游戏配置获取算法流程

1. 爬虫流程：
   1. 从京东网站爬取电脑的url链接以及商品的属性
   2. 根据爬取到的url链接，爬取每件商品的评论，及评论对应的型号信息（注：一件商品可能会包含多个型号），每件商品对应的评论不超过1000条。总计4946件商品，1757020条评论，
2. 文本预处理
   1. 筛选出商品名称中包含“游戏" 的商品对应的评论
   2. 去除长度不超过20字的评论
   3. 随机选择75000条评论
   4. 利用jieba分词
   5. 去停用词
   6. 利用将评论转化为bow表示（sklearn，countvectorizer，参数默认）
3. LDA对评论聚类
   1. 训练LDA

lda = LatentDirichletAllocation(n\_topics=8,

max\_iter=1000,

learning\_method='batch',

verbose=1,

random\_state=2018)

1. 输出每个主题中概率最高的前10个词，人工分析主题

# Topics in LDA model:

# Topic #0屏幕 散热: 有点 声音 屏幕 散热 硬盘 风扇 键盘 不错 感觉 固态

# Topic #1游戏: 游戏 办公 不错 流畅 电脑 玩游戏 不卡 运行 配置 一点

# Topic #2物流: 鼠标 快递 包装 键盘 物流 京东 没有 收到 很快 东西

# Topic #3客服: 客服 电脑 耐心 问题 态度 非常 真的 服务 特别 解决

# Topic #4品牌: 主机 笔记本 联想 机箱 品牌 戴尔 小米 显示器 华硕 支持

# Topic #5开机/运行速度: 非常 不错 速度 电脑 很快 值得 购买 满意 质量 物流

# Topic #6外观/速度: 不错 电脑 速度 开机 喜欢 很快 外观 感觉 好看 收到

# Topic #7电脑问题: 电脑 问题 系统 没有 知道 开机 软件 客服 现在 售后

1. 计算每条评论的主题分布，将概率最大的主题设置为该评论的主题
2. 利用BDC找出游戏主题的关键字
   1. 将上述带主题的75000条评论作为经过分词、去停用词后作为BDC的输入，得到每个词在每个主题上出现的次数以及对应的BDC值
   2. 一个词在哪个主题下出现的次数最多，则将该词设置为该主题
   3. 选择属于游戏主题的词，按BDC排序，得到的游戏主题词如下:

计算机生成了可选文字:
wordtopic
Tokens
守望弓皂钅呈
不在话下
不元
打打
小游戏
刺灕战汤
英雄关盟
18
3
5
39
14
69
33
103
39
28
79
62
30
105
20
411
39
5
38
23
4
13
45
7
23
34
115
什
54
142
59
50
39
39
144
96
54
176
39
706
66
BDC
05881
0560D
0、5278
04711
04483
04457
0、4380
04322
0、4322
04181
0、4110
04081
0、4046

可以看到守望先锋等游戏名词排名很靠前。

1. 找到评论中包含游戏主题词的商品，并统计这些商品的配置信息，得到对应游戏主题词的配置信息。
   1. 筛选出主题为游戏的评论
   2. 筛选出包含游戏主题词的评论
   3. 统计这些评论对应的商品的配置信息（CPU，内存，显卡），现在评论的商品型号中找，如果没找到再在商品属性中找。选择出现次数最多的配置作为游戏主题词最终的配置。
   4. 人工筛选主题词，分别找到属于游戏和属于办公的主题词，得到游戏配置列表和办公配置列表。