CS6370 - Natural Language Processing

Project Proposal

Team-13

1. Abhishek Kumar - CS21M002
2. Ganesh Jatavath - CS21M019
3. Mohammed Safi Ur Rahman Khan - CS21M035

Introduction

The work that was done as part of the assignments was building a simple search engine that takes in query and document text, does basic preprocessing on it (like lemmatization, stop word removal, sentence segmentation, etc.), and then finally builds a simple index that is then used to find the relevant documents and return them in order of their relevance.
We have also evaluated our search engine and reported the results.

Motivation

While working on building a search engine using a simple vector space model, we noticed many discrepancies in the output thereby pointing us to the various limitations of implementing our search engine using a just simple vector space model.
Some of these limitations are:

1. This model fails to take the order of words into account. Although this issue may be solved by using the tri-grams approach (in general n-grams) or other preprocessing techniques like phrase collection.
2. This model assumes an orthogonal relationship between words. I.e., it assumes that words are not related at all. This is a grave mistake as in the real world, there are many relationships that occur between words (like polysemy, synonymy, hyponymy, etc). Because of this assumption, we miss out on many related documents.
3. The words in the query must precisely match the words in the documents. This limitation is resolved to some extent by using lemmatization but it still exists.
4. It relies on blind matching of words if query and words of documents and never consider the semantic closeness and similarity of words.
5. The vector space model is very computationally expensive involving a heavy number of intensive calculations. Also, the vector space model is not very flexible in the sense that any new additions to the term vocabulary require recalculation of all vectors thereby making it computationally expensive.

In this work, we aim to try to solve some of these limitations of the vector space model thereby boosting the performance of the search engine.

Hypothesis

If we observe, a lot of limitations of the vector space model are revolving around the fact that this model fails to take into account the semantic relationship of the words (i.e., assuming orthogonality between words).
Our hypothesis is that we may be able to improve the performance of our search engine by removing the previous orthogonal assumption of words and modeling relationships between words and also by considering the context of the words (i.e., the distributional hypothesis which states that words that are used and occur in the same contexts tend to purport similar meanings)

Implementation

Ferdinand de Saussure proposed the Structural view of language (structural linguistics), which essentially put forth the idea that words don't refer to actual real-world objects rather they refer to some “abstract concepts”.
A concept can only be understood by its relationship with other concepts.
Progressing with this idea, we are proposing to introduce this notion of concept in our indexing process by using distributional semantics i.e., Latent Semantic Analysis methods (wherein, we get the relationship between the terms and the documents from the corpus itself).
Further, we also propose to model these concepts as Wikipedia articles thereby introducing some heavy knowledge into our system (i.e., Explicit Semantic Analysis). In further advancements, we also propose to include a spell checking/correction module in our search engine to increase the robustness, introducing WordNet-based concepts (i.e, Synsets) to index our documents, etc.

Evaluation

Here, we have to effectively compare two hypotheses, one of them is using the simple vector space model, and the other is using LSA.
We shall evaluate both of these methods using the same metrics as used earlier i.e., MAP, nDCG, Precision, Recall, and F-Score. Then we shall compare the two sets of results (one set for each hypothesis) and reach the conclusion of the better of the two hypotheses.

During our evaluation of the Vector space model, we came across some queries that were not performing as well as expected. We shall run these queries with both the models and compare the two hypotheses by using this result.

If required, we shall further apply more techniques to improve the performance of our Search Engine.

References:

1. https://docs.oracle.com/en/database/oracle/machine-learning/oml4py/1/mlpug/explicit-semantic-analysis.html
2. https://en.wikipedia.org/wiki/Explicit_semantic_analysis
3. https://towardsdatascience.com/latent-semantic-analysis-intuition-math-implementation-a194aff870f8
4. https://www.analyticsvidhya.com/blog/2021/09/latent-semantic-analysis-and-its-uses-in-natural-language-processing/