Topic Modeling

Natural Language Processing Laboratory
2017/05/09

What is Topic Modeling?

A method extract thematic structure from corpus

Latent Dirichlet Allocation

- An unsupervised machine learning method
- Widely used in topic discovery from corpus
- Can be used for document classification as well
- Lots of incomprehensible math
  - Dirichlet distribution, Bayesian inference, EM algorithm, Gibbs sampling, …

Related Paper

- LDA (origin)
  - http://www.cs.columbia.edu/~blei/papers/BleiNgJordan2003.pdf
- hLDA
  - https://papers.nips.cc/paper/2466-hierarchical-topic-models-and-the-nested-chinese-restaurant-process.pdf

Recent Research

- Topic2Vec
  - https://arxiv.org/abs/1506.08422
  - https://github.com/scavallari/Topic2Vec
- Topical Word Embedding
  - http://nlp.csai.tsinghua.edu.cn/~lzy/publications/aaai2015_twe.pdf
  - https://github.com/largelymfs/topical_word_embeddings
- lda2vec
  - https://arxiv.org/abs/1605.02019
  - https://github.com/cemoody/lda2vec

Useful Python Packages

- scikit-learn
  - http://scikit-learn.org/stable/auto_examples/applications/topics_extraction_with_nmf_lda.html
- gensim
  - http://radimrehurek.com/gensim/wiki.html#latent-dirichlet-allocation
- pyLDAvis
  - Topic Modeling Visualization
  - http://nbviewer.jupyter.org/github/bmabey/pyLDAvis/blob/master/notebooks/pyLDAvis_overview.ipynb

Lab 10: Topic Modeling on News

- Sports news from Yahoo News
- In Chinese
- 2920 documents
- Need segmentation (use jieba)
- Use either *scikit-learn* or *gensim*

Topics in LDA model:

Topic #0: 冠軍 球員 2017 世界 中華隊 比賽 網球 今年 美國 去年 20 球隊 生涯 籃球 選手 拉波娃 一朗 拿下 10 奪冠

Topic #1: 比賽 球隊 今日 認為 教練 nownews2017 記者 賽後 總教練 希望 相當 球員 對於 因此 綜合 受傷 隊友 能夠 表現 球迷

Topic #2: 本季 先發 敲出 投手 全壘打 支安打 三振 局下 今日 比賽 nownews2017 綜合 擊出 打擊 表現 失分 記者 二壘 安打 局上

Topic #3: 兄弟 中信 職棒 富邦 中華 悍將 統一 中央社 桃猿 記者 中職 lamigo 獅隊 盧彥 2017 球迷 棒球場 桃園 主場 開球

Topic #4: 季後賽 nba 籃板 勇士 助攻 騎士 領先 爵士 馬刺 巫師 火箭 暴龍 攻下 系列賽 比賽 公牛 表現 快艇 取得 塞爾

Topic #5: 比賽 拿下 21 頭殼 晉級 擊敗 冠軍 隊伍 勝利 2017 英雄 選手 newtalk 對手 閃電 直落 新加坡 聯盟 戴資穎 超級

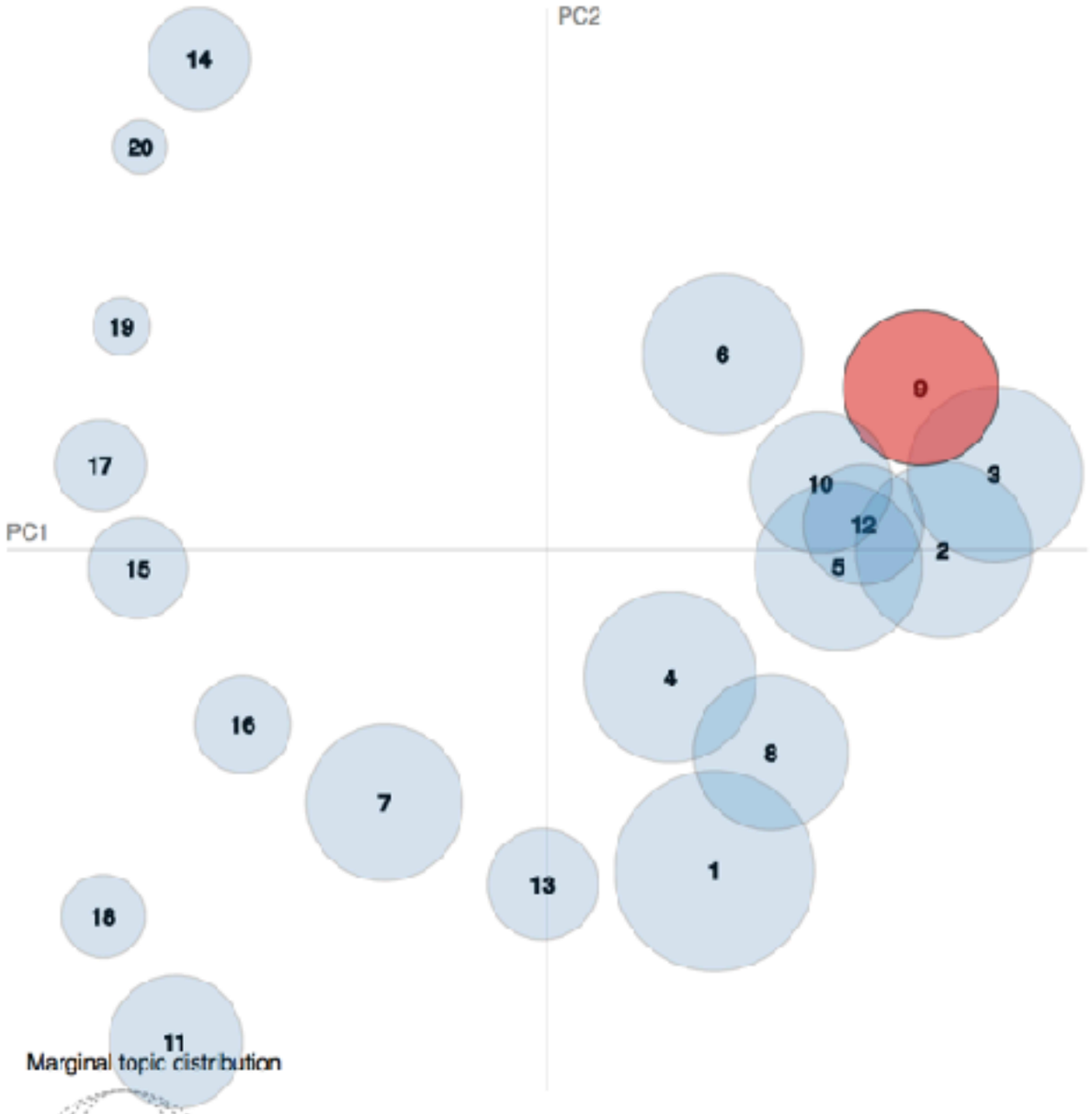
Topic #6: 聯盟 胡智為 光芒 投手 職棒 美國 棒球 先發 偉殷 2017 比賽 3a 出賽 機會 上場 牛棚 名單 登板 去年 生涯

Topic #7: 選手 運動 2017 體育 比賽 參加 戴資穎 全國 協會 賽事 金牌 活動 世界 大運 羽球 記者 今年 國際 希望 舉辦

Topic #8: 偉殷 林魚 全壘打 比賽 費城 無安打 水手隊 水手 大都會 先發 投手 魚隊 本季 退場 殷仔 滿貫 保送 聯盟 失分 2017

Topic #9: 選手 並列 打出 標準桿 高爾夫 成績 低於 領先 公開賽 桿的 日本 俱樂部 golf 排名 2017 綜合 球場 回合 冠軍 世界排名

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 9 (6.5% of tokens)