**Introduction**

Computational text processing has offered many benefits. Across areas ranging from cognition, advertising, politics to psychology. The fields in which they can applied seem to be ever growing. Machine learning(ML) algorithms are also used to provide insight and help us understand text, words and documents. Currently there are a plethora of ML algorithms that are being used to perceive the interpretation of texts.

The most widely used methods in text analyses includes word to vector, LDA(Latent Dirichlet Allocation), Latent Similarity Index(LSI). All of these methods use a procedure where, the words present in documents are converted to vectors and the algorithm determines the distance or proximity of words. After which the texts are compared to measure the similarity. Apart from the widely used algorithms, Word Mover’s Distance is also an approach that is starting to be used in the field of NLP.

Using these approaches to compare and contrast which method performs best is the idea of this project. When models are trained for certain corpora, how LDA, LSI and the WMD approaches vary. Recently there also has been a merge of the popular methods in, LDA2vec making it a combinatronic approach. This involves using data structures which are most suited for storing and iterating through the data.

Since most similarity indexes are based on heuristics and is subjective and the analyses is done on the corpus, the results have a comparative metric. The corpora that was used involved self tagged opinion pieces which was freely available online. Also we used technical papers from tagged journals. This gave us the ability to test if the methods were specific to the type of corpus or if it was independent of the corpora that was being used.

Finally we created a tool which could be used to input corpora after which the comparative study would provide results on the similarity indexes. This tool was implemented in python and used the characteristics of a web project which involved using a server-client architectured system.

<<Also defining and architecturing the system to have flask, OOPS principles and how plug and play can be constructed to have users choose the method for the corpus check>>

<<What are the advantages of using the integration and the using a plug and play kind of similarity comparison - performance, benchmarking, availability,..>>

**Related works**

There are different ways in which text analysis can be performed. They can broadly be categorized into – singular approaches and hybrid approaches.

Singular approaches involve the main idea being the use of one method to understand, predict and extract features from text. Using approaches such as paragraph vectors, Latent Dirichlet Allocation, TF-IDF. The biggest drawbacks of using singular approaches is they seem to be catered to relationships formed with words and documents. If there are larger subsets – sentences, paragraphs, phrases the classification and the performance of these algorithms does not match up.

Hybrid approaches on the other hand involve consolidating the approaches together. Which is where these approaches fill in the gaps that singular methods leave behind. Using lda2vec takes the approach of using word prediction using an allocation while converting the words into vectors. Paragraph vectors looks at trying to create vectors from paragraphs and is then used to predict the following words.

Each of the individual approaches is discussed in chapter 2 below.

<<An image of comparative study between singular approach and hybrid approach >>

**Background**

This chapter details the methods used to find similarity across sets of documents.

1. Paragraph Vectors

This paper talks about using paragraph vectors instead of word or document vectors. To emphasize, the fact that the method can be applied to variable-length pieces of texts, anything from a phrase or sentence to a large document. This method is eventually used to predict the next word in the document. The algorithm uses unlabeled data and thus can be used for tasks which does not involve labeled information. The datasets used are a Stanford Sentiment Dataset and an IMDB Dataset. Though this method involved using structured datasets, the computation for it is deemed high since it is an unsupervised algorithm.

1. LDA2Vector

This paper combines the approaches of LDA and word2vec generating relationships between documents and topics. The ability to classify documents and obtain features which creates the relationships is useful for performance and for word prediction. Using hybrid methods seems to perform better when it is with regard to classification and discrimination. Newsgroup datasets were used to train LDA and word2vec to extract latent topics and word vectors. Performance and error rates were calculated via benchmarking and to contrast other methods such as SVMs and TF-IDF.

1. Hyperlink based Semantic Measure

This paper evaluates a new method to find similarity. Wikipedia Link-based Measure uses hyperlinks on wikipedia to create connections and find out similarity across articles. Using heuristic approaches and statistical commonness, articles are identified and earmarked for relatedness. Also using benchmarking against TF-IDF a google similarity distance. As against using word similarity which is more expensive, using WLM is less computationally expensive.

1. WordNet based Measures

This paper gives a highlight of three types of methods which measure semantic-relatedness.

The first is a theoretical examination of a proposed measure for those mathematical properties thought desirable, such as whether it is a metric, whether it has singularities, whether its parameter- projections are smooth functions.

The second kind of evaluation is comparison with human judgments. Insofar as human judgments of similarity and relatedness are deemed to be correct by definition, this clearly gives the best assessment of the “goodness” of a measure.

The third approach is to evaluate the measures with respect to their performance in the framework of a particular application. If some particular NLP system requires a measure of semantic relatedness, we can compare different measures by seeing which one the system is most effective with, while holding all other aspects of the system constant.

**Bibliography**

1. Distributed Representations of Sentences and Documents - Quoc Le, Tomas Mikolov
2. A Hybrid Document Feature Extraction Method Using Latent Dirichlet Allocation and Word2Vec - Zhibo Wang, Long Ma, and Yanqing Zhang
3. An Effective, Low-Cost Measure of Semantic Relatedness obtained from Wikipedia Links - David Milne, Ian H. Witten
4. Evaluating WordNet-based Measures of Lexical Semantic Relatedness - Alexander Budanitsky, Graeme Hirst