基于在线新闻的影视投资数据分析

Data Analysis for Film Investment using Online News

王挺
Ting Wang

上海外国语大学
Shanghai International Studies University

个人简介

王挺

利物浦大学计算机博士，清华大学智能技术与系统国家重点实验室博士后，计算机高级工程师

现任：

- 上海外国语大学中国国际舆情研究中心研究员，专任教师
- 上海拟合智能科技有限公司, CIO
- 中国计算机学会青年计算机论坛无锡分论坛学术委员
- 科技部火炬中心中国创业导师
- IEEE 会员

曾任：

- 清华大学无锡应用技术研究院大数据中心主任助理
- 美国好莱坞Raleigh Studios亚洲区副总
- 无锡国家数字电影产业园影视服务中心副主任
- 江苏华莱坞投资发展有限公司电影发展部副部长
- 解放军总参谋部第56研究所研发工程师

研究兴趣：

- 人工智能
- 机器学习
- 大数据
- 机器视觉
- 自然语言处理

今日主题

今日关键词：

网络数据分析，在线新闻，影视，投资数据分析

关注的问题：

是否可以通过分析影视新闻数据来预测票房，为投资做决策分析？

需要探讨的问题：

1. 为什么要做数据分析？
2. 什么是网络数据分析？
3. 网络数据分析的主要技术路线是什么？
4. 如何将网络新闻数据分析应用在影视投资决策领域？

Why data analysis?

为什么要做数据分析？

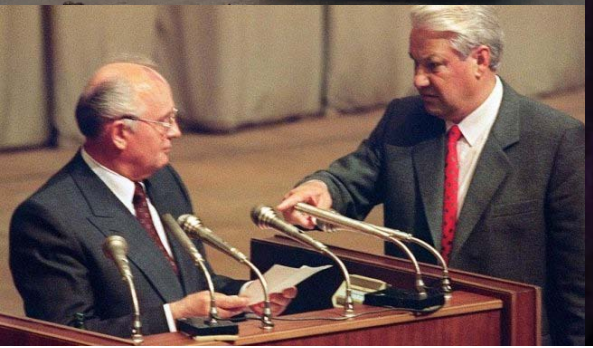
为什么要做数据分析？

为什么要做数据分析？

布雷顿森林体系与
美元国际货币体系

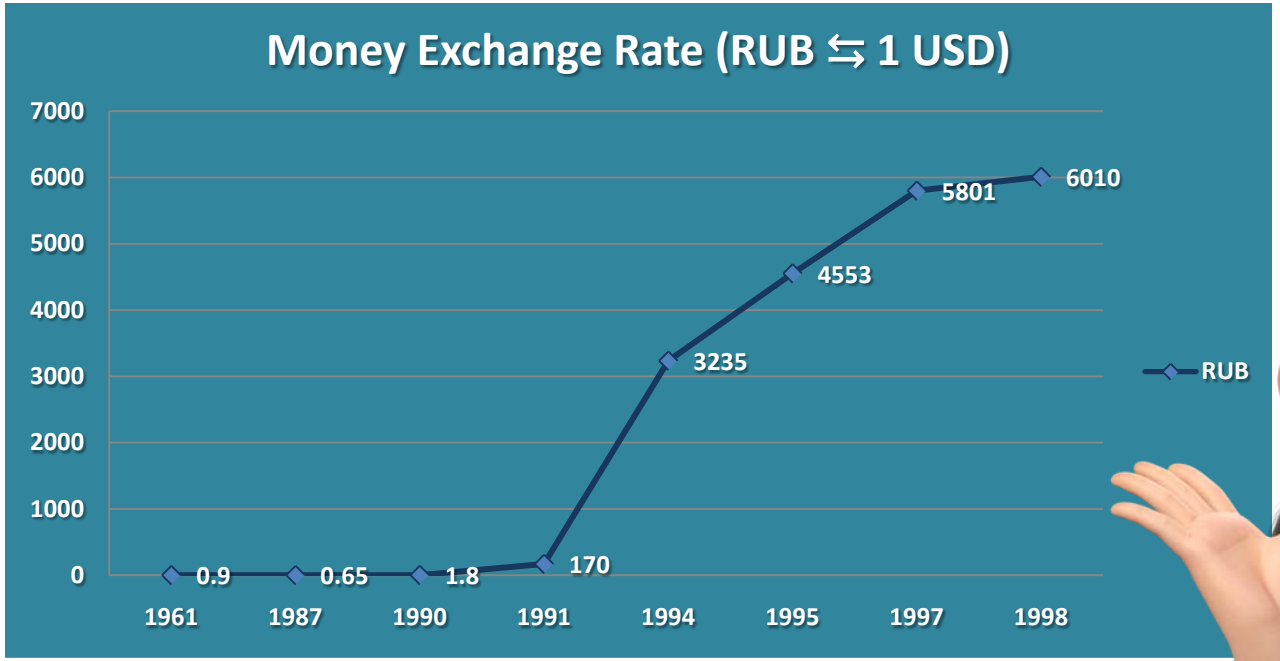
为什么要做数据分析？

政治事件影响经济走势

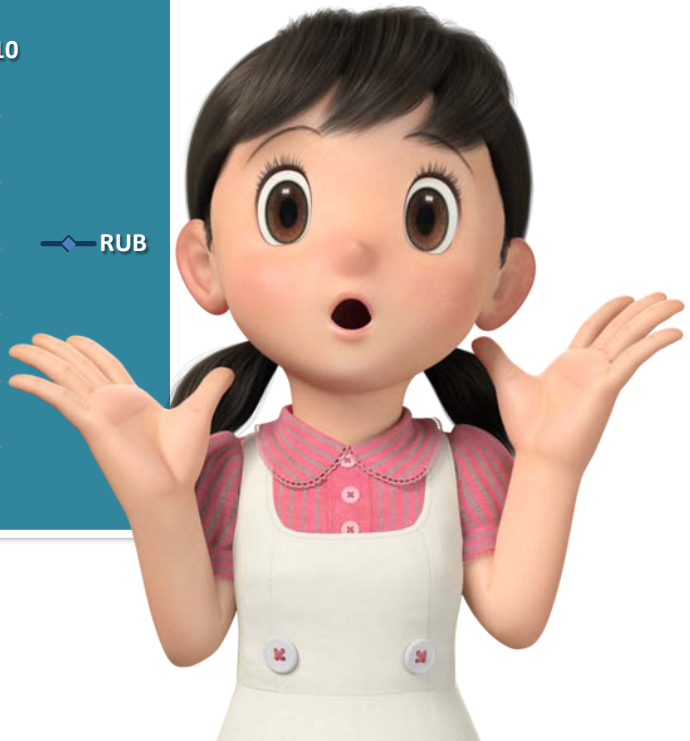


苏东剧变，1990s

为什么要做数据分析？



美元 VS 卢布



为什么要做数据分析？

http://cn.investing.com/

卢布贬值引发代购潮，俄罗斯华人大淘便宜货

由于近期卢布的连续贬值，用

销售范围主要围绕化妆、日用

网易新闻 网易首页 > 新闻中心 > 滚动新

卢布大贬值 快去俄罗斯海

2014-12-28 02:35:41 来源：华商报(西安)

又到岁末，人民币依然维持全年坚挺走势
是著名的金砖国家俄罗斯的卢布跌幅更是大。
少呢。

血拼乐淘：在莫斯科的高档商场Atrium
名牌商品，那里的迪奥、香奈儿、兰蔻等大牌
专柜的雅诗兰黛面霜小棕瓶50ml为例，在莫斯科卢布。
价格为950元；兰蔻一款在我国内地售价1460元的小
约880元，在莫斯科专柜价格不到600元。很多美妆和奢侈品的价格已经便宜过香港，部分品牌
至是全欧洲最便宜的。

腾讯教育 高考 考研 公务员 校园 留学 外语 中小学

卢布贬值震惊全球 中国留学生做起奢侈品

留学 浙江在线-钱江晚报[微博] 2014-12-18 09:35 我要分享

浙江在线12月18日讯：（钱江晚报记者 钟卉）在2014年的最后
场空前巨大的危机中。卢布在本周一暴跌超10%，创下1998年来最大
斯央行一口气加息650个基点，从10.5%提高至17%，仍拦不住卢布的
贬值25%。

受国际油价持续走低、西方对俄施加经济制裁、俄国际储备处
合作用，今年以来卢布贬值幅度已经超过50%。危机当前，一惯强
布“寒潮”中的俄罗斯民众已无法冷静。俄罗斯街头出现这样的场
数字不停地变化，银行提款机经常被提取一空，俄罗斯人冲进商场

MTS RUS 12:24 100%

个人信息 我的二维码

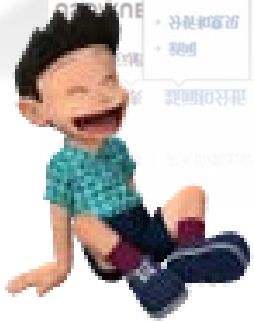
小柒海外正品代购

俄罗斯

扫一扫上面的二维码图案，加我微信

为什么要做数据分析？

新闻对投资的重要性



为什么要做数据分析？

为什么要做数据分析？


网易新闻 网易首页 应用 网易考拉

英国脱欧民意暂领先3%(组图)

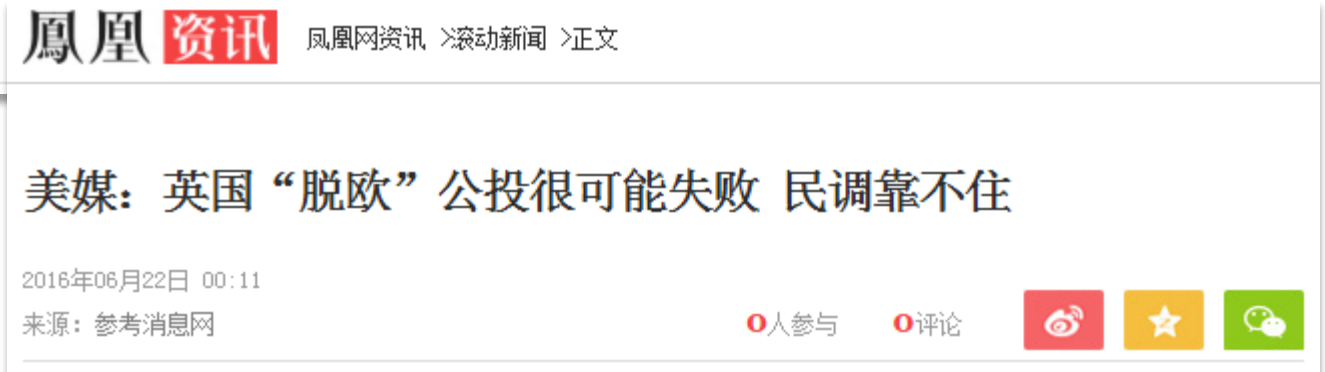
2016-06-07 04:37:00 来源:广州日报(广州)


凤凰资讯 凤凰网资讯 >滚动新闻 >正文

距离英国公投还剩4天 “留欧”民意支持率反超“脱欧”

2016年06月20日 04:41

来源：新快报


凤凰资讯 凤凰网资讯 >滚动新闻 >正文

美媒：英国“脱欧”公投很可能失败 民调靠不住

2016年06月22日 00:11

来源：参考消息网

0人参与 0评论

为什么要做数据分析？

FUSION 2016
Heidelberg July 5-8

为什么要做数据分析？

基于新闻的欧元人民币汇率的实时数据分析
http://finance.qq.com/zt2016/gongtou/index.htm

实时投票中间结果对汇率的影响

Time	News	Rate
06-24 06:17	Gibraltar (IN)	1:7.52
06-24 06:31	New Castle (IN)	1:7.51
06-24 07:28	Sunderland (OUT)	1:7.47
06-24 09:23	Oxford (IN), North Ireland (IN)	1:7.43
06-24 10:01	49.79% (IN), 50.21% (OUT)	1:7.30
06-24 11:07	48.95% (IN), 51.05% (OUT)	1:7.25
06-24 11:40	Wales (OUT)	1:7.22
06-24 12:13	48.28% (IN), 51.72% (OUT), 339/382 Regions	1:7.24

为什么要做数据分析？

我的最终货币兑换结果：

(7.52-7.24)*2200=616 RMB

一顿德国猪肘
就这样被省出来了~
么么哒！(づ￣ 3￣)づ

为什么要做数据分析？

数据分析的重要性：

1. 获取新知识新信息
2. 扩大收益
3. 规避风险

数据分析的主要目的：

1. 发现新的知识和信息
2. 得出新结论
3. 决策支撑

What is online data analysis?

什么是网络数据分析？

什么是网络数据分析？

定义：

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

Analysis of data *is a process of inspecting, cleaning, transforming, and modeling data with the goal of* ***discovering useful information***, ***suggesting conclusions***, *and* ***supporting decision-making***.

网络数据分析：基于网络数据的数据分析。

什么是网络数据分析？

两种不同的类型:

- 定量分析
  - 预测
  - 回归
- 定性分析
  - 分类
  - 聚类

什么是网络数据分析？

信源：网络数据分析的数据来源

1. 物联传感网数据分析
2. 互联网数据分析
3. 移动网络数据分析

信源的重要性：优质的信源是成功的一半

1. 覆盖面广
2. 噪声数据少

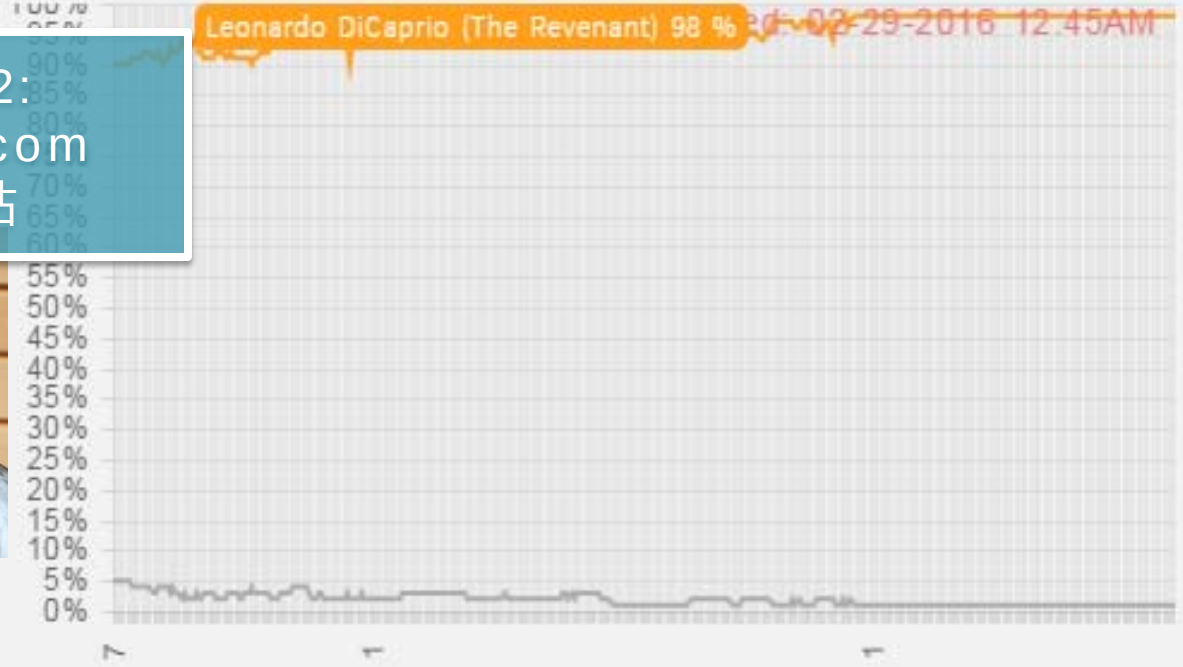
什么是网络数据分析？

Leading Actor

Leonardo DiCaprio (The Revenant)

Eddie Redmayne (The Danish Girl)

EXAMPLE 2:
predictwise.com
一个预测网站



什么是网络数据分析？

David Rothschild

PhD of Wharton School of Business at the University of Pennsylvania

Microsoft researcher at Microsoft Research in New York City

成功案例：

2012年成功预测全部51张总统选举人票中的50个。

2013年成功预测全部24个奥斯卡奖里的20个。

2014年成功预测全部15场世界杯足球赛结果。

涉及领域：

- 政治
- 体育
- 娱乐
- 经济金融

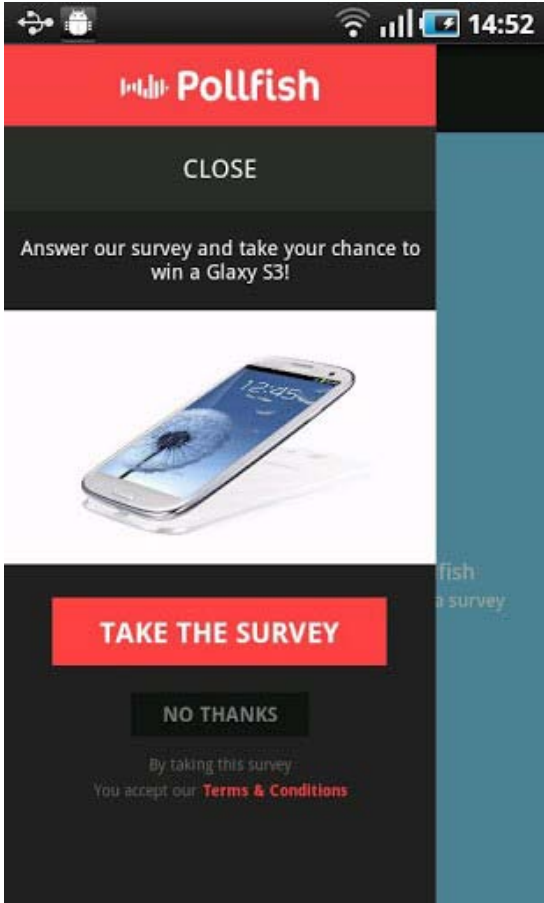
什么是网络数据分析？

技术方法

- 数据采集:
  Pollfish, MSN, Xbox

- 数据分析:
  基于历史数据的统计分析

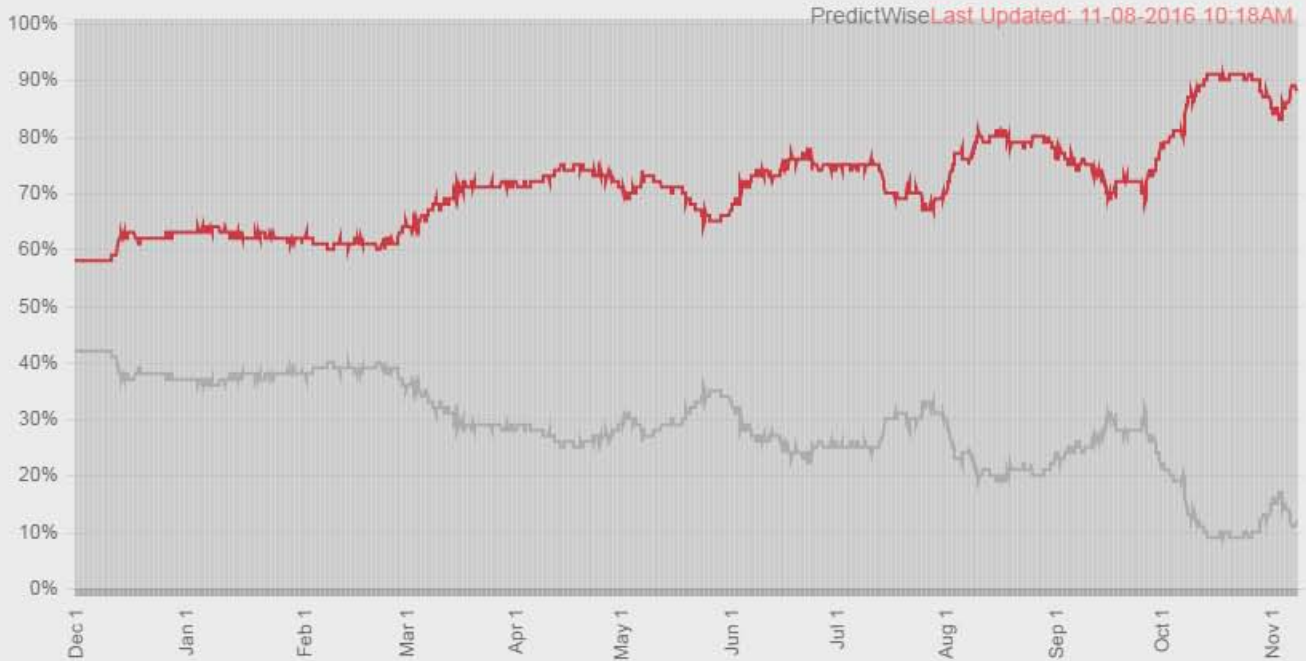
http://predictwise.com/



什么是网络数据分析？

2016 PRESIDENT - GENERAL ELECTION

Democratic	88 %
Republican	12 %



什么是网络数据分析？

Politics

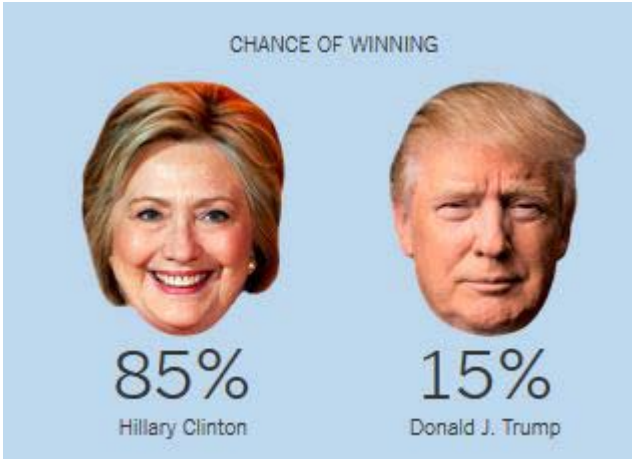
2016 PRESIDENT - GENERAL ELECTION

Republican	98 %
Democratic	2 %

什么是网络数据分析？

1. 覆盖面：Examine data quality - in this election polls were not reaching all likely voters

2. 偏见：Beware of your own biases: many pollsters were likely Clinton supporters and did not want to question the results that favored their candidate. For example, Huffington Post had forecast 98% chance of Clinton Victory.

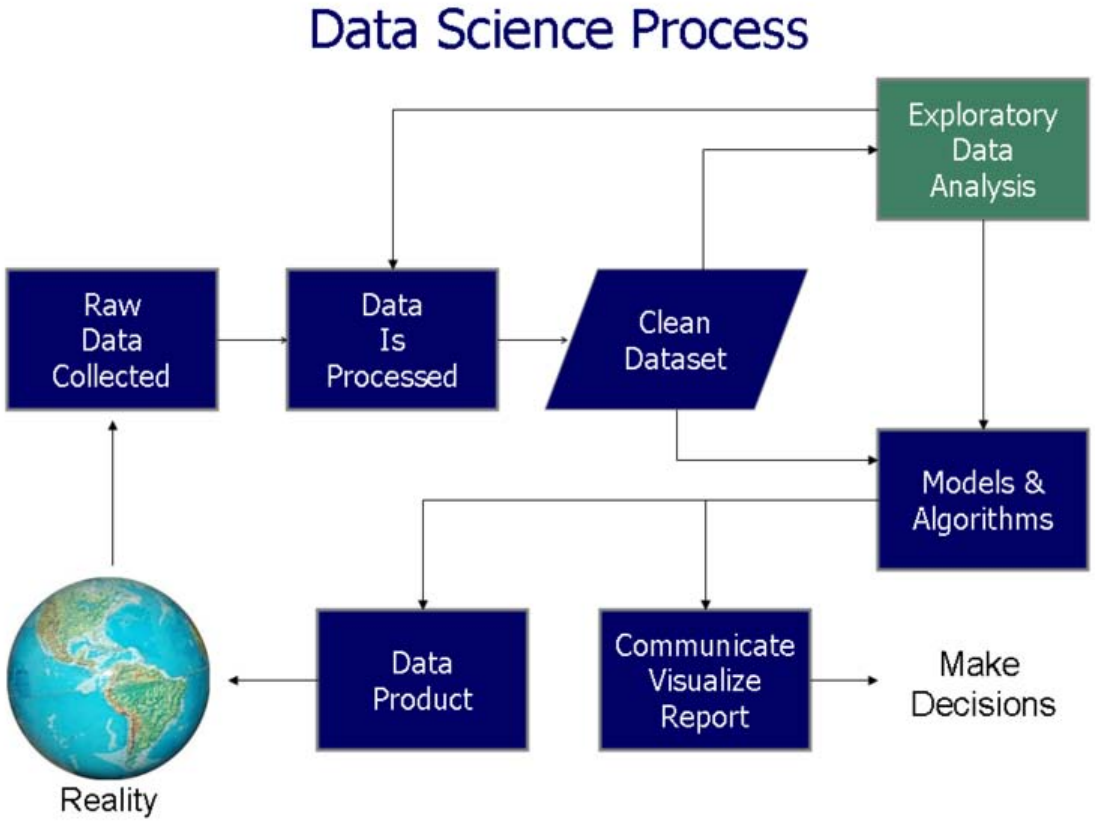


http://www.kdnuggets.com/2016/11/trump-shows-limits-prediction.html

How to analysis with online data?

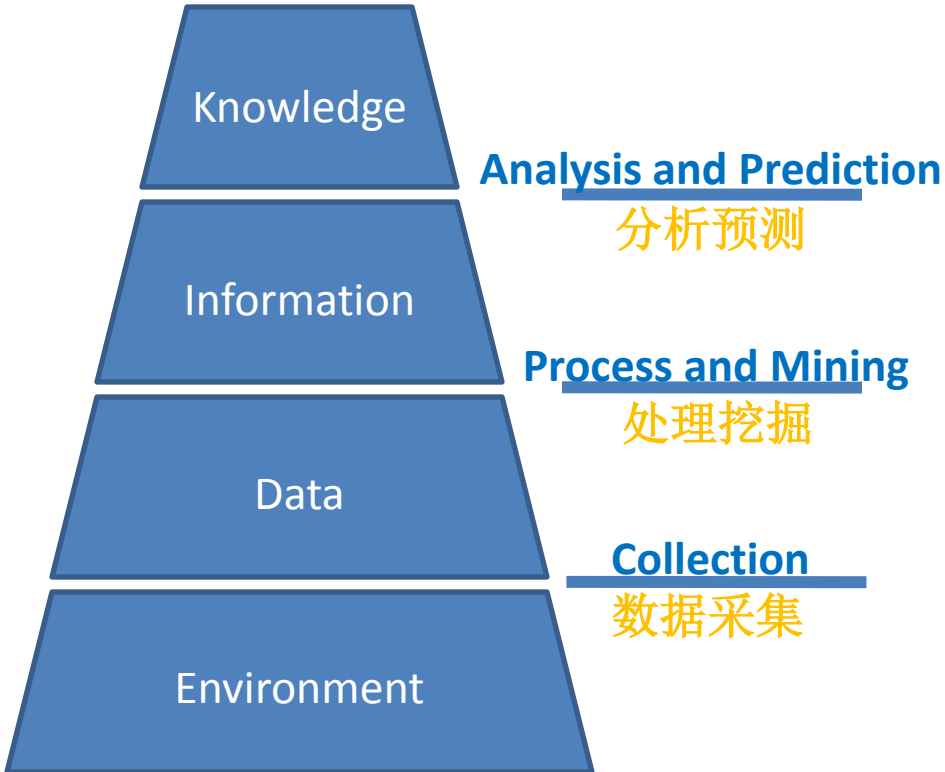
如何进行网络数据分析？

如何进行网络数据分析？


Data Science Process
Exploratory Data Analysis
Raw Data Collected
Data Is Processed
Clean Dataset
Models & Algorithms
Reality
Data Product
Communicate Visualize Report
Make Decisions

如何进行网络数据分析？

数据 、信息和知识之间的关系



如何进行网络数据分析？

技术路线

1	2	3	4
信息获取	数据清洗与信息检索	知识融合与信息更新	预测及分类等
网络数据	网络信息结构化处理	数据更新与特征分析	个性化产品与服务

抓　处理　分析　卖

How to analysis film news data?

如何进行影视新闻数据分析？

问题描述

自2008年以来，中国电影产业发展十分迅速，票房收入逐年提高。但是也存在不少问题。每年70%的影视节目亏损，20%保本，仅有10%能盈利。市场缺乏科学论证体系，盲目乐观，导致投资亏损严重。

电影产业的发展需要数据分析和决策支持。电影产业的发展归根到底是如何获得更高的票房收益，电影产业的决策支持归根到底是如何准确预测未来票房。

预测票房需要建立相应的数学模型，并且有相应的考察维度和数据。

问题假设

假设1：社会舆论对于影片最终票房有影响。

假设2：同一篇文章中所提到的影片应该具有相关性。

假设3：历史相关数据对预测未来有帮助。

数据准备

新闻, 24100 篇文章

http://www.entgroup.cn/

2007.11-2016.11

电影票房数据, 1893 部电影

http://58921.com/

2008.1-2016.11

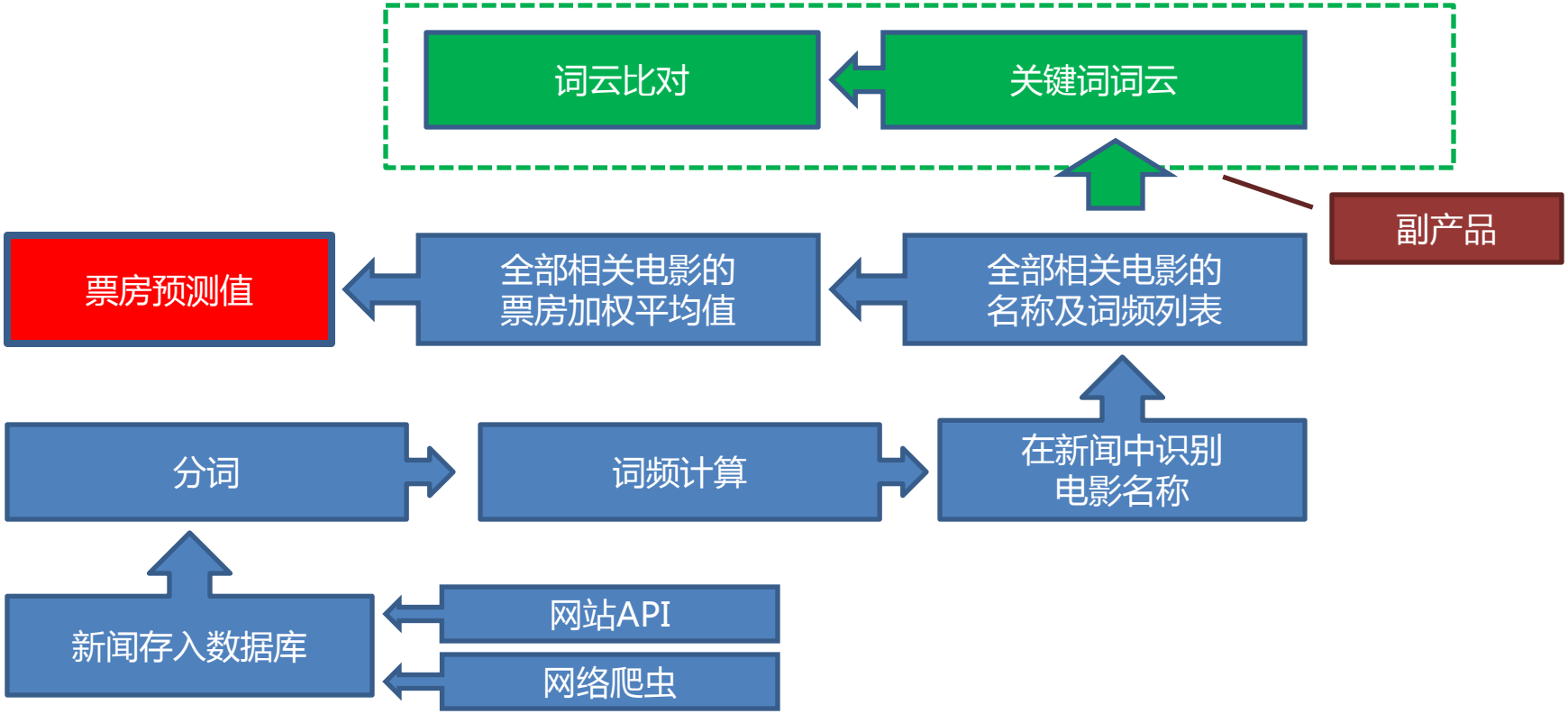
词典

149921 中文单词

508 停用词

技术路线

数据处理流程


词云比对
关键词词云
副产品
票房预测值
全部相关电影的
票房加权平均值
全部相关电影的
名称及词频列表
分词
词频计算
在新闻中识别
电影名称
新闻存入数据库
网站API
网络爬虫

票房预测

《长城》
2016-2017贺岁档

实际票房：
11.73亿人民币

预测结果：
11.64亿人民币

参数设置与数据来源：

Please input the Film Name:
长城

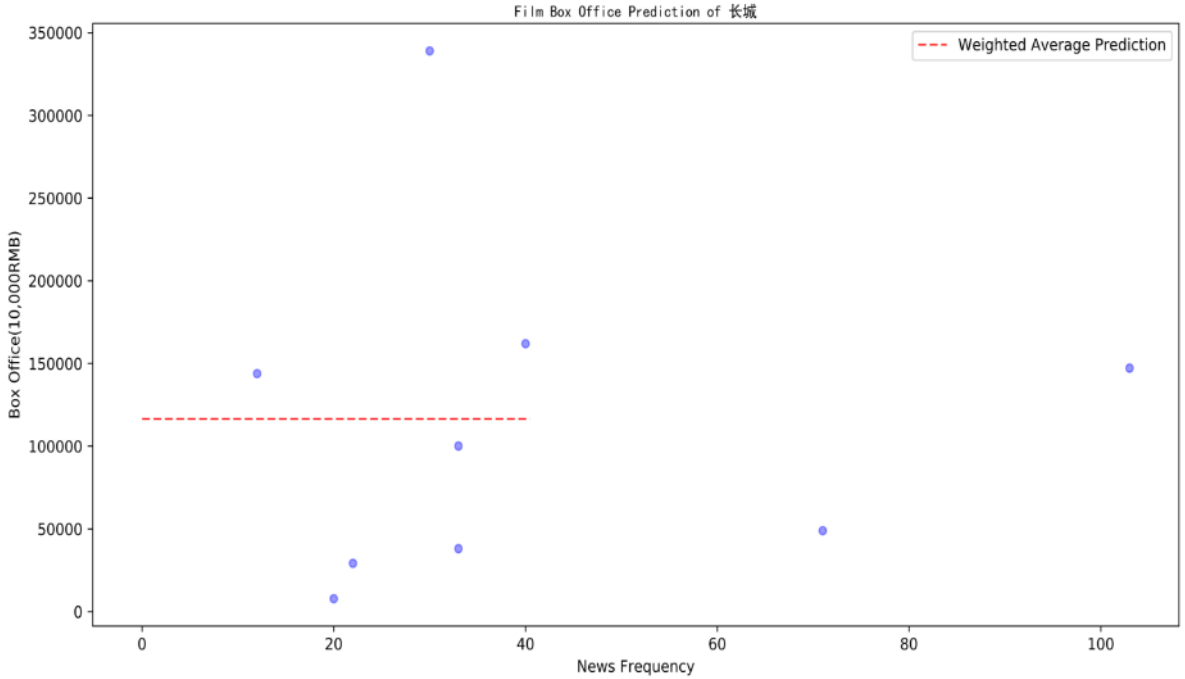
Please input the Frequency of Keyword:
10

Start Date - End Date of News
2015-1-1 2016-12-1

Predict

Home

Film Box Office of 长城 : 116401.04395604396(x10,000) RMB



节目和影视作品的相似度评价

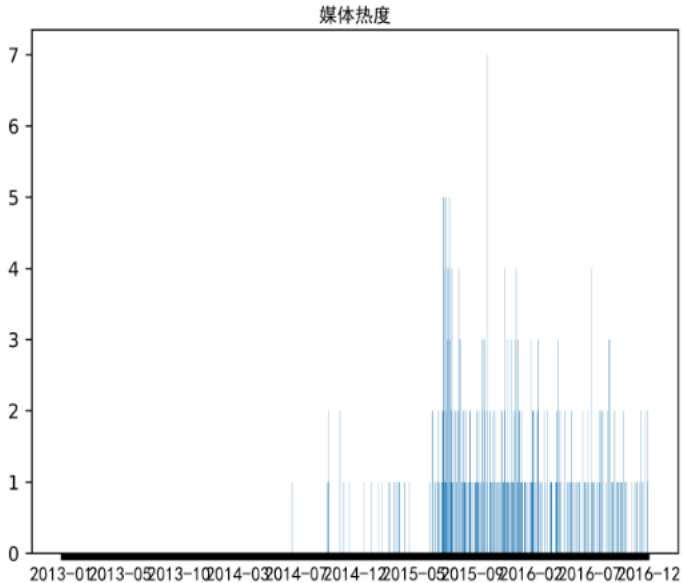
《芈月传》和《甄嬛传》的相似度计算结果对比

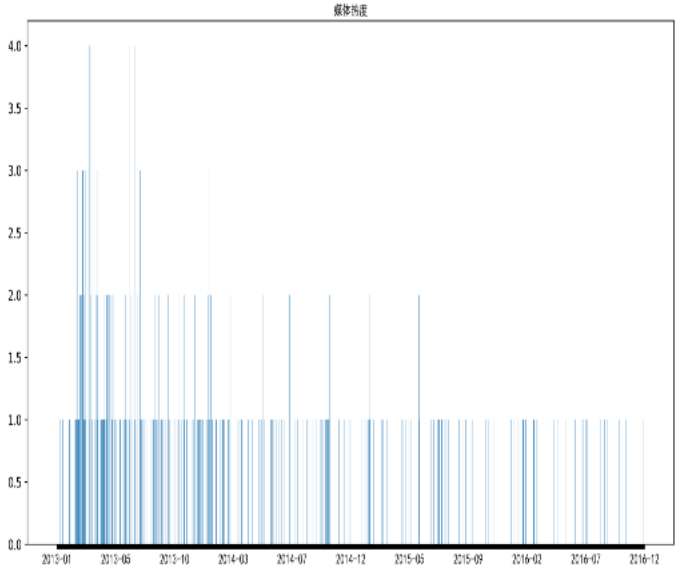
Similarity: 59.74025974025974%

Key Word 1:芈月传	Key Word 2:甄嬛传
	

节目和影视作品的相似度评价

《芈月传》和《甄嬛传》的媒体热度





影视基地的相似度评价

无锡国家数字电影产业园 “华莱坞”和青岛万达东方影都的相似度对比

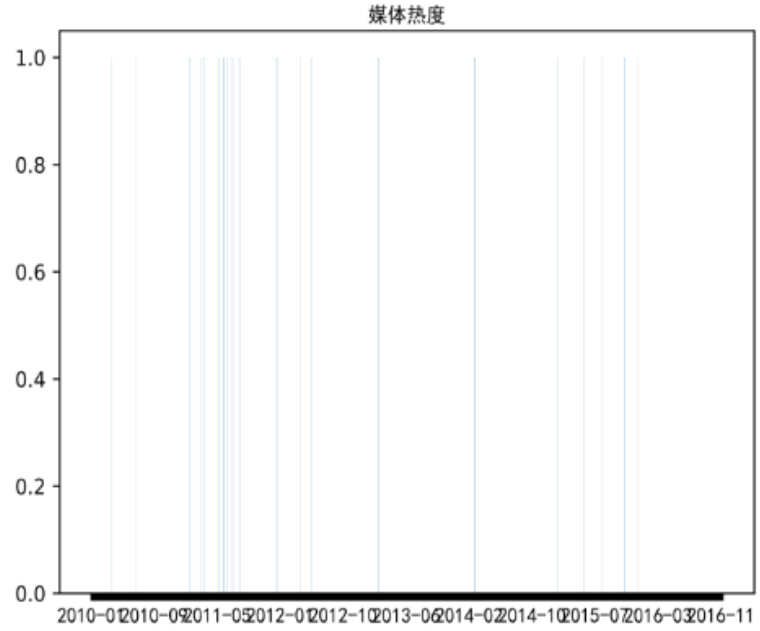
Similarity: 27.777777777777778%

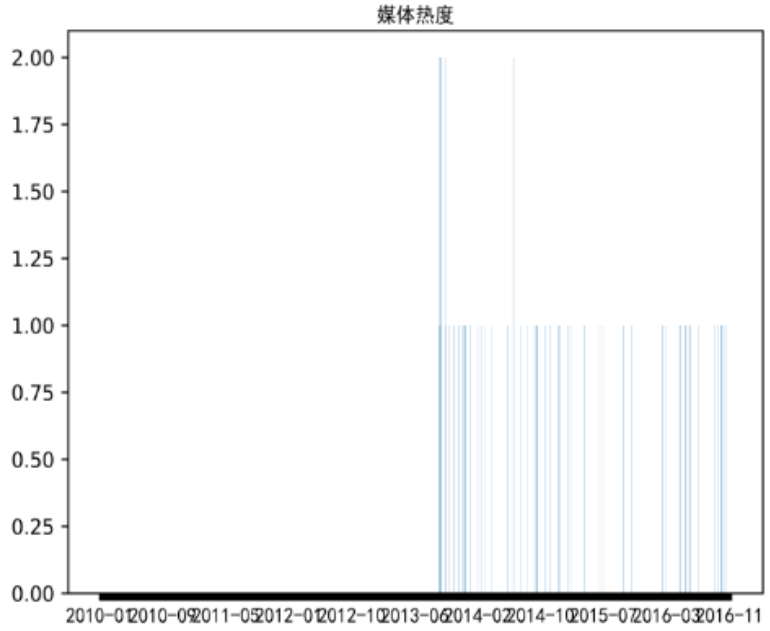
Key Word 1:华莱坞

Key Word 2:东方影都

影视基地的相似度评价

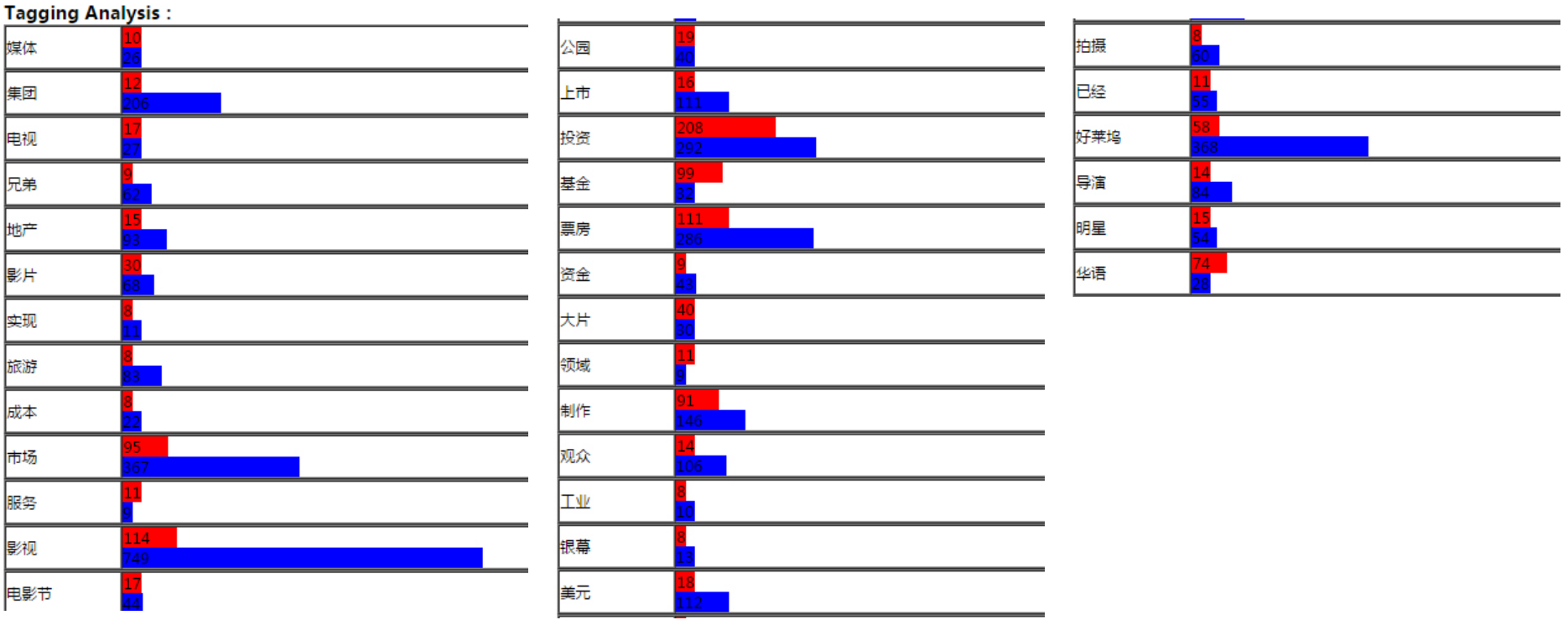
无锡国家数字电影产业园“华莱坞”和青岛万达东方影都的媒体热度对比





影视基地的相似度评价

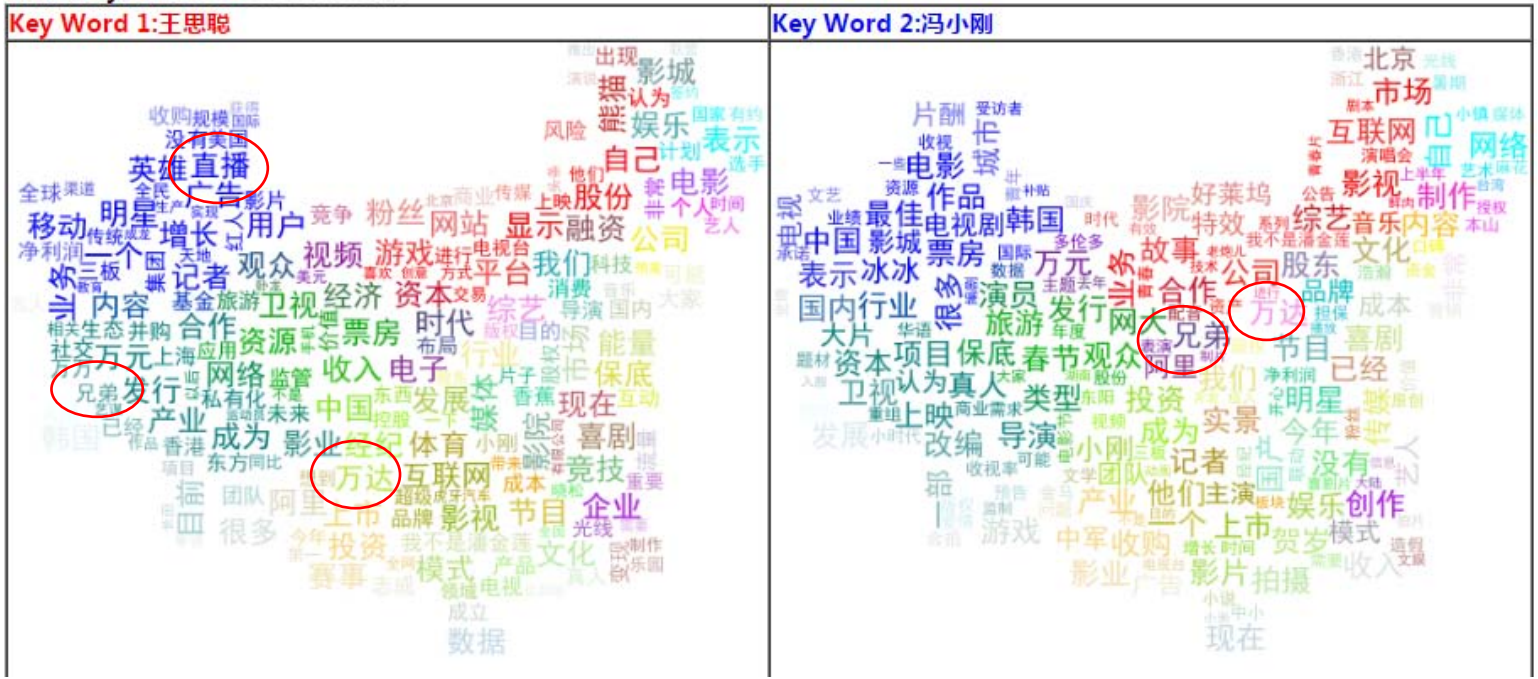
无锡国家数字电影产业园“华莱坞”和青岛万达东方影都的共有关键词标签对比（节选）



影视人物评价

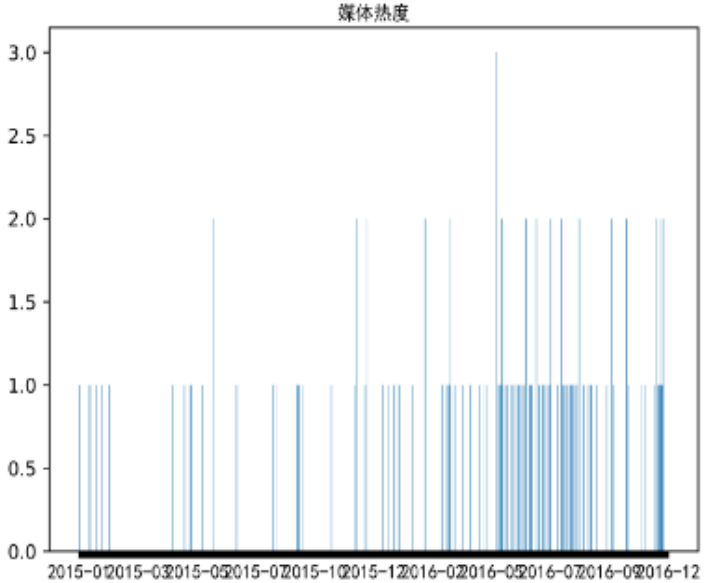
“王思聪”和“冯小刚”的相似度对比

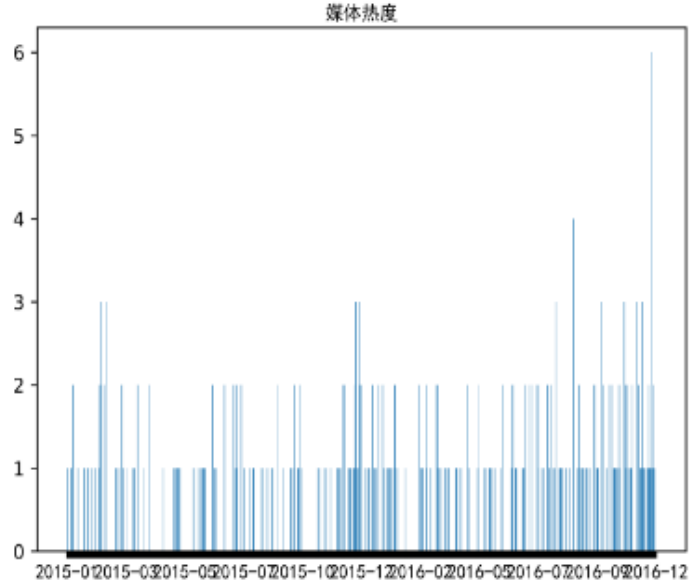
Similarity: 41.66666666666667%



影视人物评价

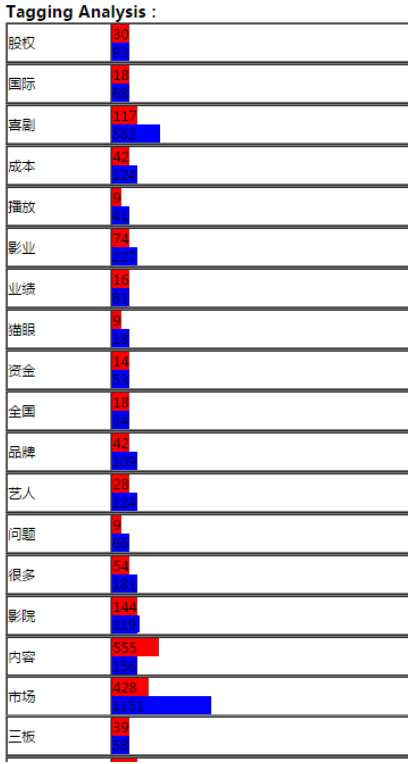
“王思聪”和“冯小刚”媒体热度对比

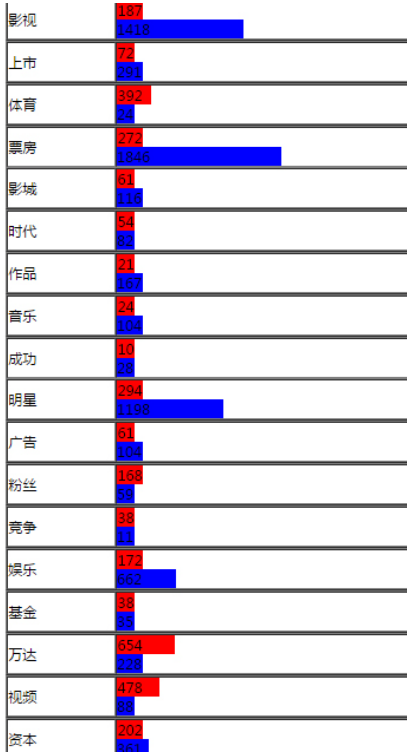


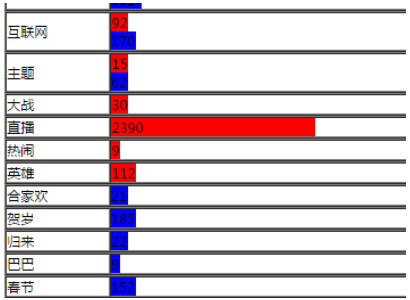


影视人物评价

“王思聪”和“冯小刚”关键词对比（节选）



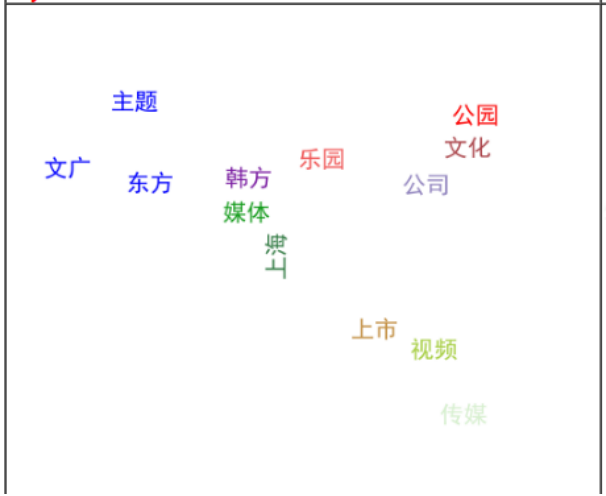




教育科研机构相似度评价

“上海交通大学”和“北京电影学院”的相似度对比

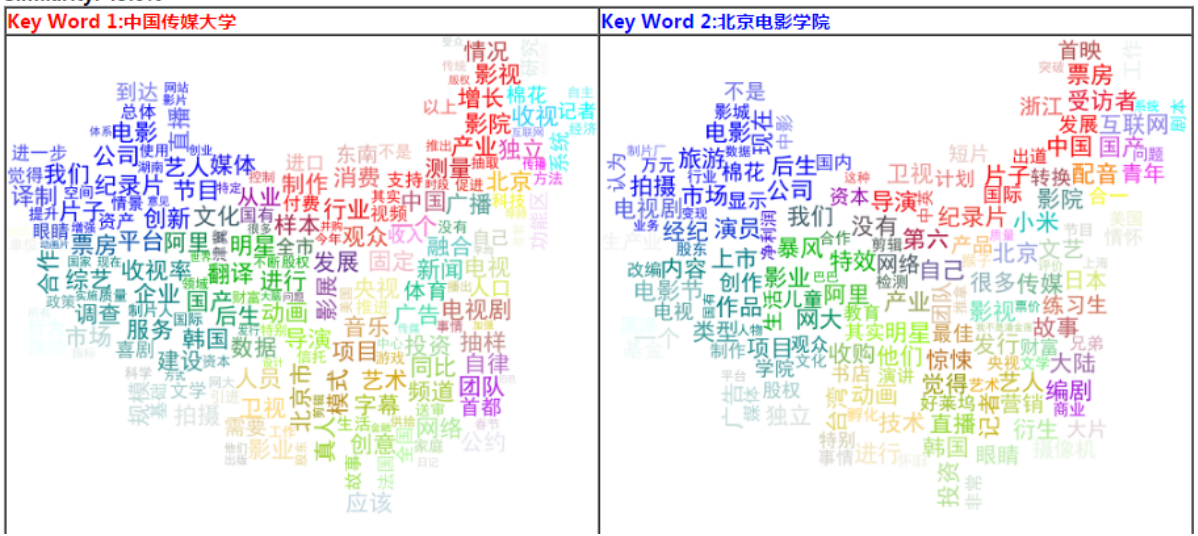
Similarity: 7.142857142857142%

Key Word 1:上海交通大学	Key Word 2:北京电影学院
	

教育科研机构相似度评价

“中国传媒大学”和“北京电影学院”的相似度对比

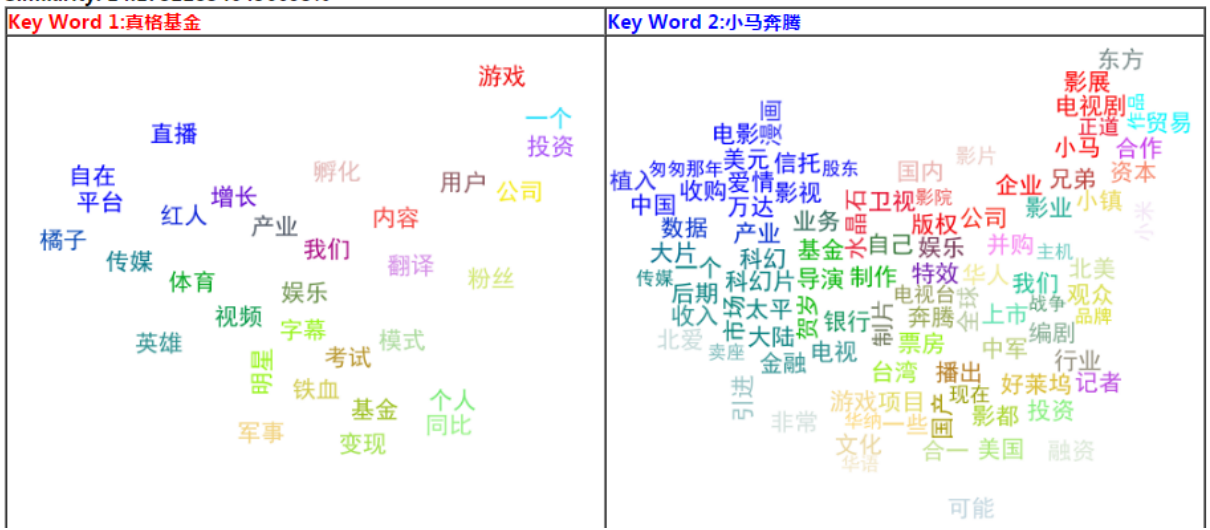
Similarity: 45.0%



投资机构评价比对

中国知名投资公司“真格基金”和“小马奔腾”的相似度对比

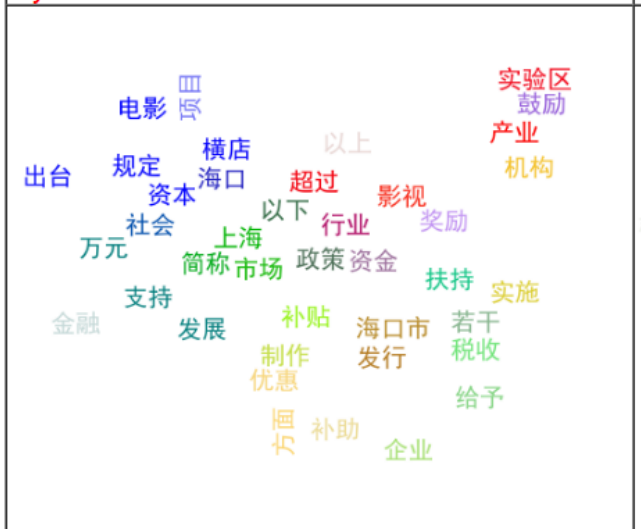
Similarity: 14.173228346456693%



政策评价对比

《关于进一步加快横店影视产业实验区发展的若干意见》和《关于促进上海电影发展的若干政策》的相似度对比

Similarity: 53.52112676056338%

Key Word 1:关于进一步加快横店影视产业实验区发展的若干意见	Key Word 2:关于促进上海电影发展的若干政策
	

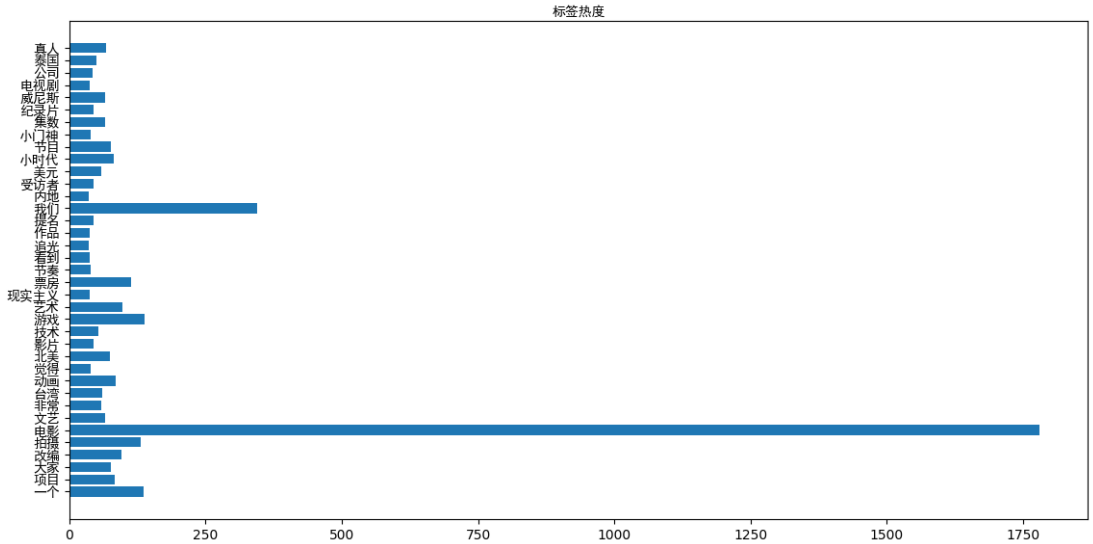
广电影视行业知识表现

电影术语“长镜头”的相关知识词汇词频

127.0.0.1:5000/KeywordTagging

Home

Key Word:长镜头 Key Word Frequency:35 Start Date - End Date:2010-1-1 - 2016-12-1



小结

结论

通过影视新闻和历史相关数据，不仅可以有效预测票房，还可以对比考察影视作品、基地、人物、机构、政策、专业知识等相关信息，未相关决策做好支撑。

未来工作

- 加入影视相关命名实体，使计算更准确
- 除词频外，再加入其它相关维度，使票房预测更精准
- 融入社交媒体的数据，丰富预测的情感计算结果

结 束

谢谢

http://www.wangting.ac.cn