# 小组作业——文本聚类

**【作业内容】**

**一、实践练习**

**1、 对于附件中所提供的文本进行聚类，尝试至少两种方法。**

- 每个数据集中包含 2 万条数据，均为电子商务网站中产品相关问题数据。
- 共 3 个数据集，不同组的同学使用的数据集不同。
  请大家在【数据集分配表】中查看本组所需要使用的数据集，并使用该数据集完成文本聚类任务。
- 数据示例：

| productId | 类别 | questionId | questions |
|---|---|---|---|
| 1505732 | 美容彩妆 | 54532269 | 是真的吗 |
| 1617542 | 美容彩妆 | 20372295 | 有没有包装盒啊 |
| 1235397 | 美容彩妆 | 27153048 | 这个可以直接当成纸膜湿敷吗 |
| 1350256 | 美容彩妆 | 24578152 | 过敏皮肤能用吗 |
| 1429200 | 美容彩妆 | 22483707 | 有包装盒和礼品袋吗 |
| 1353385 | 美容彩妆 | 28514540 | 孕妇可以使用吗 |
| 1353385 | 美容彩妆 | 65026830 | 一瓶可以用多久和芙丝丽芳那个性价比更高 |
| 1317419 | 美容彩妆 | 16806412 | 日期在哪怎么没找到 |

**2、 将聚类结果可视化，对比不同聚类方法的聚类效果。**

**3、 为每条数据贴上聚类得到的类别标签【label】，如下表所示：**

| productId | 类别 | questionId | questions | label |
|---|---|---|---|---|
| 1505732 | 美容彩妆 | 54532269 | 是真的吗 | 0 |
| 1617542 | 美容彩妆 | 20372295 | 有没有包装盒啊 | 1 |
| 1350256 | 美容彩妆 | 24578152 | 过敏皮肤能用吗 | 2 |
| 1429200 | 美容彩妆 | 22483707 | 有包装盒和礼品袋吗 | 1 |
| 1317419 | 美容彩妆 | 16806412 | 日期在哪怎么没找到 | 3 |

将该结果输出并保存至文件：output.csv

**4、 聚类结果分析：**

- 以词云图等可视化形式展示每个类别的文本内容
- 分析每个类别的中心话题

**二、问题分析**

你认为在聚类任务中，本例的文本聚类有什么特点和困难，你们采取了什么分析和解决方式？

**三、问题思考**

简要回答以下问题：

1、层次聚类、k 均值聚类、基于密度的聚类 DBSCAN 三种聚类方法中，哪种需要指定类间距？哪种需要先确定聚类个数？为什么？

2、如果一个数据集非常大，而且分析前不知道合适的聚类数，你认为应该如何实施聚类分

析？你在本例中，如何确定的聚类个数？

3、对比不同聚类方法在本例中的聚类效果，说明各个方法的优缺点、适用情况。

4、是否任何一个数据集都适合进行聚类？如果不是，什么样的数据集不适合进行聚类分析？

5、 如何评价聚类结果的好坏？（提示：主观指标和客观指标）

评价一下本例中自己的聚类结果。

**【作业提交】**

**提交 3 个文件：**

**1、 聚类结果**

文件一：output.csv

**2、 作业报告（10 页以内，精要，写你认为重要的内容）**

文件二：组号.doc（例：1 组.doc）

内容包含：

- 小组成员及分工
- 实验（实验流程说明、使用的模型方法、实验结果分析）
- 问题分析
- 问题思考

**3、 源代码**

文件三：code.py 或 code.ipynb

- **打包发送至邮箱【bigdata_xu@126.com】**
- **压缩包、邮件均命名为【组号 队长姓名】（例：1 组 王**）**
- **截止日期【5 月 18 日（周二）晚上 12 点前】**

---

**此次作业在作业评阅完毕后统一答疑，过程中遇到问题同学们可以互相讨论，查阅资料，填写问题统计问卷 https://www.wjx.cn/vj/eeF2lLd.aspx。**

**阅读以下内容可以帮助同学们更好地完成该作业：**

**【中文文本聚类一般流程】（baseline）**

数据清洗，如去重、去除一些无意义符号、文本等。

**1、分词**

常用第三方库（jieba） pip install jieba

**2、去停用词**

可自行搜索常用停用词库，stopword.txt

参考：

https://github.com/goto456/stopwords

https://blog.csdn.net/qq_33772192/article/details/91886847

**3、生成 tfidf 矩阵**

可自行编写代码或

使用工具（如 from sklearn.feature_extraction.text import TfidfVectorizer

**4、聚类**

常用 sklearn

**5、为每个评论文本贴上类别标签，并描述每类中心话题**

有余力的同学可以尝试使用 word2vec、bert 等词向量表示工具提取文本特征，进行文本聚类，或使用一些 NLP 开源工具包进行尝试，或针对本数据集文本的特点进行模型设计改进，或尝试使用主题抽取模型等进一步探索挖掘文本内容……尝试有没有更好的效果。

**【作业说明】**

**1、对文本聚类**

- 本质：计算文档间的距离（本作业中每个文档对应每条问题）
- 聚类方法：K-Means、层次聚类、高斯混合模型（根据经验，在文本聚类上一般比 Kmeans 性能更好）等。

**2、给聚类结果的簇贴标签，试图描述每个簇对应文本的中心话题，参考方法：**

- 用聚类结果的质心向量来描述当前簇谈论的话题：
  文档集合中的每篇文档用一个特征向量来描述，如果特征向量的权重是 TFIDF 权重，该权重代表了一个特征描述一篇文档的能力。那么，文档聚类的结果获得多个簇，每个簇的质心我们定义为一个话题（图 1 中白色的点），抽取质心向量的 top k 个词项，就描述了这个话题。

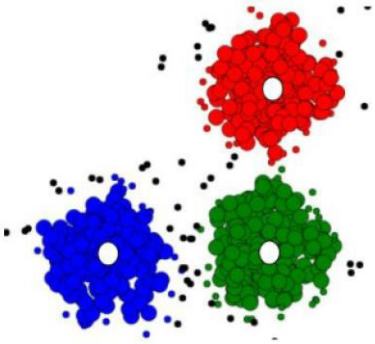

图 1

- 特征选择的方法，如使用互信息或卡方挑选出当前簇的特征作为当前簇的话题（或标签）。使用质心向量中的高权重词项作为簇的标签会有个问题，它会挑选出一些无意义的词。特征选择是挑选那些可以将当前簇和其他簇差异化的词项，但也容易挑选到一些稀有的词项。将特征选择结合对稀有词项进行惩罚的方法往往可以得到较好的簇的标签。

---

**有需要的同学可以参考以下案例，以进一步理解文本聚类。**

**【经典案例练习】路透社文章的文本数据分析与可视化**

基于 kmeans 聚类来检测话题，从路透社 R8 语料库 r8-train-no-stop-id.txt 检测八个话题。每个话题是 Top 10 的词项列表。

参考链接（仅供参考，可自行搜索语料库并练习）：

https://www.cnblogs.com/panchuangai/archive/2020/10/07/13779895.html

得到各个话题的中心词如下（结果不唯一）：

1. mln dlrs net year loss reuter cts shr profit kly
2. cts loss net mln shr reuter profit revs qtr hmy
3. oil opec prices dlrs mln crude bpd reuter pct avialable

4. shares dlrs stock company pct share reuter mln common cvt

5. offer usair american company gencorp dlrs twa group reuter terminating

6. pct bank billion year rate reuter banks rates mln midrate

7. trade japan japanese reuter agreement foreign year countries states steeply

8. company reuter corp dlrs mln unit sale merger acquisition ofp

**【参考案例】**

**使用 python 对微博评论进行分词、文本聚类**

https://blog.csdn.net/weixin_43873702/article/details/111931428

**中文短文本聚类**

https://www.cnblogs.com/chen8023miss/p/11977212.html