Retrieving 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

Modes of information communication in online social networks

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

Introduction

- Text is a hugely rich but unstructured information source
- Social media offers large, real-time corpus of spontaneous communication and expression
- Retrieval depends on bursty and ambiguous search terms
- Simple word frequency matrix methods, also rich latent structure
- NLP offers methods to discover structure and help retrieval
- Twitter‘s communication model makes it especially useful

Natural language on twitter

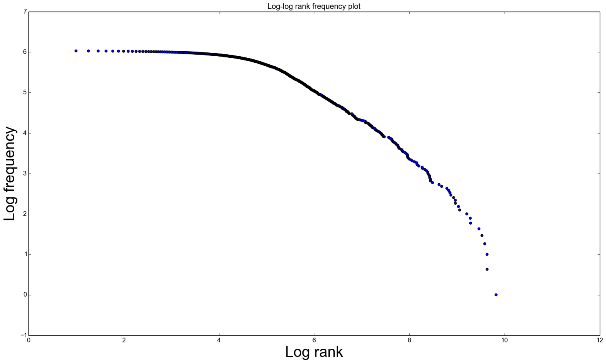
- Text is the principal mode of communication broadcast on twitter
- Limit on post length causes some issues, but fixable [1]
- Simple statistical linguistics can aid retrieval
- Linguistic structure can be identified with parsing

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

Zipf's laws

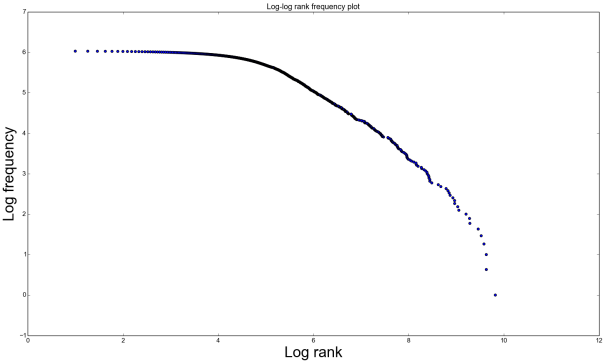
- In natural languages, word frequencies have a very heavy-tailed distribution
- Zipf's Law (1935): The frequency of a word is inversely proportional to its rank in the frequency table
- Zipf (1945): The more frequent a word is, the more senses it is likely to have
- frequent search terms give high recall,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

Rank frequency of terms



Data 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



Data 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"]`

Discovering query terms with a classifier

- 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 terms of interest

Example Naive Bayes classifier

- Train
- Simple bag-of-words model
-

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