

项目文档说明介绍

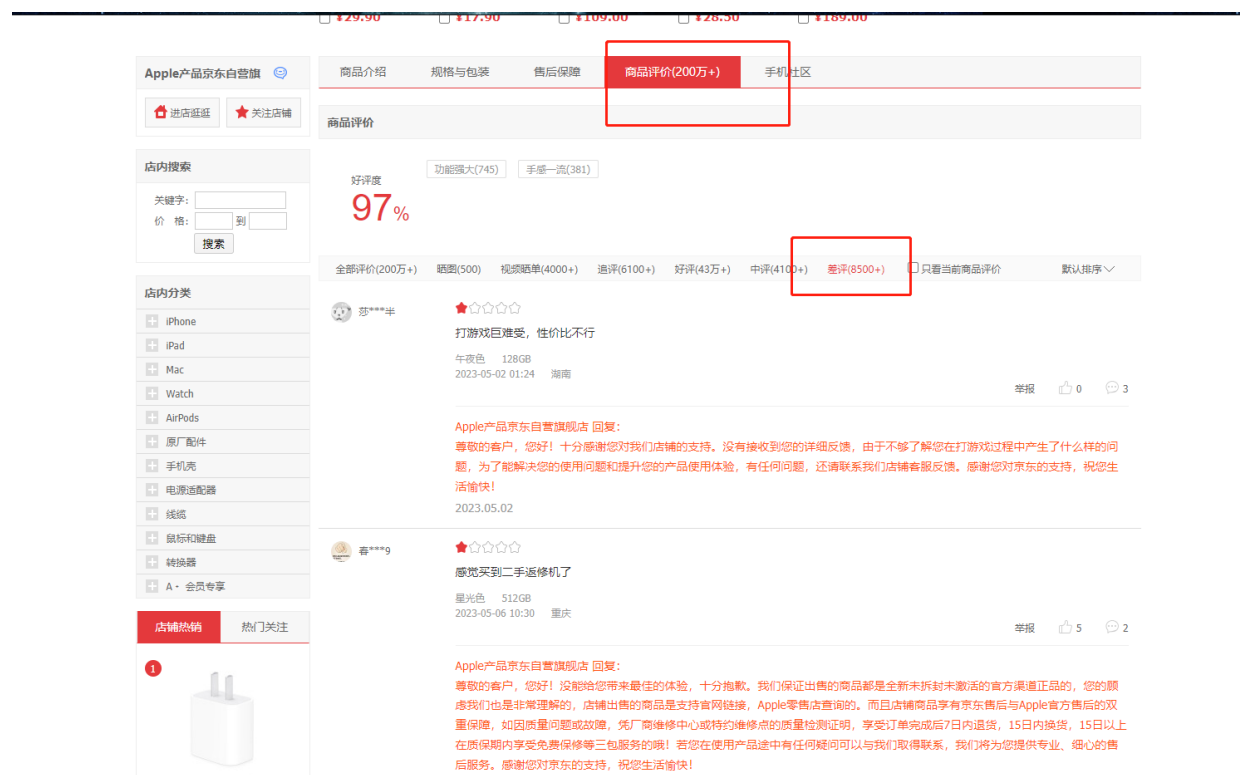
该项目一共分为6步：

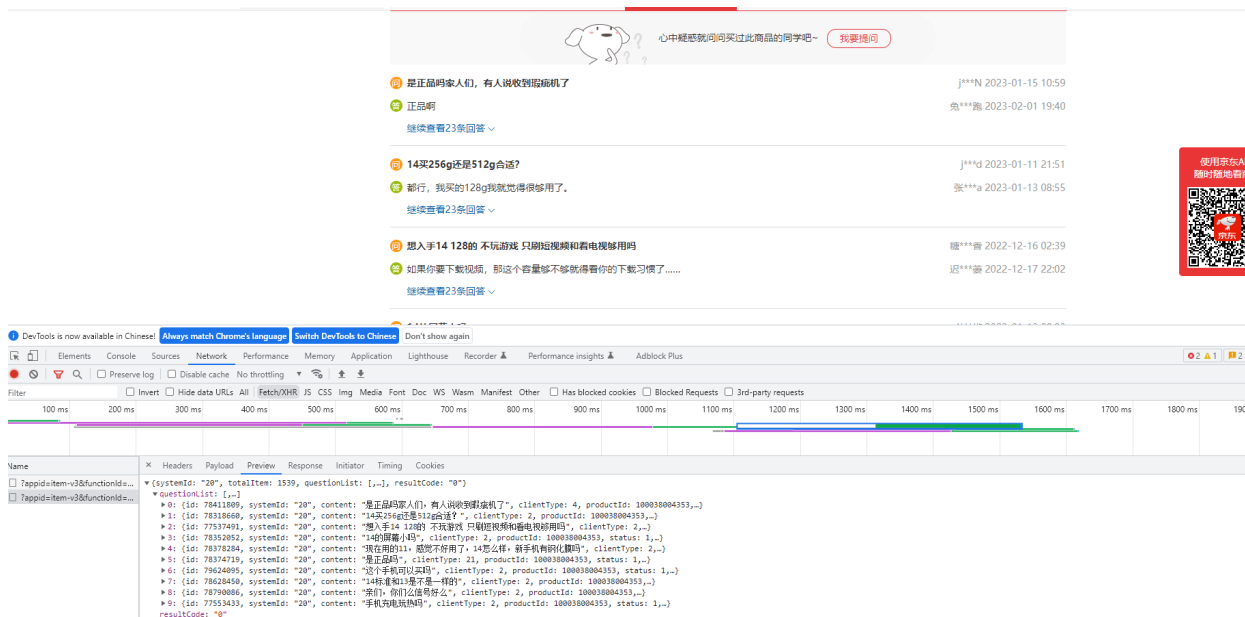
1. 数据获取，获取是获取京东商店里面的评论信息，获取评论里面的差评信息
2. 数据预处理，这个步骤是把一些无效数据给去除掉，为后面的分析做铺垫
3. 词性分析，该步骤是处理名词，形容词，以及TF-IDF算法的处理，其目的是为了对文本进行深度挖掘探讨，了解文本的整体方向，并且绘制出相应的图和词云
4. LDA主题建模，该主题建模主要是用于识别大规模文档集或语料库汇总的隐藏的主题信息，以概率分布的形式给出来，从而通过分析一些文档抽取它们的主题分布后，便可以根据主题分布进行主题或文本分类
5. 网络语义分析，网络语义的作用是实现信息共享，知识推理，数据融合等功能，从而便于进行语义分析、主题提取、信息检索等任务处理
6. 逻辑回归，向量机分类任务，通过这个两大机器学习的模型，我们可以进行一些对应的分类任务，从而在应对庞大的数据集的时候，可以通过该模型去进行分类任务，从而减速我们对数据进行分类的时间

数据获取

数据获取来源于京东：<https://global.jd.com/>

其关键词：生鲜，数码，家电







通过读取对应的JSON文件来获取相应的数据，然后通过Python中的requests的get请求来获取对应的文本内容，接着根据pandas中来进行数据保存

其主要的代码是这一块

```
def get_json(html,item):
    content = html.json()
    try:
        content = content['comments']
        df = pd.DataFrame()
        for c in content:
            df['id'] = [c['id']]
            df['anonymousFlag'] = [c['anonymousFlag']]
            df['content'] = [c['content']]
            df['creationTime'] = [c['creationTime']]
            df['class'] = [item]
            df.to_csv('data.csv', encoding='utf-8-sig', mode='a+', index=False,
            header=False)
            time.sleep(0.2)
        except:
            pass
```

在数据获取完之后，保存为data文件，其主要代码和文件为下面这两个

 crawl.py	2023/5/8 9:52	JetBrains PyChar...	6 KB
 data.csv	2023/5/8 9:52	Microsoft Excel ...	913 KB

[illegible]

5928	1.88E+10	1 这降价速度和降价价	2023/2/11 23:17	数码				
5929	1.88E+10	1 体验非常差买了20	2023/2/11 20:49	数码				
5930	1.88E+10	1 京东服务超差, 刚	2023/2/12 12:02	数码				
5931	1.83E+10	1 垃圾东西, 第一天就	2022/10/27 19:30	数码				
5932	1.91E+10	1 手机背面有瑕疵, 中	2023/5/2 15:04	数码				
5933	1.88E+10	1 辣鸡京东, 地区歧视	2023/2/16 6:36	数码				
5934	1.88E+10	1 电池非常耐用, 上班	2023/2/23 13:29	数码				
5935	1.87E+10	1 东西不支持七天无	2023/1/6 16:48	数码				
5936	1.87E+10	1 太重, 那镜头太大、	2023/1/17 7:35	数码				
5937	1.88E+10	1 手机信号不好, 很	2023/2/16 1:04	数码				
5938	1.88E+10	1 两天一个价两天一	2023/2/6 15:30	数码				
5939	1.88E+10	1 刚过就降价, 真是	2023/2/5 1:36	数码				
5940	1.88E+10	1 花的全新未激活的	2023/2/9 11:55	数码				
5941								

- 标点符号过滤：去除文本中的一些不影响语义的标点符号，如逗号、句号等。
- 停用词过滤：去除文本中的一些没有实际意义的词，如“的”、“了”、“呢”等。
- 中文识别：判断文本是否为中文，可以使用正则表达式或者编码方式来实现。
- 机械压缩：去除文本中的一些重复出现的词，如“进行”、“处理”等。
- 分词：将文本中的连续字符切分成有意义的词，可以使用jieba分词等工具来实现。
- 词语过滤：去除文本中的一些单个字或者无意义的词，只保留两个词以上的内容。

如下是主要的文件和源码:

词频分析一共分为三步：

3. 词性形容词，通过处理形容词，我们可以大体知道对应不同分类中，情感的趋势是什么，更方便让我们了解卖家对于物流的一个整体的态度

对应的文档如下：

名称	修改日期	类型	大小
LDA主题	2023/5/8 9:52	文件夹	
词性数据	2023/5/8 9:52	文件夹	
网络语义	2023/5/8 9:52	文件夹	
crawl.py	2023/5/8 9:52	JetBrains PyChar...	6 KB
data.csv	2023/5/8 9:52	Microsoft Excel ...	913 KB
lda主题建模.py	2023/5/8 9:52	JetBrains PyChar...	6 KB
new_data.csv	2023/5/8 9:52	Microsoft Excel ...	468 KB
score.csv	2023/5/8 9:52	Microsoft Excel ...	1 KB
stopwords-cn.txt	2023/5/8 9:52	文本文档	15 KB
词性处理.py	2023/5/8 9:52	JetBrains PyChar...	5 KB
机器学习分类.py	2023/5/8 9:52	JetBrains PyChar...	4 KB
数据预处理.py	2023/5/8 9:52	JetBrains PyChar...	2 KB
网络语义.py	2023/5/8 9:52	JetBrains PyChar...	6 KB

主题分析

潜在狄利克雷分配，即LDA模型（Latent Dirichlet Allocation，LDA）是由Blei等人在2003年提出的生成式主题模型[⑩ Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3:2003.]⑩。生成模型，即认为每一篇文档的每一个词都是通过“一定的概率选择了某个主题，并从这个主题中以一定的概率选择了某个词语”。LDA模型也被称为三层贝叶斯概率模型，包含文档（d）、主题（z）、词（w）三层结构，能够有效对文本进行建模，和传统的空间向量模型（VSM）相比，增加了概率的信息。通过LDA主题模型，能够挖掘数据集中的潜在主题，进而分析数据集的集中关注点及其相关特征词。

LDA模型采用词袋模型（Bag Of Words，BOW）将每一篇文档视为一个词频向量，从而将文本信息转化为易于建模的数字信息。

定义词表大小为L，一个L维向量(1,0,0,...,0,0)表示一个词。由N个词构成的评论记为。假设某一商品的评论集D由M篇评论构成，记为。M篇评论分布着K个主题，记为。记a和b为狄利克雷函数的先验参数，q为主题在文档中的多项分布的参数，其服从超参数为a的Dirichlet先验分布，f为词在主题中的多项分布的参数，其服从超参数b的Dirichlet先验分布。LDA模型图如图6所示

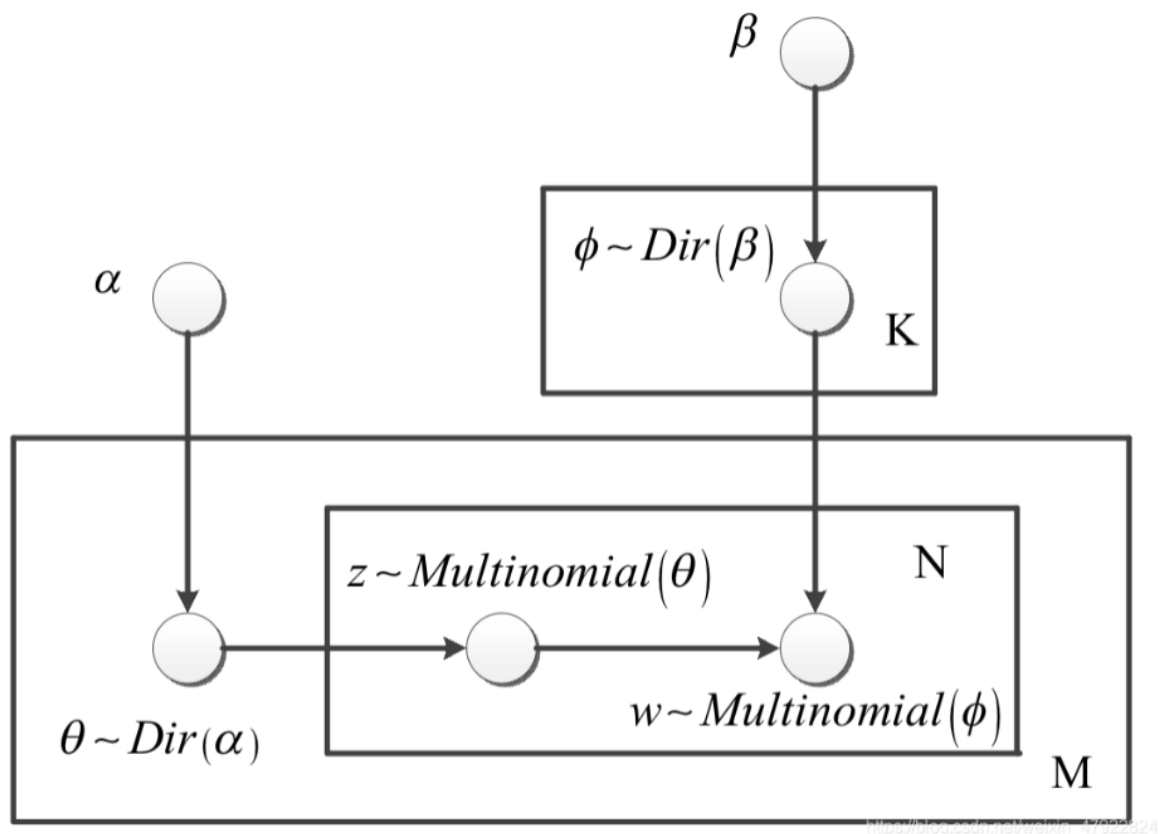


图6 LDA模型结构示意图

LDA模型假定每篇评论由各个主题按一定比例随机混合而成，混合比例服从多项分布，记为式（1）。

$$z \mid \theta = \text{Multinomial}(\theta) \quad (1)$$

而每个主题由词汇表中的各个词语按一定比例混合而成，混合比例也服从多项分布，记为式（2）。

$$w \mid z, \phi = \text{Multinomial}(\phi) \quad (2)$$

在评论 d_j 条件下生成词 w_i 的概率表示为式（3）。

$$P(w_j \mid d_j) = \sum_{s=1}^K P(w_i \mid z = s) \times P(z = s \mid d_j) \quad (3)$$

其中， $P(w_i \mid z = s)$ 表示词 w_i 属于第 s 个主题的概率，表示第 s 个主题在评论 d_j 中的概率。

DA主题模型是一种无监督的模式，只需要提供训练文档，它就可以自动训练出各种概率，无需任何人工标注过程，节省大量人力及时间。它在文本聚类、主题分析、相似度计算等方面都有广泛的应用，相对于其他主题模型，其引入了狄利克雷先验知识。因此，模型的泛化能力较强，不易出现过拟合现象。

LDA主题模型可以解决多种指代问题，例如：在热水器的评论中，根据分词的一般规则，经过分词的语句会将“费用”一词单独分割出来，而“费用”是指安装费用，还是热水器费用等其他情况，如果简单的进行词频统计及情感分析，是无法识别的，这种指代不明的问题不能购准确的反应用户情况，运用LDA主题模型，可以求得词汇在主题中的概率分布，进而判断“费用”一词属于哪个主题，并求得属于这一主题的概率和同一主题下的其他特征词，从而解决多种指代问题。

主题参考文献：<https://zhuanlan.zhihu.com/p/75222819>

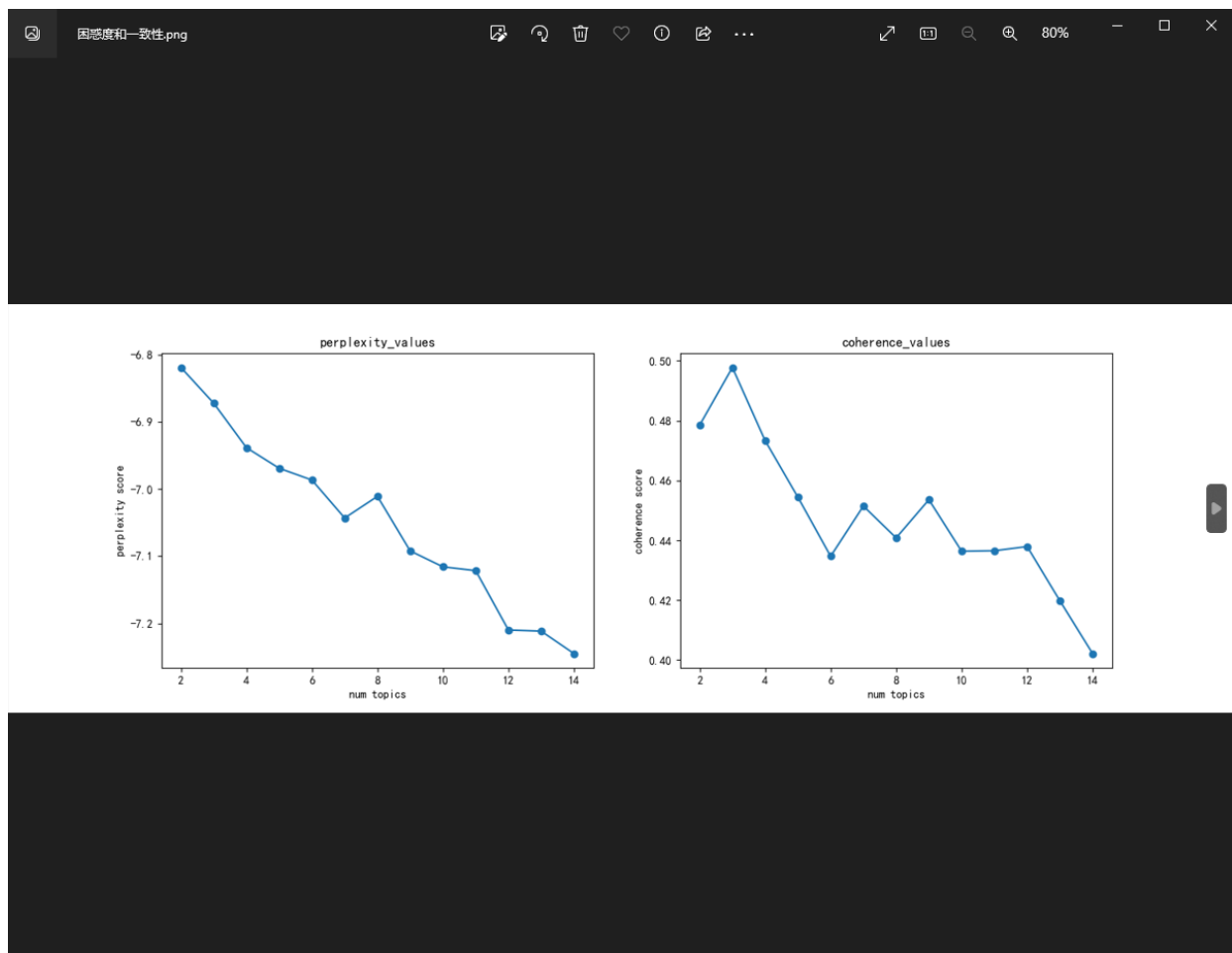
我们寻优的方法是采用困惑度和一致性

主题困惑度一致性是一种用来评价主题模型好坏的指标¹。主题困惑度表示的是对于一篇文章来说，我们有多不确定它是属于某个主题的。主题困惑度越低，说明模型越能够准确地划分主题²。主题一致性表示的是一个主题中的词语是否有语义上的连贯性。主题一致性越高，说明模型越能够提取有意义的主题³。主题困惑度一致性可以结合使用，以找到最合适的主题个数¹。

(1) 困惑度、主题一致性，lda模型找出主题相关词 - CSDN博客. https://blog.csdn.net/stay_foolish12/article/details/127240167.

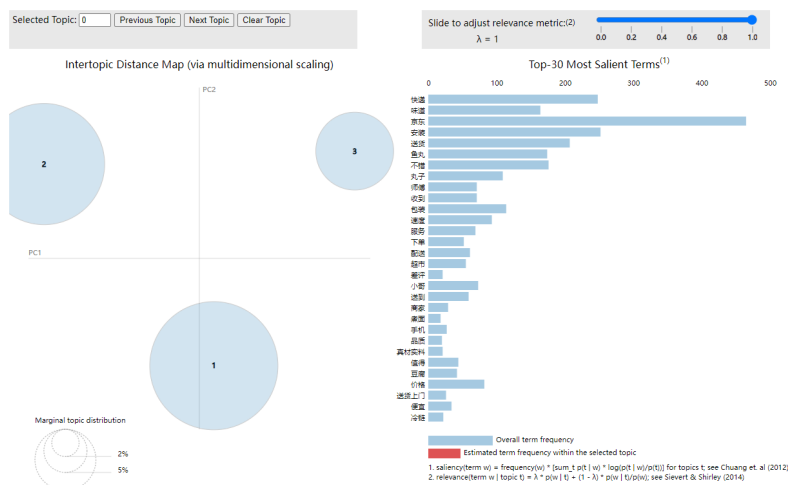
(2) 主题模型 (三) : LDA主题个数选择 - 知乎. <https://zhuanlan.zhihu.com/p/106982034>.

(3) 困惑度(perplexity)的基本概念及多种模型下的计算 (N-gram <https://zhuanlan.zhihu.com/p/114432097>.



根据一致性的最高点，我们选择主题数为3

而后生成的建模如下：



主要代码和文档如下：

名称	修改日期	类型	大小
LDA主题	2023/5/8 9:52	文件夹	
词性数据	2023/5/8 9:52	文件夹	
网络语义	2023/5/8 9:52	文件夹	
crawl.py	2023/5/8 9:52	JetBrains PyChar...	6 KB
data.csv	2023/5/8 9:52	Microsoft Excel ...	913 KB
lda主题建模.py	2023/5/8 9:52	JetBrains PyChar...	6 KB
new_data.csv	2023/5/8 9:52	Microsoft Excel ...	468 KB
score.csv	2023/5/8 9:52	Microsoft Excel ...	1 KB
stopwords_cn.txt	2023/5/8 9:52	文本文档	15 KB
词性处理.py	2023/5/8 9:52	JetBrains PyChar...	5 KB
机器学习分类.py	2023/5/8 9:52	JetBrains PyChar...	4 KB
数据预处理.py	2023/5/8 9:52	JetBrains PyChar...	2 KB
网络语义.py	2023/5/8 9:52	JetBrains PyChar...	6 KB

网络语义分析

基于的共线矩阵的网络语义分析是一种用于表示词与词之间的语义关系的一种网络理论，由美国人工智能专家司马贺在1973年提出的¹。其原理就是以词语为网络的结点，以沟通结点的共现次数表示词语之间的语义关系，构成一个彼此相互联系的网络，以达到理解自然语言句子的语义关系²。这种分析方法可以用于揭示某一学科某一领域某一主题的研究热点与趋势、知识结构与演化等³。基于的共线矩阵的网络语义分析可以用Python实现，具体步骤如下¹²：

- 首先要对文本进行分词，可以使用jieba等工具进行中文分词。
- 然后要对分词进行词频统计，并选取一定数量的关键词作为构建共现矩阵的对象。
- 接着要使用numpy设置共现语义的初始矩阵，矩阵的行列都是关键词，矩阵的元素是关键词之间的共现次数。
- 共现次数的计算规则是，如果两个词在同一个文本单元（如一句话或一篇文档）中出现，并且它们之间的位置距离不超过一定范围（如1或2），则认为它们共现了一次。
- 最后要使用networkx和matplotlib等工具绘制共现语义网络图，图中的节点是关键词，节点的大小表示词频，节点之间的连线表示共现关系，连线的粗细表示共现次数。

相关文献：

- (1) Python 实现文本共现网络分析 卖山楂啦prss的博客-CSDN博客: https://blog.csdn.net/qg_42374697/article/details/113060314.
- (2) Python实现共现语义网络语义网络图_饕餮&化骨龙的博客 https://blog.csdn.net/m0_45827246/article/details/121241477.
- (3) 如何理解格雷马斯的语义矩阵？矩阵分析文本有何意义？- 知乎. <https://www.zhihu.com/question/489907723>.

相关的文档如下：

名称	修改日期	类型	大小
LDA主题	2023/5/8 9:52	文件夹	
词性数据	2023/5/8 9:52	文件夹	
网络语义	2023/5/8 9:52	文件夹	
crawl.py	2023/5/8 9:52	JetBrains PyChar...	6 KB
data.csv	2023/5/8 9:52	Microsoft Excel ...	913 KB
lda主题建模.py	2023/5/8 9:52	JetBrains PyChar...	6 KB
new_data.csv	2023/5/8 9:52	Microsoft Excel ...	468 KB
score.csv	2023/5/8 9:52	Microsoft Excel ...	1 KB
stopwords_cn.txt	2023/5/8 9:52	文本文档	15 KB
词性处理.py	2023/5/8 9:52	JetBrains PyChar...	5 KB
机器学习分类.py	2023/5/8 9:52	JetBrains PyChar...	4 KB
数据预处理.py	2023/5/8 9:52	JetBrains PyChar...	2 KB
网络语义.py	2023/5/8 9:52	JetBrains PyChar...	6 KB

机器学习分类任务

逻辑回归分类是一种基于概率的分类方法，它假设数据服从逻辑分布，然后用极大似然估计法来求解模型参数。逻辑回归分类的优点是可以直接得到分类的概率，而且损失函数是凸函数，容易求解。逻辑回归分类的缺点是容易受到噪声和异常值的影响，而且不能处理非线性可分的数据。

SVM分类是一种基于间隔的分类方法，它试图找到一个最优的超平面来划分数据，使得两类数据之间的间隔最大。SVM分类的优点是可以处理线性不可分的数据，通过使用核函数来映射到高维空间，而且对噪声和异常值有较强的鲁棒性。SVM分类的缺点是求解过程比较复杂，需要选择合适的核函数和参数，而且不能直接得到分类的概率。

这是处理的步骤

逻辑回归分类是通过最大化样本输出到正确分类的概率来减少错误率，相应的损失函数是负对数似然。逻辑回归分类的举措包括：

- 对数据进行预处理，如标准化、归一化、缺失值处理等；
- 选择合适的特征变量，如使用多项式回归来增加非线性特征；
- 选择合适的正则化项，如L1或L2，来防止过拟合或欠拟合；
- 选择合适的优化方法，如梯度下降、牛顿法、拟牛顿法等，来求解模型参数；
- 评估模型的性能，如使用准确率、召回率、F1值、AUC等指标；

SVM分类是通过寻找最佳划分超平面来减少错误率，相应的损失函数是hinge loss。SVM分类的举措包括：

- 对数据进行预处理，如标准化、归一化、缺失值处理等；

- 选择合适的核函数，如线性核、多项式核、高斯核等，来处理线性可分或者非线性可分的数据；
- 选择合适的正则化项，如L1或L2，来控制模型的复杂度和容错性；
- 选择合适的优化方法，如坐标下降法、序列最小优化法、内点法等，来求解对偶问题或者原始问题；
- 评估模型的性能，如使用准确率、召回率、F1值、AUC等指标；

相关参考文献：

- (1) 【机器学习】逻辑回归（非常详细） - 知乎 - 知乎专栏. <https://zhuanlan.zhihu.com/p/74874291>.
- (2) 分类模型(1)——逻辑回归和SVM_svm分类模型_Fran OvO的博客-CSDN博客. https://blog.csdn.net/weixin_52640021/article/details/128296360.

这里有一点是需要注意的这里相关模型并没有采用AUC指标，因为是多分类的任务所以无法采用AUC指标，AUC和ROC这两个指标只能用于二分类任务，对于多分类任务是行不通的

下面是对应的源码和文件；

名称	修改日期	类型	大小
LDA主题	2023/5/8 9:52	文件夹	
词性数据	2023/5/8 9:52	文件夹	
网络语义	2023/5/8 9:52	文件夹	
crawl.py	2023/5/8 9:52	JetBrains PyChar...	6 KB
data.csv	2023/5/8 9:52	Microsoft Excel ...	913 KB
lda主题建模.py	2023/5/8 9:52	JetBrains PyChar...	6 KB
new_data.csv	2023/5/8 9:52	Microsoft Excel ...	468 KB
score.csv	2023/5/8 9:52	Microsoft Excel ...	1 KB
stopwords_cn.txt	2023/5/8 9:52	文本文档	15 KB
词性处理.py	2023/5/8 9:52	JetBrains PyChar...	5 KB
机器学习分类.py	2023/5/8 9:52	JetBrains PyChar...	4 KB
数据预处理.py	2023/5/8 9:52	JetBrains PyChar...	2 KB
网络语义.py	2023/5/8 9:52	JetBrains PyChar...	6 KB