The First Data Mining Experimental Report

一、实验任务

1、预处理文本数据集，并且得到每个文本的 VSM 表示。

2、实现 KNN 分类器，测试其在 20Newsgroups 上的效果。

Dataset

- The **20 Newsgroups dataset** is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups.
- 20news-18828.tar.gz - 20 Newsgroups; duplicates removed, only "From" and "Subject" headers (18828 documents)

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

Deadline: 2018.11.5, 23:00 http://qwone.com/~jason/20Newsgroups/

二、实验过程

1、 data_process: 对于数据集进行预处理，分词：主要分为采用正则表达式及 split 进行分词、并将所有非字母符号作为分割符号；单词处理：去掉长度小于 3、去掉停用词、复数变单数、去掉文档中 frequency 小于等于 15 的单词；统计单词 tf、idf 等数据，并利用倒排索引进行 knn 加速运行；测试数据文本中出现单词为基础，建立相应词典 word_idf、word_doc_tf、doc_word_tf 三个词典，建立测试文本，所有出现该文本中单词的训练文本的 vsm 表示。

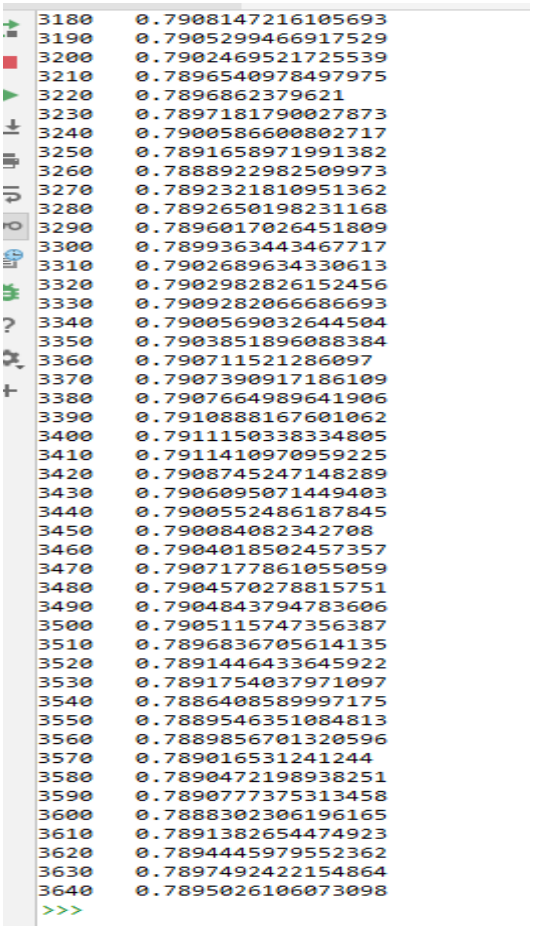
2、 knn：利用欧几里得距离计算相似性，采取倒排索引进行 knn 加速。

三、实验过程

取不同的 K 值结果将改变，实验结果如下：

当 K 取 15 时，运行 3640 个测试数据结果稳定在 0.79 左右。

```
Python 3.6.7 |Anaconda, Inc.| (default, Oct 24 2018, 09:45:24) [MSC v.1912 64 bit (AMD64)] on win32
>>> import main
>>> main.compute_acc_without_reload()
Prepare already Finished!
10   0.8888888888888888
20   0.7894736842105263
30   0.6896551724137931
40   0.7435897435897436
50   0.7959183673469388
60   0.7627118644067796
70   0.782608695652174
80   0.759493670886076
90   0.7752808988764045
100   0.7575757575757576
110   0.7706422018348624
120   0.7647058823529411
130   0.7751937984496124
140   0.7841726618705036
150   0.7919463087248322
160   0.7987421383647799
170   0.8047337278106509
```



四、心得体会

初学 python 感触良多，编辑简化优点很多，但如何实验 knn 的分类对我而言尤其是预处理部分通过查阅资料、同门的帮助才得以顺利开展，今后要更细致的去学习、去思考。