NATURAL LANGUAGE PROCESSING

Gayane Petrosyan
McGill University

Luys DAP 2012

Text classification

- Spam classification
- Authorship classification
- Age/Gender classification
- Sentiment analysis
- Subject identification
- Language Identification

Text Classification
Supervised learning

- Input
  - Document **d**
  - A fixed set of classes $\boldsymbol{C = \{c_1, c_2, \ldots c_n,\}}$
  - A training set **m** hand-labeled documents $\boldsymbol{(d_1, c_1), \ldots (d_m, c_m)}$
- Output
  - A learned classifier **y:d->c**

Naïve Bayes

- Document is bag of words(only count of words is considered)
- Bayes, because Bayes rule is used
  - $P(c/d) = \frac{P(d/c)P(c)}{P(d)}$
  - $\mathrm{C} = \mathrm{argmax}\, P(c/d) = argmax\, \frac{P(d/c)P(c)}{P(d)} = argmax\, P(d/c)P(c)$
- Naïve, because
  - $\mathrm{C} = argmax\, P(d/c)P(c) = argmax\, P(x_1 \dots x_n/c)P(c) \approx argmax\, P(x_1/c) \dots P(x_n/c)P(c)$
  - $P(c) = \frac{count(c)}{N}$
  - $P(w/c) = \frac{count(w,c)+1}{count(c)+|V|}$

Google search engine

- Spiders/Googlebot/Google’s Web Crawler
- Building index
- PageRank

Before Indexing

- Tokenize (split words)
- Stop words removal (a, the, …)
- Stemming (friends, friendship->friend,)
- Normalization (U.S.A->USA)

Search Index

- Term-document matrix

	D1	D2	…	Dn
W1				
W2				
…				
Wm				

- Posting list
  - Word1 -> d1, d2, … dj
- Positional indexes
  - Word1 -> d1(p11,…pi1), d2(p12…pi2), … dj(p1j…pij)

Ranking

- The more the term appear in the doc the higher the score
- Best known: TF-IDF
  - TF: term frequency of the document
  - Df: document frequency of the term

- How long the Web page has existed?
- Is the term in the title
- Is the term in the URL

- Google uses: The number of other Web pages that link to the page in question (PageRank)
- http://www.youtube.com/watch?v=BNHR6IQJGZs

Personalization

- “Zuckerberg Said,

‘A Squirrel Dying In Your Front Yard May Be More Relevant To Your Interests Right Now Than People Dying In Africa’”

- Google uses 57 signals like
  - What computer you are using?
  - What browser you are using?
  - Where are you located ?
  - …

- http://www.ted.com/talks/lang/en/eli_pariser_beware_online_filter_bubbles.html