主题模型数据分析

Yuhang Ge

June 9, 2021

1 数据预处理

通过读取所有 Json 文件，获取每个 Json 数据的人物简介字段，对人物简介字段进行中文文本分词、去停用词、去数字等预处理。下面是一个样例：

预处理前：

> 周晓文，*1954* 年出生于北京市，中国内地导演、编剧、摄影师、制作人，毕业于北京电影学院摄影系。*1986* 年，执导战争片《他们正年轻》，从而开启了他的导演生涯 *[1]*。*1988* 年，自编自导犯罪片《疯狂的代价》*[2]*，他凭借该片获得夏威夷国际电影节荣誉奖。*1991* 年，凭借执导的剧情片《青春无悔》获得上海大学生电影节最佳导演奖。*1994* 年，执导剧情片《二嫫》，该片获得洛迦诺国际电影节评审团大奖 *[3]*。*1998* 年，执导古装宫廷剧《吕后传奇》*[4]*。*2000* 年，执导古装武侠剧《天龙八部》。*2002* 年，担任古装剧《大脚马皇后》的导演 *[5]*。*2006* 年，执导都市冒险剧《逃亡香格里拉》*[6]*。*2009* 年，执导现代情感剧《晚婚》*[7]*。*2011* 年，担任剧情片《百合》的导演和制作人，该片获得第 *14* 届电影华表奖优秀数字电影奖 *[8]*。

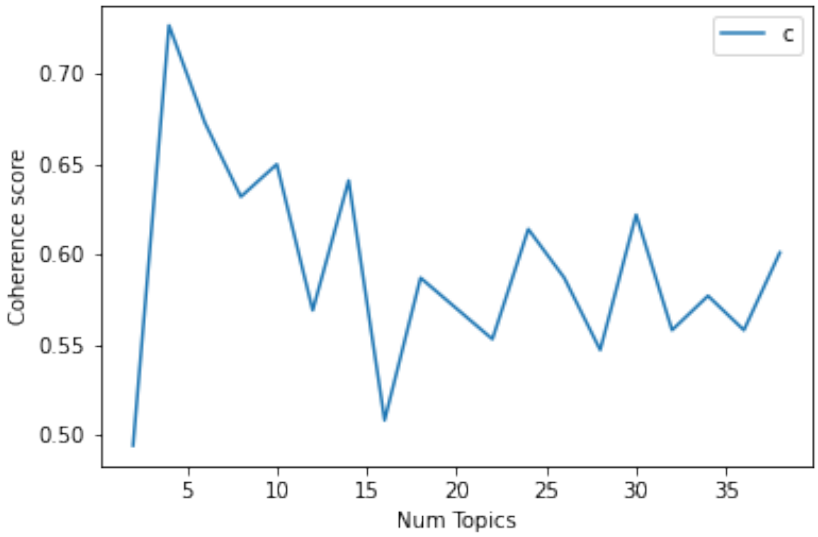
预处理后：

> 周晓文 年出 生于 北京市 中国 内地 导演 编剧 摄影师 制作 毕业 北京电影学院摄影系 执导 战争片 年轻 开启 导演 生涯 自编 自导 犯罪 疯狂 代价 该片 获得 夏威夷 国际 电影节 荣誉奖 执导 剧情片 青春 无悔 获得 上海 大学生 电影节 最佳 导演奖 执导 剧情片 该片 获得 洛迦诺 国际 电影节 评审团 大奖 执导 古装 宫廷 吕后 传奇 执导 古装 武侠剧 天龙八部 担任 古装剧 大脚 马皇后 导演 执导 都市 冒险 逃亡 香格里拉 执导 现代 情感 晚婚 担任 剧情片 百合 导演 制作 该片 获得 电影 华表奖 优秀 数字 电影

2 LDA 主题模型分析

2.1 最优主题数寻找

通过分别计算区间 $K = \{2 - 40\}$ 内主题的主题一致性（coherence value），绘制出下图，可以找到最优主题数为 $k = 6$。



将处理后的数据通过 LDA 主题模型进行数据分析，设置最优主题数为 6，并且每个主题的主题词数为 20，得到主题分布如下：

(0, ’0.025*” 主演” + 0.018*” 出演” + 0.016*” 电视剧” + 0.015*” 饰演” + 0.012*” 电影” + ’ ’0.010*” 获得” + 0.010*” 中国” + 0.009*” 生于” + 0.008*” 同年” + 0.007*” 参演” + 0.006*” 古装” ’ ’+ 0.006*” 毕业” + 0.006*” 日出” + 0.006*” 都市” + 0.005*” 最佳” + 0.005*” 参加” + ’ ’0.005*” 个人” + 0.005*” 内地” + 0.005*” 情感” + 0.005*” 爱情”’),

(1, ’0.027*” 主演” + 0.026*” 电影” + 0.016*” 获得” + 0.015*” 中国” + 0.011*” 最佳” + ’ ’0.009*” 电视剧” + 0.008*” 生于” + 0.007*” 执导” + 0.007*” 出演” + 0.007*” 饰演” + ’ ’0.006*” 同年” + 0.006*” 电影节” + 0.005*” 参演” + 0.005*” 导演” + 0.005*” 爱情” + ’ ’0.005*” 担任” + 0.005*” 上映” + 0.005*” 日出” + 0.005*” 毕业” + 0.004*” 公主”’),

(2, ’0.015*” 中国” + 0.012*” 获得” + 0.011*” 专辑” + 0.008*” 发行” + 0.007*” 电视剧” + ’ ’0.007*” 生于” + 0.006*” 最佳” + 0.006*” 主演” + 0.006*” 电影” + 0.006*” 音乐” + 0.005*” 同年” ’ ’+ 0.005*” 演员” + 0.005*” 公元前” + 0.005*” 参加” + 0.005*” 日出” + 0.004*” 出演” + ’ ’0.004*” 担任” + 0.004*” 成为” + 0.004*” 推出” + 0.003*” 个人”’),

(3, ’0.011*” 中国” + 0.010*” 获得” + 0.008*” 主演” + 0.006*” 生于” + 0.005*” 皇后” + 0.005*” 电影” ’ ’+ 0.005*” 最佳” + 0.005*” 电视剧” + 0.004*” 韩国” + 0.004*” 出演” + 0.004*” 担任” + ’ ’0.004*” 皇帝” + 0.004*” 元年” + 0.003*” 导演” + 0.003*” 时期”

+ 0.003*” 日本” + 0.003*” 爱情” ’ ’+ 0.003*” 大赏” + 0.003*” 日出” + 0.003*” 女演员”’),

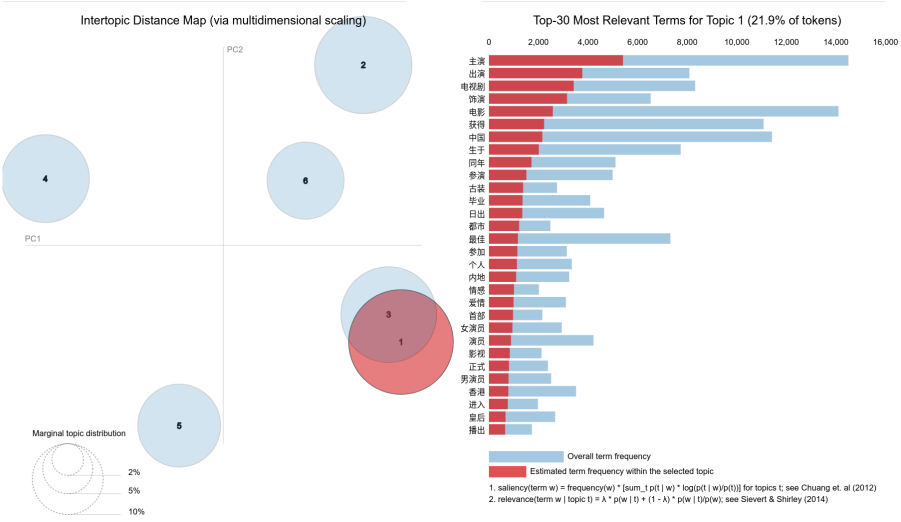
(4, ’0.022*” 电影” + 0.010*” 出演” + 0.009*” 参演” + 0.009*” 饰演” + 0.008*” 生于” + 0.007*” 中国” ’ ’+ 0.007*” 主演” + 0.007*” 获得” + 0.007*” 最佳” + 0.007*” 香港” + 0.006*” 电视剧” + ’ ’0.005*” 演员” + 0.005*” 执导” + 0.004*” 日出” + 0.004*” 时期” + 0.004*” 成为” + 0.003*” 导演” ’ ’+ 0.003*” 谥号” + 0.003*” 毕业” + 0.003*” 大臣”’),

(5, ’0.017*” 电影” + 0.012*” 获得” + 0.011*” 最佳” + 0.010*” 中国” + 0.009*” 主演” + 0.009*” 生于” ’ ’+ 0.006*” 香港” + 0.006*” 演员” + 0.005*” 同年” + 0.005*” 参演” + 0.005*” 个人” + ’ ’0.004*” 台湾” + 0.004*” 日出” + 0.004*” 电视剧” + 0.004*” 毕业” + 0.004*” 成为” + ’ ’0.004*” 专辑” + 0.004*” 音乐” + 0.003*” 出演” + 0.003*” 男演员”’)

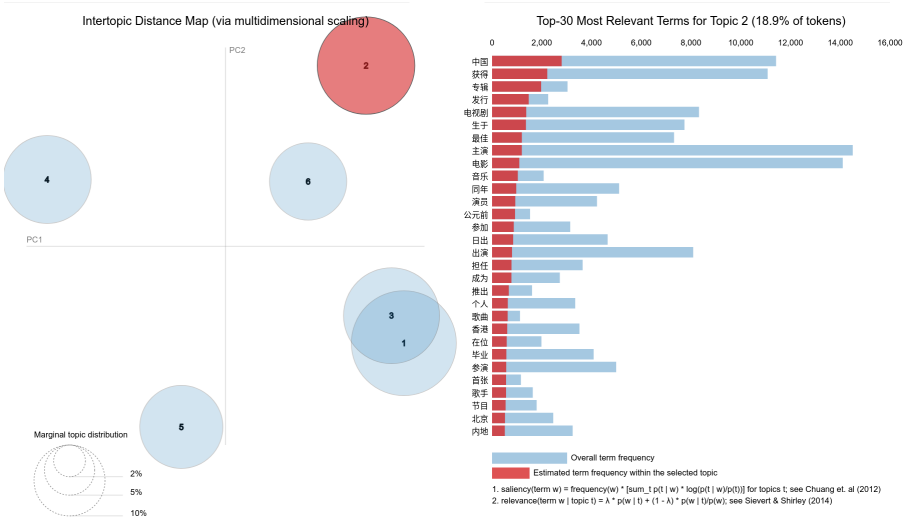
3 主题可视化

下面是分别 6 个主题分布的可视化图:

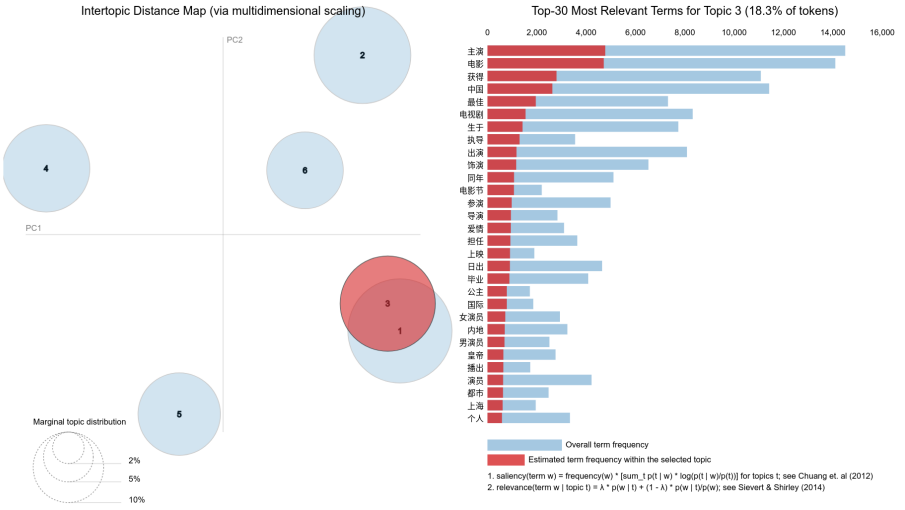
主题 1:



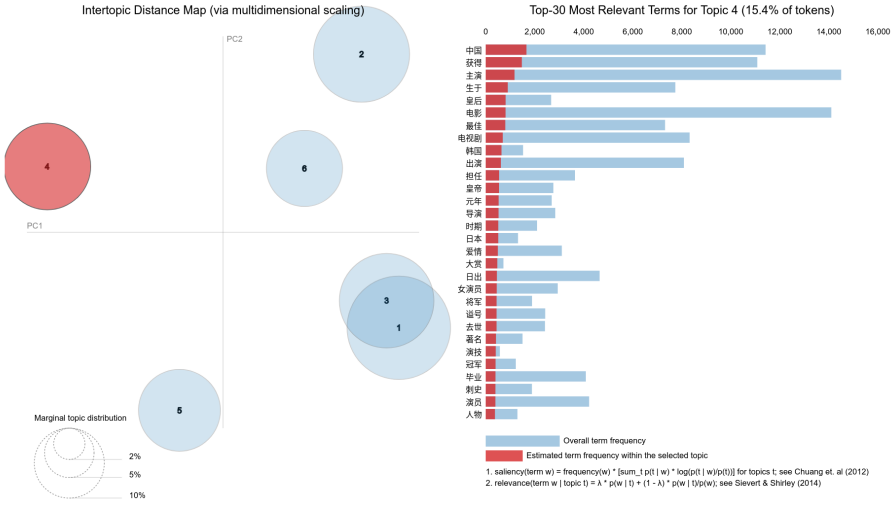
主题 2:


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
2
4
6
3
1
5
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 2 (18.9% of tokens)
0
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
中国
获得
专辑
发行
电视剧
生于
最佳
主演
电影
音乐
同年
演员
公元前
参加
日出
出演
担任
成为
推出
个人
歌曲
香港
在位
毕业
参演
首张
歌手
节目
北京
内地
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

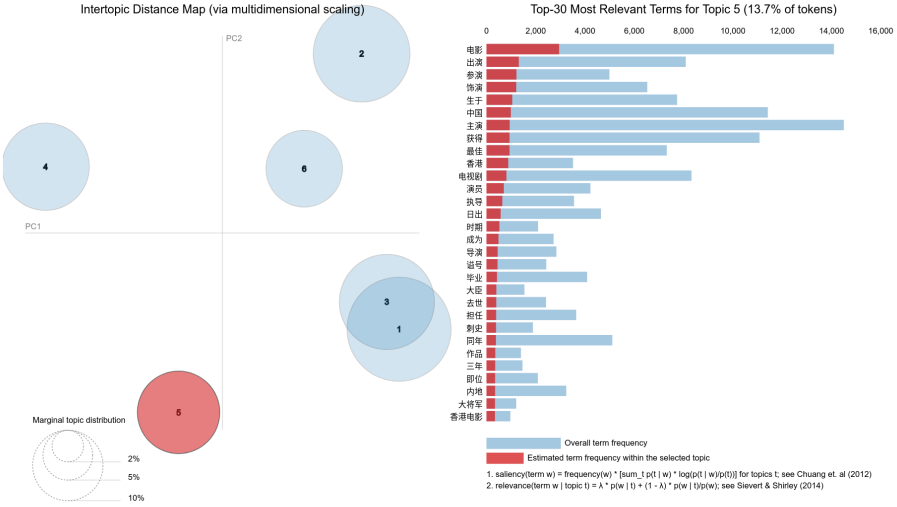
主题 3:


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
2
4
6
3
1
5
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 3 (18.3% of tokens)
0
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
主演
电影
获得
中国
最佳
电视剧
生于
执导
出演
饰演
同年
电影节
参演
导演
获得
担任
上映
日出
毕业
公主
国际
女演员
内地
男演员
皇帝
播出
演员
都市
上海
个人
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

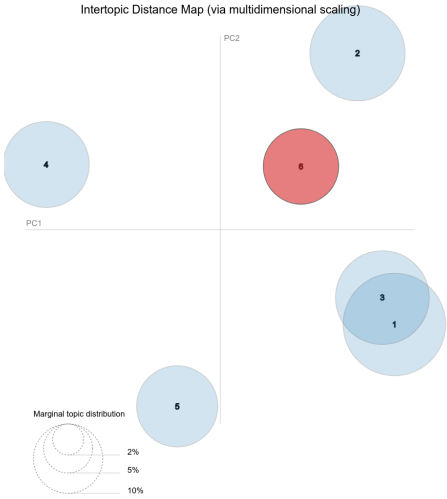
主题 4:

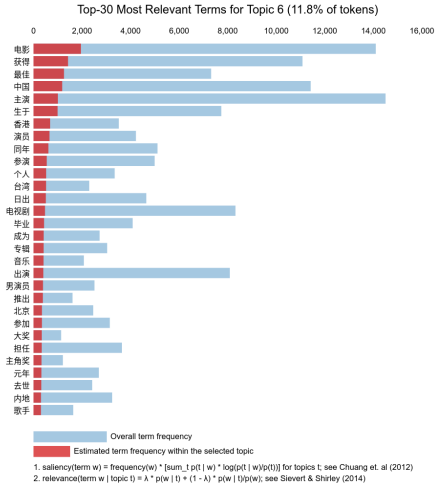

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
2
4
6
3
1
5
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 4 (15.4% of tokens)
0
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
中国
获得
主演
生于
皇后
电影
最佳
电视剧
韩国
出演
担任
皇帝
元年
导演
时期
日本
爱情
大赏
日出
女演员
将军
谥号
去世
著名
演技
冠军
毕业
制史
演员
人物
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

主题 5:


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
2
4
6
3
1
5
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 5 (13.7% of tokens)
0
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
电影
出演
参演
饰演
生于
中国
主演
获得
最佳
香港
电视剧
演员
执导
日出
时期
成为
导演
谥号
毕业
大臣
去世
担任
刻史
同年
作品
三年
即位
内地
大将军
香港电影
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

主题 6:


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
2
4
6
3
1
5
Marginal topic distribution
2%
5%
10%


Top-30 Most Relevant Terms for Topic 6 (11.8% of tokens)
0
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
电影
获得
最佳
中国
主演
生于
香港
演员
同年
参演
个人
台湾
日出
电视剧
毕业
成为
专辑
音乐
出演
男演员
推出
北京
参加
大奖
担任
主角奖
元年
去世
内地
歌手
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)