Architecture

Directory and file structure

- Src
    - Bolts
        - Parse.py
        - Wordcount.py
    - Spouts
        - Tweets.py
- Topologies
        - EX2tweetwordcount.py
- createDBAndTable.py
- finalresults.py
- histogram.py

Description of Architecture

Spout:

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

Parse:

extract 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, ("\"?><,\'.:;!~[]()&%$*/=+-_^$\\")

Wordcount:

Interacts with DB and table about word count.

postgres :

Tcount: database is created

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)

1. run: sparse quickstart EX2Tweetwordcount and update 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 EX2Tweetwordcount

python createDAAndTable.py
```

3. under EX2Tweetwordcount run: sparse run

4. python finalresults.py <inputword>

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

Error:

1. The steps provided on ISVC wall did not fix the connection error in testing Twitter app when running the sample python script, not sure if this is causing "2"

2. Storm gives error without logs, unable to successfully debug yet:

```
8366 [Thread-35] ERROR backtype.storm.task.ShellBolt - Halting process: ShellBolt died.

java.lang.RuntimeException: backtype.storm.multilang.NoOutputException: Pipe to subprocess seems to be broken! No output read.

IOError: [Errno 2] No such file or directory: u'/tmp/05c5238e-8d39-4cb1-82df-bbd8744c6016/supervisor/stormdist/EX2tweetwordcount-1-1460363533/resources/"/root/EX2Tweetwordcount/logs"/streamparse_EX2tweetwordcount_count-bolt_3_9508.log'
```