Unit 7
Applications of NLP

Natural Language Processing (NLP)
MDS 555

Objective

- Text Vectorization
- TF-IDF
  - Algorithm
  - Implementation

Text Vectorization

- Text Vectorization is the process of converting text into numerical representation.
- A technique for converting text into finite length vectors
  - Bag-of-Words
  - TF-IDF
  - Word2Vec
- Code the text into the numeric values

IF-IDF

- Term Frequency — Inverse Document Frequency (TF-IDF)
  - A technique for converting text into finite length vectors
  - Gives insights about the less relevant and more relevant words in a document
  - The importance of a word in the text is of great significance in information retrieval

Term Frequency

- It is a measure of the frequency of a word (w) in a document (d).
  - TF is defined as the ratio of a word's occurrence in a document to the total number of words in a document.
  - The denominator term in the formula is to normalize since all the corpus documents are of different lengths.

$$TF(w,d) = \frac{occurences\ of\ w\ in\ document\ d}{total\ number\ of\ words\ in\ document\ d}$$

Term Frequency (TF)

- The initial step is to make a vocabulary of unique words and calculate TF for each document.
- TF will be more for words that frequently appear in a document and less for rare words in a document.

Documents	Text	Total number of words in a document
A	Jupiter is the largest planet	5
B	Mars is the fourth planet from the sun	8

Words	TF (for A)	TF (for B)
Jupiter	1/5	0
Is	1/5	1/8
The	1/5	2/8
largest	1/5	0
Planet	1/5	1/8
Mars	0	1/8
Fourth	0	1/8
From	0	1/8
Sun	0	1/8

Inverse Document Frequency (IDF)

- It is the measure of the importance of a word.
  - Term frequency (TF) does not consider the importance of words.
  - Some words such as’ of’, ‘and’, etc. can be most frequently present but are of little significance.
  - IDF provides weightage to each word based on its frequency in the corpus D.

Inverse Document Frequency (IDF)

- IDF of a word (w) is defined as
  - $\ln = \log_e$

$$IDF(w, D) = \ln\left(\frac{Total\ number\ of\ documents\ (N)\ in\ corpus\ D}{number\ of\ documents\ containing\ w}\right)$$

Inverse Document Frequency (IDF)

- In our example, since we have two documents in the corpus, N=2.

Words	TF (for A)	TF (for B)	IDF
Jupiter	1/5	0	ln(2/1) = 0.69
Is	1/5	1/8	ln(2/2) = 0
The	1/5	2/8	ln(2/2) = 0
largest	1/5	0	ln(2/1) = 0.69
Planet	1/5	1/8	ln(2/2) = 0
Mars	0	1/8	ln(2/1) = 0.69
Fourth	0	1/8	ln(2/1) = 0.69
From	0	1/8	ln(2/1) = 0.69
Sun	0	1/8	ln(2/1) = 0.69

TF-IDF

- It is the product of TF and IDF.
  - TFIDF gives more weightage to the word that is rare in the corpus (all the documents).
  - TFIDF provides more importance to the word that is more frequent in the document.

$$TFIDF\ (w,d,D) = TF(w,d) * IDF(w,D)$$

TF-IDF

Words	TF (for A)	TF (for B)	IDF	TFIDF (A)	TFIDF (B)
Jupiter	1/5	0	ln(2/1) = 0.69	0.138	0
Is	1/5	1/8	ln(2/2) = 0	0	0
The	1/5	2/8	ln(2/2) = 0	0	0
largest	1/5	0	ln(2/1) = 0.69	0.138	0
Planet	1/5	1/8	ln(2/2) = 0	0.138	0
Mars	0	1/8	ln(2/1) = 0.69	0	0.086
Fourth	0	1/8	ln(2/1) = 0.69	0	0.086
From	0	1/8	ln(2/1) = 0.69	0	0.086
Sun	0	1/8	ln(2/1) = 0.69	0	0.086

Why Ln in the IDF?

- TFIDF is the product of TF with IDF.
- Since TF values lie between 0 and 1,
- Not using ln can result in high IDF for some words, thereby dominating the TFIDF. We don't want that, and therefore
- We use ln so that IDF should not completely dominate the TFIDF.

Disadvantage of TFIDF

- It is unable to capture the semantics.
- For example, funny and humorous are synonyms, but TFIDF does not capture that.
- Moreover, TFIDF can be computationally expensive if the vocabulary is vast.

Thank you