CS6370 Project Proposal

- cs17b034 Mooizz, cs17b005 Subhash

- Introduction and the goal of the project

    1. We developed an information retrieval system based on the vector space model on Cranfield Dataset. In this vector space model. In this model, we represented Documents and Queries as vector forms of their terms. The weights of these vector forms are taken from the tf-idf table of the terms. Later using these vector representations, we calculated the cosine similarity of Document and Query to get retrieval results.
    2. As you know this is the basic retrieval model using vectors and the goal of this project is to make the retrieval system more sophisticated and a better version that can be used and appreciated in practical applications.

- Limitations of the basic vector space model
    1. Documents with similar content but different vocabularies are not retrieved properly. This is because we are not considering Synonymy property of the terms.
    2. Documents with different content but same vocabularies are not retrieved properly. This is because we are not considering Polysemy property of the terms.
    3. The context of the terms is not modelled properly because of the unigram version of the model.
    4. Order of terms is not considered because it is a bag of words model.

- Hypotheses for addressing the above limitations

    In this project, we are mainly trying to make the system better by considering the Semantics of the Terms. So we mainly implement LSA and ESA to address the Synonymy and Polysemy limitations of the Vector Space Model.

1. A model with LSA is a better retrieval system than the basic Vector Space Model on Cranfield Dataset.
2. A model with ESA is a better retrieval system than the basic Vector Space Model on Cranfield Dataset.
   We are currently referring to this paper for the implementation of LSA and reference paper for Explicit semantic analysis mentioned in class, ie this. We want to update the implementation if we find a better resource in the future.

On top of the above implementations. We would like to make some changes to the basic vector space model to improve its efficiency and performance if time permits. Namely, we would like to include titles in the similarity measure. We would also like to implement basic spellcheck(only for words that are not present in a dictionary).

- Implementations to realize the above Hypotheses

1. In LSA, Then, the high-dimensional and sparse term-document matrix is reduced by SVD (Singular Value Decomposition) and transformed into the low-dimensional vector space, namely the space representing the latent semantic meanings of the words. In the above-mentioned paper, we also try to improve the retrieval performance of the vector space model (VSM) by utilizing user-supplied information of those documents that are relevant to the query in addition to LSA. This information is already present in the Cranfield dataset. To include this information we try to develop a supervised model. To experiment we divide the Cranfield dataset into Training and Test Set.

2. In ESA, we use a pre-trained machine learning model to map the terms/documents of corpus to a weighted sequence of Wikipedia concepts of d-dimension (parameter) ordered by their relevance to the input. Using these Wikipedia synset vectors and a suitable similarity measure (cosine metric) we find relatedness among queries and documents.

To include Titles in the similarity measure we plan to consider them as individual documents and while calculating similarity we give more weightage to the Title vector. For Spell Check, we preferably want to use a library.

- Evaluation

We want to use the **Precision, Recall, F-score, Average Precision, Mean Average Precision, nDCG** metrics to measure the performance of the systems. **Precision vs Recall** curve will be mainly used to compare the systems.