CSCI 6515 Assignment 2 – part 1

Due Date: Dec 18.

In this assignment you will practice some text mining algorithms. You will work with the Enron Email Dataset [1]. All the necessary algorithms and operations are available in the Apache Spark. You can find documentation about Spark's LDA and K-means implementations here: [2] [3]. You are expected to use Apache Spark for all the steps. You don't have to use Python API of Apache Spark, you can use R, Java or Scala, at your convenience.

- **Data Preprocessing**
  - You should remove header information in each file, also don't forget the headers of previous messages in context!
  - You should remove attachment information in each file, if it exists
  - You should remove stopping words
  - You should apply stemming
  - You need to upload your data to Hadoop or Amazon S3 to use it with your Spark cluster. *(it might be a good idea to start initially working on a subset of the data on your machine)*

- **Data Discovery**
  - Provide summary/statistics about each user's data,
    - e.g. Name, Email Address, Number of Emails, average length of each email, number of people he/she had contact with etc.
  - Discover how many users they communicated with and represent this data as a directional graph:
    - Each node will represent a user
    - Each edge will represent a collection of emails between users in a directional manner.
    - e.g "Alice sent 15 emails to Bob" & "Alice received 20 emails from Bob"

- **Data Analysis – Task 1**
  - For all users (as a parameter to your program), you are expected to run:
    - K-means over email body & subject on the set of all emails this user has received or sent. The result will allow you to understand the clustering nature of emails. Change the parameters and measure the quality of your clusters (use the method discussed in class) to find out optimal 'k' value.
      - Please check TF-IDF [4]. Using TF-IDF with K-means might improve your results.
    - LDA on the same data as k-means, to understand the topic model of emails. You may use the value of k that you have determined works best with k-means on the same data.
      - You are expected to use the topic information to label the most likely communication between users. See next data analysis task for details.

- Justify and discuss your input parameters for both algorithms. What values you used and how did you decide?
- Compare the results of LDA and K-means. Are the document distributions across clusters similar? What are the execution times? Did you try different input parameters if so how does it affect the execution times?
- After LDA, print most popular topics and their terms. Do those terms provide you with knowledge regarding the user, e.g. user's activities, roles, department, etc.? Discuss the results.

- **Data Analysis – Task 2**
  - You are expected to write another Spark application that creates a labeled communication graph for a given user as an argument, as follows:
    - Use output of LDA from previous task to label the edges of directed graph you created during the data discovery task. Each edge of the graph will contain most likely topics of users communication.
    - Feel free to enrich the graph with additional information.
    - You are not expected to provide a graph visualization, but a logical representation of it.

- **Technical Solution**:
  - Please provide high level description of your technical solution.
  - What kind of technical problems you encountered with and how did you solve?
  - Any other technical details you want to share?

- **Submission**
  - You are expected to upload a zip file (in format of CSCI6515_A2_{YOUR_BANNER_ID}.zip) contains following items;
    - *README.txt file*: this file should contain step by step installation and execution information of your project. If you used any third party libraries then you should list them here.
    - *Source code*
    - *CSCI6515_A2_{YOUR_BANNER_ID}.pdf* : This file will contain your report. You should cover the questions stated in this document and your explain/justify your workflow.

[1] https://www.cs.cmu.edu/~./enron/enron_mail_20150507.tgz
[2] http://spark.apache.org/docs/latest/mllib-clustering.html
[3] https://spark.apache.org/docs/latest/ml-clustering.html
[4] http://spark.apache.org/docs/latest/ml-features.html#feature-extractors