Course Project CS 242

Information Retrieval & Web Search: Winter 2018

Part A

Trevor Van Meter

tvanm001

860835689

February 9, 2018

**Project Overview**

In this era of Technology, massive amounts of data and text are being generated online every second. This vast quantity of data contains information that can potentially be highly valuable if gathered, filtered, and searched efficiently. It is an invaluable resource that can be mined to generate useful business insights, trends, current moods, locations-based events, and much more. The difficulty lies in the fact that extracting, mining, and analyzing all of this data and text is not easy because of the lack of structure and the many language fluctuations that exist. Luckily, there are tools that exist that can aid in this adventure through Natural Language Processing, Text Mining, and Text Indexing. This project is broken down into 2 parts, Part A: gathering, mining, storing, and indexing the data, and Part B: building a graphical user interface to search the queried data.

This current work is for Part A of the project and will complete the following tasks:

1. **Collecting the data** – using Tweepy to query and stream tweets from the Twitter API
2. **Storing the data** – extracting the data from the Twitter API request response, parse into JSON format, and storing the results in a CSV formatted file
3. **Indexing the results** – utilize Lucene indexing to index the tweets and special Twitter characteristics such as hashtags (“#hashtag”), and mentions (“@mention”)
4. **Search the index** – following the building of the Lucene index, utilize the index to search for the top matching documents for a particular term

The system was written in Python and is dependent on the Python 3 libraries, Tweepy python library, as well as PyLucene (the Python wrapper for Lucene).

**Collaboration Details**   
  
I attempted multiple times to reach out to my teammate however never got a response. Due to the timeline approaching, I decided to work on this project alone. For Part A, I have selected to stream Twitter data utilizing the Twitter Streaming API. I will implement the entire system in Python utilizing the PyLucene libraries. My applications structure will provide the end user the capability for searching Twitter for a specific term, or streaming the latest Tweets. The options will allow for providing a pre-existing file or generating a new file of data containing the Tweets. The system will allow for searching for a list of terms within the PyLucene index as well and deliver the results to the console. I designed and implemented the entire system by myself which includes: the Twitter query mechanism, the Twitter saving mechanism, integration of Lucene indexing, and searching of Lucene index.   
  
Overview – Crawler:

Overview – Lucene Index:

Limitations:

Obstacles & Solutions:

Deployment Instructions:

Lucene Instructions: