数据挖掘与商务分析-第 2 次平时作业

姓名 + 学号

请于 11 月 26 日晚 10 点前提交

1 新闻文本聚类

利用新闻文本数据 (news_data.csv，字段描述见表1)，进行新闻文本聚类。

表 1: 新闻文本数据字段描述

字段	描述
news_id	新闻唯一标识
title	新闻标题
content	新闻内容

具体任务包括但不限于：

1. 使用中文分词工具[1]，对新闻内容进行文本预处理。
2. 利用向量空间模型，采用 TF-IDF 权重，对预处理后的新闻内容进行向量化表示。
3. 运用 K 均值方法对新闻向量进行聚类分析，选择适当的聚类数量。
4. 根据聚类中心的代表性字词，归纳各组新闻的特点。

请同时提交 *.ipynb 和 *.html 文件，其中 *.html 是将 *.ipynb 导出后的版本。

[1]结巴中文分词: https://github.com/fxsjy/jieba