Project Proposal

CS6370: Natural Language Processing

JAIN DEVANSH RAKESH [CH17B050]	K SANTOSH [CH17B053]

- ## Limitations of the Vector Space Model:

Bag of words: The order in which the words appear in the document is lost in the vector space representation. For example, “John is the father of Johnson” and “Johnson is the father of John” have the same vector space representation.

Term Independence: All the terms in a vector space model are assumed to be independent. There is no relation between terms which is not always the case.

Polysemy and Synonymy: Same term may be used to express different things in different contexts. But the vector space model could fetch irrelevant documents which share some words from the query which would affect the precision. Similarly, different terms may be used to express the same thing. Here, the similarity of relevant documents with the query may be low because they do not share the same words. This affects the recall.

- ## Hypotheses:

Improvising over Bag of words: As the bag of words model is an order-less representation of words, it fails to capture spatial information of the given context. Hence, using n-gram indexing will retain more information on ordering of words in the text reducing the information loss in the regular vector space model.

Incorporating semantics and word relatedness: To handle synonymy and polysemy, we can identify words which are highly related to each other. In our system we missed out some documents because the word was different but the word meaning was the same. If we use word relatedness, we can link different queries and get desired documents. We can modify the vector space so that two similar words can have closer vector representations. This will improve the recall of our system.

We can compute word relatedness probabilities for each pair of words in the training corpus. It will be difficult to find the same for all words due to the large number of words, therefore we will prioritise based on frequency of words.

To ensure that semantic similarity is respected and not just word similarity, we plan to use Latent Semantic Analysis. The queries and documents can be analysed for latent meaning of the same to avoid missing out documents due to synonymy. The queries and documents will be compared in the concept space to better capture latent semantics.

Realization of Hypotheses:

n-gram indexing:

In a n-gram indexing model we treat each n consecutive words as a term in the vocabulary space. Then, we would calculate the weights of each of the terms as we did in the case of the regular vector space model using tf-idf scores. The similarity between query and document can then be evaluated using cosine similarity. The parameter n will be decided based on the corpus and performance of the model.

Latent Semantic Analysis:

To overcome the problem of polysemy and synonymy in case of VSM where only documents are represented in the term space, LSA can be used which assumes that semantic relations between words may be present which may not be explicit but latent in a large sample of text. In most cases, the term document matrix is sparse, has a high dimension and is noisy. We can find a lower rank approximation of this matrix using truncated singular value decomposition technique. Truncated SVD decomposes the original matrix into three matrices:

$$A \approx U_k \Sigma_k V_k^T$$

Here, k is a hyperparameter which denotes the number of concepts we want to generate out of the text. Similarities between vectors can be calculated using cosine similarity measure by first transforming the vector using the decomposed matrices.
This technique helps in easily getting similarity between the following:

- Terms with terms
- Documents with documents
- Terms with documents

Word relatedness indexing using wordnet:

SInce the number of words are too high in the corpus, we will prioritise words based on frequency and compute word relatedness index for each pair of words. Feasibility and effectiveness of such prioritization is to be determined. The index will be computed as proportional to word-word distance in wordnet. For implementation, we might not have an indexing for all words in a given query, but only for most recurring ones. Queries will be mapped to documents with the query words and their related words (setting a threshold for index). This ought to increase the recall. The practicality of this hypothesis for our case will have to be analysed.

- ## Evaluation of the IR systems :

Methods mentioned above will be used to make different IR systems (various hybrid models will be tried). We will evaluate our system on a representative test corpus. We will use measures like MAP, F-score averaged over queries, nDCG, etc. to evaluate a given system.

The baseline for comparison would be the simple vector space model. We will evaluate the performance measures of all the models by taking a finite number of random samples from a bag of queries which will give us a series of performance measures for a given model. We will use paired t hypothesis testing to check if the model is better in an overall sense than the baseline.

Apart from the above, we will also try spell checks in corpus and queries. We will detect spelling errors through dictionary spellings and use edit distance to find the corrected word. To evaluate spell check models we can use mean reciprocal ranking as a measure. We will also try to implement auto completion of queries. Methods include: using past query data, matching algorithm to match incomplete query with a complete query in standard set (will need a standard set of query a priori), probabilistic model trained on standard queries predicting next words, etc.