Social Media Data

L715/B659

Dept. of Linguistics, Indiana University

Fall 2016

The Plan

1. Challenges with social media data
2. NLP additional processing (normalization, etc.)
3. Acquiring social media data

Much of the material is based on Farzindar & Inkpen (2015), *Natural Language Processing for Social Media*

- If you want to know more, consider taking ILS-Z639, *Social Media Mining*

Challenges with social media data

- Lack of context in data (cf., e.g., 140 character limit)
- Redundancy (e.g., retweets)
- Noise (e.g., spam)
- Data sparsity for particular users or phenomena
- Non-traditional language usage (e.g., abbreviations, misspellings, nonces, ungrammaticalities)
- Subjectivity of information
- Topic drift
- Lack of & bias in true ground truth (e.g., age of users)

Opportunities with social media data

- A chance to work with large amounts of data
- Many practical applications (health care, politics, defense, advertising)
- New opportunities for analyzing language & variability

Example Twitter texts

NLP processing

We have two main questions:

1. What new NLP tools do we need to process social media data?
   - Text normalization ...
2. What adaptations do we need to make to current NLP tools?
   - Re-training on annotated social media data ...

Text normalization

1. Identify orthographic anomalies
   - Is a term in/out of the vocabulary (OOV)?
2. Normalize (“correct”) anomalies/errors

Q: To what extent is normalization necessary? To what extent can it be defined for certain terms?

Language identification

For some types of data, you may even need to confirm you're working with the language you want

- langid.py: https://github.com/saffsd/langid.py
- CLD2: https://github.com/CLD2Owners/cld2
- LangDetect: https://code.google.com/archive/p/language-detection/
- YALI: https://github.com/martin-majlis/YALI
- TextCat: http://odur.let.rug.nl/~vannoord/TextCat/

And specifically for Twitter:

- LDIG: https://github.com/shuyo/ldig

Re-training NLP tools

Questions:

1. How much annotated data can you get?
2. What type of annotated data do you need? e.g., will a mixed bag of social media data types work?
3. What can you do with lots of unannotated data?

Tokenization

Biggest issue: new kinds of tokens

- Hashtags, emoticons, etc.
- Regular expressions can typically handle such cases ...

POS tagging

Main challenge for POS tagging: OOV words

- Also: need to adapt tagsets (see next slide)

Success in:

- Using regular expressions for some Internet-specific terms (hashtags, etc.)
- Retraining on a small amount of tagged social media data
- Clustering words

POS tagset from Gimpel et al (2011)

N	common noun
O	pronoun
^	proper noun
S	nominal + possessive
Z	proper noun + possessive
V	verb (inc. copula, auxiliaries)
L	nominal + verbal, verbal + nominal
M	proper noun + verbal
A	adjective
R	adverb
!	interjection
D	determiner
P	pre- or post-position, subordinating conjunction
&	coordinating conjunction
T	verb particle
X	existential *there*, predeterminer
Y	X + verbal
#	hashtag
@	at-mention
~	discourse marker (continuation across tweets)
U	URL or email address
E	emoticon
$	numeral
,	puncutation
G	other

Chunking & parsing

Similarly, parsing works better by re-training on appropriate data

One can also employ **chunking** instead of full parsing

- Obtain non-recursive nominal structures

Tweeboparser is developed for Twitter data:
http://www.cs.cmu.edu/~ark/TweetNLP/

Acquiring social media data

Available data sets

Already collected available data sets:

- TREC Microblog Track (2011–2015): http://trec.nist.gov/data/microblog.html
  - Queries with document names relevant to query
  - Tweets2011 Corpus: http://trec.nist.gov/data/tweets/
- SemEval Task on Sentiment Analysis in Twitter (2013–2016):
  - http://alt.qcri.org/semeval2016/task4/
  - http://alt.qcri.org/semeval2015/task10/
  - http://alt.qcri.org/semeval2014/task9/
  - https://www.cs.york.ac.uk/semeval-2013/task2/
- TAC 2008 Opinion Summarization:
  - http://tac.nist.gov/2008/summarization/

Acquiring social media data

Available data sets (2)

- Making Sense of Microposts Challenges
  - Entities liked to DBpedia resources
  - e.g., http://oak.dcs.shef.ac.uk/msm2013/
- EMNLP Workshop on Computational Approaches to Code Switching
  - http://care4lang1.seas.gwu.edu/cs2/call.html
  - http://emnlp2014.org/workshops/CodeSwitch/call.html
- PAN Challenges on Author Profiling
  - http://pan.webis.de/clef16/pan16-web/author-profiling.html
- Blog Authorship Corpus:
  - http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm
- ICWSM (2009–2011) datasets:
  - http://www.icwsm.org/data/
- Stanford Large Network Dataset Collection:
  - http://snap.stanford.edu/data/

Acquiring social media data

General web scraping

Simplest condition: you know what you're looking for → use something like `urllib`

```
>>> import urllib.request
>>> with urllib.request.urlopen('http://python.org') a
...     html = response.read()
>>> html[:30]
b'<!doctype html>\n<!--[if lt IE '
```

Includes tools for CGI (form-filling), error handling, authentication handling

https://docs.python.org/3/howto/urllib2.html

Acquiring social media data

General web scraping (2)

Next: you want to crawl from a particular source point (e.g., blog host) → use something like `scrapy`

- https://scrapy.org

Basic steps:

1. Import `scrapy`
2. Create a spider which scrapes the desired type of information
3. Run with `scrapy runspider`

http://doc.scrapy.org/en/1.1/intro/overview.html

Acquiring social media data

General web scraping (3)

Also recommended is `RoboBrowser`

> *RoboBrowser is a simple, Pythonic library for browsing the web without a standalone web browser. RoboBrowser can fetch a page, click on links and buttons, and fill out and submit forms. If you need to interact with web services that don't have APIs, RoboBrowser*

https://robobrowser.readthedocs.io/en/latest/readme.html

Acquiring social media data

General web scraping (4)

I'm providing an example script from Paul Richards on using Lang-8 data (username/password required)

- Tip from Paul: "When in doubt, read, read, read the html source page."

Acquiring social media data

API collection

Next: you want data from a site with an API (application programming interface):

- https://gigaom.com/2010/10/29/using-apis-not-quite-as-hard-as-it-looks/
- Note that APIs typically have some kind of rate limit

e.g., the Twitter API

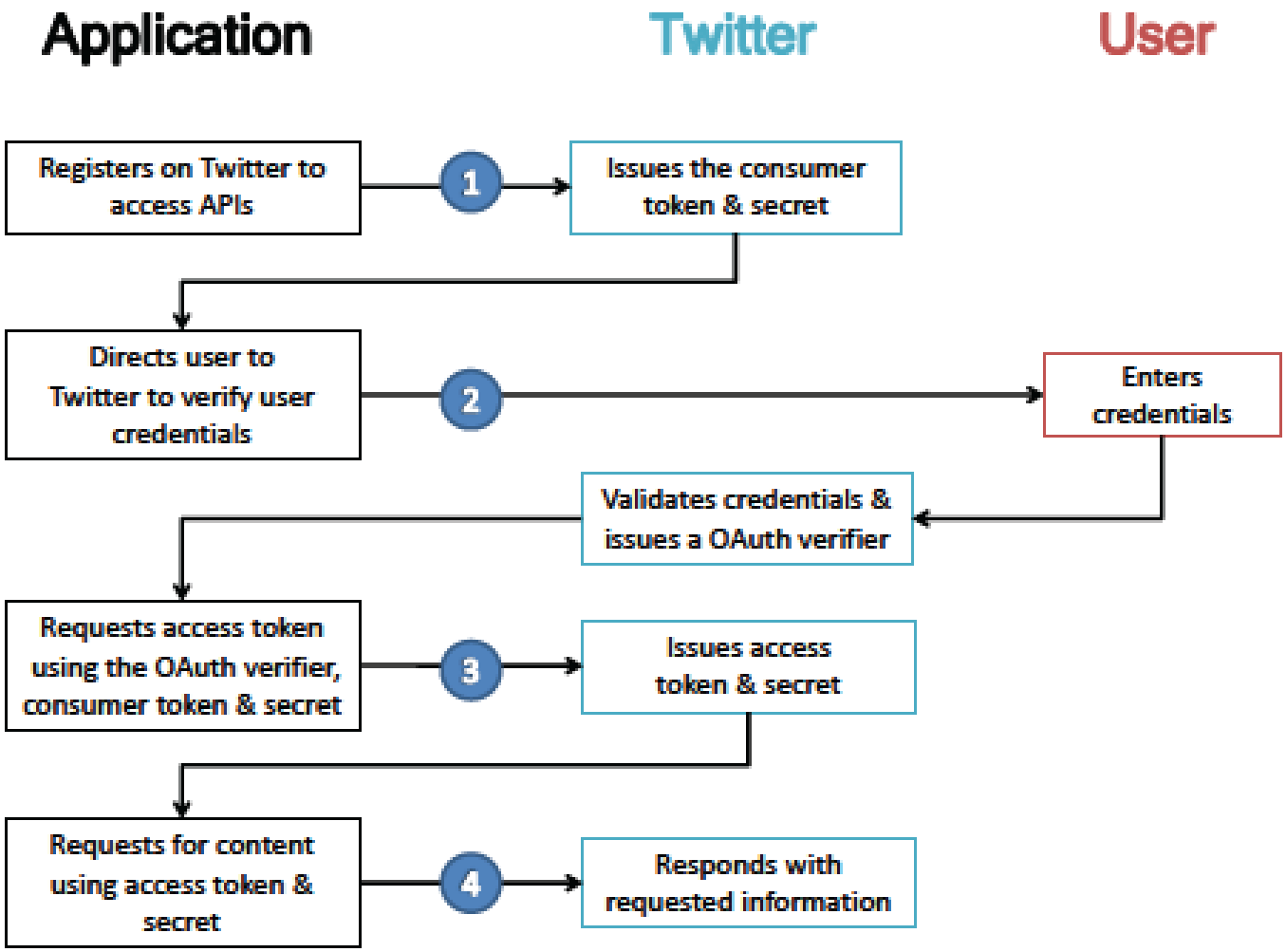
- https://dev.twitter.com/overview/api

Useful book on *Twitter Data Analytics*:

- http://tweettracker.fulton.asu.edu/tda/

Acquiring social media data

API collection: Twitter (Kumar et al 2013, fig. 2.1)



Acquiring social media data

Specific libraries

Python library for getting Twitter data:

- http://www.tweepy.org
- Requires you to have Twitter credentials (i.e., an account)
- Lots of documentation & code snippets

My Twitter Scraper gives CSV output:

- https://sourceforge.net/projects/mytwitterscraper/

Tools for TREC Microblog tasks (2011–2015):

- https://github.com/lintool/twitter-tools
- Mainly for TREC data, but plugs into Lucene Analyzer (https://lucene.apache.org)

Acquiring social media data

Spam & noise detection

Some issues:

- Language identification
  - http://www.aclweb.org/aclwiki/index.php?title=Language_Identification_Tools
- Duplicate or near-duplicate document detection (including retweets)
  - http://www.wis.ewi.tudelft.nl/duptweet/
  - http://bootcat.sslmit.unibo.it
- Informative document detection
- Spam & deception detection
  - Often requires filtering based on # of followers, posting patterns, etc.
- Metadata recovery

Acquiring social media data

Ethical issues

- Privacy policies
  - e.g., from the PAN site: “to Twitter’s privacy policy we cannot provide tweets directly, but only URLs referring to them. You will have to download them yourself. . . .”
- User misunderstandings
- Bugs allowing unauthorized access
- Lack of ethics in marketing