NRV Tweets

- Final Presentation
- VT CS4624, Blacksburg, VA
- Sponsors: Mohamed Magdy,
  Dr. Andrea Kavanaugh, Ji Wang
- Ben Roble, Justin Cheng,
  Marwan Sbitani
- 5/06/14

NRVTweets

- Actually Tweets and RSS News stories
- Take 360,000 Tweets and 15,000 RSS News stories from Virtual Town Square DB
- Use Natural Language Processing to associate tweets and news stories with topics
- Upload all data into Solr to make it searchable

Selecting a Tool/API

- Dozens available
- Commercial licenses
- Free usage limits

Free Tools/API's considered

- word2vec - inaccurate
- Weka - data import issues
- Stanford Topical Modeling Toolbox - only .csv files
- LingPipe - unclear
- Gensim - good
- PDLA C++ - lack of documentation
- MALLET - good
- Mahout - needs HADOOP environment

MALLET -

- Machine Learning for Language Toolkit
- java source - http://mallet.cs.umass.edu/
- document classification
  - Naive Bayes, Maximum Entropy, Decision Trees
- topical modeling
  - LDA, Pachinko Allocation, Hierarchical LDA

Parsing the data

- Modify into JSON
- Large file size
- Stripping stop words/symbols
- Preparing for MALLET
- Train data in MALLET
- Infer topics from data
- Export back to JSON

Topic Modeling

- Topics created from groups of tokens (words), each weighted differently
- NLP standard is 20 top tokens, but Tweets are on average 15 words [1]
- Needed to be specific
- The three most heavily weighted tokens from the most relevant topic was returned

[1] http://blog.oup.com/2009/06/oxford-twitter/

Uploading to Solr

- Add schema.xml with our json fields
- Upload to Solr
- Now searchable by topics

Results

- All tweets and RSS stories associated with topics
  - virginia tech hokies
  - county board supervisors
  - police crash vehicle
  - veterans war military
  - food pantry program
  - rotary club blacksburg
- Tweets associated with hashtags

<topic titles="big ten, marshall wood fractures foot reducing, smokey classic, big ten challenge, basketball freshman marshall wood turning heads preseason practices, georgia, join, conference, " totalTokens="13853" alpha="0.2" id="6">

<word count="2239" weight="0.1616256406554537">big</word>
<word count="857" weight="0.06186385620443225">georgia</word>
<word count="785" weight="0.05666642604490002">ten</word>
<word count="652" weight="0.04706561755576409">join</word>
<word count="383" weight="0.02764744098751173">conference</word>

<phrase count="98" weight="0.08376068376068375">big ten</phrase>
<phrase count="42" weight="0.03589743589743589">marshall wood fractures foot reducing</phrase>
<phrase count="35" weight="0.02991452991452991">smokey classic</phrase>
<phrase count="31" weight="0.02649572649572649">big ten challenge</phrase>
<phrase count="27" weight="0.02307692307692308">basketball freshman marshall wood turning heads preseason practices</phrase>

</topic>

- Dashboard
- Logging
- Core Admin
- Java Properties
- Thread Dump

collection1

- Overview
- Analysis
- Dataimport
- Documents
- Files
- Ping
- Plugins / Stats
- Query
- Replication
- Schema Browser

Request-Handler (qt)

/select

common

q

topics : spring

fq

sort

start, rows

0 10

fl

df

Raw Query Parameters

key1=val1&key2=val2

wt

json

indent

debugQuery

dismax

edismax

hl

facet

spatial

spellcheck

Execute Query

http://localhost:8983/solr/collection1/select?q=topics+%3A+spring&wt=json&indent=true


```
{
  "responseHeader": {
    "status": 0,
    "QTime": 2,
    "params": {
      "q": "topics : spring",
      "indent": "true",
      "wt": "json",
      "_": "1399225837997"
    }
  },
  "response": {
    "numFound": 1726,
    "start": 0,
    "docs": [
      {
        "id": "164912",
        "title": [
          "RT @snowshoemtn: Do Spring Break right this year. #TheBallhooter #SpringBreak2013 http://t.co/LMGA7SWr"
        ],
        "description": "NULL",
        "url": "NULL",
        "views": "0",
        "created_at": [
          "2013-02-02 05:47:26"
        ],
        "updated_at": [
          "2013-02-02 05:47:26"
        ],
        "user_id": "NULL",
        "topic_id": "-1",
        "kind": "4",
        "popularity": 0,
        "image_file_name": "NULL",
        "image_content_type": "NULL",
        "image_file_size": "NULL",
        "image_updated_at": "NULL",
```

- Dashboard
- Logging
- Core Admin
- Java Properties
- Thread Dump

collection1

- Overview
- Analysis
- Dataimport
- Documents
- Files
- Ping
- Plugins / Stats
- Query
- Replication
- Schema Browser

Request-Handler (qt)

/select

common

q

hashtags : #Jobs

fq

sort

start, rows

0 10

fl

df

Raw Query Parameters

key1=val1&key2=val2

wt

json

indent

debugQuery

dismax

edismax

hl

facet

spatial

spellcheck

Execute Query

http://localhost:8983/solr/collection1/select?q=hashtags+%3A+%23Jobs&wt=json&indent=true


```
{
  "responseHeader": {
    "status": 0,
    "QTime": 1,
    "params": {
      "q": "hashtags : #Jobs",
      "indent": "true",
      "wt": "json",
      "_": "1399224933395"
    }
  },
  "response": {
    "numFound": 82,
    "start": 0,
    "docs": [
      {
        "id": "168414",
        "title": [
          "General Manager-United States-Christiansburg http://t.co/TkOhnySY\n #Jobs"
        ],
        "description": "NULL",
        "url": "NULL",
        "views": "0",
        "created_at": [
          "2013-02-06 21:09:03"
        ],
        "updated_at": [
          "2013-02-06 21:09:03"
        ],
        "user_id": "NULL",
        "topic_id": "-1",
        "kind": "4",
        "popularity": 0,
        "image_file_name": "NULL",
        "image_content_type": "NULL",
        "image_file_size": "NULL",
        "image_updated_at": "NULL",
```

Future work

- Modify NLP tool parameters
  - Enhance topic association
- Use different NLP tool/algorithm