Final Report

By :

Ritvik Aryan Kalra
2019115002

Arihanth Srikar Tadanki
2019113004

Base Model

For the base model we had chosen a statistical approach to understand the similarities in the sentences. The approach taken was to use TF-IDF.

TF-IDF

"Term Frequency — Inverse Document Frequency" is abbreviated as TF-IDF. This is a method for calculating the number of words in a collection of documents. We usually assign each word a score to indicate its prominence in the document and corpus.

The formula for TF-IDF is as follows:

$$\text{TF-IDF} = \text{Term Frequency}(\text{tf}) \times \text{Inverse Document Frequency}(\text{idf})$$

We'll start by explain what **Term Frequency** and **Inverse Document Frequency** are.

Term Frequency

This metric determines how often a word appears in a manuscript. This is greatly dependent on the document's length and the word's generality; for example, a relatively common word like "was" can appear many times in a document.

Remember that we need to vectorize the document at the end. When vectorizing papers, we must take into account more than just the words that appear in the content. If we do that, the vector lengths for both texts will be different, making it impossible to compute the similarity. So, we vectorize the documents based on the vocabulary. The corpus's vocabulary is a list of all possible universes.

The formula is

$$\text{tf(t,d)} = \text{count of t in d} / \text{number of words in d}$$

Document Frequency

This metric assesses the value of texts throughout the corpus as a whole. The sole distinction between TF and DF is that TF is a frequency counter for a term t in document d, whereas DF is a count of term t occurrences in the document set N. DF stands for the number of documents in which the word appears.

$$\mathrm{df(t)} = \mathrm{occurence\ of\ t\ in\ N\ documents}$$

Inversed Document Frequency

The IDF is the inverse of the document frequency, which assesses the informativeness of word t. When we calculate IDF, the most often occurring words, such as stop words, will have a very low value (since they are present in almost all of the texts, and N/df will give that word a very low number). Finally, we have what we're looking for: a relative weighting.

$$\mathrm{idf(t)} = N/\mathrm{df}$$

At query time, when the word is not present in is not in the vocab, it will simply be ignored. But in few cases, we use a fixed vocab and few words of the vocab might be absent in the document, in such cases, the df will be 0. As we cannot divide by 0, we smoothen the value by adding 1 to the denominator.

$$\mathrm{idf(t)} = \log(N/(\mathrm{df} + 1))$$

The final formula for TF-IDF comes out to be

$$\mathrm{tf\text{-}idf(t,\ d)} = \mathrm{tf(t,\ d)} \times \log(\mathrm{N}/(df + 1))$$

Stuff left to add, please head to this link

https://rvk7895.notion.site/Final-Report-4e33b9f9ff1a4c4b943faf28473aba68