Application Idea

The idea of this application is to stream data or tweets from Twitter, parse the tweets into words, and count the number of unique words. The word and count data is saved in a postgresql database as long as the stream application is running. This allows us to run queries against the database as the data is incoming.

Architecture description

The storm application uses a tweet spout to gather the data from twitter using the tweepy api. Each tweet is then emitted to the parse bolt that removes extra characters and splits the tweet into words. The words are then emitted to the wordcount bolt where each word is added to the postgresql database and the count is updated by one. The database `tcount` collects the words and counts in a table `tweetwordcount`. This table can then be interacted with using python scripts to get results from the data or by directly calling sql queries against it.

File dependencies

This application depends on the `UCB MIDS W205 EX2-FULL` AMI and uses the builtin `python` version `2.7.3`.

These additional python libraries are required for this to run:

- psycopg2
- tweepy

Please see the `readme.md` for full setup and running instructions.

File structure:

```
.
├── Architecture.md
├── Architecture.pdf
├── Plot.png
├── extweetwordcount
│   ├── README.md
│   ├── config.json
│   ├── fabfile.py
│   ├── project.clj
│   ├── src
│   │   ├── bolts
│   │   │   ├── __init__.py
│   │   │   ├── parse.py
│   │   │   └── wordcount.py
│   │   └── spouts
│   │       ├── __init__.py
│   │       └── tweets.py
│   ├── tasks.py
│   ├── topologies
│   │   └── extweetwordcount.clj
│   └── virtualenvs
│       └── wordcount.txt
├── readme.md
├── requirements.txt
├── scratch
│   ├── exploration.ipynb
│   └── top-20.csv
├── screenshots
│   ├── screenshot-extractResults.png
│   ├── screenshot-stormComponents-postgresql.png
│   ├── screenshot-stormComponents-topology.png
│   └── screenshot-twitterStream.png
└── scripts
    ├── create_db.py
    ├── create_table.py
    ├── finalresults.py
    └── histogram.py
```

The python requirements can be installed with the `requirements.txt` file. The storm application is located in the `extweetwordcount` directory. The topology is located in the `extweetwordcount/topologies/` directory. The `parse.py` and `wordcount.py` bolts are located in `/extweetwordcount/src/bolts/` and the spout `tweets.py` is located in `/extweetwordcount/src/spouts/`. The database creation scripts are in the `scripts` folder along with the output python scripts. See `readme.md` for further instructions.