IT412: Natural Language Processing

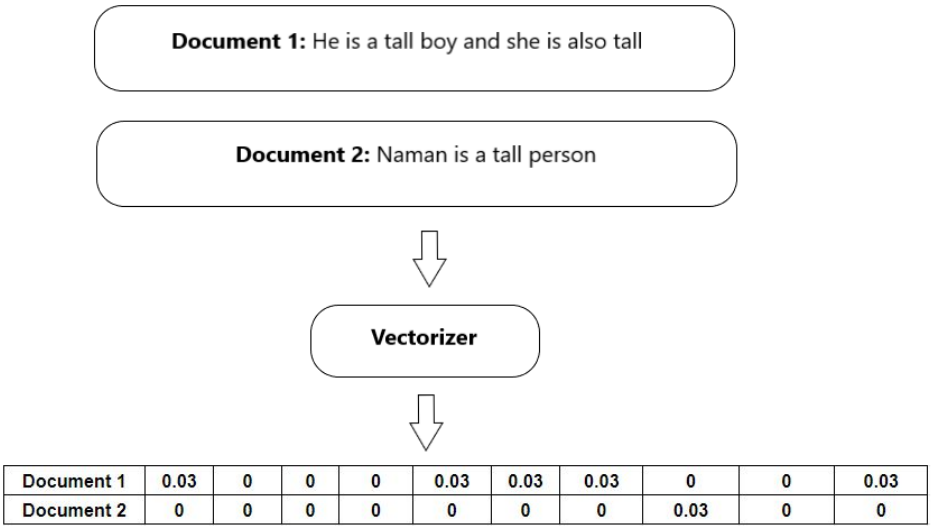
Assignment 1: Document Representation using TF-IDF

Instructor: Prasenjit Majumder

Learning Outcome: After this assignment you will learn how to represent document using tf-idf. How to index large set of corpus using Terrier. 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

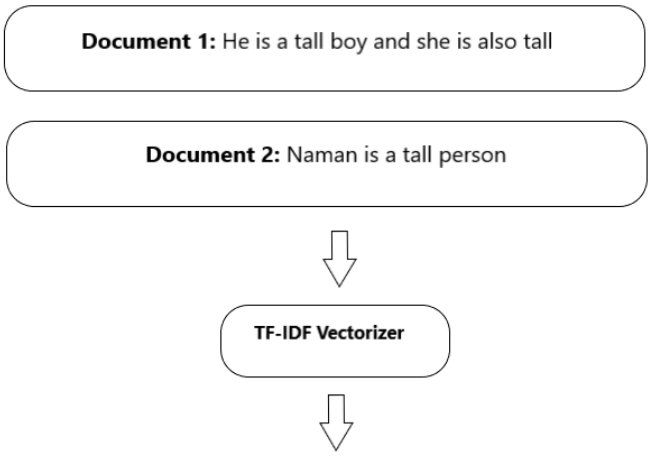
$$tf_{t,d} = \frac{\text{Number of times t appears in d}}{\text{Total number of terms in d}} \tag{1}$$

- 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 news articles, which is in XML format. It contains news on different categories for the year 2004 to 2007. You can download the dataset from this link: FIRE Dataset
- The Queries are in "en.topics.76-125.2010". The query is of the format shown in Figure 1. Use the sentences enclosed in desc tag for framing your query vector
- The "en.qrels.76-125.2010.txt" contains the documents that are relevant to a query. The format of a qrel is such: Query_No Q0 Document ID Relevance score.
- Relevance score is binary 0 or 1. 1 is for relevant, 0 is for otherwise.
- The documents in the dataset is in the format shown in Figure 2.

```
<top lang='en'>
<num>76</num>
<title>Clashes between the Gurjars and Meenas</title>
<desc>
Reasons behind the protests by Meena leaders against the
inclusion of Gurjars in the Scheduled Tribes.
</desc>
<narr>
The Gurjars are agitating in order to attain the status of a
Scheduled Tribe. Leaders belonging to the Meena sect have
been vigorously opposing this move. What are the main reasons
behind the Meenas' opposition? A relevant document should
mention the root cause(s) behind the conflict between these
two sects.
</narr>
</top>
```

Figure 2: Query Format

3.2 Exercise

1. Index the documents of the corpus using Terrier. Use TF-IDF for representing the documents.
2. Use the text in <title> in the queries for getting the TF-IDF representation of the queries.

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

Figure 3: News Format

3. Calculate the cosine similarity between the query and document vectors.
4. For each query retrieve the top 10 documents based on cosine similarity.
5. Evaluate this using the Mean Average Precision (MAP)

4 References

- For knowing about Terrier: http://terrier.org/
- For installation of PyTerrier python version of Terrier on colab: https://pyterrier.readthedocs.io/en/latest/installation.html
- For examples on using PyTerrier: https://github.com/terrier-org/pyterrier
- For reading about Mean Average Precision: https://towardsdatascience.com/breaking-down-mean-average-precision-map-ae462f623a52
- An Introduction to Information Retrieval: Christopher D.Manning ,Prabhakar Raghavan, Hinrich Schütze

4.1 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 23rd August 2021 at 11 PM**