2017-7-9

应用集成第三次作业报告

基于语义分析的电影分类

编写人：141250131 汪文藻

一. 分工

模型训练，电影分类、结果展示应用程序编写	汪文藻 141250131
文本预处理	孙浩 141250115

二. 目的

淘票票、时光网、美团电影平台的电影分类具有不一致、模糊、粗略的现象，对于同一部电影标签分类具有同义不同词或者不同义的现象。本次作业目的拟通过对电影简介的语义分析，对电影重新分类打标签。

三. 分类过程

本系统基于语义分析对电影打标签

文本：

爬取了豆瓣 5 月档期的 21 部电影的简介和影评，一共 1430 篇文章，约 653500 字。

预处理：

需要对每篇文章进行分词和过滤，去除无实际意义的词如“阿、呀、是、的”等。我们使用 thulac 对文本进行分词和过滤，thulac 作为分词工具具有正确率高，可以判断词语词性，我们利用这个特点过滤了语气词助词之类的虚词。

训练：

本语义分析使用 LDA 算法，是一种基于词频的算法，使用 python 的 Lda 数学库实现了该算法，间 buildmodel.py。根据淘票票原本的电影分类，我通过归纳总结预先确定了电影分为 9 类。将 1430 篇经过预处理的文本作文训练数据进行无监督的训练得到分类模型，topic.txt 为类别模型。

编号	定义类别	关键词（取前 30 个）
0	喜剧	闹剧 喜剧 演绎 笑料 拔节 得到 超然 导致 仅 拥有 有幸 出 领导 很 来到 大勺 家庭 飞船 椿 角色 可 新 贴 去 可以 轮盘 趣味 爆笑
1	纪实、自然、情感	自然 细节 动物 纪录 系列 跳 真人 收尾 是 开始 特效 关系 很 也 诟病 成名 感人 刻意 到 深度 选择 女性 也 无意间 真 略 一直 仅
2	校园、青春	头发 课堂 请 女生 过程 清纯 初恋 作品 得到 是 号称 名字 奶奶灰 故事 阵容 明确 会 表白 话题 熟悉 学校 结束 细节 学会 注定 稳稳当当 影片 水平 成功 毒舌

3	动作、惊险	飞车 犯罪 寻找 频道 人 平静 名字 看 开始 一样 角色 突然 人 都 反转 不 嚼 普通人 重拾 观众 着实 精明 不 是 是 看 导演 就 车祸 最佳
4	文艺	Topic 4: 流逝 咖喱 生活 美丽 难道 故事 爱情 可能 学会 就 去 情趣 也 契约 将 发现 名字 绝不 演绎 开启 精明 举行 梦想 椿 或许 来到 平凡 打造 深度 向上
5	冒险	Topic 5: 冒险 发现 好奇 背叛 仅 手持式 催泪弹 意外 生物 是 得到 救 拥有 寻找 水 也 选择 都 咖喱 不 研究 也 变成 事情 泾流 水平 话语 梦 兄长 过瘾
6	爱情	Topic 6: 故事 必须 伤害 习惯性 声音 到 剧情 爱 女性 清楚 戏骨 单身 为 是 个性 巨大 飞船 伤害 约 进 会 是 让 朋友 终止 角色 生活 小 为 留
7	科幻、怪兽	Topic 7: 异形 救 外星 画面 学会 屠杀 发现 苏醒 催泪弹 比如 变成 主持人 人 去 是 偏见 联想 也 不 送 结婚 直接 过程 难 名字 很 人设 粉 赞同 导致
8	惊悚、悬疑	Topic 8: 恐惧 探险 失踪 亡灵 心理 闹鬼 亮 意义 领导 恐惧 深度 有幸 鬼刷 不 成名 超然 是 也 海底 一样 大志 辗转 可是 不 故事 也 是 网

电影类别判断：

使用已经建立的类别模型，根据电影的简介，对电影进行分类，结果存于 doc-topic.txt

中，映射结果如下

电影	类别	对应 topic 结果
麻辣学院	喜剧	0
重返狼群	纪实、情感、自然	1
速度与激情 8	动作、惊险	3
迷失 Z 城	冒险	5
诡异酒楼	悬疑、惊悚	8
荡寇风云	动作、惊险	3
美好的意外	文艺	4
神奇女侠	科幻、怪兽	7
神农溪之恋	爱情	6
加勒比海盗 5：死无对证	悬疑、惊险	8
内心引力	纪实、情感、自然	1
哆啦 A 梦：大雄的南极冰冰凉大冒险	动作	3
异星觉醒	科幻、怪兽	7
异形：契约	科幻、怪兽	7
异兽来袭	科幻、怪兽	7
新木乃伊	科幻、怪兽	7
摔跤吧!爸爸	喜剧	0
借眼	悬疑、惊悚	0
梦幻佳期	爱情	6
此情此刻	文艺	4
碟仙之毕业照	校园、青春	2

四． 结果对比分析

与淘票票原分类的对比：

电影	本系统	淘票票
麻辣学院	喜剧	搞笑、爆笑、喜剧
重返狼群	纪实、情感、自然	情感
速度与激情 8	动作、惊险	动作
迷失 Z 城	冒险	恐怖、惊悚
诡异酒楼	悬疑、惊悚	惊悚、恐怖
荡寇风云	动作、惊险	历史、战争
美好的意外	文艺	情感、明星
神奇女侠	科幻、怪兽	奇幻、动作
神农溪之恋	爱情	爱情
加勒比海盗 5：死无对证	悬疑、惊险	奇幻、冒险
内心引力	纪实、情感、自然	纪实
哆啦 A 梦：大雄的南极冰冰凉大冒险	动作	动画、冒险

异星觉醒	科幻、怪兽	科幻
异形：契约	科幻、怪兽	科幻
异兽来袭	科幻、怪兽	科幻
新木乃伊	科幻、怪兽	科幻、汤姆克鲁斯
摔跤吧!爸爸	喜剧	喜剧
借眼	悬疑、惊悚	恐怖
梦幻佳期	爱情	爱情
此情此刻	文艺	文艺
碟仙之毕业照	校园、青春	恐怖、惊悚

本次基于语义的分类，基本上是准确的，除了《碟仙之毕业照》被分错了，这是因为这部电影介绍中充满了学生，毕业，同学，朋友，年轻之类常出现在校园电影里的词。

此外，本次分类符合类别需要抽象的概念，比如在《新木乃伊》这种科幻、怪兽篇种，去掉了“汤姆克鲁斯”这种不是电影分类的关键词。却掉了分类中冗余、语义重复的词，如“恐怖、惊悚”被我改为了“悬疑、惊悚”，“搞笑、爆笑、喜剧”精炼为“喜剧”。修正淘票票不准确的分类，如《美好的意外》是一部爱情片，可是淘票票将它分的过于笼统，情感包含很多，不一定专指爱情，明星这一类别显然不能为观众所理解。本系统对于类别的名称与时俱进，近 10 年，随着《生化危机》、《撕裂人》等电影的流行，科幻片与怪兽片正在日益结合，所以我认为，把这两种词放在一起，可以满足大部分要求。而且在 topic7 中既出现了“外星”、“异形”这种代表科幻的词，又出现了如“屠杀”代表怪兽灾难片的词，所以我将这两种词放在一起代表一类电影，从结果来看，对于异星觉醒、异形：契约、异兽来袭是贴切的。