



New Media Data Analytics and Application

Lecture 13: System Development Case Study

Ting Wang

Outlines

- Systems Thinking for Product Designing
- A Case Study: Film Box Office Prediction
- To Be A Good Data Analyst





上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

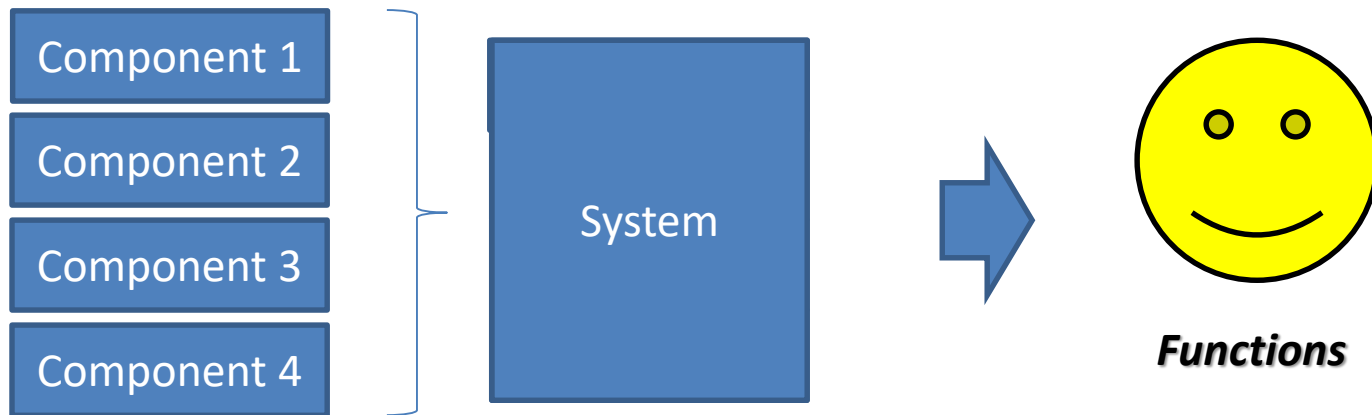
circulating development for your goals

Systems Thinking for Product Designing

Systems Thinking for Product Designing

What is a System?

In computer science and information science, system is a software system which has components as its structure and observable inter-process communications as its behavior.

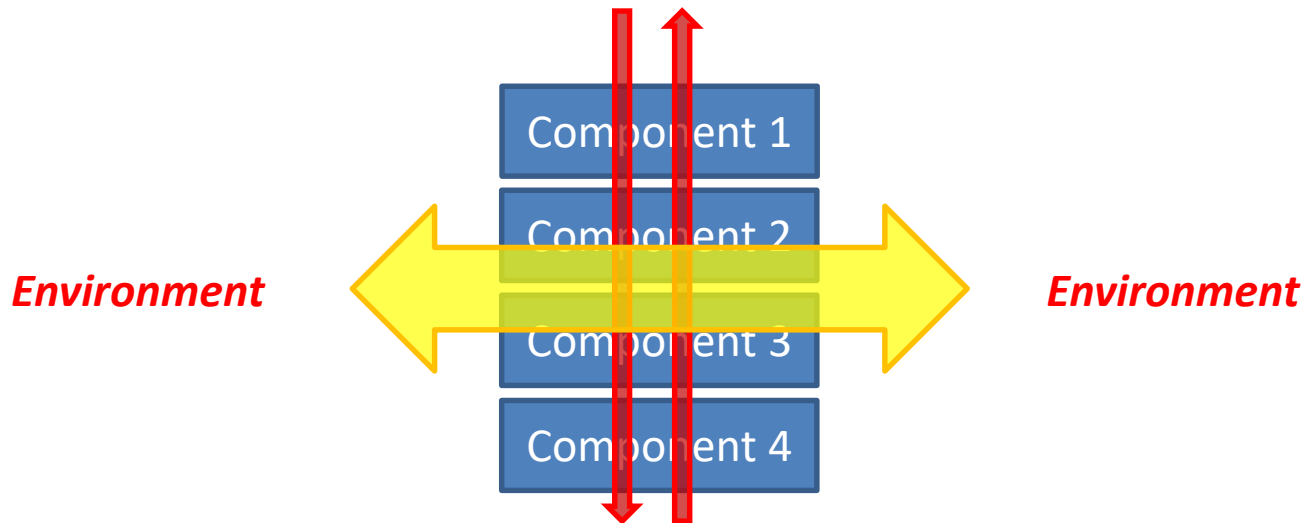


Systems Thinking for Product Designing

What is Systems Thinking?

Global, Optimal, and Integrated thinking methodology for software development and operation.

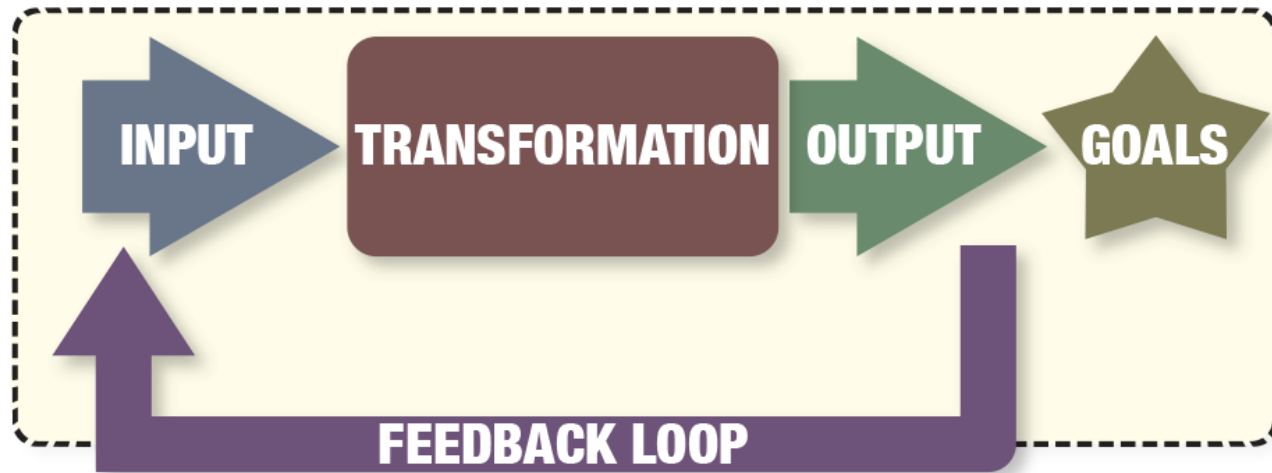
- Interactions between system and its components
- Interactions between system and its environment



Systems Thinking for Product Designing

Two recommended Systems Thinking Approaches

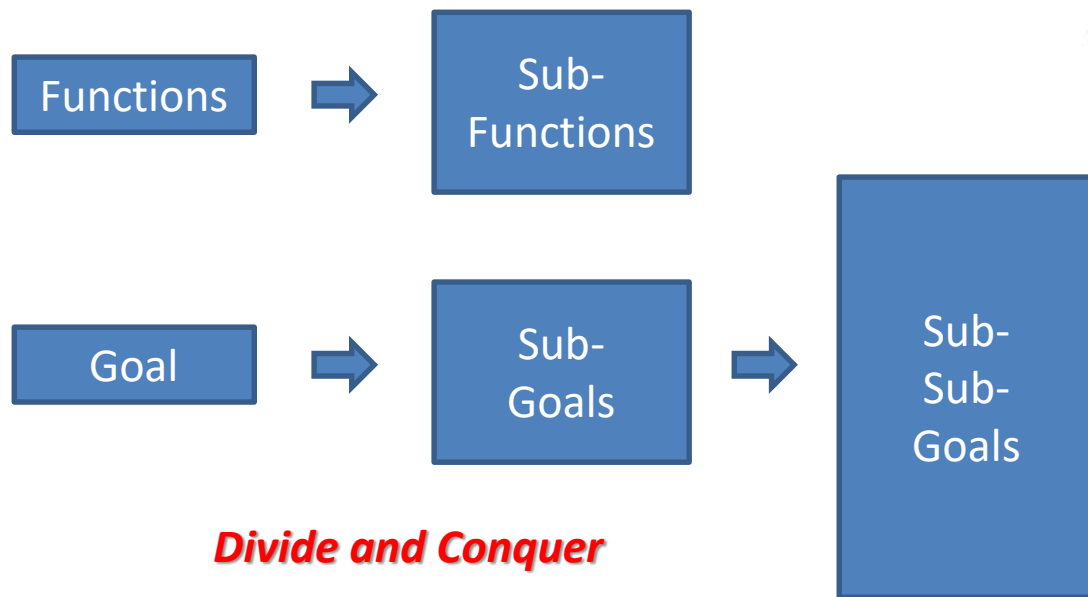
- Goal Seeking
- Input and output



Systems Thinking for Product Designing

Goal Seeking (Global optimization) 全局最优

a global optimization of a function or a set of functions according to some criteria



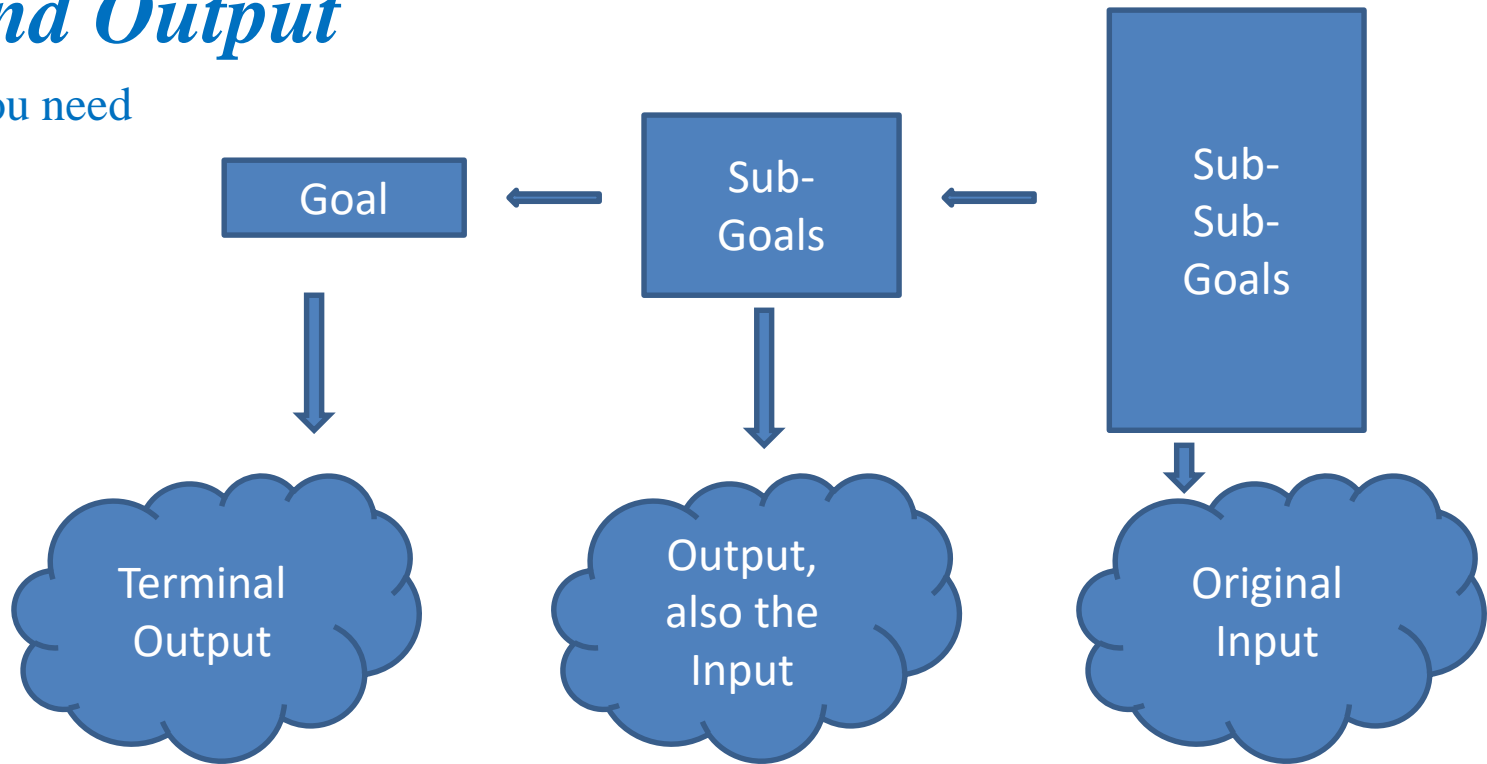
先定一个能达到的小目标



Systems Thinking for Product Designing

Input and Output

all for what you need



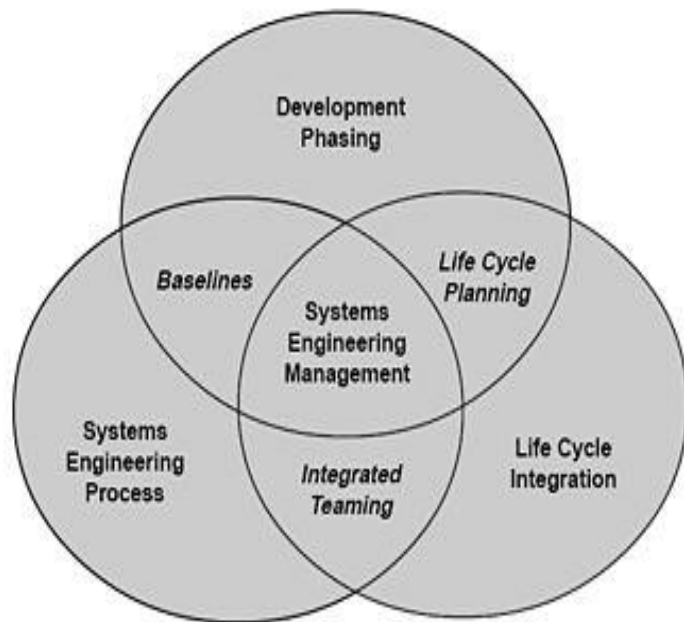
Systems Thinking for Product Designing

System Engineering 系统工程

ensures all likely aspects of system are considered, and integrated into a whole product.

Software Engineering

(in software and information industry)





上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

a case study

Film Box Office Prediction

EXAMPLE 1: Films

Film Box Office Prediction

Case Description

Film Box Office Prediction

- is crucial to film investment
- is significant to the market without Completion Bond
- can be done by a number of approaches

In this case, film box office prediction will be computed based on the information collected by online film news reports.



Software Analysis



Film Box Office Prediction

Terminal Goal

To make a decision:

whether a film is worth of being invested or not.



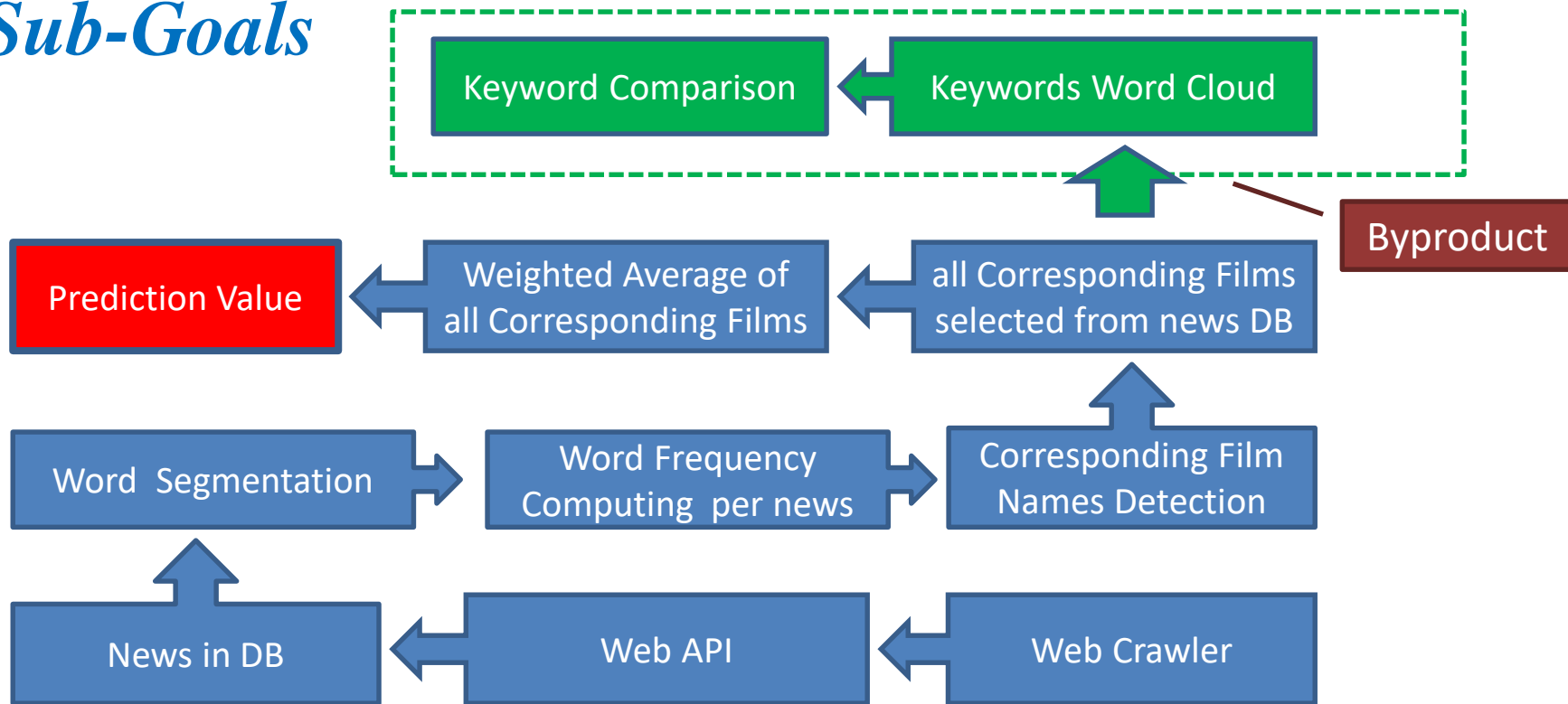
Final Output

This depends on the prediction value of the box office of the potential film project.



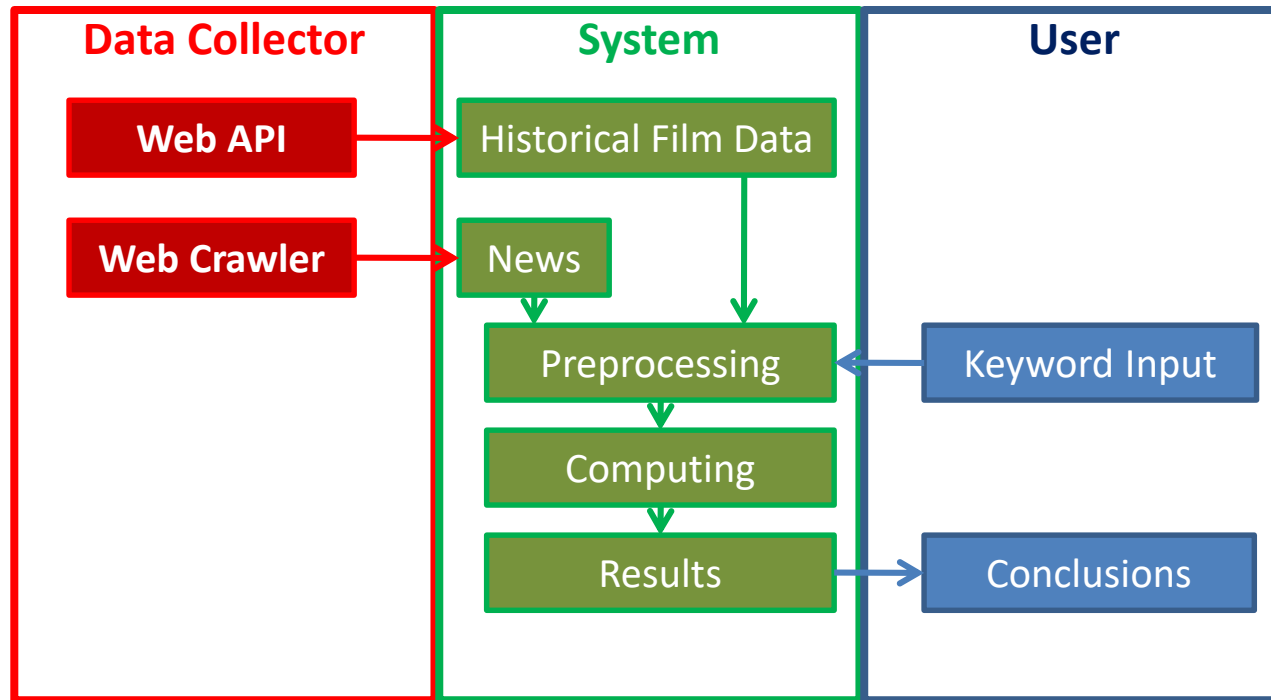
Film Box Office Prediction

Sub-Goals



Film Box Office Prediction

Activity Diagram



Film Box Office Prediction

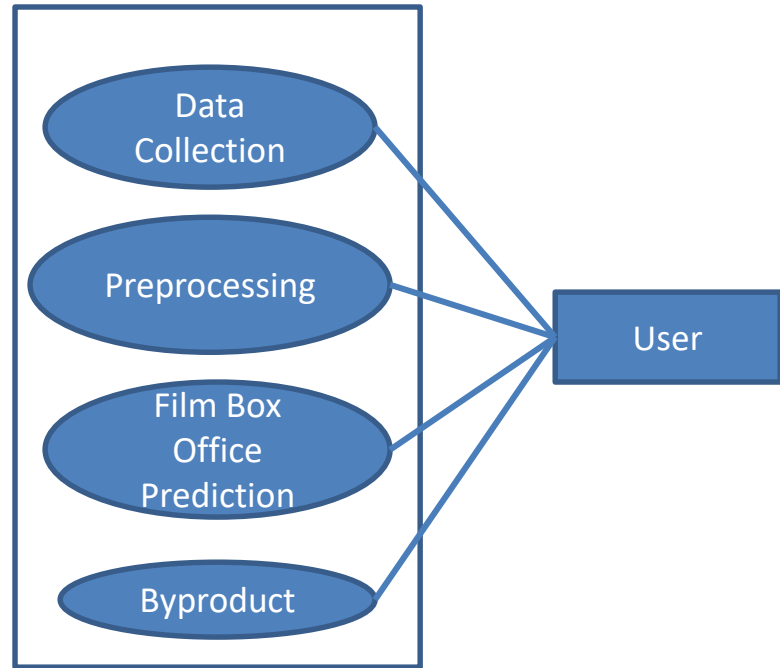
Functions

1. Film Box Office Prediction
2. Byproduct: Keyword Comparison
 - Word Cloud
 - Media Attention
 - Feature Comparisons



Film Box Office Prediction

Use Case Diagram



Film Box Office Prediction

Input and Output

Input: Keywords of film name

- Byproduct: Keywords
- Other conditions: Word Frequency, Periods,...

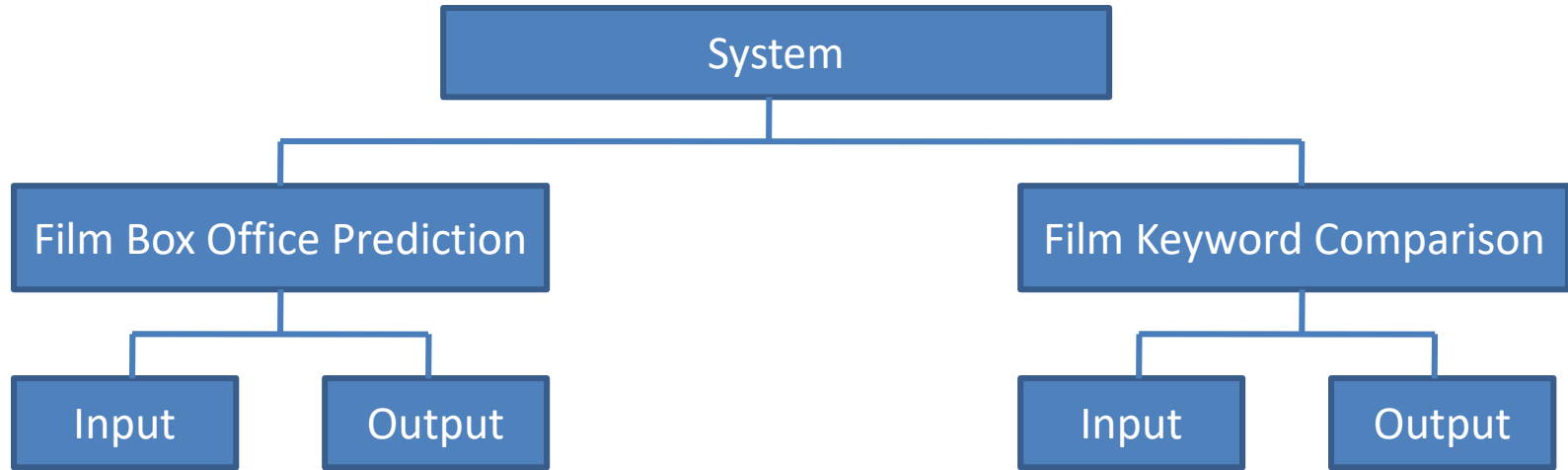
Output: Prediction value of film box office

- Word Cloud,
- Media Attention,
- Word Frequency Comparison



Film Box Office Prediction

System Design



Film Box Office Prediction

System Architecture

*Weighted Average Computing
Word Cloud, Media Attention*

*Historical Film Box Office Records
Statistical Computing of News Report*

Flask, Word Frequency Computing

Word Dictionaries

My SQL

Python

Film Box Office Prediction

Byproduct

Corresponding Film Detection

News Analysis

Keyword Feature Selection

Keyword Input

Word Segmentation

Preprocessing

Database

Web Crawlers

Web APIs



Film Box Office Prediction

Databases

Word_Dictionary

News

Stop_Word

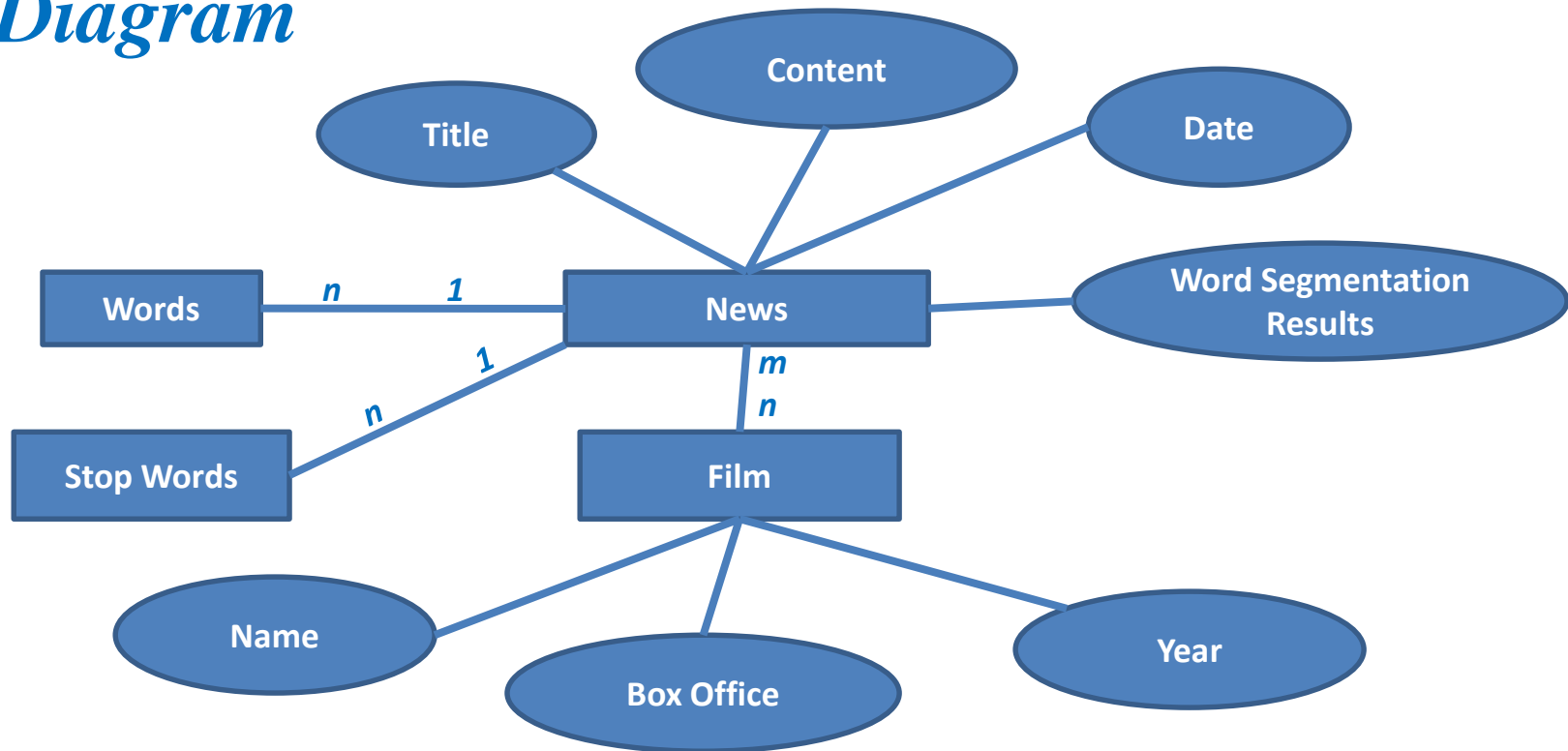
Historical_Film_Box_Office

Tips: Film names also can be used for word segmentation.



Film Box Office Prediction

ER Diagram



Computing Steps



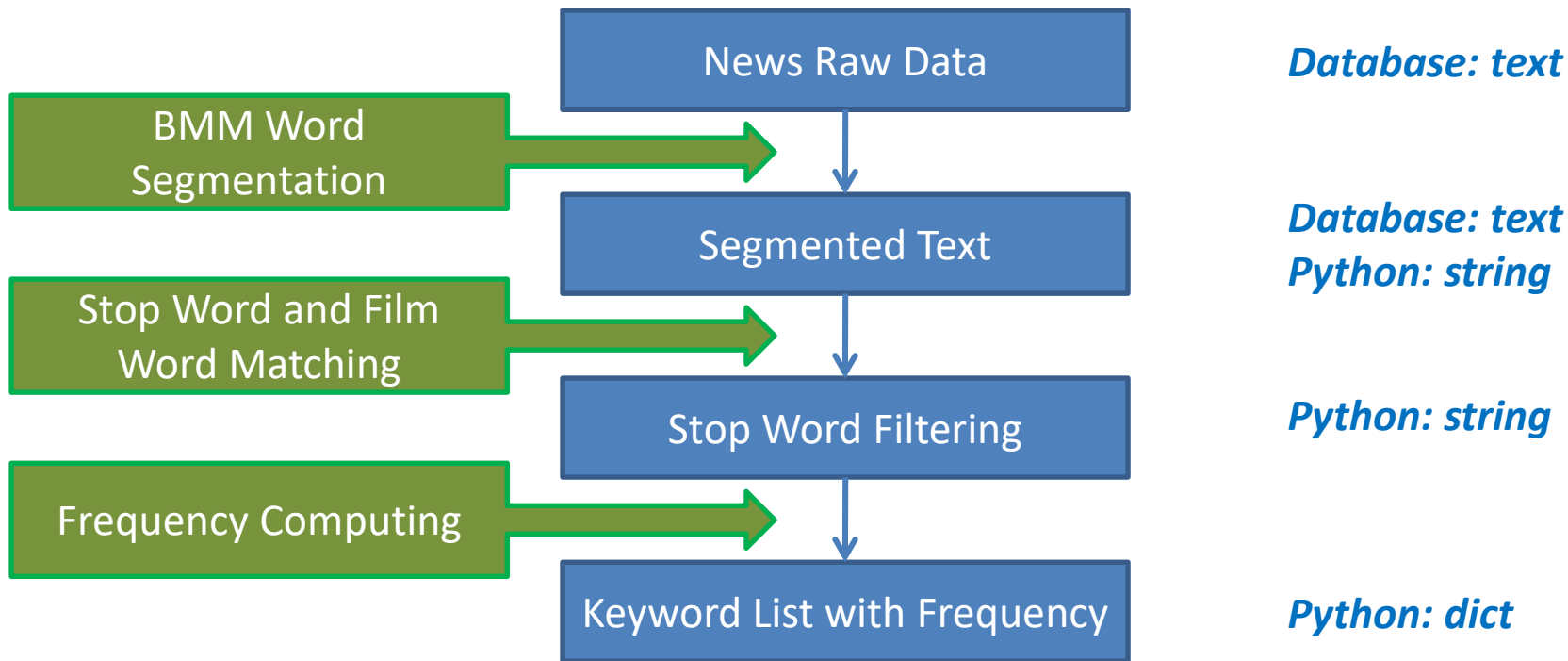
Film Box Office Prediction

Data Collection



Film Box Office Prediction

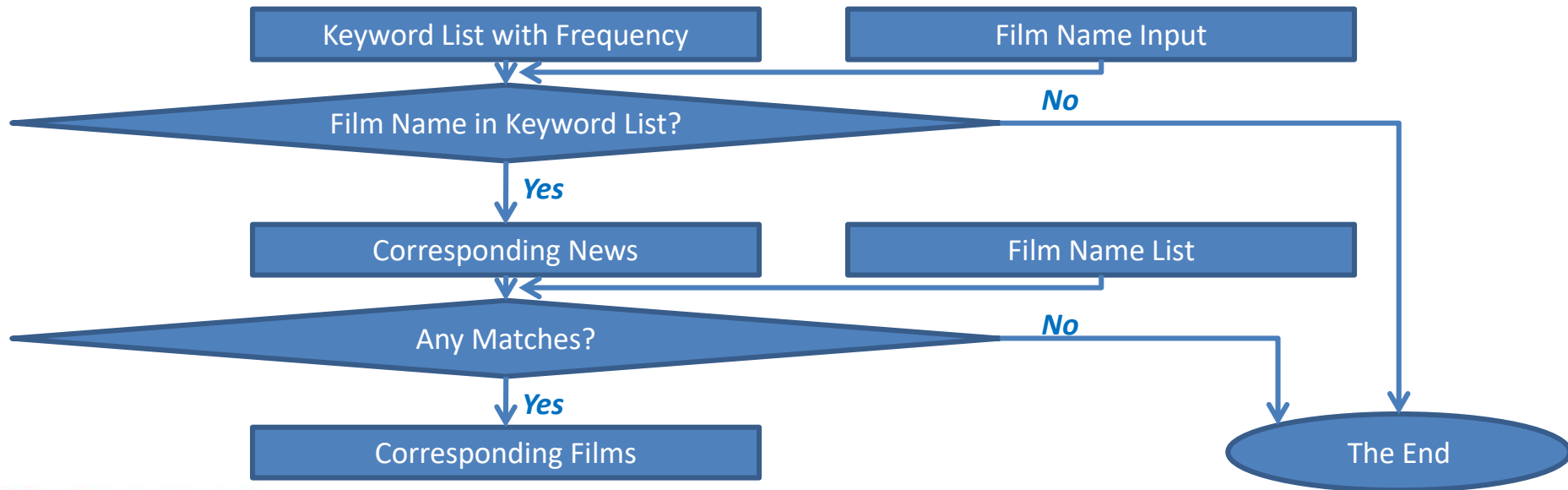
Data Transformation



Film Box Office Prediction

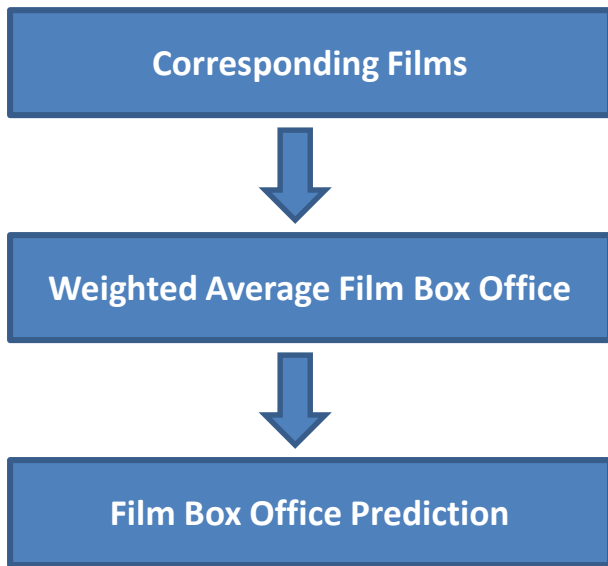
Information Acquisition (From Data to Info.)

For Film Box Office Prediction



Film Box Office Prediction

Prediction and Data Visualization

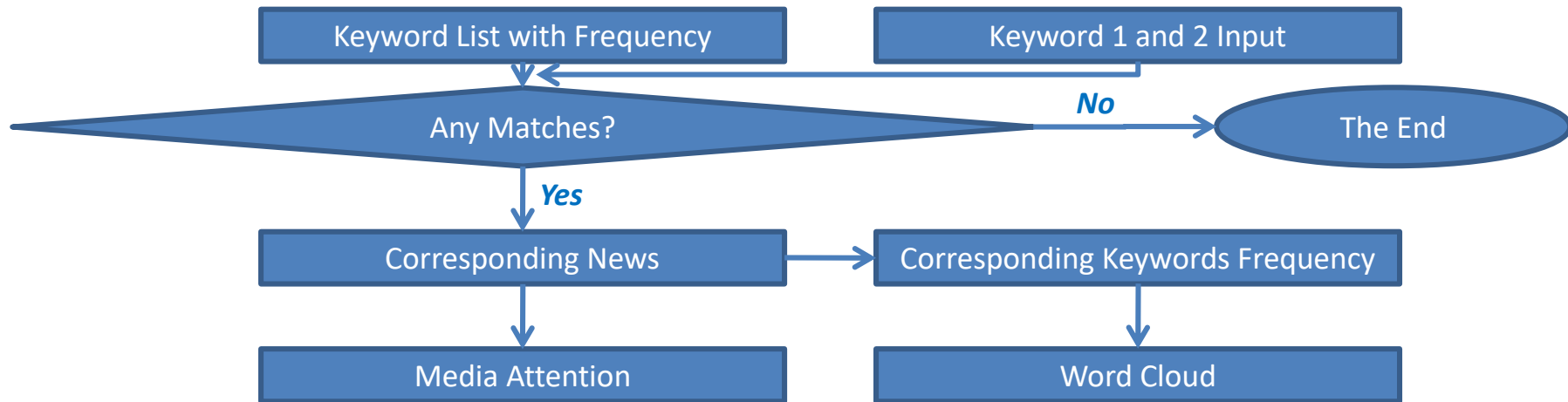


$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_k f_k}{n}$$

Film Box Office Prediction

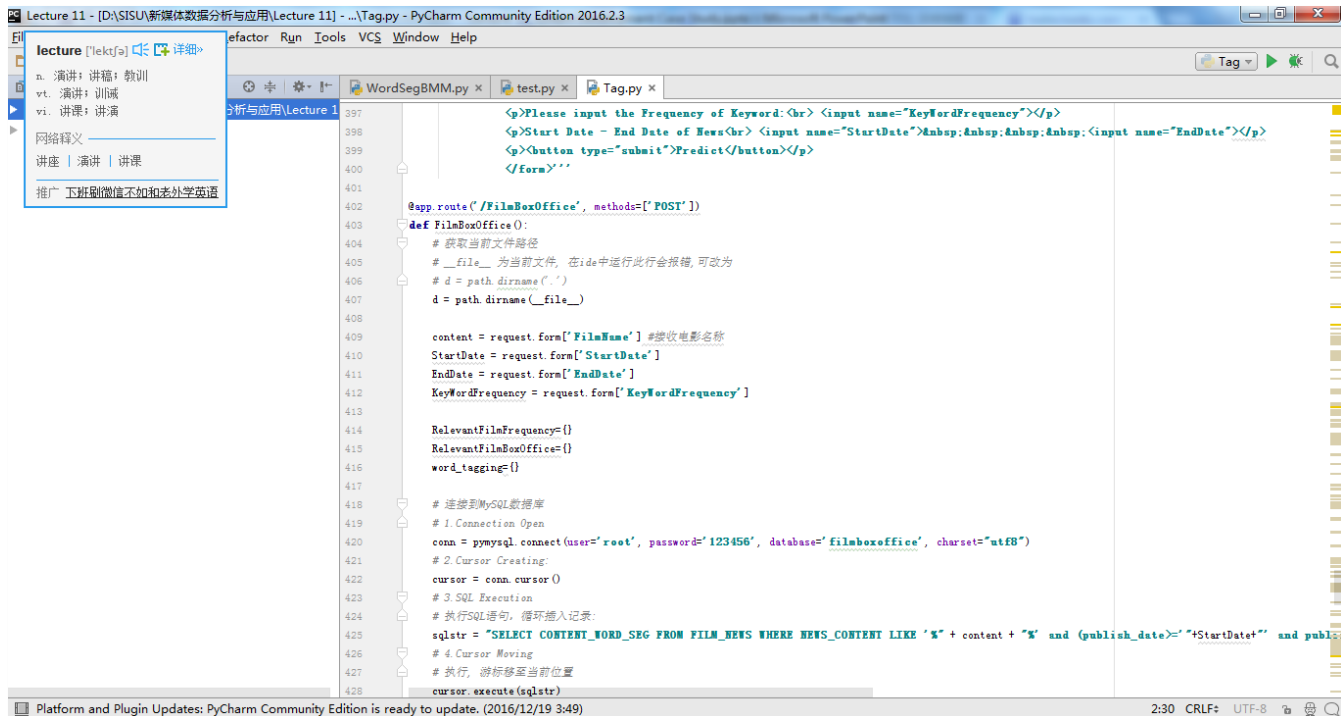
Text Mining

For Byproduct, Keyword Comparison



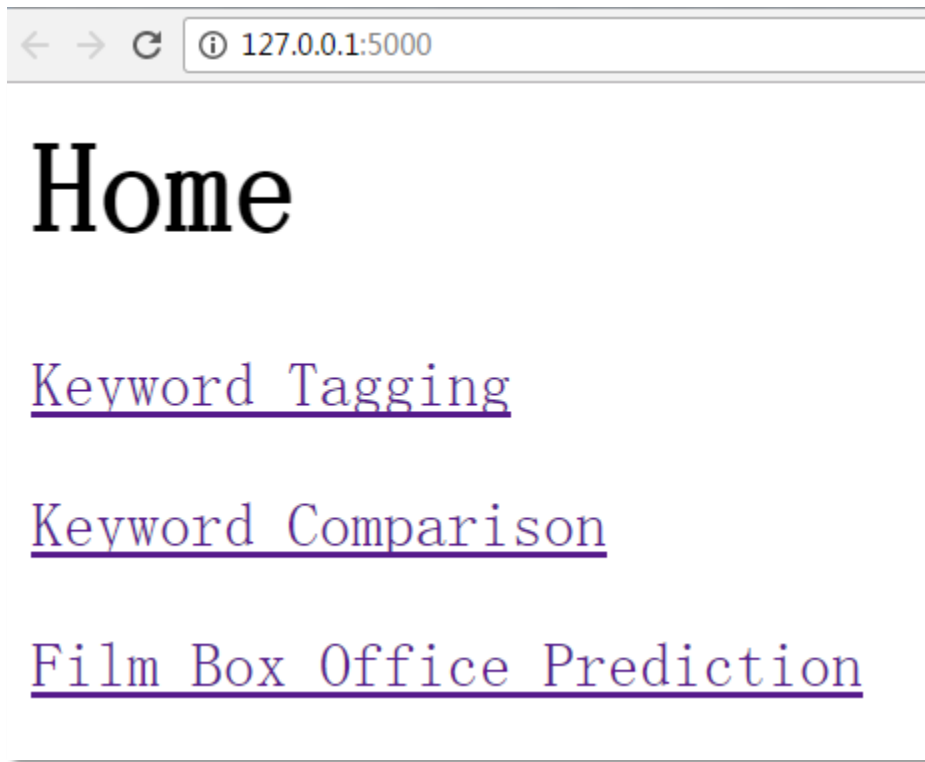
Software Development

Python
PyCharm
Flask
MySql



Film Box Office Prediction

Testing



Film Box Office Prediction

Input for Keyword Comparison

Please input the Keywords:

Please input the Frequency of Keyword:

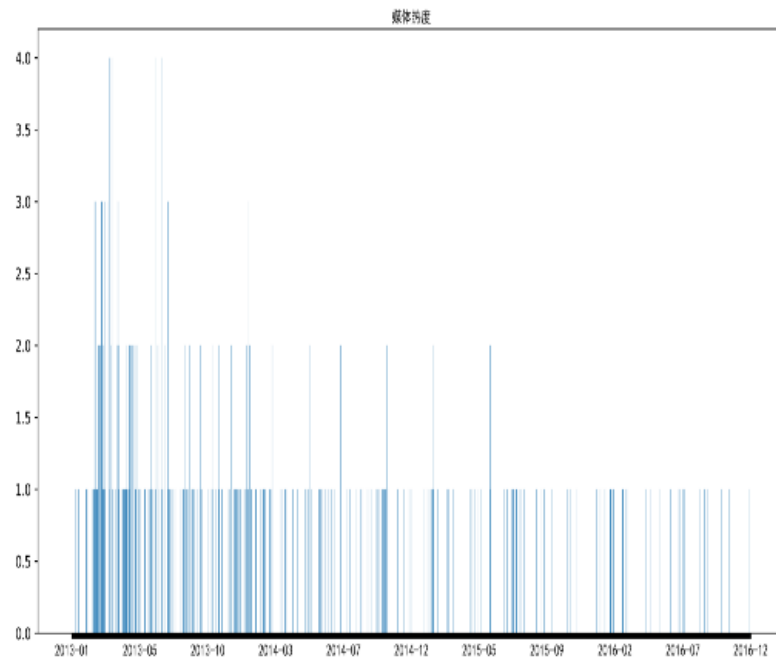
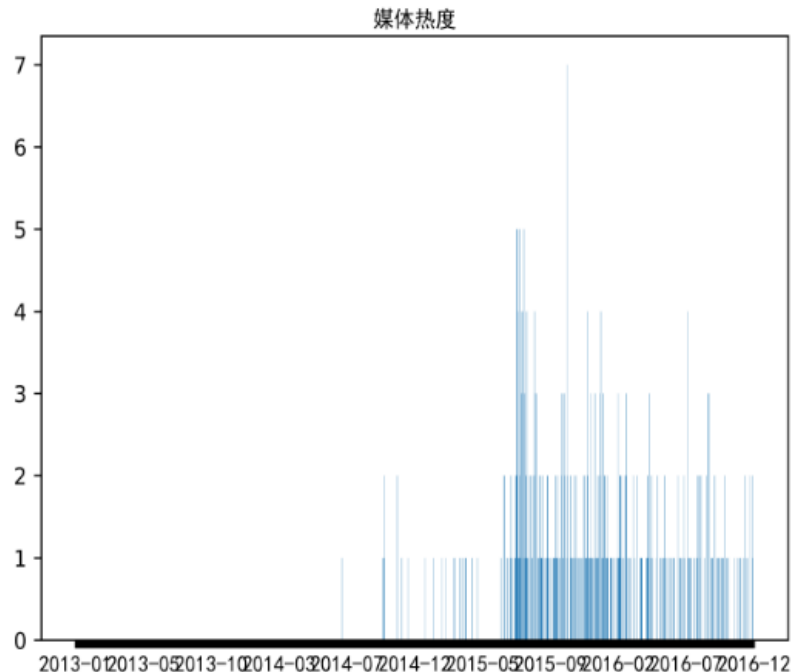
Start Date - End Date



Key Word 1: 半月传

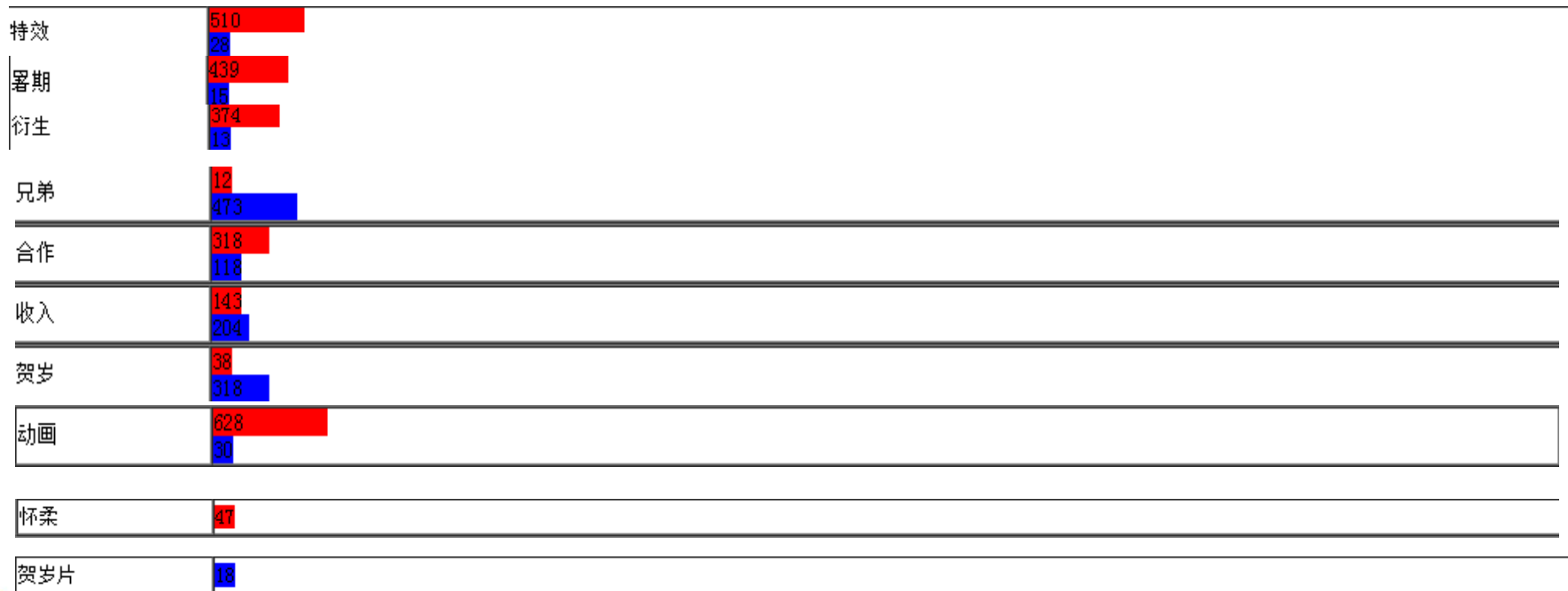


Film Box Office Prediction



Film Box Office Prediction

- Keyword Comparison



Film Box Office Prediction

← → ↻ ⓘ 127.0.0.1:5000/FilmBoxOffice

Please input the Film Name:

长城

Please input the Frequency of Keyword:

2

Start Date – End Date of News

2016-1-1

2016-12-1

Predict



Film Box Office Prediction

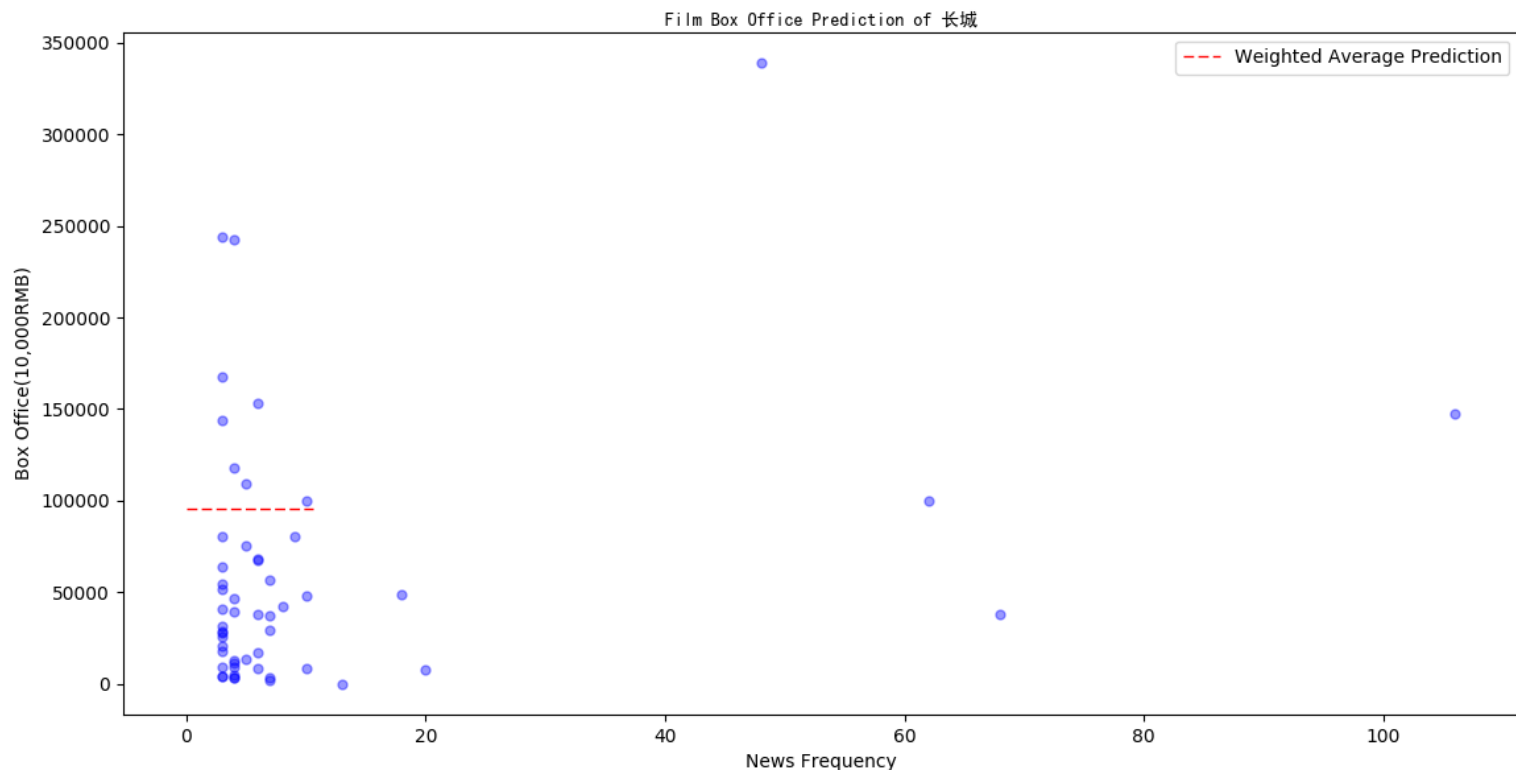


[Home](#)

Film Box Office of 长城 : 95428.38819320215(x10,000) RMB



Film Box Office Prediction



Film Box Office Prediction

• 《芳华》

[Home](#)

Film Box Office of 芳华: 151097.2136392405(x10,000) RMB

Please input the Film Name:

Please input the Frequency of Keyword:

Start Date - End Date of News

Predict



芳华 (2017)

Youth

累计票房
142241.5万

类型: 战争/剧情/爱情

片长: 136min

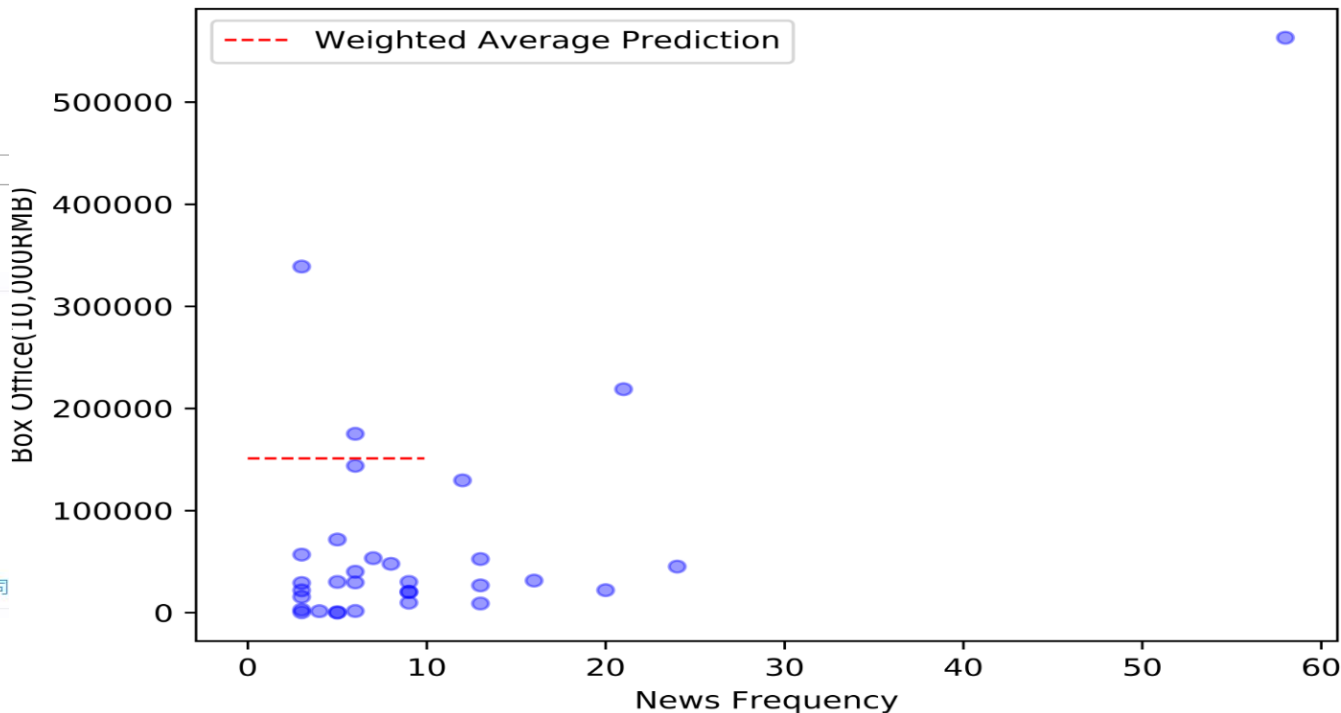
上映时间: 2017-12-15 (中国)

制式: 2D/IMAX

国家及地区: 中国

发行公司: 华谊兄弟电影有限公司

Film Box Office Prediction of 芳华



上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

Film Box Office Prediction

- 《芳华》

[Home](#)

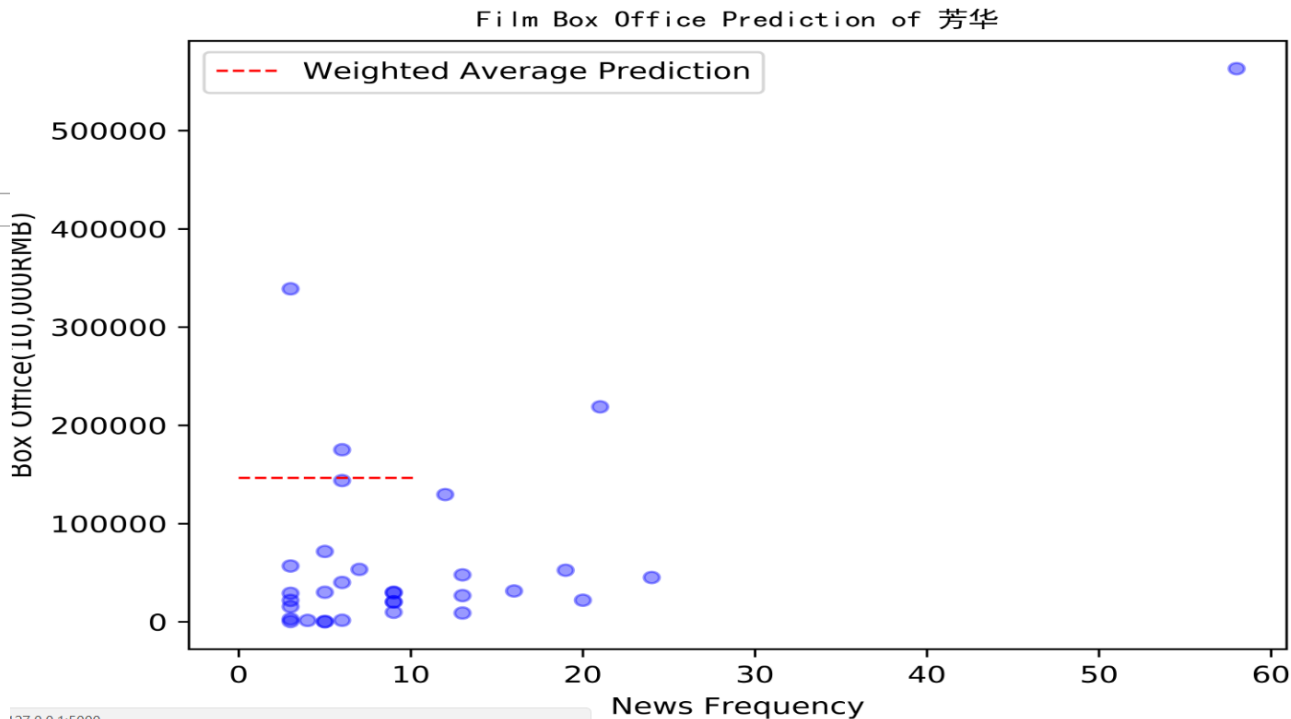
Film Box Office of 芳华 : 146639.15003030302(x10,000) RMB

Please input the Film Name:

Please input the Frequency of Keyword:

Start Date - End Date of News

Predict



Conclusions



Film Box Office Prediction



Ask A Question

What are the shortages of this system?

Do you have any ideas about developing a better one?





上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

tips for your career

To Be A Good Data Analyst

To Be A Good Data Analyst

Tip 1

- You have opinions, so do data
- How to read and interpret these data is very important, it depends on your opinions
- Sometimes, GUESS is important, a hypothesis is crucial to the problem



Guess for Hypothesis



EXAMPLE 2:
Film Stars

Guess for Hypothesis

哪种关系更稳定？ What kind of relationship is more steady between Male and Female?

- 不是东风压倒西风，就是西风压倒东风 One Strong, One Weak
- 两种风差不多强劲 Equal

Take Films Stars as an example:

Hypothesis

男女之间，不是东风压倒西风，就是西风压倒东风，你待她太好，她未必会投桃报李。

——司溟 《鸠之媚》

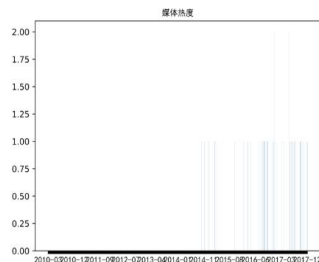
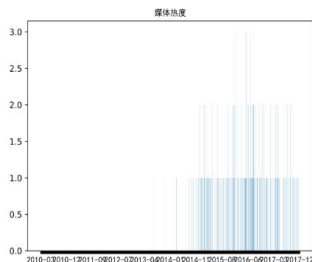
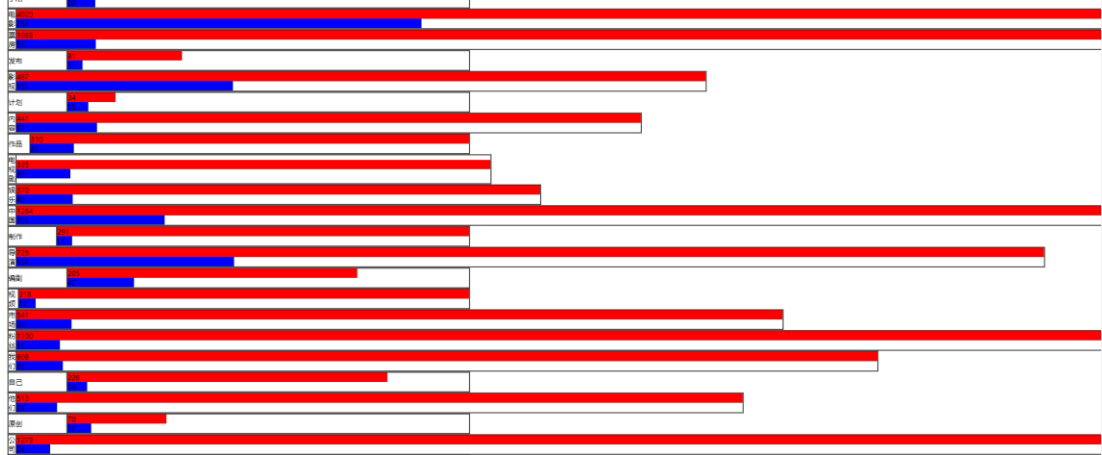


Guess for Hypothesis

- 鹿晗 关晓彤;
- 孙俪 邓超;
- 佟丽娅 陈思诚;
- 李小璐 贾乃亮



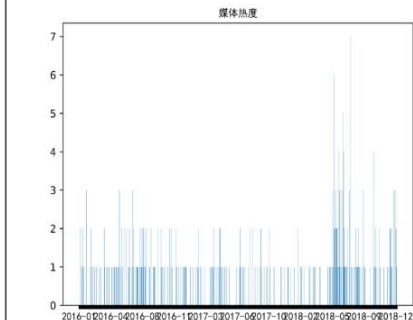
• 鹿晗 关晓彤 (2018)



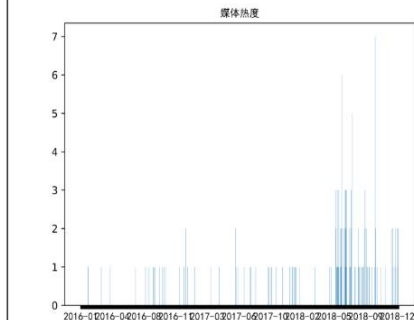
• 鹿晗 关晓彤 (2019)

Similarity: 54.00576368876081%

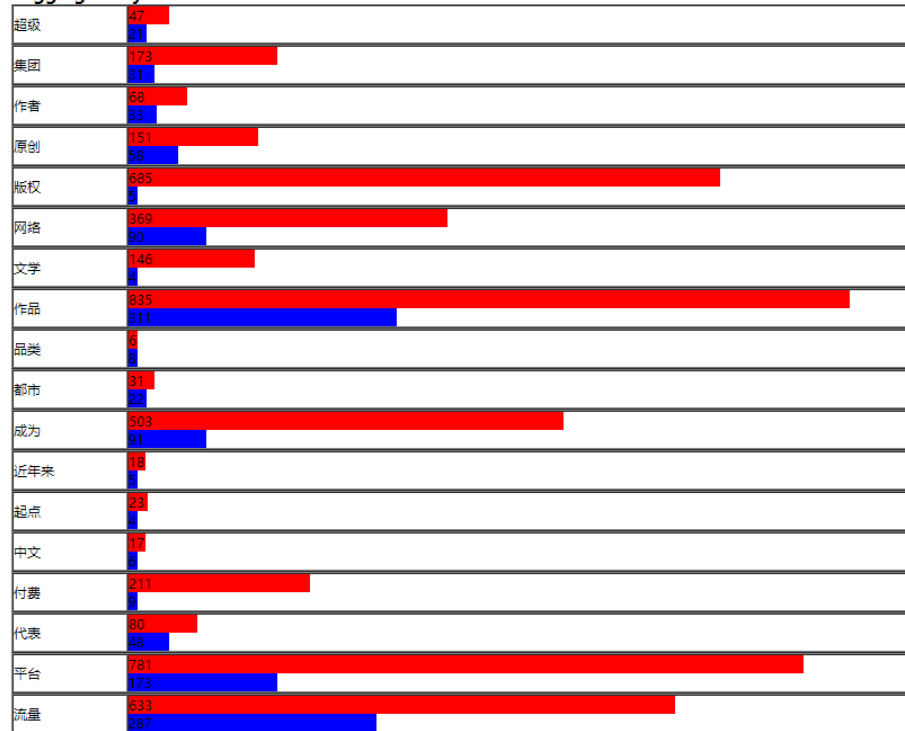
Key Word 1: 鹿晗



Key Word 2:关晓彤



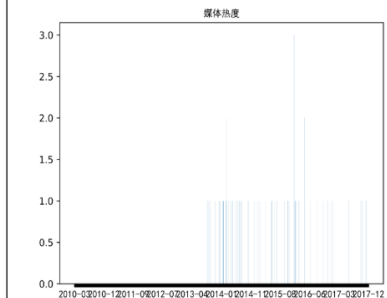
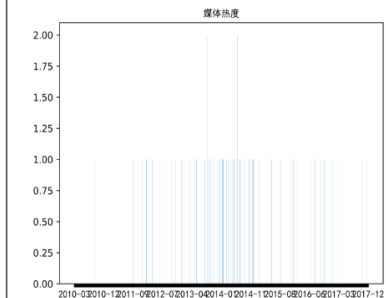
Tagging Analysis:



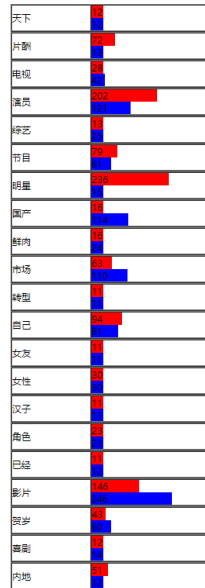
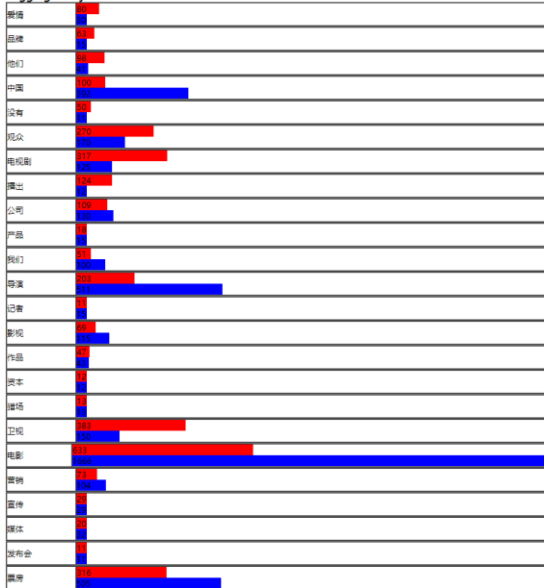
- 佟丽娅 陈思诚(2018)

Similarity: 45.76802507836991%

Key Word 1:佟丽娅



Tagging Analysis:

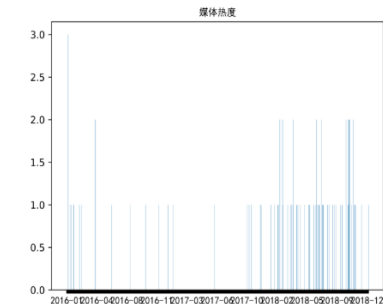
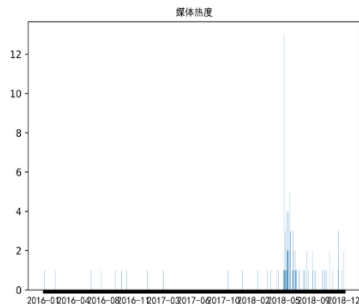


- 佟丽娅 陈思诚(2019)

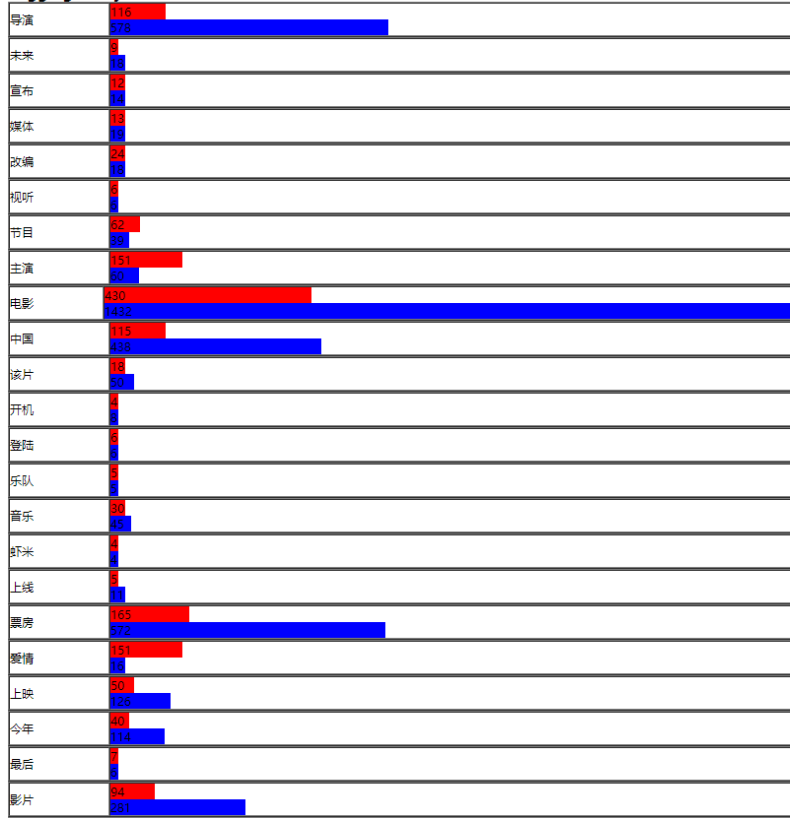
Similarity: 41.10091743119266%

Key Word 1:佟丽娅

Key Word 2:陈思诚



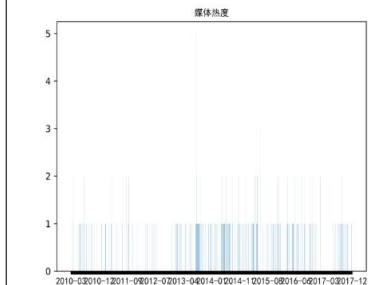
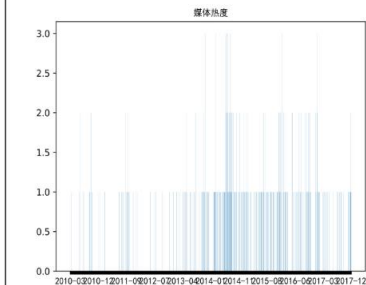
Tagging Analysis:



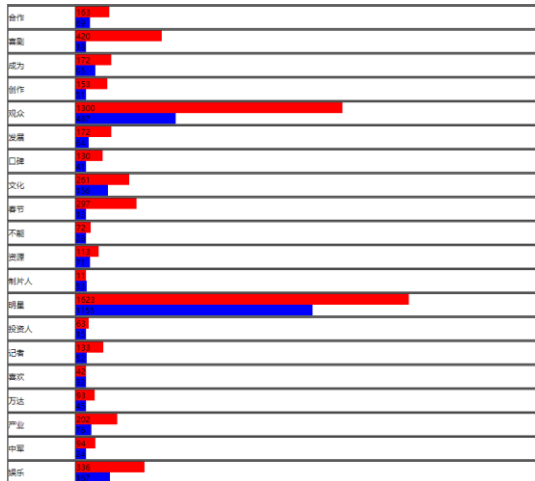
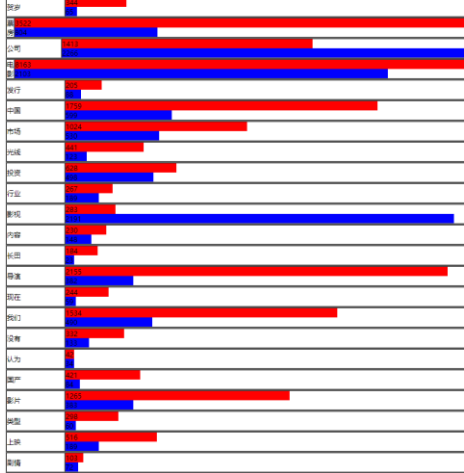
- 邓超 孙俪(2018)

Similarity: 52.14408233276158%

Key Word 1:邓超



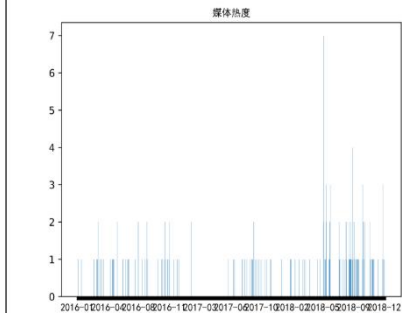
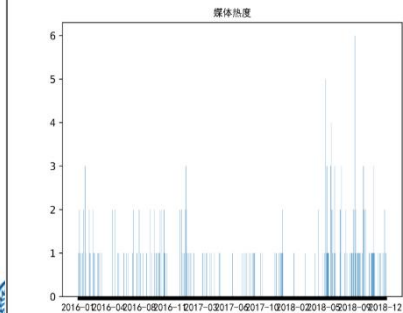
Tagging Analysis:



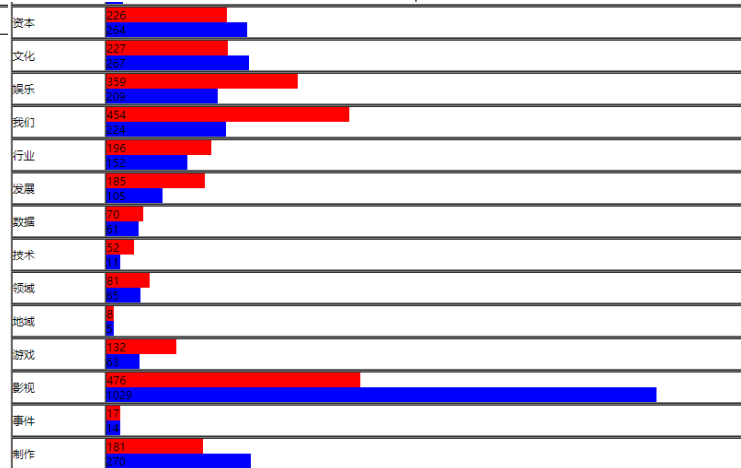
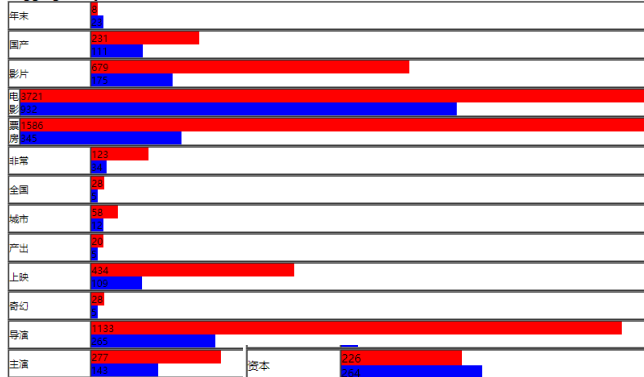
- 邓超 孙俪(2019)

Similarity: 59.7268588770865%

Key Word 1: 邓超



Tagging Analysis:

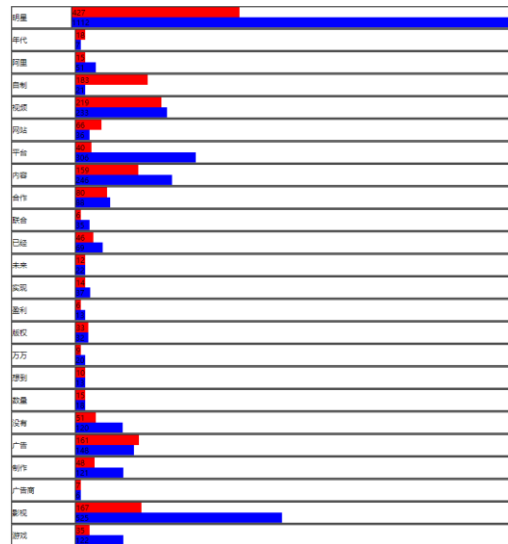
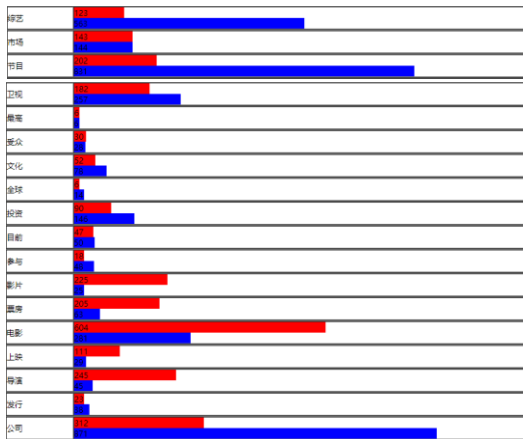
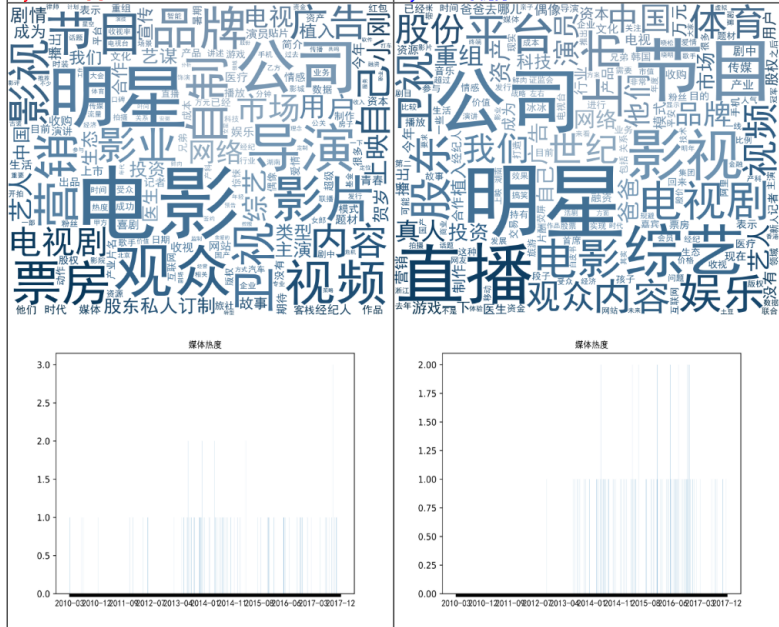


- 李小璐 贾乃亮 (2018)

Similarity: 55.70247933884298%

Key Word 1: 李小璐

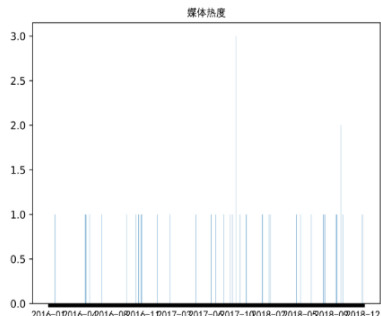
Key Word 2:贾乃亮



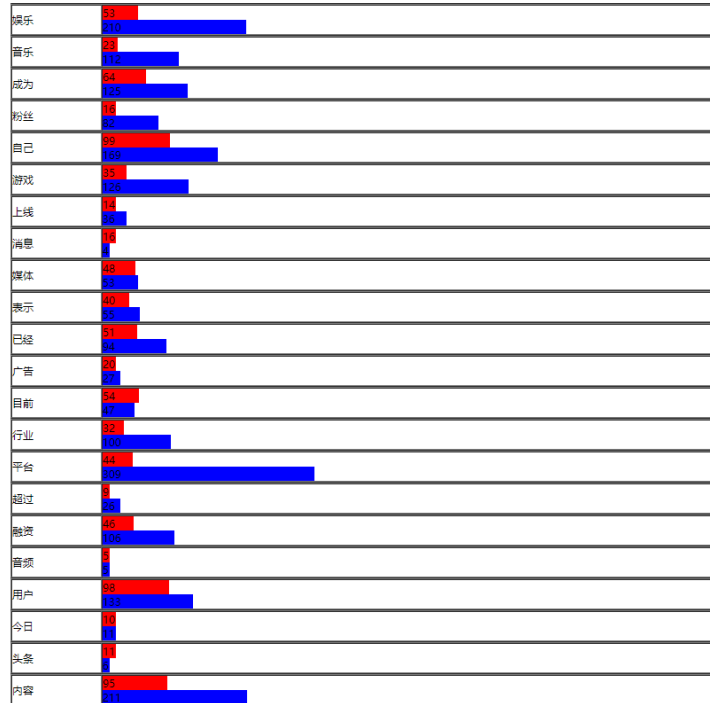
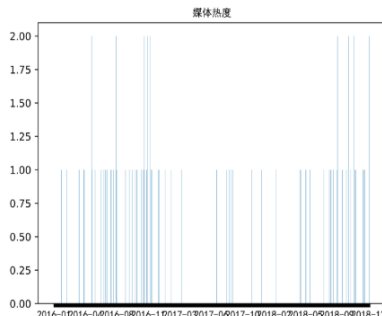
• 李小璐 贾乃亮 (2019)

Similarity: 50.21173623714459%

Key Word 1: 李小璐



Key Word 2: 贾乃亮





Similarity +++



Similarity ---

Now, what is your conclusion?



To Be A Good Data Analyst

Tip 2

- Data Quality is always the most important
- Precise Prediction needs good data quality



To Be A Good Data Analyst

Tip 3

- Data Analysis is not the end, but a new start.
Decision Support is more important.



To Be A Good Data Analyst

Tip 4

- To know more about your business, which is more important than to know more algorithms and mathematic models.



To Be A Good Data Analyst

Tips 5

- Conclusions that are not correct, feasible or applicable are useless
- Conclusions will change, if some elements, such as hypothesis, time, and place are changed





上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

Reference

Books and Chapters (1)

<https://item.jd.com/11983227.html>

Chapter 1-2

Machine Learning Package Installation

Machine Learning Theory Foundations



Reference

Books and Chapters (2)

<https://item.jd.com/11803260.html>

Chapter 5

Data Mining Essentials

Online Reference:

<http://www.public.asu.edu/~huanliu/>



Books and Chapters (3)

<https://item.jd.com/11676691.html>

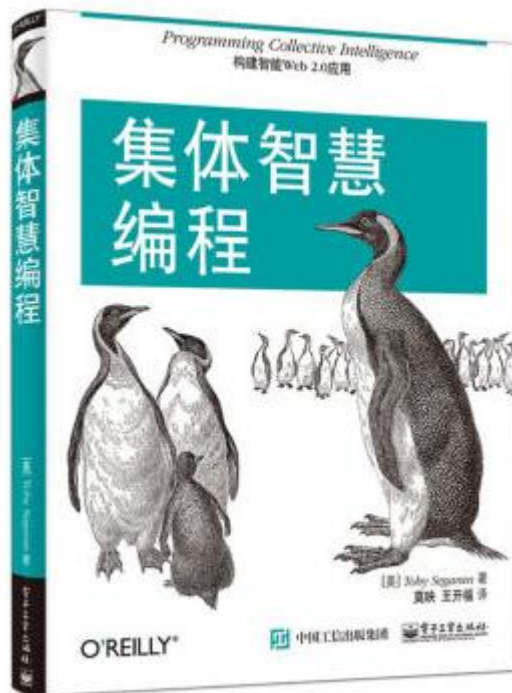
Python Data Visualization



Books and Chapters (4)

<https://item.jd.com/11667512.html>

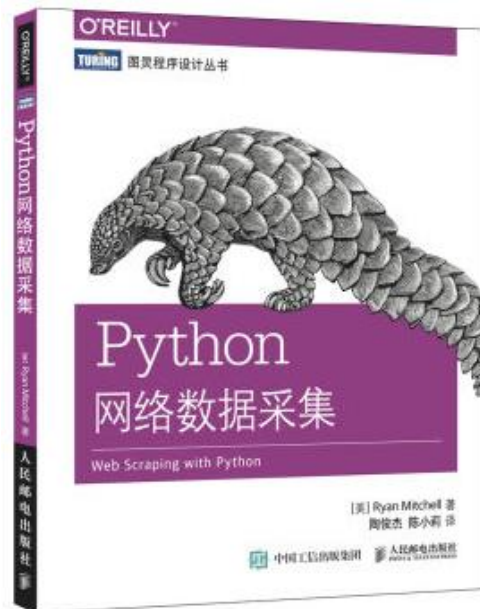
Programming Collective Intelligence



Books and Chapters (5)

<https://item.jd.com/11896401.html>

Python网络数据采集



All References for this Course:

- 张海藩.软件工程导论(第六版)[M].北京:清华大学出版社.2013年
- Meliir Page-Jones.UML面向对象设计基础[M].北京:人民邮电出版社.2012年
- 王珊、萨师煊.数据库系统概论（第5版）[M].北京:高等教育出版社.2014年
- 廖雪峰的官方网站.Python教程[OL].<http://www.liaoxuefeng.com/wiki/0014316089557264a6b348958f449949df42a6d3a2e542c000>.2016年
- Microsoft Virtual Academy.使用Python编程简介[OL].https://mva.microsoft.com/zh-cn/training-courses/-python--8360?l=EK9zuOO8_2604984382.2016年
- Ryan Mitchell. Python网络数据采集[M].北京:人民邮电出版社.2016年
- 宗成庆.统计自然语言处理（第2版）[M].北京:清华大学出版社.2013年
- Steven Bird, Ewan Klein, Edward Loper. Python自然语言处理[M].北京:人民邮电出版社.2014年
- Reza Zafarani, Mohammad Ali Abbasi, Huan Liu. 社交媒体挖掘[M].北京:人民邮电出版社.2015年
- 范淼, 李超.Python机器学习及实践：从零开始通往Kaggle竞赛之路[M].北京:清华大学出版社.2016年
- Igor Milovanovic.Python数据可视化编程实战[M].北京:人民邮电出版社.2015年
- Toby Segaran.集体智慧编程[M].北京:电子工业出版社.2009年





The End of Lecture 13

Thank You

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

