题目　English Text Data Processing

姓名 骆克云 学号 MG1633052 邮箱 streamer.ky@foxmail.com 联系方式18115128082

(南京大学 计算机科学与技术系, 南京　210093)

1　实现细节

1.1　语料库类名、文件名、路径映射（projectutil.py）

将数据文件统一放置在一个目录下，将类别名、文件名和文件路径一一映射，并使用 OrderedDict 按类别名将类名排序，实现有序存储。本处理的目的是可以通过类别名获取某一类别下的所有文件，通过文件名得到文件的绝对路径。

1.2　分词，提取骨干(assignment1/EnglishTextDataProcessing/preprocess.py)

1.2.1　分句，分词

对于每一个打开的文件，首先载入 NLTK 的 sentence_tokenizer 分句模块并小写化，分句后，使用正则表达式进行分词，分词模式为缩略词、数字百分号、单词（具有中间横线）以及标点符号。然后将单词列表扁平化，过滤掉非字母，特殊处理文献中大量存在的“.”和“-”。

1.2.2　词干提取

使用 NLTK 中的 nltk.stem.lancaster.LancasterStemmer() 算法进行词干提取，使用 nltk.stem.WordNetLemmatizer 进行词干还原，使用 nltk.corpus.wordnet. morphy 进行同义词匹配。具体如下：对于单词列表中的每一个单词，首先进行词干还原，然后进行同义词匹配，最后提取骨干，过滤掉那些单词中字符全部相同的单词，将结果列表返回。

1.3　停用词选择（data/StopWords/english）

使用 NLTK 语料库中提供的 english 停用词，共有 153 个。

经过上述步骤后，便可以得到一个文件的单词列表。

1.4　TF-IDF类(assignment1/EnglishTextDataProcessing/tfidf.py)

该类主要成员变量为：文件名-单词列表字典，单词对应的文件名集合，经过排序的所有单词。

1.4.1　读取/添加语料：（def add_doc(self, file)）

本方法通过文件名添加单词，同时计算 TF，记录单词所出现的文件，为 IDF 计算做准备。

TF = 该单词在文件中出现的次数/文件中单词数量

1.4.2　计算 TF-IDF：（def tf_idf(self, file)）

本方法对于一个给定的文件名，遍历该文件中的单词，计算 TF*IDF(保留 10 位小数)，返回结果字典。

IDF = log(文件数量/包含该单词的文件数量)

tf-idf = TF*IDF

1.4.3　生成向量数据

保存所有文件的单词列表，并进行排序。以类别为单位，对类别文件中的每一个文件产生一行记录：获得该文件的 TF-IDF 字典，对单词列表中每一个单词的索引和单词，如果该单词 tfidf 值大于 10^-11,即非 0.0，单词存在于该文档中，将索引和 tf-idf 值作为元组保存。最后将结果写入文件中。

1.5 产生数据集

生成 TFIDF 类的一个实例，添加所有文件，保存所有有序的单词及其索引到文件 word_list.txt 中，产生所有类别的向量文件，文件名为类别名称。

2 结果

2.1 实验设置

数据来源：Corpus about papers from International Conference on Machine Learning 2014/2015
数据预处理：解压到 data 目录下，在生成数据集过程中会自动进行预处理

2.2 实验结果

1. 在 Windows 下的运行情况如下：

```
D:\workspace\dataminingcource>python runAssignment1.py
====开始进行处理=====
语料库集合单词个数：31016
共运行时间：147 秒
======结束处理=======
```

2. 在 Linux 下运行情况如下

```
➜  dataminingcource (master) ✗ python3 runAssignment1.py
====开始进行处理=====
语料库集合单词个数：31016
共运行时间：137 秒
======结束处理=======
```

3. 查看结果文件

```
➜  result (master) ✗ tree

├── 10. Ranking
├── 11. Reinforcement Learning
├── 12. Supervised Learning
├── 13. Theory
├── 14. Unsupervised and Semi-Supervised Learning
├── 15. Others
├── 1. Active Learning
├── 2. Applications
├── 3. Bayesian Learning and Graphical Model
├── 4. Deep Learning
├── 5. Ensemble and Crowdsourcing
├── 6. Feature Learning
├── 7. Kernel Methods
```

├── 8. Online Learning
├── 9. Optimization
└── word_list.txt

4 向量文件查看

➜ result (master) ✗ cat 7.\ Kernel\ Methods | more

[66 : 0.0003263809,74 : 0.0003301392,90 : 0.0010528731,110 : 0.0036554499,118 : 0.0003325065,125 : 0.0007106901,126 : 9.95189e-05,133 : 0.0006425161,137 : 0.0016269568,151 : 0.0009887176,162 : 0.0001185383,167 : 0.0001711317,190 : 7.0635e-05,218 : 0.0007969578,284 : 6.87891e-05,338 : 0.0003228391,355 : 0.0014497266,421 : 0.0002478651,439 : 0.000352703,573 : 9.197e-05,622 : 5.7607e-05,633 : 0.0002213909,659 : 0.000280023,662 : 3.971e-06,673 : 0.0001233687,694 : 0.0003105527,758 : 0.000166863,762 : 0.0003853522,788 : 3.77179e-05,819 : 0.0013460536

5 结果反思

数值过小，可能有精度损失。

注意：

1. 最终提交的报告最好保存为 pdf 格式
2. 压缩格式为 zip 格式，请勿使用需要安装特定软件才能打开的压缩方式
3. 作业的文件夹目录请按照网页要求，代码、结果放在不同子文件夹中。作业网页上给出的数据不需要再次提交