COEN 242 (BIG DATA) PROJECT OUTLINE

GROUP: 4

OBJECTIVE:

Analyzing the performance of Naïve Bayes Algorithm on Weka, Mahout and Hadoop framework tools using the Enron email dataset.

TEAM MEMBERS:

1)	Ajay Videkar	W1111605
2)	Harsha Teja Kanikicherla	W1071900
3)	Harshkumar Pandya	W1114569
4)	Jillian Carleton	W0355725
5)	Lakshitha Raj Vasanadu	W1115006
6)	Pratham Vasa	W1136169
7)	Seema Sardesai	W1008598
8)	Shrividya Manmohan	W1026992
9)	Spandana Namburu	W1060542
10)	Srinivas Reddy	W1136100

PROJECT IDEA:

We will be using the Enron email dataset to perform the following analysis:

1) Identification of the author/owner of the email.

2) Detection of SPAM (erroneous and useless) emails.

3) Calculating the number of email messages (frequency) per author (members of the organization).

NAIVE BAYESIAN CLASSIFICATION ALGORITHM:

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. In machine learning, Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. It is a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features.

INFORMATION ABOUT THE DATASET:

It was a scandal that happened in October 2001 that eventually led to the bankruptcy of Enron Corporation, an energy company with its headquarters in Texas, United States of America. Some workers from within the company had to do with the downfall of the company. Higher executive authority from the company decided to make the emails of all the employees publicly available in the internet so that the culprit can stop down its malicious activity. We will be analyzing these emails and implement the selected algorithm on this dataset. The data set consists of 1,227,255 emails with 493,384 attachments covering 151 custodians.

TASK DIVISION

TASK TOPIC	MEMBERS
Studying the dataset information and cleansing the dataset.	• Spandana Namburu • Shrividya Manmohan
Converting the dataset into Weka and Mahout compatible formats and implementing the algorithm on Weka and Mahout platforms.	• Jillian Carleton • Ajay Videkar • Harshkumar Pandya • Pratham Vasa
Implementation of Mapper and Reducer functions for Hadoop framework for the Naïve Bayesian classification algorithm.	• Harsha Teja Kanikicherla • Lakshitha Raj Vasanadu • Seema Sardesai • Srinivas Reddy
Performance analysis of the algorithm on different frameworks.	• Entire Team