o   Collaboration Details: Description of contribution of each team member.

o   Overview of system, including (but not limited to):

  Architecture

  Index Structures.

  Search Algorithm.

o   Limitations of system.

o   Instructions on how to deploy the system.

  Ideally, you should include an indexer.bat (Windows) or indexer.sh (Unix/Linux) executable file that takes as input all necessary parameters e.g.

**Example:** [user@server] ./indexer.sh <output-dir>

  A web-application (e.g. Web Archive) that can be deployed to a webserver like Tomcat.

o   Screenshots showing the system in action.

CS172 Crawler Project Part A

Group: Spencer Lee and Jeffrey Chen

## Collaboration Details

Spencer handled development of the project due to his prior experience and familiarity with Java. Jeffrey handled the initial setup and documentation.

Overview

Libraries

Json.simple:

Jsoup:

Lucene-analyzers:

Lucene-core:

Lucene-queryparser:

Twitter4j-core: Provides a Java library that allows easy communication with the Twitter API.

Twitter4j-stream: Extends the core library to allow easy communication with the Twitter Streaming API.

Gson: Google’s JSON interpreter that is used to decode JSON that Twitter uses. It is also used to reencode strings into JSON format.

Jsoup: Provides a handful of classes that allow HTML to be downloaded and parsed into human-readable strings. The crawler uses it to get the title of the webpage and store it as part of Tweet information.

## Indexer Runtime Overview

1. When run, the Indexer accesses all tweet files found in the same directory the program was run in.
2. The Indexer then converts all tweet JSON data into Tweet objects, storing all Tweets in memory.
3. The Tweet objects are then broken down with screenname indexed

## Search Runtime Overview

## Limitations of the System

The system works best in English language searches. However, due to the fact that Twitter data is global in scope, searching through multiple languages is not possible at this time.

## Instructions on Usage of the System

## System in Action