1. Directory and file structure for Twitter application:

The submission contains two sub-directories: "EX2Tweetwordcount" and "screenshots". The "EX2Tweetwordcount" folder contains all python files and the streamparse system to execute EX2Tweetwordcount twitter application. The "screenshots" contains three png figures of an end-to-end execution of your application of your choice

.

EX2Tweetwordcount folder:

The most two important subfolders under "EX2Tweetwordcount" are "src" and "topologies". Under the topology, the spout and bolt components are defined in the EX2Tweetwordcount.clj. Under the "src" folder, there are two additional subfolders, bolts and spouts. They are created to store three python class files: tweets.py parse.py and wordcount.py.

Screenshots folder:

screenshot-twitterStream.png

screenshot-storm-components.png

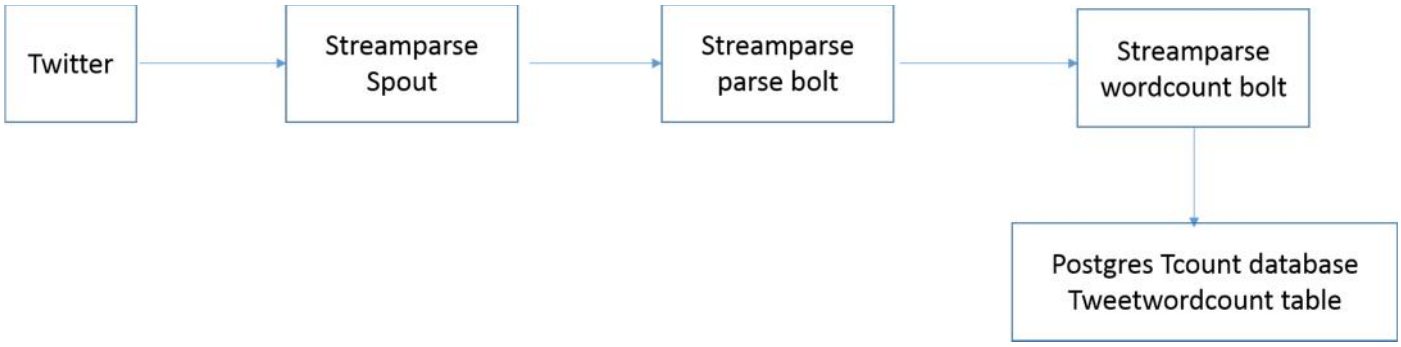
screenshot-extract-results.png

Additional files:

finalresults.py: returns words and their total number of word occurrences in the stream
histogram.py: returns all the words that their total number of occurrences in the stream is more or equal than k1 and less or equal than k2.
Architecture.pdf : A complete documentation (maximum of four pages) of your Twitter application
Readme.txt: a file that shows the step-by-step instructions
plot.png: a bar chart showing the top 20 words in your Twitter stream.
barplot_top20.py: the python codes used to generated the plot.png

2. Application idea and architecture

In this study, we utilize the Python Tweepy library to connect to Twitter Streaming API and extract tweets. We further use the Streamparse package to parsed those live stream tweets, counts the number of each word in the stream of tweets, and writes the final results back to the Postgres database. We also write serval additional python programs to query the generated postgres database and display the outcome on the screen or the plot.



4. File dependencies:

For tweets.py, we import the following python packages:

Import itertools, time

import tweepy, copy

import Queue, threading

from streamparse.spout import Spout

For parse.py, we import the following python package:

import re

from streamparse.bolt import Bolt

For wordcount.py, we import the following package:

from collections import Counter

from streamparse.bolt import Bolt

import psycopg2

from psycopg2.extensions import ISOLATION_LEVEL_AUTOCOMMIT