Retrieving and Parsing Linguistic Expressions of Political Attitudes

Paul Nulty

Department of Methodology,
London School of Economics and Political Science,

QUANTESS ERC Project

Computational Social Science, ECCS 2014
24th September 2014

Social information modes on online networks

- Network structure (friend, follower, subscriber)
- Simple actions: like, retweet, mention, favorite
- Multimedia: links, animations, videos, images
- Linguistic (text): Posts, comments, tweets

Linguistic communication

- Social media offers large, real-time broadcast text corpus of spontaneous communication and expression
- Retrieval depends on bursty and ambiguous search terms
- NLP offers methods to discover structure and help retrieval

Natural language on twitter

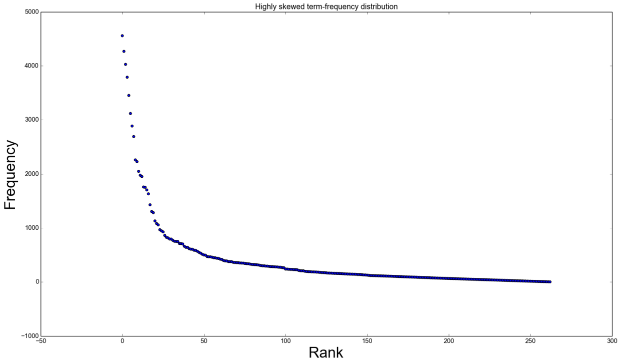
- twitter language is non-standard, but can be normalized [1]
- Simple statistical linguistics can aid retrieval
- Syntactic structure of statements can be extracted
- Applications: twitter as sensor for public health, natural disasters, sentiment

[1] Syntactic normalization of twitter messages, Kaufman and Kalita 2010)

Zipf's laws

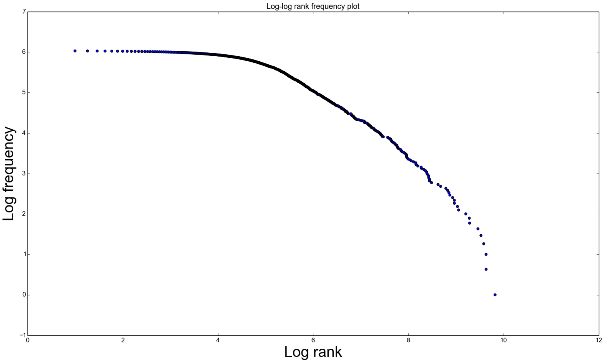
- In natural languages, word frequencies have a very heavy-tailed distribution
- Zipf's Law (1935): The frequency of a word is inversely proportional to its rank in the frequency table
- Zipf (1945): The more frequent a word is, the more senses it is likely to have
- frequent search terms give high recall, but low precision

Rank frequency of terms



From 260,619 tweets (no retweets), from twitter 'gardenhose' api on Scottish referendum day, containing any of these terms: `["#indyref", "salmond", "cameron", "scotland", "scottish", "referendum", "vote", "voted", "voting"]`

Log-Log Rank frequency



From 260,619 tweets (excluding retweets, 1.02M total), from twitter ‘gardenhose’ api on Scottish referendum day, containing any of these terms: `["#indyref", "salmond", "cameron", "scotland", "scottish", "referendum", "vote", "voted", "voting"]`

Precision vs Recall

- Initially, prefer recall over precision
- Hone search terms by learning association between terms and concept of interest
- e.g. Initially search for "vote", "cameron", "indyref"
- learn which terms co-occur with precise concept of interest

Example: Naive Bayes classifier for Scottish referndum opinion

- Treat #bettertogether and #voteyes as training labels
- Simple bag-of-words model (without hashtag labels, 800 most common terms)
-
- predicted 1866 out of 2367 tweets correctly (79%) (out of sample)
- Most informative features identify useful terms for further search.

terms predictive of 'no'(#bettertogether and #nothanks)

term	Direction	Ratio
kingdom	no	20.7 : 1.0
stupid	no	16.1 : 1.0
united	no	15.0 : 1.0
stay	no	8.9 : 1.0
#no	no	8.6 : 1.0
#uk	no	6.6 : 1.0
#votenoscotland	no	6.1 : 1.0
#voteno	no	5.5 : 1.0
union	no	5.2 : 1.0
sense	no	4.8 : 1.0
enough	no	4.8 : 1.0
uk	no	4.4 : 1.0
britain	no	4.3 : 1.0
leave	no	4.3 : 1.0

terms predictive of 'yes'(#voteyes and #yesscotland)

term	Direction	Ratio
#freedom	yes	29.8 : 1.0
#letsdothis	yes	20.9 : 1.0
#voteaye	yes	20.2 : 1.0
#savvy	yes	18.0 : 1.0
imagine	yes	9.8 : 1.0
opportunity	yes	9.4 : 1.0
fairer	yes	9.2 : 1.0
#hopeoverfear	yes	8.1 : 1.0
hands	yes	7.6 : 1.0
society	yes	7.0 : 1.0
#independence	yes	6.7 : 1.0
excited	yes	6.3 : 1.0
atsymb_nicolasturgeon	yes	5.9 : 1.0
brave	yes	5.5 : 1.0

Structured Natural language processing

- Recent methods in NLP have moved beyond bag-of-words model
- Named entity recognition, co-reference resolution,
- Typed dependency parsing
- Detect specific expressions rather than term mentions
- Stanford Core NLP toolkit [2]

[2] (Manning et al, ACL 2014)

Stanford Core NLP: Named entities

```
[u'the',
  u'Lemma': u'the',
  u'NamedEntityTag': u'O',
  u'PartOfSpeech': u'DT'}],
[u'NHS',
  u'Lemma': u'NHS',
  u'NamedEntityTag': u'ORGANIZATION',
  u'PartOfSpeech': u'NNP'}],
[u'budget',
  u'Lemma': u'budget',
  u'NamedEntityTag': u'O',
  u'PartOfSpeech': u'NN'}],
[u'in',
  u'Lemma': u'in',
  u'NamedEntityTag': u'O',
  u'PartOfSpeech': u'IN'}],
[u'Scotland',
  u'Lemma': u'Scotland',
  u'NamedEntityTag': u'LOCATION',
  u'PartOfSpeech': u'NNP'}],
[u'is',
  u'Lemma': u'be',
  u'NamedEntityTag': u'O',
  u'PartOfSpeech': u'VBZ'}],
[u'100',
  u'Lemma': u'100',
  u'NamedEntityTag': u'PERCENT',
  u'NormalizedNamedEntityTag': u'%100.0',
  u'PartOfSpeech': u'CD'}],
```

'atsymb_YesScotland But the NHS budget in Scotland is 100% under Scottish Parliament control so privatisation is not an issue.'

Stanford Core NLP: typed dependency parse

```
[{dependencies': [[root', ROOT', gives'],
                  [csubj', gives', voting'],
                  [dobj', voting', Yes'],
                  [iobj', gives', hashsymb_Scotland'],
                  [det', position', a'],
                  [amod', position', better'],
                  [dobj', gives', position'],
                  [prep_in', position', Europe/UK'],
                  [det', cases', all'],
                  [prep_in', Europe/UK', cases']]
```

'voting Yes gives hashsymbScotland a better position in Europe/UK in all cases'

Classification with dependency relations as features

- Treat #bettertogether and #voteyes as training labels
- Without relations involving hashtag labels, 60 most common dependencies)
- 76% accuracy out of sample
- Need a much bigger corpus for complex features

Most informative dependency relations (yes side)

term	Direction	Ratio
dobj_make_history	yes	6.9 : 1.0
dobj_do_this	yes	5.6 : 1.0
advmod_important_most	yes	5.1 : 1.0
nsubj_vote_we	yes	5.1 : 1.0
nsubj_do_we	yes	4.5 : 1.0
nsubj_do_'s	yes	4.3 : 1.0
nn_Scotland_luck	yes	3.9 : 1.0