# Architecture

# Directory and file structure

**EXTWOTweetwordcount**

**Src**

* + Bolts
    - Parse.py
    - Wordcount.py
  + Spouts
    - Tweets.py

**Topologies**

* + - EX2tweetwordcount.py

createDBAndTable.py

finalresults.py

histogram.py

# Description of Architecture

**Spout:**

Tweepy Twitter API, it streams any tweets that filtered out special characters

**Parse:**

Extracts words by splitting the tweets. And each word is checked if it starts with special characters, if not, it will be added to valid words array. (#, @, rt, http, ("\"?><,\'.:;!~[]()&%$\*/=+-\_^$\\"). This is to avoid counting the retweets/mention/hashtag etc.

**Wordcount:**

Interacts with DB and table about word count. ​

**Postgres​ data storage:**

**Tcount**:​ database

**Tweetwordcount**: table with two columns: word (text, primary key), count (int)

# File dependencies

python 2.7

Streamparse

Tweepy

psycopg2

Apache Storm, Postgres

# Execution Steps (also in README.txt)

1. run: sparse quickstart EXTWOTweetwordcount and update corresponding files with the ones in this Git Repo ( API credentials contained in tweets.py)

2. create db and table

mount /dev/xvdf /data

cd /data

./start\_postgres.sh

cd ~

cd EXTWOTweetwordcount

python ./createDBAndTable.py

3. under EXTWOTweetwordcount run: sparse run #takes a while to get more data

4. python finalresults.py <optional inputword>

5. python python histogram.py <lowlimit> <upperlimit>

Screenshot (picture file in Screenshots folder):