



New Media Data Analytics and Application

Lecture 8: Natural Language Processing
A Brief Introduction

Ting Wang

- Data Analysis for Online Journalism
- Natural Language Processing
- Semantic Resource for NLP
- Keyword Analysis





some analysis approaches for data journalism

Data Analysis for Online Journalism

Data Analysis for Online Journalism



Ask A Question



Now, We have data.

What shall we do in the next step for online journalism analysis?



Data Analysis for Online Journalism

1. How to extract information from data?

Data Clean and Preprocessing

2. How to measure the information of news?

Quantitative Modules

3. How to analysis these information?

Comparison, Classification and Clustering



Data Analysis for Online Journalism

EXAMPLE 1:
papi酱



围脖关键词

围脖关键词利用自然语言处理的关键词抽取技术，分析用户近期发表微博内容，提取代表用户兴趣的关键词，并采用文档可视化技术对关键词进行可视化，便于用户快速了解自己、好友、主题等的关键词。

使用以下账号登录



<http://app.thunlp.org/>



Data Analysis for Online Journalism

April, 2016

November, 2016



papi酱



Data Analysis for Online Journalism

罗辑思维退出papi酱令人深思 网红经济不行了吗？

3685

我要评论

2016-11-25 15:23 来源：新京报



罗辑思维退出papi酱，网红经济不行了吗？

对个人形象的克制使用和保持健康，才是网红经济不成为一锤子买卖的关键。

据报道，罗辑思维已与著名网红papi酱分手。记者调查发现，早在今年8月29日，papi酱所在的公司春雨听雷在股东一栏里就去掉了罗辑思维的投资经营主体北京思维造物投资管理有限公司。

罗辑思维退出papi酱的背后：互联网风口关闭

2016年11月26日 09:59

创事记

微博

作者：maomaobear

我有话说(5人参与)

A⁻

A⁺

+ 订阅



欢迎关注“创事记”的微信订阅号：sinachuangshiji



Data Analysis for Online Journalism

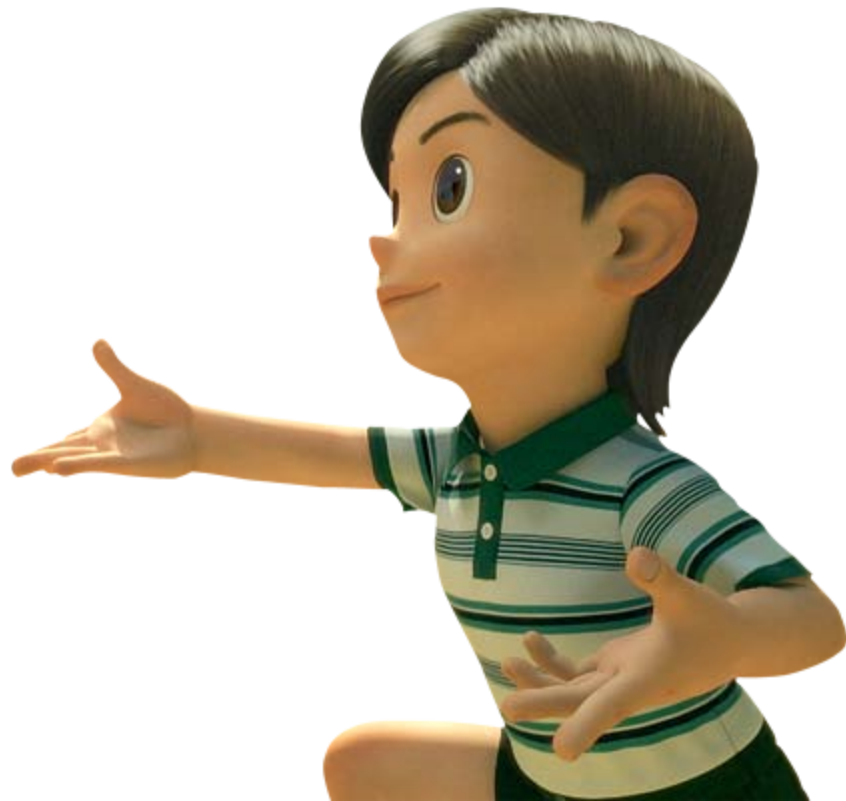
Technical Approaches

1. Keyword Extraction and Tag Analysis
2. News Tracking
3. News Alignment and Comparison
4. Location-based News Analysis
5. ...



Data Analysis for Online Journalism

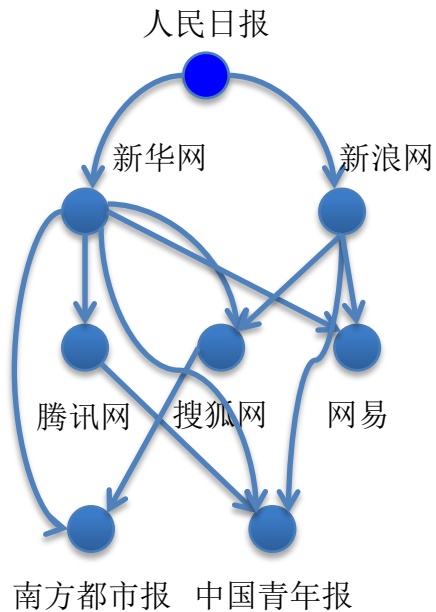
1. Keyword Extraction and Tag Analysis



Data Analysis for Online Journalism

2. News Tracking

人民日报评论《“小”了时代，窄了格局，矮了思想》传播路径图



Communication Efficiency Analysis base on Big Data

媒体	传播效率
新华网	71.67
新浪网	31.0
搜狐网	16.2
中国青年报	9.22
南方都市报	6.41



Data Analysis for Online Journalism

3. News Alignment and Comparison

韩国推迟与日本签署军事协定 被指外交失礼

2012年06月29日 17:37:31
来源: 中国新闻网

【字号: 大 中 小】 【打印】

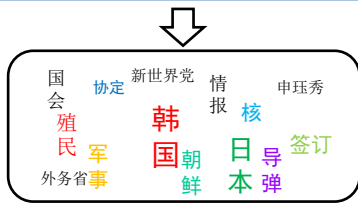
【纠错】

据韩联社报道, 韩国政府决定推迟原定在当地时间29日下午4点签署《韩日军事情报综合保护协定》(GSOMIA)。

韩国外交消息人士表示, 根据朝野要求, 决定在正式签署前先向国会进行说明。今后的日程不得而知。

韩国政府秘密推进签署军事情报协定引发了舆论非议, 因此新世界党要求政府推迟或取消签署协定。这虽然属于外交失礼, 但韩国政府接受了新世界党的提议。韩国驻日大使申珏秀则向日本外务省传达了韩国政府的立场。

《韩日军事秘密保护协定》将是韩国摆脱日本殖民统治后与日本之间签订的第一个军事协定。如果该协议签订, 韩日今后有望互通有关朝鲜军队、朝鲜社会动向、朝核以及导弹问题等方面的情报。(中新网6月29日电)



South Korea to Sign Military Pact With Japan

By CHOE SANG-HUN
Published: June 28, 2012

SEOUL, South Korea — In a significant step toward overcoming lingering historical animosities with its former colonial master, the South Korean government has unexpectedly announced that it will sign a treaty with [Japan](#) on Friday to increase the sharing of classified military data on what analysts cite as two major common concerns: [North Korea's](#) nuclear and missile threats and [China's](#) growing military might.

The announcement set off a political firestorm in South Korea, where resentment of Japan's early 20th-century colonization remains entrenched and any sign of Japan's growing military role is met with deep suspicion. The opposition accused President [Lee Myung-bak](#) of ignoring popular anti-Japanese sentiments in pressing ahead with the treaty, the first military pact between the two nations since the end of colonization in 1945.

Connect With Us on Twitter
Follow @nytimesworld for international breaking news and headlines.
Twitter List: Reporters and Editors

FACEBOOK
TWITTER
GOOGLE+
EMAIL
SHARE
PRINT
REPRINTS

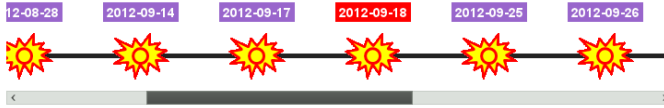
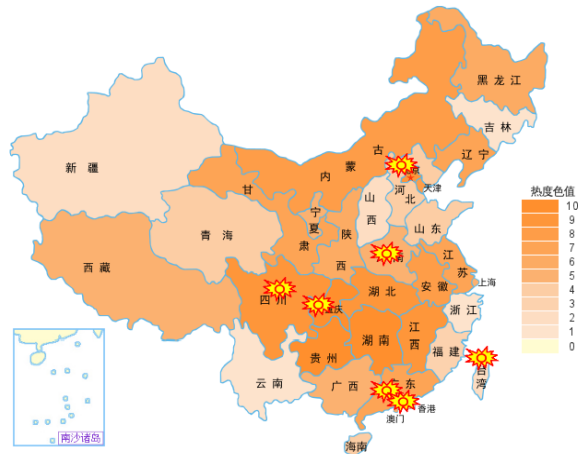
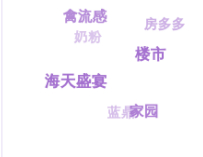
RUBY SPARKS COMING SOON



*Between
Different
Languages,
Websites,
Nations,
and
People*

- Notes : 1. Words in the same color have the same meaning in translation
2. The size of the word represents the importance of the word, the larger, the more important





大陆: 钓鱼岛事件波及日本在华大型企业。9月17日，包括佳能、松下和精工等在内的企业相继传来工厂生产暂停的消息。劳拉消息称，9月17日，华谊商谈、优衣库、7-11等日资零售企业也在这两天暂时停业。相关企业表示，这两天是为了保障员工的安全。记者昨晚从国防部新闻事务局有关负责人获悉，美国防部长莱昂·帕内塔已于昨晚抵达北京，开始对中国进行为期4天的正式访问。受钓鱼岛事件影响，近日中国多城市出现反日示威游行。

华人: 惠普国际评级周二表示，该机构认为，如果中日钓鱼岛争端升级并持续，日本大型汽车制造商和科技制造公司的评级可能承压。中日钓鱼岛争议持续升温，中国各地周末爆发反日示威游行。受此影响，日本多家汽车生产企业决定暂停在中国大陆工厂的生产。中国国防部长梁光烈周二上午与到访的美国防部长帕内塔举行会谈，在中日爆发钓鱼岛争端之际，帕内塔访华备受外界关注。

外媒: The 81st anniversary of a Japanese invasion brought a fresh wave of anti-Japan demonstrations in China on Tuesday, with thousands of protesters venting anger over the colonial past and a current dispute involving contested islands in the East China Sea. China's Warnings of Further Actions as Anti-Japan Protests Resume At least two of 11 Chinese ocean surveillance and fishery patrol ships sailing near East China Sea islets claimed by both Tokyo and Beijing have entered what Japan considers its territory.

4. Location-based News Analysis

- (1). Sentiment Analysis
- (2). Trend Analysis

博然般若 企业微博口碑分析

关于 语言

关注程度 关联转发 公众微博

总数: 84141

2013年 五月 1日 至 2013年 五月 31日

地区: 华东 华中 华南 华北 东北 西南 西北 东南 西南 西北

247 15,113

地图数据源自Google Chart Tools API, 商海通高平提供。

盛大 很好 稳定性 用户 都在 小球 品牌 你对 性也商务 电脑 zhouzi 暗恋 富士通 她的 主机 所有 生产 质量 我心 了解 性能 积极 本土 安全 日本 了我 我 陶醉 定位 非常好 西方 说谈 叶子 深博 外观

阿奇哥来红美
最近入手索尼录音笔D50，真是移动办公神器啊，录音录得实在，嗓子有点疲劳，睡了，明天继续

2013-5-31 23:59:51 来自 WeicoPro

Data Analysis for Online Journalism

柴静

EXAMPLE 2:
Under the Dome

穹顶之下

同呼吸共命运

一位资深记者的道义良心 一个普通母亲的社会责任

The influence by *Under the Dome*, made by Jing Chai, February 28, 2015

Data Analysis for Online Journalism

Data Description

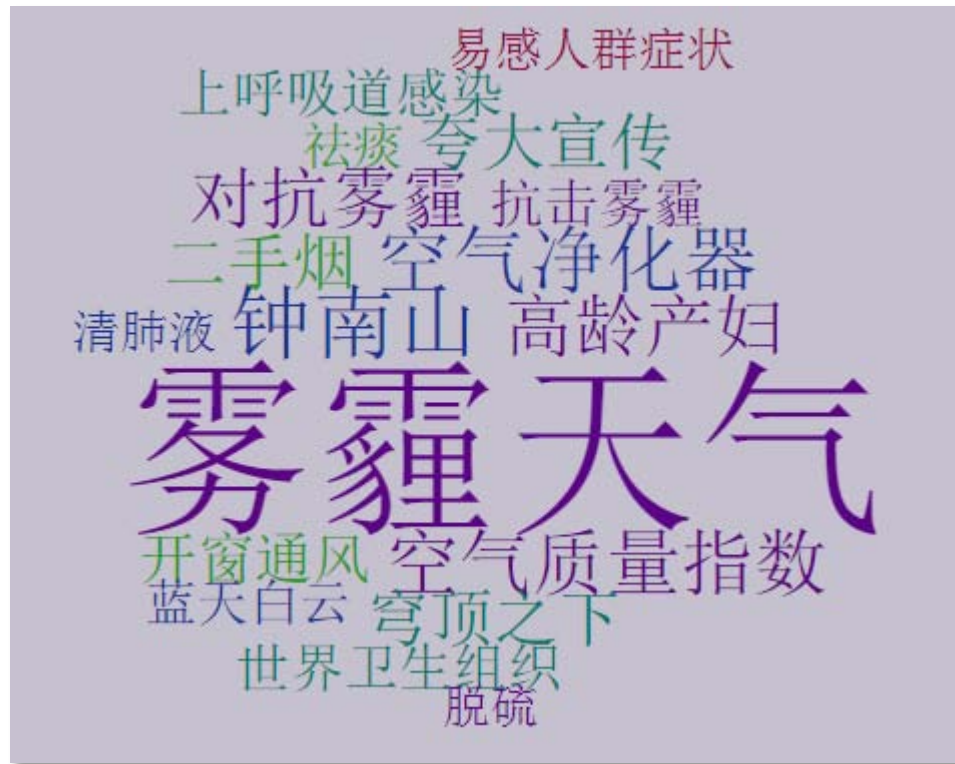
Keyword Category	Keywords	Number of Weibo	Number of Weibo without repetition
呼吸系统疾病	呼吸系统疾病 支气管炎 哮喘 咳嗽感冒 胃肠型感冒、咽炎、支气管肺炎、上呼吸道感染、尘肺、结核病、鼻炎、咽喉炎、鼻窦炎、扁桃体炎	11321	6648
肿瘤	肿瘤 新生儿肿瘤 肺癌 肺肿瘤	10655	7005
汽油	汽油质量	16	14
发电	发电、电力行业	1620	1032
净化	净化器、清新剂	9887	4770
煤	燃煤、煤炭	6587	4613
	总 计	40086	24082



Data Analysis for Online Journalism

Keyword Extraction Based on Weibo

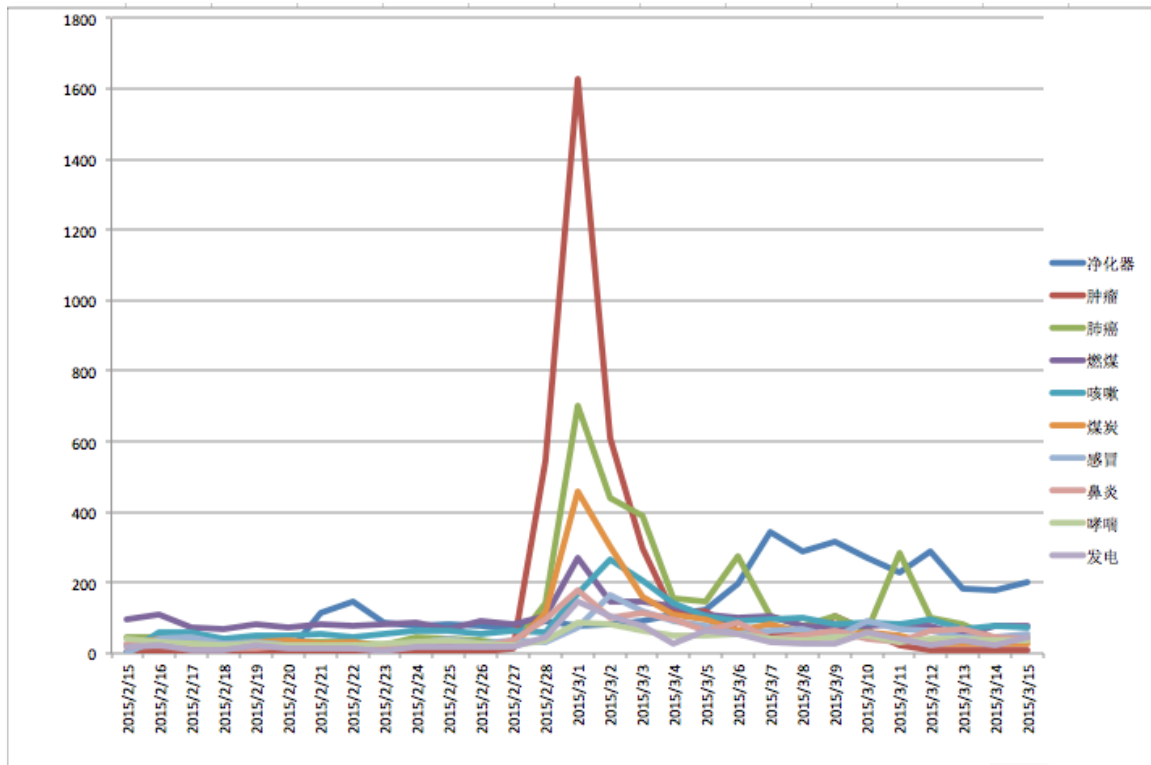
- What can you find in this graph?
- What will you do for your group?



Data Analysis for Online Journalism

Keyword Analysis Based on Weibo

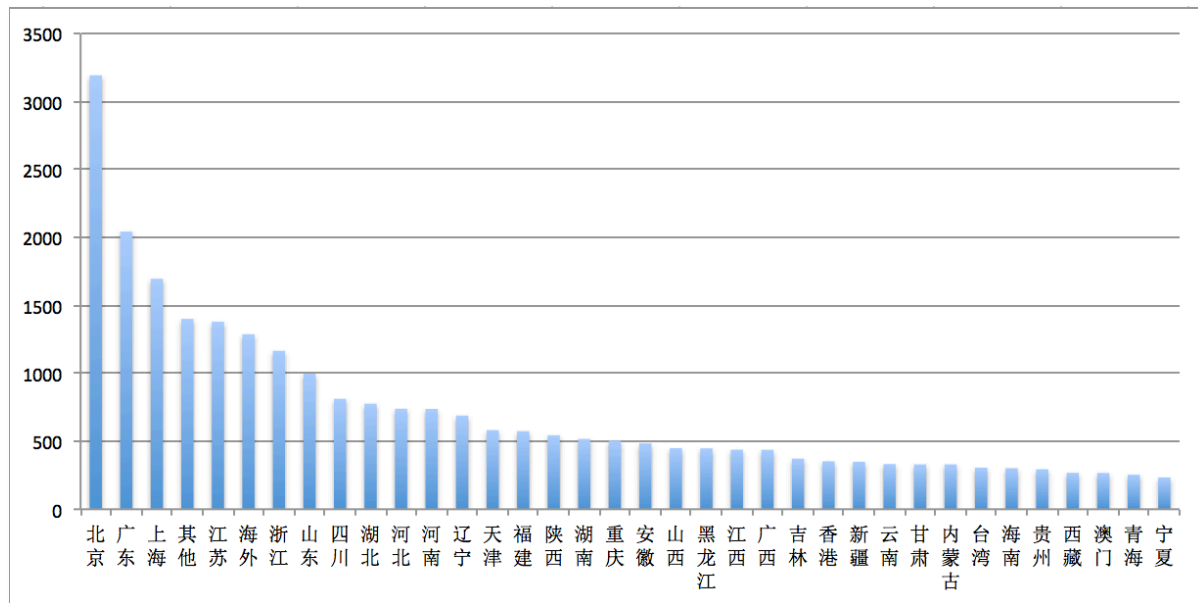
- What can you find in this graph?
- What will you do for your group?



Data Analysis for Online Journalism

Keyword Analysis Based on Weibo

- What can you find in this graph?
- What will you do for your group?



Data Analysis for Online Journalism

Conclusions:

- Keyword is an abstract of online media
- The frequency of using keywords is important



Data Analysis for Online Journalism

Correlative Scientific Technologies

Natural Language Processing

Statistics *Machine Learning*

ARTIFICIAL INTELLIGENCE

Machine Translation **Psychology**

Computer Science **Linguistics**



上海外国语大学

SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

a brief introduction to natural language processing

What is Natural Language Processing

Natural Language Processing

What is NLP

Natural language processing is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages.

- NLP is related to the area of human–computer interaction.
- NLP involves natural language understanding and natural language generation.
- Also called as Computational Linguistics



Natural Language Processing

Objectives

Let your computer know you, and let you know the world

Qualitative Data



Quantitative Data



Natural Language Processing

Tasks (1)



Automatic summarization 自动总结

Machine translation 机器翻译

Named entity recognition 命名实体识别

Natural language generation 自然语言生成

Natural language understanding 自然语言理解

Optical character recognition 光学字符识别

Part-of-speech tagging 词性标注

Parsing 语法解析

Question answering 问答系统

Relationship extraction 关系提取 (主体之间的关系)



Natural Language Processing

Tasks (2)



Sentence breaking 断句 (文言文, 语音)

Sentiment analysis 情感分析

Speech recognition 语音识别

Speech segmentation 语音切分 (词汇)

Topic segmentation and recognition 主题切分与识别

Word segmentation 分词 (中日韩等)

Word sense disambiguation 词汇歧义削减

Information retrieval 信息检索、信息过滤

Information extraction 信息抽取

Speech processing 语音处理 (文字语音互转)



Natural Language Processing

Approaches

1. Symbolicism 符号主义

- Regulation 规则（显性规则）：决策树DT
- Statistics 统计（隐形规则）：贝叶斯、HMM、PCA

2. Connectionism 链接主义

- Neural Networks 神经网络：Deep Learning

3. Actionism 行为主义

- Evolutionism 进化主义：遗传算法GA、PSO



Natural Language Processing

NLP using Python

NLTK (<http://www.nltk.org/>)

- Current Version : NLTK 3
- Installation (<http://www.nltk.org/install.html>)
- `import nltk`

Natural Language Processing

Foundations of NLP:

Semantic Resource 语义资源

- Dictionary 字典
- Stop Word List 停用词表
- Knowledge Graph 知识图谱
 - Knowledge Base 知识库
 - Semantic Networks 语义网络
- Regulation Base 规则库
- Corpus 语料库





上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

the foundation of NLP

Semantic Resource

Dictionary 字典

	ID	Chinese	English	Memo
1	1	α 值	Alpha	在股票收益方面, α 值衡量某种证券或基金经风险调整后的回报。α 值是代表证券收益率超出风险/收益模型所...
2	2	《美国破产法》第七章	Chapter 7	《美国破产法》第七章是关于非自愿清盘的法规, 债权人据此请求法庭颁令判决债务人破产。该章赋予由法庭...
3	3	《美国破产法》第十一章	Chapter 11	按《美国破产法》第十一章的安排, 无力偿债的债务人若成功申请破产保护, 将可保住企业的财产及经营的控...
4	4	3A等级	Triple A Rated	参见AAA/Aaa (3A等级)。 债券
5	5	3A等级 (最高信用评级)	AAA	给予优质债券的最高评级。由标准普尔、穆迪和惠誉国际等主要评级机构评定。参见Credit Rating (信用评级...
6	6	J-曲线	J-Curve	经济学上的一个概念, 指一个变量在受到某种刺激时, 有时可能会继续按原先的方向发展, 然后才出现明显的...
7	7	Vega值	Vega	量度期权标的资产价格波动率的变动如何影响期权价值的指标。参见Option (期权)。The measure of change...
8	8	β 值	Beta	贝塔系数是量度股票投资系统风险的指标。所谓系统风险是指股票投资中没有办法通过分散投资来减低的风险...
9	9	⊖ 系数	NULL	参见: ⊖ 值
10	10	θ 值	Theta	量度期权价值如何随着期权有效期的缩减而变动的指标。期权的价值会随着时间过去——即期权日益接近到期...
11	11	λ 值	Lambda	指量度期权杠杆水平的一个比率, 显示标的资产的价格每变动一个百分点, 可导致期权价格变动的百分比。标的...
12	12	阿尔法系数	NULL	参见: α 值
13	13	阿拉伯石油输出国组织	OAPEC	英文Organization of Arab Petroleum Exporting Countries的缩写。该组织的宗旨是促进阿拉伯产油国之间...
14	14	阿历山大过滤器	Alexander's Filter	指技术分析的一种方法, 以涨跌百分比来衡量特定时间内价格上涨或下跌的速度。升速很快为买进讯号, 反之...
15	15	阿姆斯特丹	ARA	英文Amsterdam/Rotterdam/Antwerp 的缩写。石油货品若称cost and freight ARA, 是指可将阿姆斯特丹/鹿特...
16	16	艾略特波浪理论	Elliott Wave Theory	技术分析的一种理论, 认为市场走势不断重复一种模式, 每一周期由5个上升浪和3个下跌浪组成。艾略特波浪...
17	17	安特卫普地区	NULL	参见: 阿姆斯特丹
18	18	按比例偿债基金	Pro Rata Sinking Fund	偿债基金是在债券到期前提前偿还部分债务的一种安排。债券发行时若附设偿债基金条款, 发债人必须定期将...
19	19	按揭证券	MBS	英文Mortgage-backed Security的缩写, 由一篮子住房抵押贷款提供担保的证券, 该抵押贷款组合每月收到的还...
20	20	按面值	At Par	指证券的售价与其面值相等。When a security is selling at a price that is equal to face value. 期货...



Stop Word List

停用词表

- Punctuation 标点
- Symbol 符号
- Function Word 虚词
- Interjection 叹词
- Empty Word 无意义的词
- Ambiguous Word 引起歧义的词

SW_ID	WORD_NAME	WPS_ID
35	'	1
36	/	1
37	\	1
38	\n	1
39	'	1
40	"	1
41	\t	1
42	:"	1
43	:“	1
44	啊	1
45	阿	1
46	哎	1
47	哎呀	1
48	哎哟	1
49	唉	1
50	按	1
51	按照	1
52	吧	1
53	吧哒	1
54	把	1



Knowledge Graph 知识图谱

	WIKI_WORD_ID	WEB_ID	WIKI_WORD
1	1	1	Alpha
2	2	2	Chapter_7
3	3	3	Chapter_11
4	4	4	AAA
5	5	5	Vega
6	6	6	Beta
7	7	7	Theta
8	8	8	Lambda
9	9	9	OAEPEC
10	10	10	ARA
11	11	11	Elliott_Wave_Theory
12	12	12	MBS
13	13	13	All_Ordinaries
14	14	14	G8
15	15	15	Paris_Club
16	16	16	MONEP
17	17	17	Pibor
18	18	18	COB
19	19	19	Basel_Committee
20	20	20	White_Knight



The screenshot shows the Wikipedia article for 'Alpha'. It includes the title 'Alpha', a subtitle 'From Wikipedia, the free encyclopedia', and a large image of the Greek letters 'Αα'. Below the image are sections for 'Greek alphabet', 'Other characters', and 'Greek dactilo'. The page also contains a 'References' section with several entries.



	WIKI_RELATIVE_WORD_ID	WEB_ID	WIKI_WORD	RELATION
1	1	1	Greek_alphabet	0
2	2	1	Beta	0
3	3	1	Gamma	0
4	4	1	Omicron	0
5	5	1	Epsilon	0
6	6	1	Zeta	0
7	7	1	Sigma	0
8	8	1	Eta	0
9	9	1	Tau	0
10	10	1	Theta	0
11	11	1	Upsilon	0
12	12	1	Iota	0
13	13	1	Kappa	0
14	14	1	Lambda	0
15	15	1	Omega	0
16	16	1	Digamma	0
17	17	1	Qoppa	0
18	18	1	Sampi	0
19	19	1	Greek_diacritics	0
20	20	1	Wikisource	0



Regulation Base 规则库

X and Y are couples -> Y and X are couples

X and Y are couples, and X is a male-> X is Y's husband

X is Y's husband -> Y is X's wife

	Condition	Result
1	X <and> Y [be] {couple}	Y <and> X [be] {couple}
2	(X <and> Y [be] {couple}) <and> (X [be] {male})	X [be] Y {husband}
3	X [be] Y {husband}	Y [be] X {wife}

Corpus 语料库

- <http://www.cncorpus.org/>
- <http://www.corpus4u.org/>
- <http://bcc.blcu.edu.cn/>
- <http://corpus.byu.edu/coca/>
- http://www.sogou.com/labs/resource/list_yuliao.php



Semantic Resource

- 1.中央研究院近代汉语标记语料库：http://www.sinica.edu.tw/Early_Mandarin/
- 2.中央研究院汉籍电子文献（瀚典全文检索系统）<http://www.sinica.edu.tw/ftms-bin/ftmsw3>
- 3.国家现代汉语语料库：<http://124.207.106.21:8080/>
- 4.国家语委现代汉语语料库：<http://www.clr.org.cn/retrieval/index.html>
- 5.树图数据库：<http://treebank.sinica.edu.tw/>
- 6.LIVAC共时语料库：<http://www.livac.org/s>
- 7.北京大学中国语言学研究中心，简称CCL语料库检索系统http://ccl.pku.edu.cn/Yuliao_Content.Aspx
- 8.北京大学《人民日报》标注语料库：<http://www.icl.pku.edu.cn>
- 9.北京语言大学的语料库：<http://www.blcu.edu.cn/kych/H.htm>
- 10.清华大学的汉语均衡语料库TH-ACorpus：<http://www.lits.tsinghua.edu.cn/ainlp/source.htm>
- 11.山西大学语料库<http://www.sxu.edu.cn/homepage/cslab/sxuc1.htm>
- 12.香港城市大学的LIVAC共时语料库：<http://www.rcl.cityu.edu.hk/livac/>或<http://www.LIVAC.org>
- 13.浙江师范大学的历史文献语料库：<http://lib.zjnu.net.cn/xueke/hyywzx/xkjj.htm>
- 14.中国科学院计算所的双语语料库：http://mtgroup.ict.ac.cn/corpus/query_process.php
- 15.中文语言资源联盟：<http://www.chineseldc.org/xyzy.htm>
- 16.红楼梦汉英平行语料库：<http://score.cripp.nie.edu.sg/hlm/index.htm#>
- 17.SKETCHENGINE多语言语料库：www.sketchengine.co.uk





using keyword to describe an article

Keyword Analysis

Keyword Analysis



Keyword Analysis

- Step 1: take keywords:
“冯小刚”，“华谊”，“万达”
- Step 2:
Get numbers of reports published every day

```
SELECT COUNT(*) AS NUMBER_COUNT, NEWS_TITLE, PUBLISH_DATE FROM  
FILM_NEWS WHERE NEWS_ID IN (SELECT NEWS_ID FROM FILM_NEWS WHERE  
NEWS_CONTENT LIKE '%万达%' AND NEWS_ID IN (SELECT NEWS_ID FROM  
FILM_NEWS WHERE NEWS_CONTENT LIKE '%华谊%' AND NEWS_ID IN (SELECT  
NEWS_ID FROM FILM_NEWS WHERE NEWS_CONTENT LIKE '%冯小刚%')) ORDER BY  
PUBLISH_DATE) GROUP BY PUBLISH_DATE
```

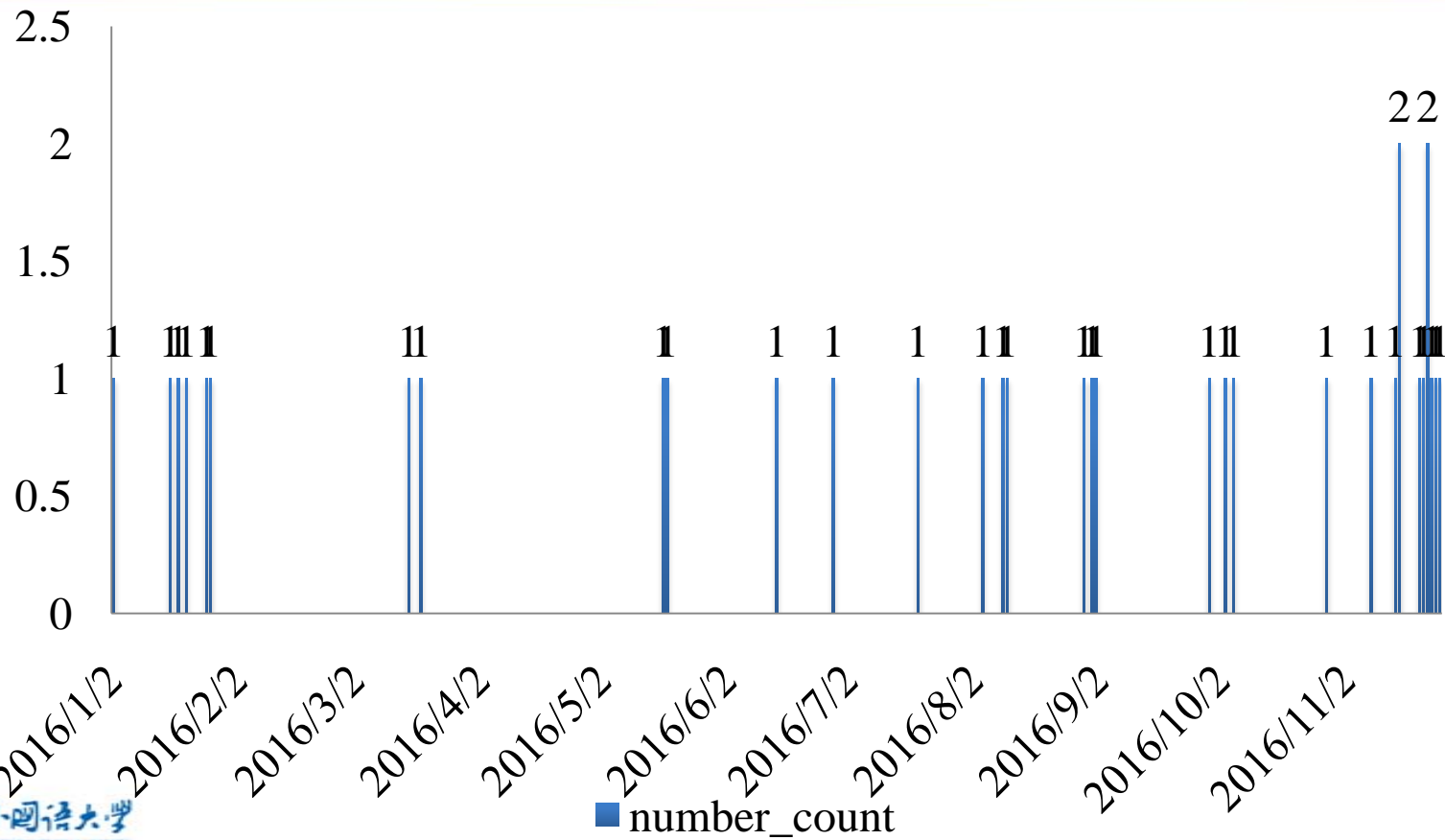


Keyword Analysis

- Data Description
 - Data Collection :
 - Web crawler gets data from Entgroup
 - The Size of the Dataset
 - About 22000 articles from 2009-2016
 - Totally, about 700 MB data



Keyword Analysis



Conclusions

- This event started from Nov. and is now still very hot in these days
- Although this event was taken placed in this Nov. but they have business much earlier.
- Summer and New Year are good seasons for film industry

Keyword Analysis



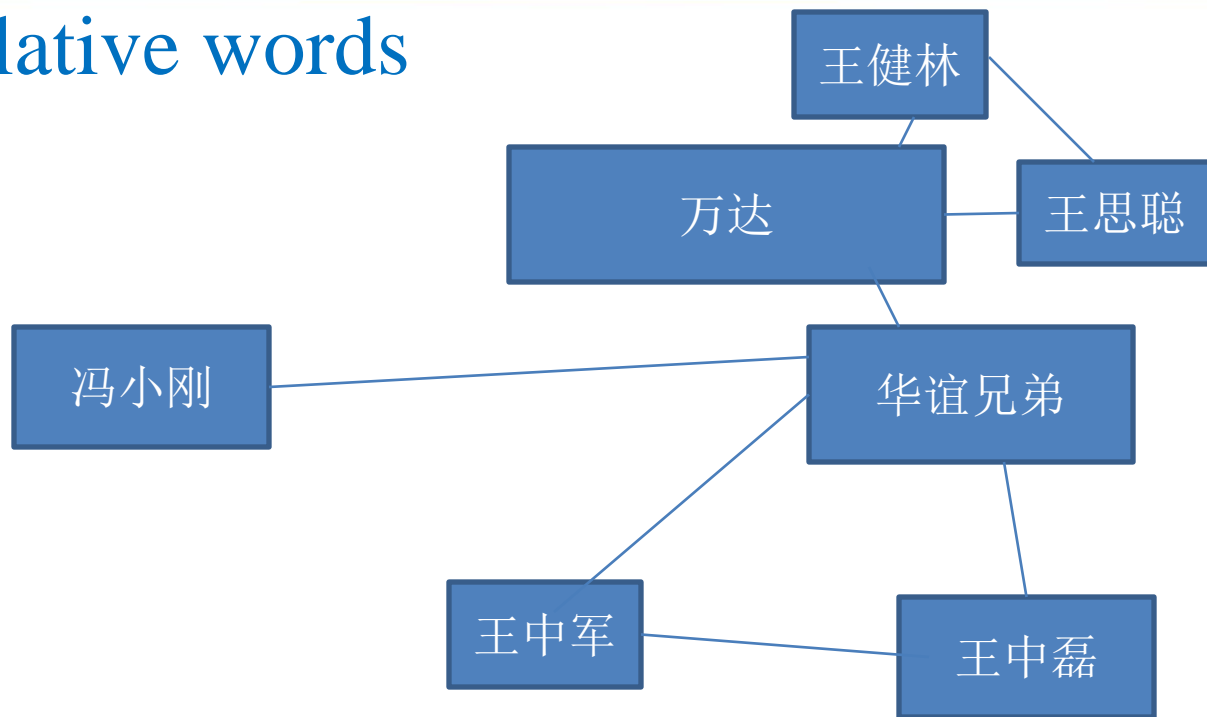
Ask A Question

Now, We have data, and results. Are these precise enough?



Keyword Analysis

- Correlative words



Natural Language Processing

- Chinese Word Segmentation 分词
 - Forward Max. matching method 正向最大匹配
 - Backward Max. matching method 逆向最大匹配
 - Statistical matching method 统计学方法



Natural Language Processing

Statistical Natural Language Processing

统计自然语言处理

- N-gram N元语法
- Hidden Markov Model(HMM) 隐马尔科夫模型
- Bayes' Theorem 贝叶斯定理



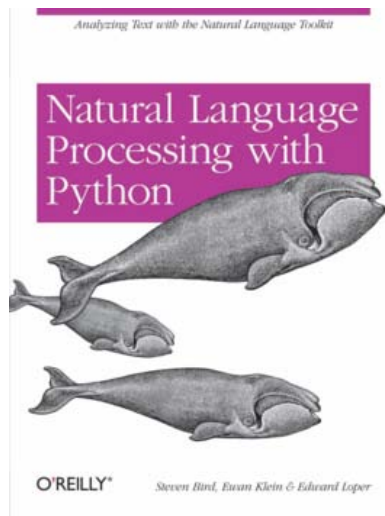


上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

Reference

NLP

https://en.wikipedia.org/wiki/Natural_language_processing





上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

Homework

Data Collection

- Finish your web crawler or API, and get data from the Internet
- Your data collection will be demonstrated with an example of data analysis application in our last lecture.

Data Analysis

- Choose a topic in current datasets
- Calculate it and demonstrate it

A Final Report of this Course

- What is your project?
- The designing of your program and database
- Your Data Collection approaches
- Your Data Analysis approaches
- A Simulation of your group homework
- Your Conclusion





上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY



The End of Lecture 8

Thank You

<http://www.wangting.ac.cn>

