DALHOUSIE UNIVERSITY

CSCI 5408
Summer 2018

PROJECT PROPOSAL

Submitted by:-
Nikhil Dhirmalani: B00775542
Khushboo Siwal: B00781497

TABLE OF CONTENTS

List of Tables

Objective

Document clustering is the act of collecting similar documents into bins, where similarity is some function on a document. Manually categorizing and grouping text sources can be extremely laborious and time-consuming, especially for publishers, news sites, blogs or anyone who deals with a lot of content. The objective of this project is to perform document clustering in order to find similarities between document based on words in the document and cluster relevant documents together using Kmeans algorithm.

The first step in the Clustering process is to create word vectors for the documents we wish to cluster. A vector is simply a numerical representation of the document, where each component of the vector refers to a word, and the value of that component indicates the presence or importance of that word in the document. The distance matrix between these vectors is then fed to algorithms, which group similar vectors together into clusters.

Value Proposition:

In this project, we will try to implement distributed Kmeans using map reduce on our own. Moreover, we will try to find hidden patterns in data and try to visualize it. We will also try to find words in the document that has impact on clustering and such hidden things that has not been explored.

Data Set :

The **20 Newsgroups** data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups each corresponding to different topic. Some of the newsgroups are very closely related to each other (e.g. comp.sys.ibm.pc.hardware / comp.sys.mac.hardware), while others are highly unrelated (e.g misc.forsale / soc.religion.christian). Below table is the list of 20 newsgroups, partitioned (more or less) according to subject matter:

Table 1: List of 20 newsgroups

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space

misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

Programming Language and Tools:

We will be using following tools and languages for extract, transform and load process:

1. Python 2.7 as the programming language
2. Pycharm IDE to run the python files
3. Numpy, Pandas, Matplotlib python packages
4. Tableau for data visualization
5. Pyspark framework for data processing

Roles and Responsibilities:

- Nikhil Dhirmalani(B00775542):

Nikhil will work with Khushboo in preparing project proposal, tracking sprint progress, presentation slides and project report. Nikhil will set up environment for apache spark, he will implement distributed Kmeans using map-reduce and use this model in order to cluster the data. Nikhil will also try to find new patterns in data from the trained model.

- Khushboo Siwal(B00781497):

Khushboo will work with Nikhil in preparing project proposal, tracking sprint progress, presentation slides and project report. Khushboo will set up environment for tableau, she will preprocess the data and try to visualize the results from trained model using different graphs in visualizing tool.

Work Breakdown Structure:

We have divided the project into five phases and each phase corresponds to a sprint. We have then added each task that we have to perform in each phase in the table given below.

Table 2: Phases of the project

Level1	**Level2**	**Level3**
1 Document Clustering using Kmeans Algorithm	1.1 Initiation	1.1.1 Determine Project Team 1.1.2 Project Team Kickoff Meeting 1.1.3 Discussion on project, dataset, algorithm, tools and programming language. 1.1.4 Review on various works used to solve the problem 1.1.5 Project work division and sprints development 1.1.6 Review each team members work and suggest changes if required 1.1.7 Develop Project Proposal and Submit. 1.1.8 Milestone: Project Proposal Approval
	1.2 Planning	1.2.1 Project Team Meeting 1.2.2 Discussion on work to be accomplished in ongoing sprint 1.2.3 Download softwares and tools used in project and environment setup on machines. 1.2.4 Data preprocessing tasks and analysis of algorithm used in project. 1.2.5 Review each team members work and suggest changes if required
	1.3 Execution	1.3.1 Project Team Meeting 1.3.2 Discussion on work to be accomplished in ongoing sprint 1.3.3 Implement the algorithm and train a model on preprocessed dataset 1.3.4 Discussion on results and methods to be implemented to improve the obtained

1 Document Clustering using Kmeans Algorithm		results
	1.4 Experiment and Visualization	1.4.1 Project Team Meeting 1.4.2 Discussion on work to be accomplished in ongoing sprint 1.4.3 Experiment with model and try to find different results. 1.4.4 Use visualization tool in order to show the results
	1.5 Closeout	1.5.1 Discuss experiments performed,results and findings 1.5.2 Presentation slides preparation and review 1.5.3 Develop Project Report and Submit 1.5.4 Project Sponsor Reviews Project Report 1.5.5 Project Report Signed/Approved

Milestone:

- 03/07/2018

  Discussion on data set, problem, proposed solution, algorithm, softwares and tools to be used in order to reach to the solution. Review of already done works on the project selected. Preparation of project proposal and review.
- 13/07/2018

  Environment set up for the project, data preprocessing, analysis of algorithm and its implementation. Tracking the ongoing sprint and reviewing team member work.

- 23/07/2018

  Trained the model on the algorithm and perform various experiments on model and visualize the results. Tracking the ongoing sprint and reviewing team member work.
- 03/08/2018

  Preparation of presentation slides, project report and other closing tasks.

REFERENCES:

I. ANON

In-text: (2018)

Your Bibliography: (2018). Retrieved from https://www.linkedin.com/pulse/nlp-text-analytics-simplified-document-clustering-parsa-ghaffari/

II. ANON

In-text: (2018)

Your Bibliography: (2018). Retrieved from https://www.linkedin.com/pulse/20141209180635-83626359-nlp-and-text-analytics-similified-document-classification/

III. HOME PAGE FOR 20 NEWSGROUPS DATA SET

In-text: ("Home Page for 20 Newsgroups Data Set", 2018)

Your Bibliography: Home Page for 20 Newsgroups Data Set. (2018). Retrieved from http://qwone.com/~jason/20Newsgroups/