CS6370 **Natural Language Processing**

Students: Akshat Meena (CS19B052) and Siddharth Singh (CS19B072)
Instructor: Sutanu Chakraborti
Semester: Jan-May 2023

Project Proposal

1.1 Limitations of the Vector Space Model

The Vector Space Model has many inherent limitations:

1. It assumes that the terms are statistically independent of each other.
2. Long documents are poorly represented because they have poor similarity values (vectors with small dot products and high dimensionality).
3. Documents with similar context but different term vocabulary won't be associated, resulting in a "false negative match".
4. Weighting is intuitive but not very formal.

1.2 Experimental Methodology

As of now, we intend to use LSA to do the information retrieval task. So, the methodology would be like:

1. First, the text corpus would be preprocessed to remove noise, stop words, and other irrelevant information. This will involve tasks such as tokenization, lemmatization and stop-word removal.
2. Then, we go on further to generate feature vectors for each document in the corpus. For LSA, we would use a matrix factorization technique such as Singular Value Decomposition (SVD) to generate a set of latent features that capture the statistical relationships between the words in the document.
3. We would test the model with different weighting techniques such as normalised weight and Glasgow model to increase effectiveness.
4. Once the final feature vectors for each document are obtained, we will rank the documents based on their similarity to the query. Many similarity measures, such as cosine similarity or Euclidean distance, can be used. But we would stick with the cosine similarity because it gives the best sense of similarity in the IR context.

1.3 Hypothesis

Let's say that the algorithm used for the project is A_1, and the algorithm used for the Assignment 2 is A_2 (which is basically a naive VSM implementation). Then, our hypothesis would be

A_1 does better than A_2 in the task of finding relevant documents given a query on the nDCG measure at some appropriate rank over the Cranfield dataset under the assumption that the semantics of our document is being governed by some hidden, or "latent" variables that not observed directly after seeing the textual material.

1.4 Evaluation

There are many metrics for evaluating an IR system, such as precision, recall, F1-score, Mean Average Precision (MAP), nDCG (normalized Discounted Cumulative Gain), and Mean Reciprocal Rank (MRR).

But we would mainly focus on the nDCG measure for our purpose since we are given the relevance scores of documents for each query (as part of the Cranfield dataset), and the relevance scores are found by consulting with a large number of domain experts and are hence a better measure for the relevance of a document to a query.

1.5 References

1. Class slides
2. Wikipedia article on Vector space model
3. Article on LSA