He	is	a	tall	boy	and	she	Naman	person	also
Document 1	0.03	0	0	0	0.03	0.03	0.03	0	0	0.03
Document 2	0	0	0	0	0	0	0	0.03	0	0

Figure 1: TF-IDF Vectorizer

3 Implementation

3.1 Dataset

- For this assignment we will use Telegraph business news articles, which is in XML format. You can download the dataset from this link: https://drive.google.com/open?id=1VR1bO9D3f3606BBYGWaS5Dt2cXfnzASU
- The file in the dataset is in the following format:

```
<DOC>
<DOCNO> </DOCNO>
<TEXT></TEXT>
</DOC>
```

3.2 Exercise

1. Perform preprocessing steps (stopword removal, stemming, text cleaning etc) on the dataset
2. Implement TF-IDF approach for the dataset. Show the size of the term-document matrix.
3. Use TF-IDF vectorizer from sklearn library to generate TF-IDF for your documents. Show the size of the term-document matrix
4. Pick the first five documents from the list and show the top five words representing each document along with their TF-IDF scores
5. Show these words for your approach as well as using TF-IDF vectorizer.

4 References

- An Introduction to Information Retrieval: Christopher D.Manning ,Prabhakar Raghavan, Hinrich Schütze

4.1 Codes

- https://github.com/mayank408/TFIDF/blob/master/TFIDF.ipynb

4.2 Submission

- You have to submit your assignment in .ipynb notebook with proper comments and explanation of your approach.
- Submit the assignment on the Google Classroom
- **Assignment is due on 9th February 2021 at 11 PM**