浙江工业大学

文本分析与挖掘

上机实验

计算机科学与技术学院

实验一、文本预处理和基本表示

Part 2. 文本的基本表示

一、 实验目的

熟悉基本词袋表示、TF-IDF 权重计算；

二、实验内容

1.英文数据词袋表示

文本样例集(可以采用其他自定样例)：

```
Doc1:The sky is blue and beautiful.
Doc2:Love this blue and beautiful sky!
Doc3:The quick brown fox jumps over the lazy dog.
```

a. 调用 Part1 中实现的预处理函数 EngPreprocess()对以上文本样例集进行预处理得到预处理后的文本数据集。

具体要求：

对 EngPreprocess()进行适当调整，使其具有以下两种格式输出：

输出格式 1：

```
['sky blue beautiful .',
 'Love blue beautiful sky !',
 'quick brown fox jumps lazy dog .',
```

输出格式 2：

```
[['sky', 'blue', 'beautiful', '.'],
 ['Love', 'blue', 'beautiful', 'sky', '!'],
 ['quick', 'brown', 'fox', 'jumps', 'lazy', 'dog', '.'],
```

b. 对以上数据集实现三种向量化表示：布尔型、词频、TF-IDF；

c. 对一个新的文档”The brown fox is quick and the blue dog is lazy!”进行同样的预处理，并用步骤 b 中得到的（基于词频）模型进行向量化表示。观察并讨论：新文档中出现了训练集中没有的词会怎么样？

d. 对保存在 nips12 文件夹中的文档数据进行预处理，并进行 TF-

IDF 向量表示。该文件夹包含了 2012 年发表在 NIPS 会议上的论文，每个文件对应一篇论文。

2. 中文词袋表示

中文样例数据(可以采用其他自定样例)

1.家乡名叫箐口村，属贵州省毕节市大方县猫场镇。张凌生于 1985 年，童年时跟一群留守的孩子玩。长到 8 岁，娘送他去村小。 2.大学录取通知单，是去猫场赶集的亲戚带回箐口村的。这在箐口村，是从来没有过的事。 3. 在这环境里他读完了六年级，接着到猫场中学读初中，再到乡政府所在地读高中。那年月，猫场镇能考上大学的极少，但张凌奇迹般地考上了。

a. 调用 ChTokenize()进行分词后，对中文语料集实现三种向量化表示：布尔型、词频、TF-IDF。