SYNOPSIS- MINOR PROJECT

TOPIC MODELING USING BIG DATA ANALYTICS

SUBMITTED BY: Farheen Nilofer(12-CSS-21), Sarah Masud(12-CSS-57)

ABSTRACT:

Topic models are a suite of algorithms that uncover the hidden thematic structure in document collections. These algorithms help us develop new ways to search, browse and summarize large archives of texts.The structure uncovered by topic models can be used to explore an otherwise unorganized collection. The models used for Topic modelling are largely computation dependent. This is where the techniques of Big Data come into use. Big data processing environments such as Hadoop are useful for the analysis of large data sets, best designed to handle unstructured data.
The project will be divided into two phases:

Phase I:

Installation of the Hadoop on a cluster of machines.

Phase II:

The general algorithms used inTopic modelling will be redefined to suit the parallel programming paradigm, this should reduce the computation time for a corpus from a day to few hours.

INTRODUCTION:

Problem: Reducing the computation time for Topic Modelling using the techniques of Big Data.

Topic Modeling: Topic models provide a simple way to analyze large volumes of unlabeled text. A "topic" consists of a cluster of words that frequently occur together. Using contextual clues, topic models can connect words with similar meanings and distinguish between uses of words with multiple meanings.

Hadoop: The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.
The core of Apache Hadoop consists of a storage part (Hadoop Distributed File System (HDFS)) and a processing part (MapReduce).
HDFS- Data in a Hadoop cluster is broken down into smaller pieces (called blocks) and distributed throughout the cluster.
MapReduce- The map and reduce functions can be executed on smaller subsets of your larger data sets, and this provides the scalability that is needed for big data processing.

PROPOSED METHOD/ALGORITHM :

In practice researchers attempt to fit appropriate model parameters to the data corpus using one of several heuristics for maximum likelihood fit. Techniques used here include singular value decomposition (SVD), the method of moments, and very recently an algorithm based upon non-negative matrix factorization (NMF). This last algorithm also generalizes to topic models that allow correlations among topics.

MapReduce is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster. Conceptually similar approaches have been very well known since 1995 with the Message Passing Interface standard having reduce and scatter operations.

PROGRAMMING ENVIRONMENT/ TOOLS:

- Hadoop Cluster (https://hadoop.apache.org/)
- Mallet (software project) (http://mallet.cs.umass.edu/)
- Stanford Topic Modeling Toolkit (http://nlp.stanford.edu/software/tmt/tmt-0.4/)
- Gensim - Topic Modeling for Humans (http://radimrehurek.com/gensim/)

REFERENCES:

1. Papadimitriou, Christos; Raghavan, Prabhakar; Tamaki, Hisao; Vempala, Santosh (1998). "Latent Semantic Indexing: A probabilistic analysis" (Postscript). Proceedings of ACM PODS.
2. Hofmann, Thomas (1999). "Probabilistic Latent Semantic Indexing" (PDF). Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval.
3. Blei, David M.; Ng, Andrew Y.; Jordan, Michael I; Lafferty, John (January 2003). "Latent Dirichlet allocation". Journal of Machine Learning Research 3: 993–1022. doi:10.1162/jmlr.2003.3.4-5.993.
4. Blei, David M. (April 2012). "Introduction to Probabilistic Topic Models" (PDF). Comm. ACM 55 (4): 77–84. doi:10.1145/2133806.2133826.
5. Sanjeev Arora; Rong Ge; Ankur Moitra (April 2012). "Learning Topic Models—Going beyond SVD". arXiv:1204.1956.
6. H. Oktay, A. S. Balkir, I. Foster, and D. Jensen(2011).”Distance estimation with MapReduce for large networks”, in *Proceedings of the Workshop on Information in Networks*, WIN, pp. 1-6.