ITCS 6190 Cloud Computing Final Project

Spring 2017

Proposed Project:

Summarizing Large Text Collection of News Articles Using K-Means Clustering and Topic Modelling based on the MapReduce framework.

Team Members:

Sampath Kumar Gunasekaran
Viseshprasad Rajendraprasad
Ajay Kumar Prathap

Tasks:

The summarization task will be performed in 4 stages:

- Document Clustering stage : K- Means Clustering technique will be employed on the multi-document text collection to create the text document clusters.
- Latent Dirichlet Allocation (LDA) : LDA topic modelling technique will be employed on each individual text document cluster to generate the cluster topics and terms belonging to each cluster topic.
- Frequent and Semantic terms: Global frequent and semantic terms are generated from the collection of multiple text documents.
- Sentence filtering: For each document, the sentences which are containing the frequent terms and semantic similar terms are selected for participation in the summarized document.

Output of this course project would be the summarized documents of large text collections.

Dataset:

The below mentioned dataset of News Articles consisting of 2225 documents from the BBC news website corresponding to stories from 2004-2005 will be used for summarization.
http://mlg.ucd.ie/datasets/bbc.html

Tools:

Java, Hadoop, MapReduce, Mallet API, WordNet Java API.

Benefits:

A summarized document helps in understanding the gist of large text documents quickly and also saves a lot of time by avoiding reading of each individual document in the large text collection.
It has a number of real life applications.

- Analyzing web search results for assisting users in further browsing.
- Generating summaries for news articles.
- Generating summaries for legal case reports, etc.

From a user's perspective, reading newspapers with lots of news content in them can be a tedious and boring task. By summarizing large paragraphs of content into fewer lines without losing the core essence of the content and important detail makes reading less painful.

Deliverables:

1 - What we will definitely accomplish:
We would conduct unsupervised clustering to cluster a variety of news articles based on content.

2 - What we are likely to accomplish:
We would want to identify the topic terms and semantic terms associated with documents in every cluster.

3 - What we would ideally like to accomplish:
We would want to generate an accurately summarized version of large text documents.

Division of responsibility:

- **Sampath Kumar Gunasekaran:**

Overall designing and modelling of the project and coding.

- **Viseshprasad Rajendraprasad:**

Setting up meetings on regular basis, environmental setup and coding.

- **Ajay Kumar Prathap:**

Coding, documentation and testing.