聚类算法作业说明

本说明包括数据文件说明以及Python脚本运行说明

编写内容

1. 修改 main.py 中 output_file 的学号
2. 修改 main.py 中 execute_cluster 的函数体，返回输入的每条文本对应的聚类编号，格式见其注释

运行环境

1. Python 3 或以上
2. sklearn 0.18.1 或以上

脚本运行

用法	作用
python3.5 main.py --train	调用 execute_cluster 为 train_tokens.json 中的每条文本分配聚类编号，并调用 calculate_nmi 进行评估
python3.5 main.py --test	调用 execute_cluster 为 test_tokens.json 中的每条文本分配聚类编号，并将结果输出到 output_file 中

数据格式

文件名	描述
train_tokens.json	训练数据文本，每行一个json格式字符串，包括文档id("docid"字段)，单词id列表("tokenids"字段)
train_topics.json	训练数据参考主题，每行一个json格式字符串，包括文档id("docid"字段)以及文档原始主题的参考值("topic"字段)
test_tokens.json	测试数据文本，每行一个json格式字符串，包括文档id("docid"字段)，单词id列表("tokenids"字段)
submit.json	提交样例，每行一个json格式字符串，包括文档id("docid"字段)，文档主题样例值("cluster"字段)
vocab.json	单词id列表中每个单词及其对应id，仅供参考，实际不使用

最终提交

最终提交以学号命名的json文件，如"10162110111.json"。提交内容格式参考"submit.json"文件。