1. INTRODUCTION

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can change when exposed to new data.

Learning from text and natural language is one of the great challenges of Artificial Intelligence and Machine Learning. One of the fundamental problems is to learn the meaning and usage of words in a data-driven fashion, that is from some given text corpus, possibly without further linguistic prior knowledge.

Classification is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. In other words, the goal of a good document classification scheme is to minimize intra-cluster distances between documents, while maximizing inter-cluster distances (using an appropriate distance measure between documents). A distance measure (or, dually, similarity measure) thus lies at the heart of document classification. In classification, the classifier learns the association between objects and classes from a so called training set.

Modelling the semantic similarity between text documents is an interesting problem for cognitive science, for both theoretical and practical reasons. Theoretically, it involves the study of a basic cognitive process with richly structured natural stimuli. Practically, search engines, text corpus visualizations, and a variety of other applications for filtering, sorting, retrieving, and generally handling text rely fundamentally on similarity measures.

MOTIVATION:

Google claims to index over 2 billion web documents, and, over 1.5 million web documents are added to the World Wide Web every day. Human back-classification of at least 2 billion web documents would be virtually impossible, and even attempting to categorize all new web documents would require unacceptable human effort. One possible solution is to force web document authors to categorize each newly created page. This has three problems:

1. Web authors cannot be relied upon to categorize their pages correctly,
2. Authors are often prone to misclassify their documents in order to increase potential web traffic
3. It does not address the problem of the (at least) 2 billion web pages that have already been created.

2. LITERATURE SURVEY

There are many other potential applications and benefits that will accrue from being able to reliably and automatically cluster and categorize corpora of documents. Much of this document clustering work is based on using either supervised or unsupervised learning techniques in order to label particular web documents as belonging to a specific category, or grouping together similar documents into clusters. This research area closely overlaps with (and in recent times indistinguishable from) research efforts known as text classification and text categorization

Various techniques have been proposed that aim to develop accurate methods for autonomous categorization. However, the research literature in this area soon reveals that almost every single such proposed technique has been tested for its categorization accuracy using different datasets; any objective, scientifically sound comparison between two categorization techniques is therefore very difficult

One of the technique is K-means which is described in [1]. The objective function of Kmeans is to minimize the average squared distance of objects from their cluster centers, where a cluster center is defined as the mean or centroid μ of the objects in a cluster C. As K-means has been implemented several times so we chose another technique i.e. probabilistic latent semantic described in [2].

Latent semantic analysis (LSA) is well-known technique which partially addresses these questions. The key idea is to map high-dimensional count vectors, such as the ones arising in vector space representations of text documents , to a lower dimensional representation in a so-called latent semantic space. As the name suggests, the goal of LSA is to find a data mapping which provides information well beyond the lexical level and reveals semantical relations between the entities of interest. In contrast to standard LSA, its probabilistic variant has a sound statistical foundation and defines a proper generative model of the data

3. CONTRIBUTION

It is important to emphasize that getting from a collection of documents to a classification of the collection, is not merely a single operation, but is more a process in multiple stages. These stages include more traditional information retrieval operations such as crawling, indexing, weighting, filtering etc. Some of these other processes are central to the quality and performance of most clustering algorithms, and it is thus necessary to consider these stages together with a given clustering algorithm to harness its true potential. We will give a brief overview of the classification process, before we begin our literature study and analysis. We have divided the offline clustering process into the four stages outlined below. The Stages of the Process of Classification:

Collection of Data includes the processes like crawling, indexing, filtering etc. which are used to collect the documents that needs to be clustered, index them to store and retrieve in a better way, and filter them to remove the extra data, for example, stopwords. Here we have used BBC news dataset.

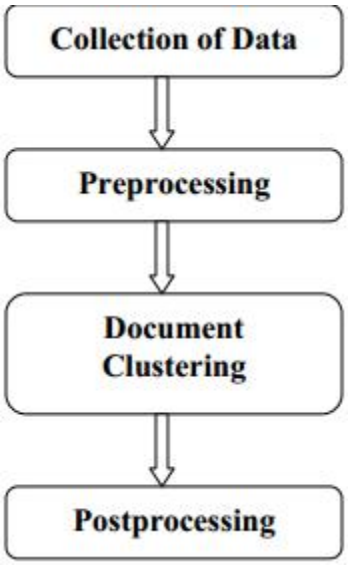

Figure.1

Preprocessing is done to represent the data in a form that can be used for classification. There are many ways of representing the documents like, Vector-Model, graphical model, etc. Many measures are also used for weighing the documents and their similarities.

Stressing simplicity first, our feature vectors were built only from text that would be seen on the screen, i.e. normal document text, image captions and link text, and no extra weight was given according to emphasis (bold typeface, italic typeface, different colours,etc). For each document, the extraction process was as follows:

- The set of all words that appeared at least once in the document was extracted.
- If stop-word removal was switched on, we removed from the set of extracted words any word that was listed in our stop-word list.
- If word stemming was switched on, we combined all the words with a similar stem, (i.e. count all occurrences of a word as a single occurrence of its stem).

Processing for finding the importance of each word or term we have used TF-IDF(*term frequency-inverse document frequency*).

The TF-IDF weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the TF-IDF weighting scheme are often used.

The importance increases proportionally to the number of times a word appears in the individual document itself--this is called Term Frequency. However, if multiple documents contain the same word many times then you run into a problem. That's why TF-IDF also offsets this value by the frequency of the term in the entire document set, a value called Inverse Document Frequency

$$idf(t, D) = log \frac{|D|}{1 + |\{d \in D : t \in d\}|}$$

- The numerator : D is infering to our document space. It can also be seen as D = d1,d2,…,dnd1,d2,…,dn where n is the number of documents in your collection. Thus for our example | D || D |, the size of our document space is 4, since we're only using 4 documents.
- The denominator : | {d∈D:t∈d} || {d∈D:t∈d} | implies the total number of times in which term t appeared in all of your document d (the d∈Dd∈D restricts the document to be in your current document space). Note that this implies it doesn't matter if a term appeard 1

time or 100 times in a document, it will still be counted as 1, since it simply did appear in the document. As for the plus 1, it is there to avoid zero division.

Document Clustering : Now that we have our matrix with the term frequency and the IDF weight, so we first converted them to vector form so that we could provide the input to the SVM classifier. But before converting it in vector form we divided our input dataset into two different categories i.e. training and testing.

.

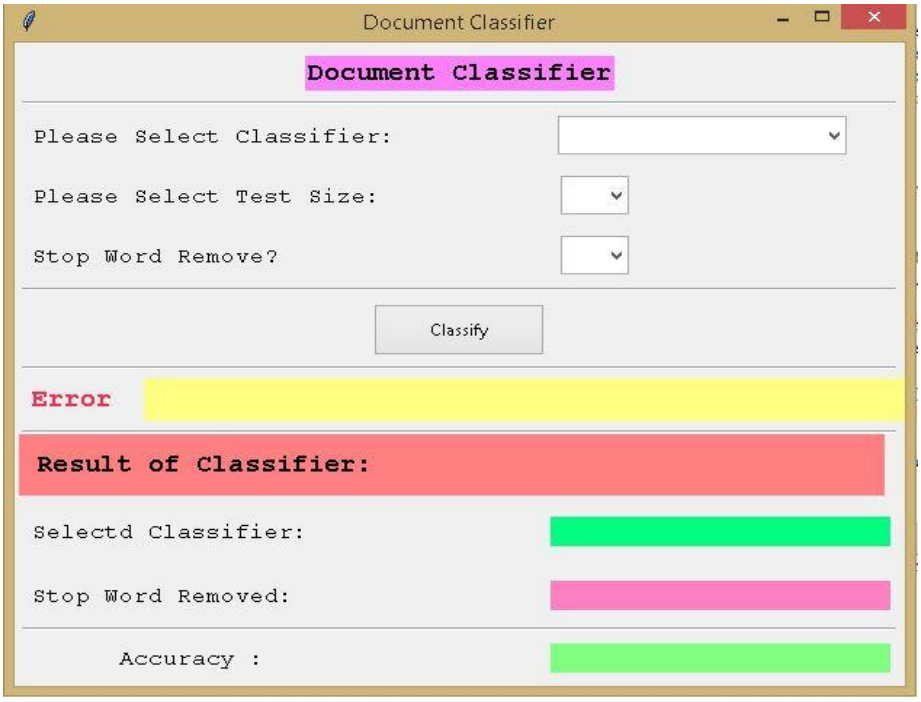

Figure.2

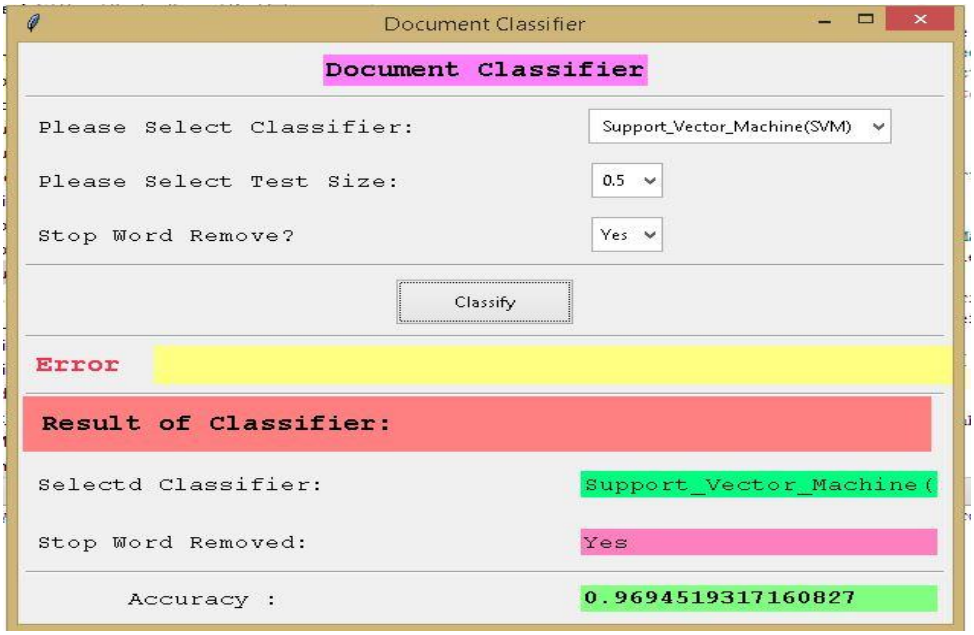

Figure.3

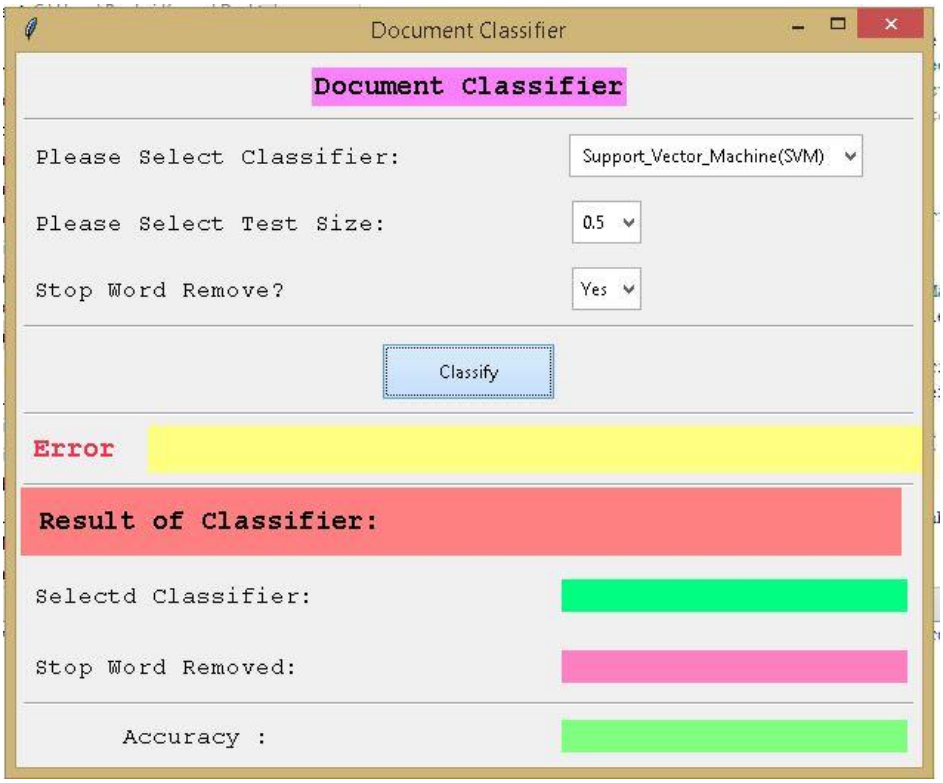

Figure.4

4. CONCLUSION

Web Document clustering is an exciting and thriving research area. In particular, unsupervised clustering of web documents so far less studied than supervised learning in this context has many future applications in organising and understanding the WWW, as well as other corpora of text. We have done data processing and document classification which is giving average 96 percent accuracy. We have done much work but still some work is left like generating title of the news with semantic analysis.

5. REFERENCES

[1] “Unsupervised Clustering” Mark P. Sinka, David W. Corne Department of Computer Science, University of Reading, Reading, RG6 6AY, UK m.p.sinka@reading.ac.uk, d.w.corne@ reading.ac.uk;pp

[2] “Probailistic Latent Semantic” Thomas Hofmann EECS Department, Computer Science Division, University of California, Berkeley & International Computer Science Institute, Berkeley, CA hofmann@cs.berkeley.edu

[3] “Semantic Similarity of Documents Using Latent Semantic Analysis” Chelsea Boling Department of Mathematics Lamar University 4400 MLK Boulevard Beaumont, TX 77710 University of Kentucky, Lexington, KY April 3-5, 2014 Faculty Advisor: Dr. Kumer Das.