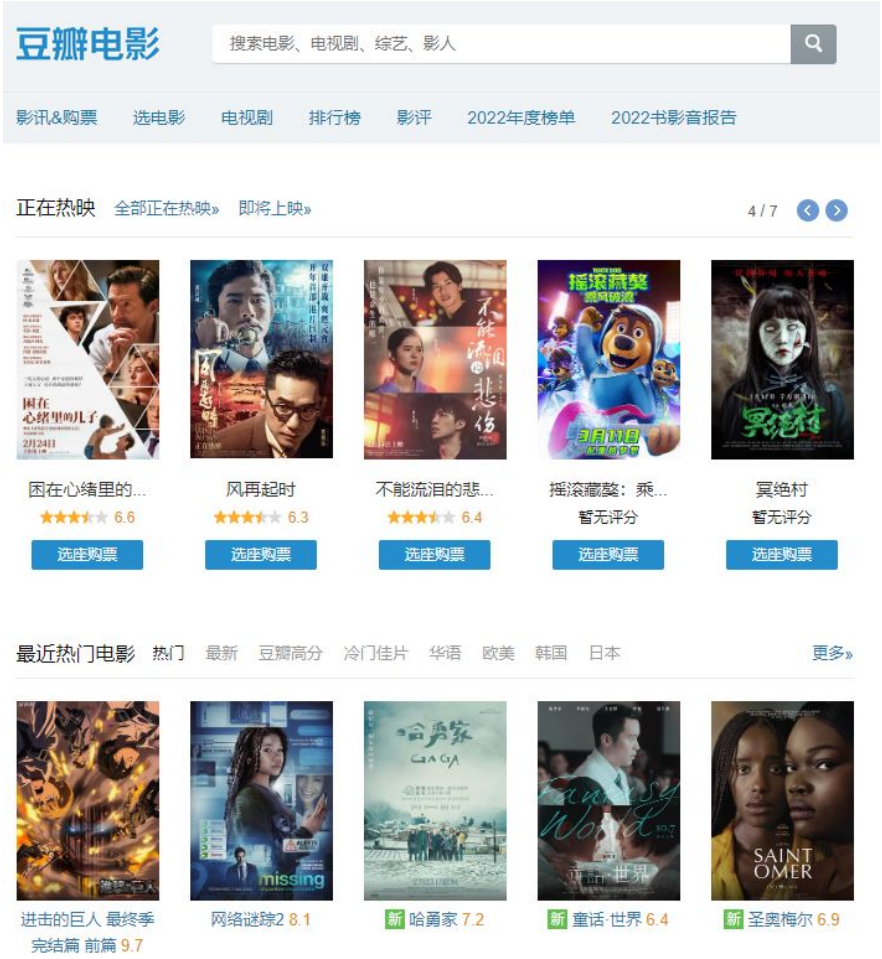
豆瓣电影评论文本分析

A Chinese NLP study based on Douban movie reviews

Daphne 丁奇

GA DSI Project

OVERVIEW 概述



关于豆瓣
What is Douban

Douban is a Chinese social networking website that offers rating, reviewing, and recommending functions for various content such as movies, books, music, and events.

- 01 PROJECT OBJECTIVE

02 DATA COLLECTION

03 DATA CLEANING & EDA

04 PREPROCESSING & MODELING

05 CONCLUSION & RECOMMENDATIONS

猫咪没有它小...... ★★★★★ 2023-01-16 15:14:01

为啥老是那么打光？bgm不要随便放好不好，烘托气氛不是这么烘托的！

为啥老是那么打光？

我在想科学边界开会，办公室谈话，在家写东西都爱不开灯的吗？全部都是窗子里透出的灯光打在脸上，好像在开什么恐怖分子大会一样，过头了，真的过头了！

Rating: ★★★★★

Comment: “Lighting so bad, background music is lousy...”

极恶非道 ★☆☆☆☆ 2023-02-08 14:46:11

太精彩了！

原著党表示非常满意，必须满分！

还原程度太高了！感谢剧组！

期待导演的下一部作品！

真的太好看了啊啊啊！❤️❤️❤️

Rating: ★☆☆☆☆

Comment: “Awesome! I’m so happy... ahhhhhhhhhhhhh”

Problem Statement

- **1. Proof of Concept:**

  Is the analysis through text EDA & NLP able to **uncover the linguistic complexity and cultural nuances of Chinese**, by a test of movie genre classification?

- **2. Insights for Douban:**

  How far apart are Douban reviews based on different movie genres?

01 PROJECT OBJECTIVE

- 02 DATA COLLECTION

03 DATA CLEANING & EDA

04 PREPROCESSING & MODELING

05 CONCLUSION & RECOMMENDATIONS

CATEGORIES

Sci-fi

3000+ reviews

Stephen Chow

3000+ reviews

DATA COLLECTION

1

2

3

	time	user	rating	comment	votes
1	2023-01-22 16:31:45	了不起的花轮君	4	原来人死后还要继续干生前的活儿...	30607
2	2023-01-22 12:01:06	电子梦	5	不必对比什么 作为里程碑他已经无敌	43863
3	2023-01-22 12:06:58	撕蓝	4	刘德华那条线有点儿意思，为这个...	24541
4	2023-01-22 13:57:05	DrunkenPeach	5	2019年看第一部的时候就想着，回...	31149

明日战记 明日戰記 (2022)

导演: 吴炫辉
编剧: 刘浩良 / 麦天枢
主演: 古天乐 / 刘青云 / 刘嘉玲 / 姜皓文 / 谢君豪 / 更多...
类型: 动作 / 科幻
制片国家/地区: 中国香港 / 中国大陆
语言: 汉语普通话 / 粤语
上映日期: 2022-08-05(中国大陆) / 2022-08-25(中国香港)
片长: 99分钟
又名: 矛盾战争 / Warriors of Future
IMDb: tt7375466

豆瓣评分
6.1 ★★★☆☆ 127189人评价
5星 7.8%
4星 20.8%
3星 46.9%
2星 19.5%
1星 4.9%

好于 69% 科幻片
好于 70% 动作片

重启地球 (2021)

导演: 林珍钊
编剧: 张圣帆
主演: 何晟铭 / 罗米 / 叶璇 / 于荣光 / 李宁 / 更多...
类型: 剧情 / 科幻
制片国家/地区: 中国大陆
语言: 汉语普通话
上映日期: 2021-09-03(中国大陆网络)
片长: 90分钟
又名: 大灾变
IMDb: tt16118262

豆瓣评分
3.7 ★★☆☆☆ 4923人评价
5星 1.6%
4星 1.6%
3星 13.7%
2星 44.2%
1星 38.9%

好于 1% 科幻片
好于 0% 剧情片

上海堡垒 (2019)

导演: 滕华涛
编剧: 滕华涛 / 韩景龙 / 江南
主演: 鹿晗 / 舒淇 / 石凉 / 高以翔 / 王宫良 / 更多...
类型: 爱情 / 科幻 / 战争
制片国家/地区: 中国大陆
语言: 汉语普通话
上映日期: 2019-08-09(中国大陆)
片长: 107分钟
又名: Shanghai Fortress
IMDb: tt6628322

豆瓣评分
2.9 ★★☆☆☆ 257614人评价
5星 1.6%
4星 1.5%
3星 6.3%
2星 20.5%
1星 70.0%

好于 0% 科幻片
好于 0% 爱情片

流浪地球2 (2023)

导演: 郭帆
编剧: 杨治学 / 龚格尔 / 郭帆 / 叶濡畅
主演: 吴京 / 刘德华 / 李雪健 / 沙溢 / 宁理 / 更多...
类型: 科幻 / 冒险 / 灾难
制片国家/地区: 中国大陆
语言: 汉语普通话 / 俄语 / 英语 / 印地语 / 法语
上映日期: 2023-01-22(中国大陆)
片长: 173分钟
又名: The Wandering Earth II / The Wandering Earth 2 / 《流浪地球》前传

豆瓣评分
8.3 ★★★★☆ 1034573人评价
5星 44.8%
4星 34.2%
3星 15.9%
2星 3.5%
1星 1.6%

好于 96% 科幻片
好于 97% 灾难片

流浪地球 (2019)

导演: 郭帆
编剧: 龚格尔 / 严东旭 / 郭帆 / 叶俊策 / 杨治学 / 吴美 / 叶濡畅 / 沈晶晶 / 刘慈欣
主演: 吴京 / 屈楚萧 / 李光洁 / 吴孟达 / 赵今麦 / 更多...
类型: 科幻 / 冒险 / 灾难
制片国家/地区: 中国大陆
语言: 汉语普通话 / 英语 / 俄语 / 法语 / 日语 / 韩语 / 印尼语
上映日期: 2019-02-05(中国大陆) / 2020-11-26(中国大陆重映)
片长: 125分钟 / 137分钟(重映版)
又名: 流浪地球：飞跃2020特别版 / The Wandering Earth

豆瓣评分
7.9 ★★★★☆ 1887905人评价
5星 34.6%
4星 36.3%
3星 21.6%
2星 4.9%
1星 2.7%

好于 93% 科幻片
好于 94% 灾难片

火星异变 (2021)

导演: 刘娜
编剧: 贺馨 / 蒋顾世
主演: 石凉 / 任重 / 素笑坤 / 刘馨棋 / 辛祚宇 / 更多...
类型: 剧情 / 科幻 / 灾难
制片国家/地区: 中国大陆
语言: 汉语普通话
上映日期: 2021-08-27(中国大陆)
片长: 75分钟
又名: Mars Anomaly / Mutation on Mars
IMDb: tt15360764

豆瓣评分
3.6 ★★☆☆☆ 3227人评价
5星 0.7%
4星 2.1%
3星 14.8%
2星 43.7%
1星 38.7%

好于 1% 科幻片
好于 0% 剧情片

No.84 豆瓣电影Top250

功夫 (2004)

导演: 周星驰
编剧: 曾瑾昌 / 陈文强 / 周星驰 / 霍昕
主演: 周星驰 / 元秋 / 元华 / 黄圣依 / 梁小龙 / 更多...
类型: 喜剧 / 动作 / 犯罪 / 奇幻
制片国家/地区: 中国大陆 / 中国香港
语言: 粤语 / 汉语普通话 / 手语
上映日期: 2004-12-23(中国大陆/中国香港) / 2015-01-15(中国大陆3D) / 2004-09-14(多伦多电影节)
片长: 100分钟(3D重映) / 95分钟(中国大陆) / 99分钟(美国)
又名: 功夫3D / Kung Fu Hustle

豆瓣评分
8.8 1085769人评价
5星 52.9%
4星 35.2%
3星 10.9%
2星 0.9%
1星 0.2%

好于 99% 喜剧片
好于 99% 动作片

No.91 豆瓣电影Top250

喜剧之王 喜劇之王 (1999)

导演: 周星驰 / 李力持
编剧: 曾瑾昌 / 周星驰 / 李敏 / 郑文辉 / 冯勉恒 / 梁嘉杰
主演: 周星驰 / 张柏芝 / 莫文蔚 / 吴孟达 / 林子善 / 更多...
类型: 剧情 / 喜剧 / 爱情
制片国家/地区: 中国香港
语言: 粤语
上映日期: 1999-02-13(中国香港)
片长: 89分钟
又名: King of Comedy

豆瓣评分
8.8 943395人评价
5星 51.6%
4星 38.5%
3星 9.2%
2星 0.6%
1星 0.1%

好于 98% 喜剧片

No.157 豆瓣电影Top250

九品芝麻官 (1994)

导演: 王晶
编剧: 王晶
主演: 周星驰 / 吴孟达 / 张敏 / 徐锦江 / 钟丽缇 / 更多...
类型: 剧情 / 喜剧 / 古装
制片国家/地区: 中国香港 / 中国大陆
语言: 粤语
上映日期: 1994-03-31(中国香港)
片长: 108分钟
又名: 九品芝麻官之白面包青天 / Hail the Judge

豆瓣评分
8.7 652636人评价
5星 46.4%
4星 43.0%
3星 10.0%
2星 0.4%
1星 0.1%

好于 98% 喜剧片

No.47 豆瓣电影Top250

大话西游之月光宝盒 西遊記第壹佰零壹回之月光寶盒 (1995)

导演: 刘镇伟
编剧: 刘镇伟 / 吴承恩
主演: 周星驰 / 吴孟达 / 罗家英 / 蓝洁瑛 / 莫文蔚 / 更多...
类型: 喜剧 / 爱情 / 奇幻 / 古装
制片国家/地区: 中国香港 / 中国大陆
语言: 粤语 / 汉语普通话
上映日期: 2014-10-24(中国大陆) / 1995-01-21(中国香港)
片长: 88分钟
又名: 西游记101回月光宝盒 / 齐天大圣东游记 / 大话东游之一 / A Chinese Odyssey Part One - Pandora's Box

豆瓣评分
9.0 1191763人评价
5星 60.2%
4星 31.7%
3星 7.3%
2星 0.5%
1星 0.2%

好于 99% 喜剧片
好于 99% 爱情片

No.20 豆瓣电影Top250

大话西游之大圣娶亲 西遊記大結局之仙履奇緣 (1995)

导演: 刘镇伟
编剧: 刘镇伟 / 吴承恩
主演: 周星驰 / 吴孟达 / 朱茵 / 蔡少芬 / 蓝洁瑛 / 更多...
类型: 喜剧 / 爱情 / 奇幻 / 古装
制片国家/地区: 中国香港 / 中国大陆
语言: 粤语 / 汉语普通话
上映日期: 2014-10-24(中国大陆) / 2017-04-13(中国大陆重映) / 1995-02-04(中国香港)
片长: 95分钟 / 110分钟(重映版)
又名: 西游记完结篇仙履奇缘 / 齐天大圣西游记 / 大话东游之二 / 大话西游之大圣娶妻 / A Chinese Odyssey Part Two

豆瓣评分
9.2 1496123人评价
5星 67.0%
4星 26.5%
3星 5.8%
2星 0.5%
1星 0.3%

好于 99% 喜剧片
好于 99% 爱情片

武状元苏乞儿 武狀元蘇乞兒 (1992)

导演: 陈嘉上
编剧: 陈建忠 / 陈嘉上
主演: 周星驰 / 张敏 / 吴孟达 / 徐少强 / 林威 / 更多...
类型: 喜剧 / 动作 / 武侠 / 古装
制片国家/地区: 中国香港
语言: 粤语
上映日期: 1992-12-17(中国香港)
片长: 96分钟(台湾) / 101分钟(香港) / 93分钟(英国)
又名: King of Beggars
IMDb: tt0100963

豆瓣评分
8.1 371788人评价
5星 29.4%
4星 48.0%
3星 21.4%
2星 1.1%
1星 0.2%

好于 96% 喜剧片

Sci-fi	Stephen Chow
3120 reviews	**3120 reviews**
Years **2019-2023**	Years **1992-2004**
Production & Casting **By different people**	Production & Casting **Stephen Chow movies**
Stories **Set in outer space**	Stories **Varied**
Rating **Low**	Rating **High**

01 PROJECT OBJECTIVE

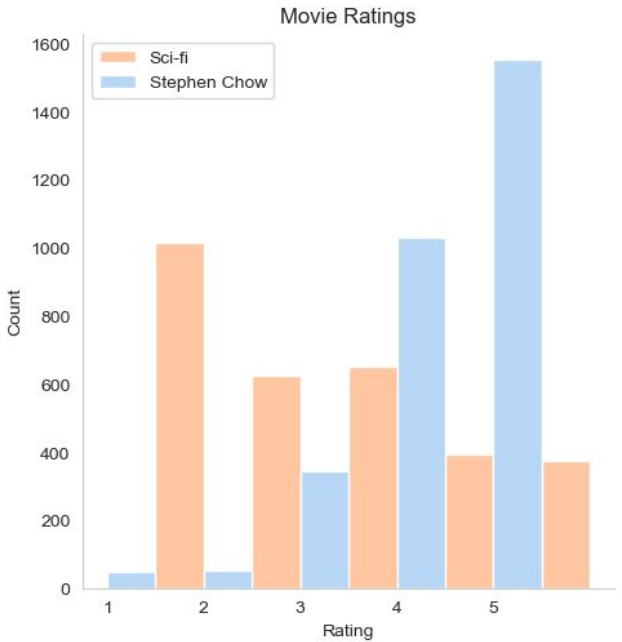
02 DATA COLLECTION

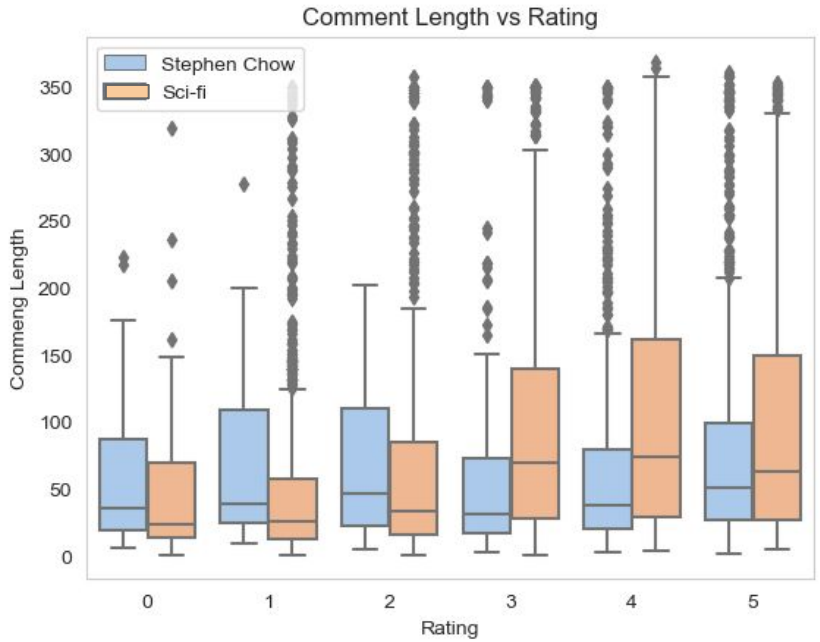
- 03 DATA CLEANING & EDA

04 PREPROCESSING & MODELING

05 CONCLUSION & RECOMMENDATIONS

Sci-fi Movies vs Stephen Chow Movies





Tokenization

So it is true people still work after they die

So | it | is | true | people | still | work | after | they | die

原 | 来 | 人 | 死 | 后 | 还 | 要 | 继 | 续 | 干 | 生 | 前 | 的 | 活 | 儿 | 是 | 真 | 的

原来 | 人 | 死后 | 还要 | 继续干 | 生前的 | 活儿 | 是真的 → **jieba.cut (结巴中文分词)**

Stopwords


“shift＋space”
Traditional Chinese
“～” “?”
Stopwords iso Chinese (ZH)
People’s Names
“科幻片” “喜剧片”
'p', 's', 'o'
Movies’ Titles

Text Visualisations

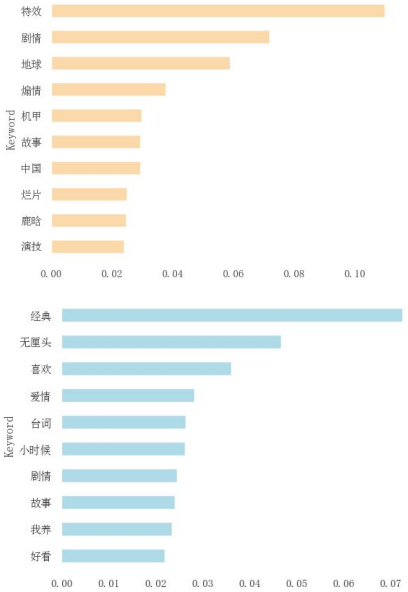
Word Cloud

Sci-fi

Stephen Chow

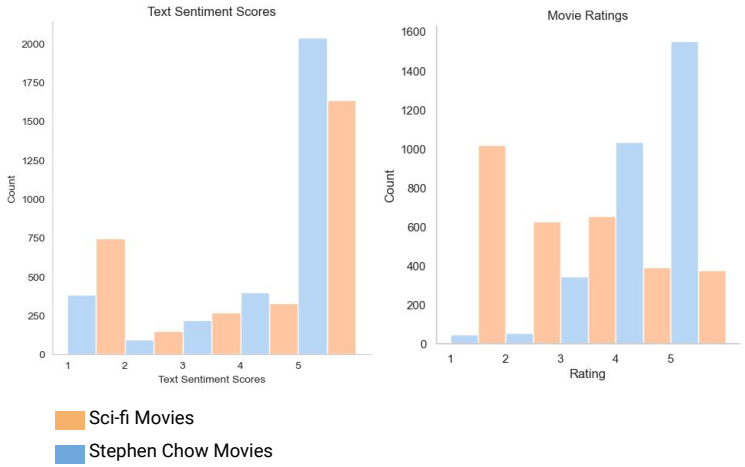
word density

TextRank



word connectivity

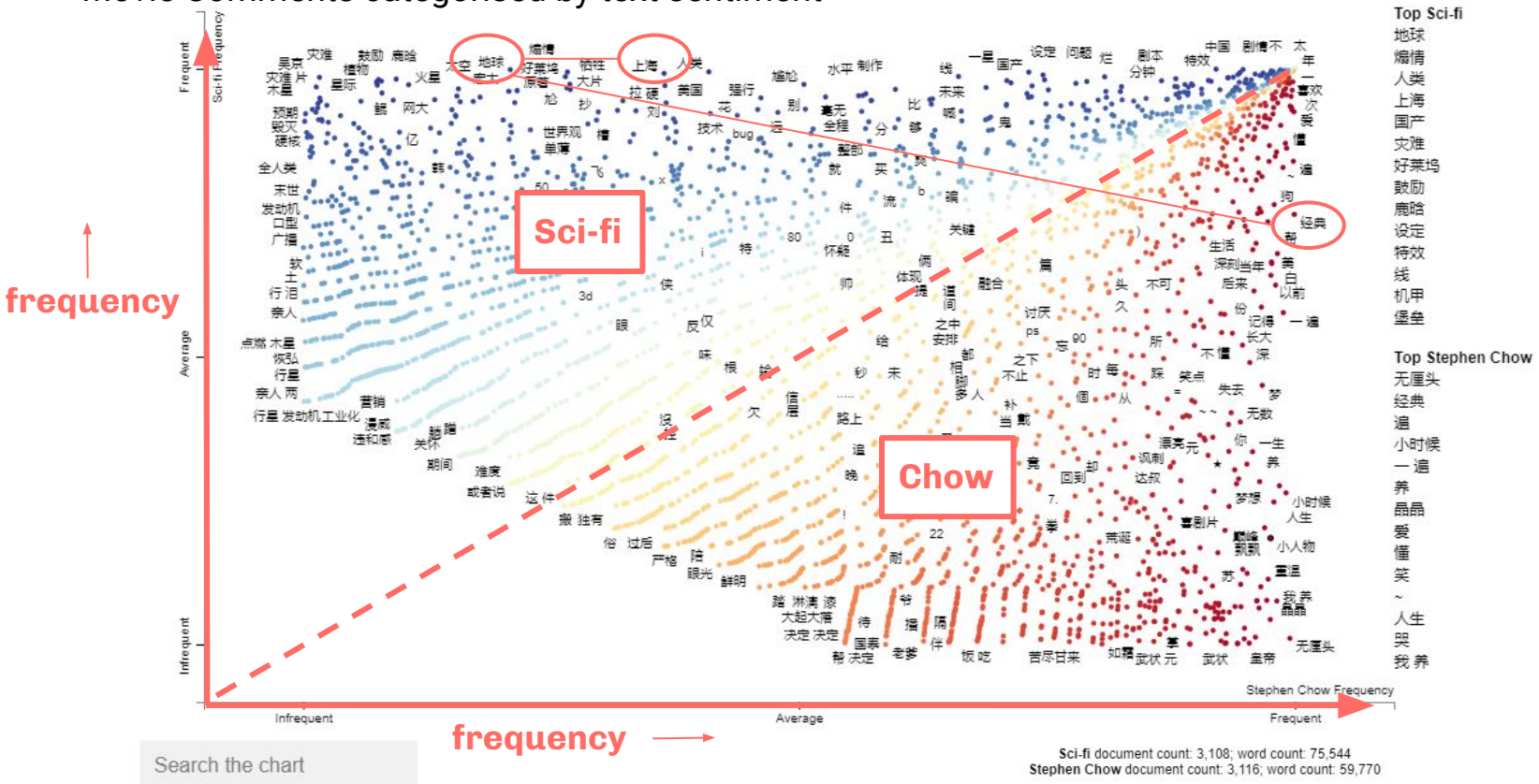
SnowNLP



word tone

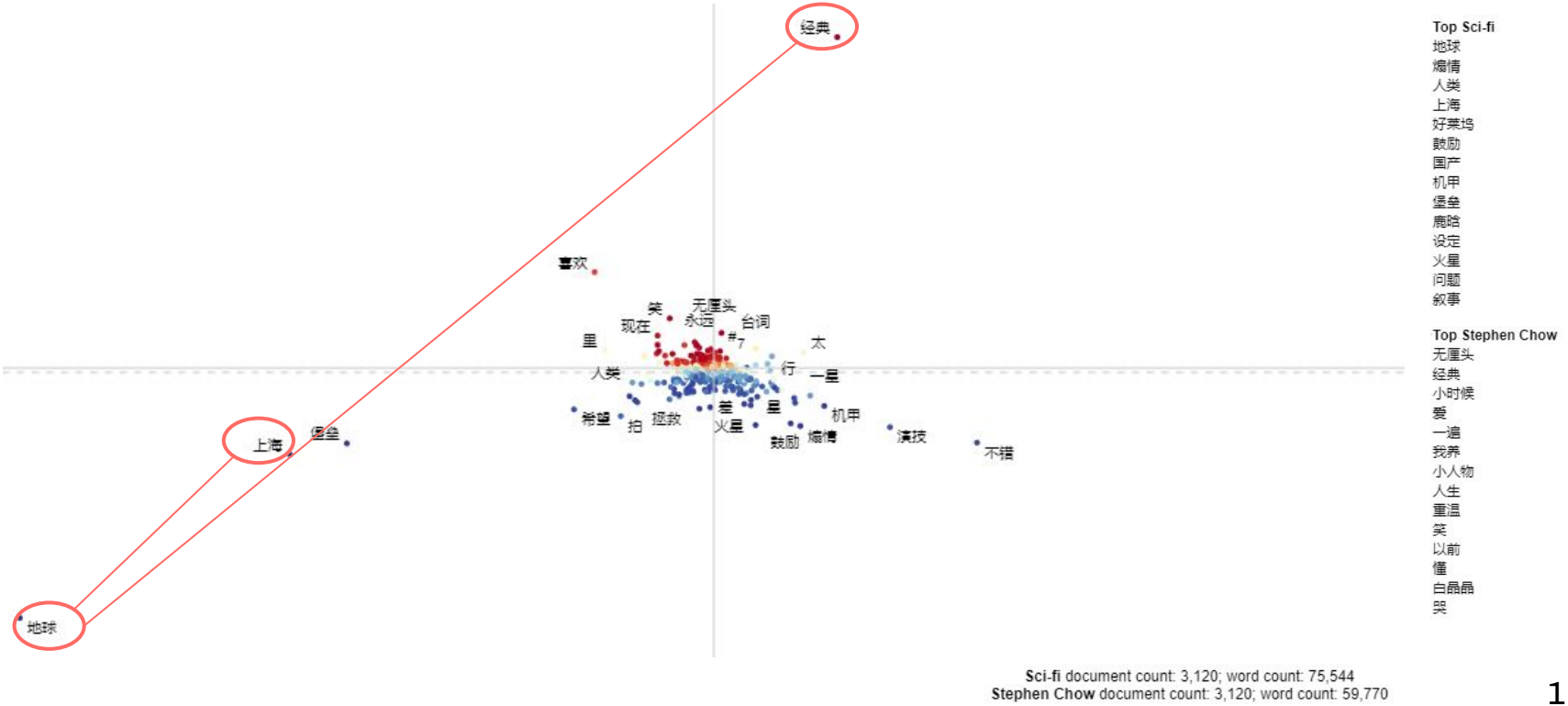
ScatterText

Movie Comments categorised by text sentiment



ScatterText

Movie Comments categorised by text sentiment using TF-IDF transformer



01 PROJECT OBJECTIVE

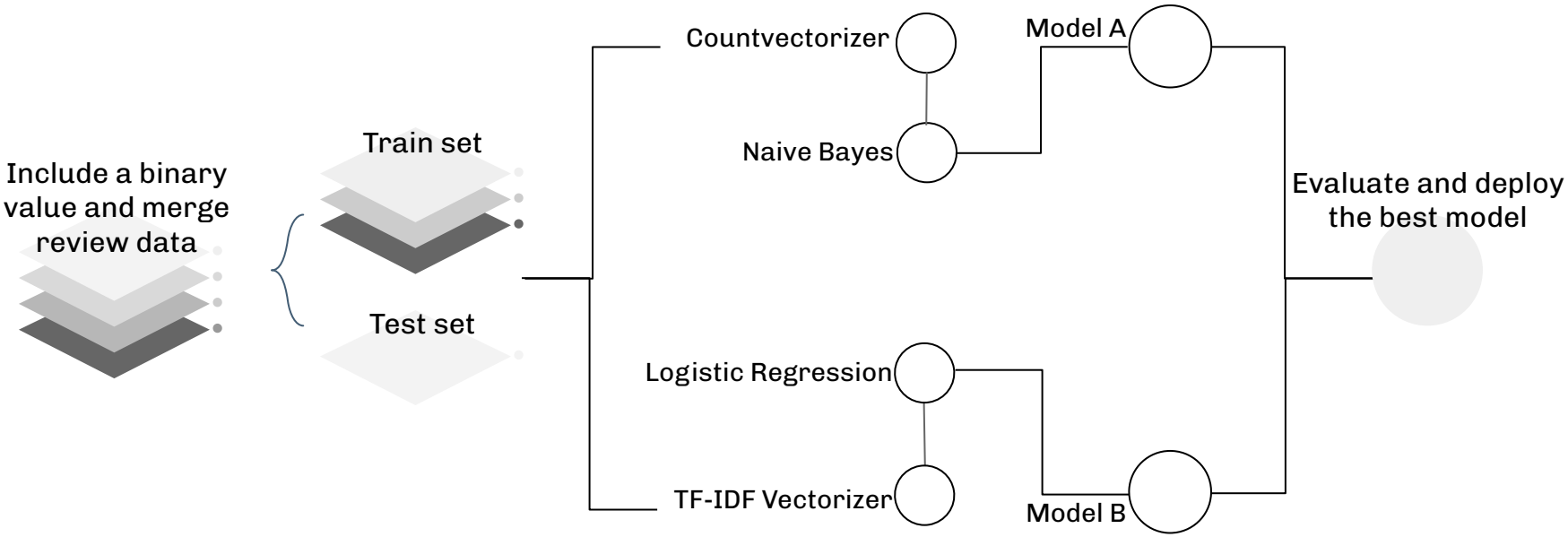
02 DATA COLLECTION

03 DATA CLEANING & EDA

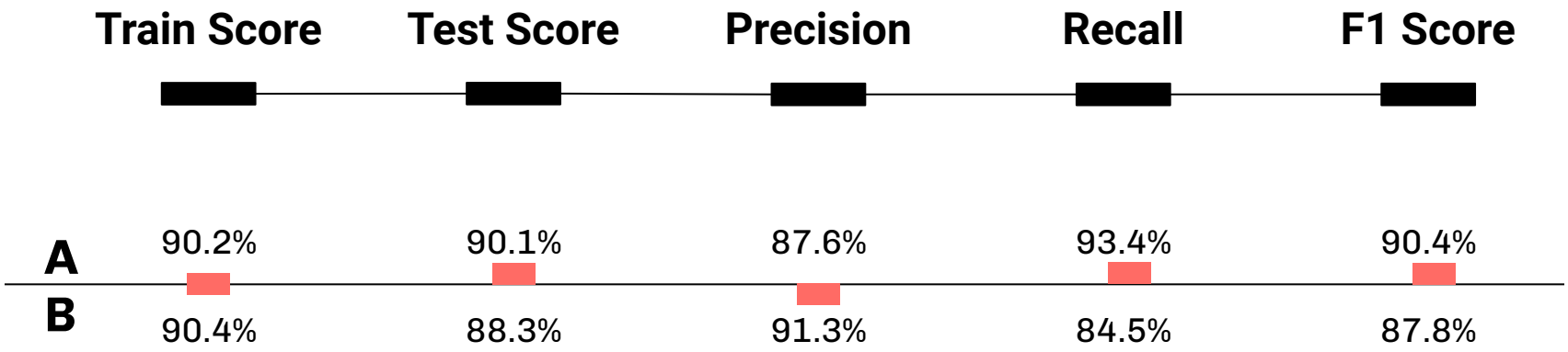
- 04 PREPROCESSING & MODELING

05 CONCLUSION & RECOMMENDATIONS

Framework


Include a binary value and merge review data
Train set
Test set
Countvectorizer
Naive Bayes
Logistic Regression
TF-IDF Vectorizer
Model A
Model B
Evaluate and deploy the best model

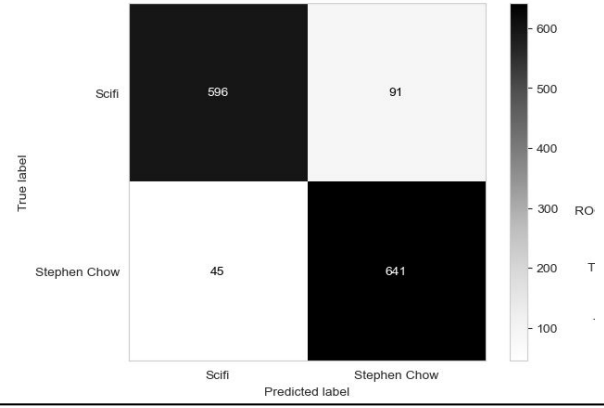
Model Comparison



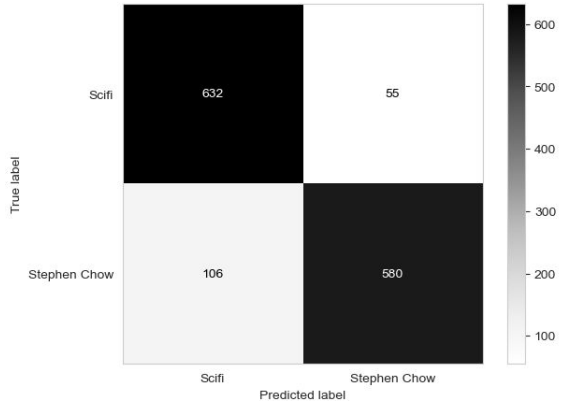
Model A: Countvectorizer + Naive Bayes
Model B: TF-IDF Vectorizer + Logistic Regression

Model Comparison

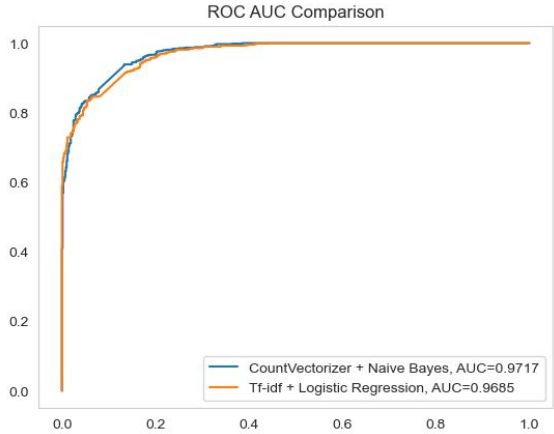
A



B

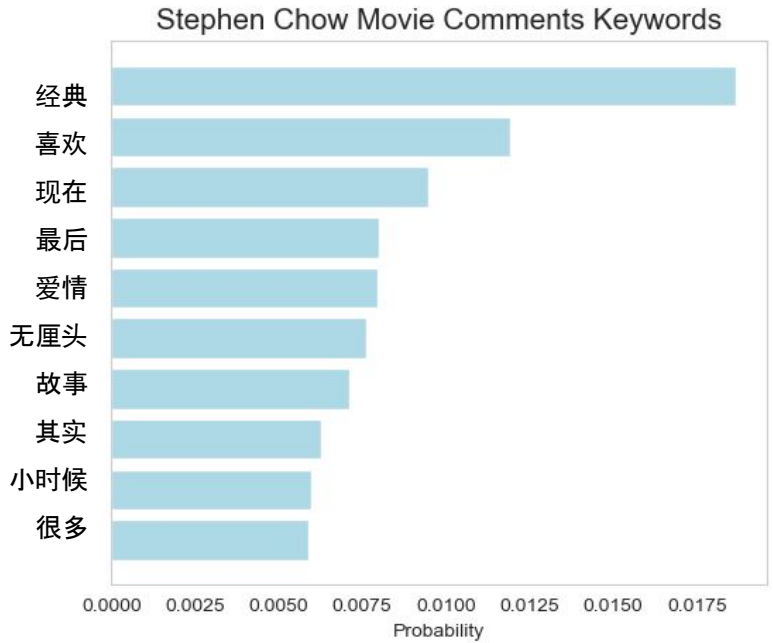


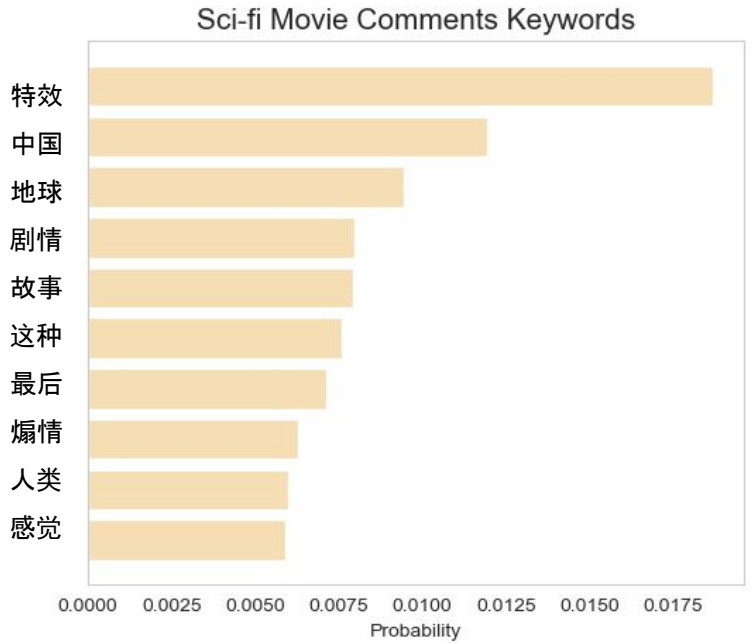
	Precision	Recall
A	87.6%	93.4%
B	91.3%	84.5%



Model A: Countvectorizer + Naive Bayes
Model B: TF-IDF Vectorizer + Logistic Regression

Coefficients





Misclassified comments

Predicted: Stephen Chow
Actual: Sci-fi

'尬到骨髓'

Tokenized
尬 骨髓

Predicted: Sci-fi
Actual: Stephen Chow

'摒除当下“保护主义”盛行所附加的“降低标准”、“一星鼓励”，这就是个三星偏低水准的作品。制作层面在当下中国电影工业而言是很亮眼的，大概达到了好莱坞二十年前的水平。剧本故事完整，三幕结构加双线，很标准的路数，最大的欠缺是人物扁平，没有能让人记住的角色，（演员表现比较糟糕）也让整部片子显得有些闷。导演很粗暴也很潦草，“饱满”的配乐听上去大都是高仿的美国货，鼓点都差不多，当然这都是行业惯例，不过直男煽情还是得多练习一下。最后从科幻层面而言，这部片子“硬科幻”定义的有待商榷。'

Tokenized
摒除 当下 “ 保护主义 ” 盛行 附加 “ 降低标准 ” “ 一星 鼓励 ” 三星 偏低 水准 作品 制作 层面 当下 中国 电影工业 亮眼 大概 达到 好莱坞 二十 年 前 水平 剧本 故事 完整 三幕 结构 加 双线 标准 路数 最大 欠缺 人物 扁平 记住 角色 演员 表现 比较 糟糕 整部 显得 闷 导演 粗暴 潦草 “ 饱满 ” 配乐 听 上去 大都 高仿 美国 货 鼓点 差不多 行业 惯例 直 男 煽情 练习 一下 最后 层面 “ 硬 ” 定义 有待 商榷

Model Limitations

Data	Process	Naive Bayes Model	Text Sentiment
● Data size & quality	● Model selection ● Parameters	● Lack of context awareness ● Bag-of-words feature selection	● Unable to utilise ratings as an additional feature
What can be done?			
● Data collection & cleaning	● Test different model combi ● GridSearch	● Explore Neural Networks	● Food for thought

01 PROJECT OBJECTIVE

02 DATA COLLECTION

03 DATA CLEANING & EDA

04 PREPROCESSING & MODELING

- 05 CONCLUSION & RECOMMENDATION

Conclusion

- **Proof of Concept:**

  While NLP can certainly be used to analyze the linguistic complexity of Chinese, it is **NOT effective in uncovering all of the cultural nuances of the language**. This is because:

- Chinese is a tonal language, which means that a single word can have multiple meanings depending on the tone in which it is spoken.
- Cultural references to literature can also be challenging for NLP systems to comprehend.

Recommendations for Douban

- **The ratings and languages are very different for reviews of Sci-fı and Stephen Chow movies.**
- Douban may consider the use of NLP model for content review or content recommendation system given its capability in comprehending Chinese text data.
- However, it is important to consider the limitations of NLP and to supplement it with cultural context and human interpretation.

THANKS!

Do you have any questions?

REFERENCES

- https://medium.com/@jjsham/nlp-tokenizing-chinese-phases-3302da4336bf
- https://stackoverflow.com/questions/2718196/find-all-chinese-text-in-a-string-using-python-and-regex
- https://medium.com/the-artificial-impostor/preview-developing-modern-chinese-nlp-models-60d4774ebfa7
- https://zhuanlan.zhihu.com/p/27198713
- https://towardsdatascience.com/chinese-natural-language-pre-processing-an-introduction-995d16c2705f
- https://towardsdatascience.com/beginners-guide-to-sentiment-analysis-for-simplified-chinese-using-snownlp-ce88a8407efb
- https://zhuanlan.zhihu.com/p/273399469