实验一：文本数据集的读写和简单处理

PPT制作：罗茂权，杨覃娟
出题人　：庞健辉，李祥圣

文件读写

C++：

http://blog.csdn.net/kingstar158/article/details/6859379/

Java：

http://blog.csdn.net/jiangxinyu/article/details/7885518/

Python：

http://www.cnblogs.com/allenblogs/archive/2010/09/13/1824842.html

字符串分割

C++：

http://blog.csdn.net/glt3953/article/details/11115485

Java：

http://blog.sina.com.cn/s/blog_b7c09bc00101d3my.html

Python：

http://blog.sina.com.cn/s/blog_81e6c30b01019wro.html

数据集

文本编号	词列表					
训练文本1	少年	救出	溺水	男童	男童	
训练文本2	老人	参加	高考			
训练文本3	男童	救出	溺水	老人	溺水	救出

不重复词向量/词汇表

救出	男童	高考	老人	参加	溺水	少年

one-hot矩阵

One-hot：使用一个向量表示一篇文章，向量的长度为词汇表的大小。
向量中的1表示存在对应的单词，0表示不存在。

数据集

文本编号	词列表					
训练文本1	少年	救出	溺水	男童	男童	
训练文本2	老人	参加	高考			
训练文本3	男童	救出	溺水	老人	溺水	救出

one-hot矩阵

	救出	男童	高考	老人	参加	溺水	少年
训练文本1	1	1	0	0	0	1	1
训练文本2	0	0	1	1	1	0	0
训练文本3	1	1	0	1	0	1	0

TF矩阵

TF（ term frequency ）：向量的每一个值标志对应的词语出现的次数归一化后的频率。

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

TF矩阵

	救出	男童	高考	老人	参加	溺水	少年
训练文本1	1/5	2/5	0	0	0	1/5	1/5
训练文本2	0	0	1/3	1/3	1/3	0	0
训练文本3	1/3	1/6	0	1/6	0	1/3	0

输出

```
[ 0.2  0.4  0.   0.   0.   0.2  0.2]
[ 0.          0.          0.33333333  0.33333333  0.33333333  0.          0.        ]
[ 0.33333333  0.16666667  0.          0.16666667  0.          0.33333333  0.        ]
```

TF-IDF矩阵

IDF：逆向文件频率：

$$\mathrm{idf_i} = \log \frac{|D|}{|\{j : t_i \in d_j\}|} \qquad idf_i = log\frac{|D|}{1 + |\{j : t_i \in d_j\}|}$$

TF矩阵

	救出	男童	高考	老人	参加	溺水	少年
训练文本1	1/5	2/5	0	0	0	1/5	1/5
训练文本2	0	0	1/3	1/3	1/3	0	0
训练文本3	1/3	1/6	0	1/6	0	1/3	0

IDF

救出	男童	高考	老人	参加	溺水	少年
log|3/2|	log|3/2|	log|3/1|	log|3/2|	log|3/1|	log|3/2|	log|3/1|

TF-IDF矩阵

$$\mathrm{tfidf}_{i,j} = \mathrm{tf}_{i,j} \times \mathrm{idf}_i$$

TF-IDF矩阵

	救出	男童	高考	老人	参加	溺水	少年
训练文本1	(1/5)*log|3/2|	(2/5)*log|3/2|	0	0	0	(1/5)*log|3/2|	(1/5)*log|3/1|
训练文本2	0	0	(1/3)*log|3/1|	(1/3)*log|3/2|	(1/3)*log|3/1|	0	0
训练文本3	(1/3)*log|3/2|	(1/6)*log|3/2|	0	(1/6)*log|3/2|	0	(1/3)*log|3/2|	0

输出

```
[ 0.03521825  0.0704365   0.          0.          0.          0.03521825  0.09542425]
[ 0.          0.          0.15904042  0.05869709  0.15904042  0.          0.        ]
[ 0.05869709  0.02934854  0.          0.02934854  0.          0.05869709  0.        ]
```

稀疏矩阵三元顺序表

one-hot矩阵

	救出	男童	高考	老人	参加	溺水	少年
训练文本1	1	1	0	0	0	1	1
训练文本2	0	0	1	1	1	0	0
训练文本3	1	1	0	1	0	1	0

三元顺序表

	3	md	
	7	nd	
	11	td	
0	0	0	1
1	0	1	1
2	0	5	1
3	0	6	1
4	1	2	1
5	1	3	1
6	1	4	1
7	2	0	1
8	2	1	1
9	2	3	1
10	2	5	1
	i	j	v

输出

```
[3]
[7]
[11]
[0, 0, 1]
[0, 1, 1]
[0, 5, 1]
[0, 6, 1]
[1, 2, 1]
[1, 3, 1]
[1, 4, 1]
[2, 0, 1]
[2, 1, 1]
[2, 3, 1]
[2, 5, 1]
```

矩阵加法运算（选做）

```
[3]
[7]
[11]
[0, 0, 1]
[0, 1, 1]
[0, 5, 1]
[0, 6, 1]
[1, 2, 1]
[1, 3, 1]
[1, 4, 1]
[2, 0, 1]
[2, 1, 1]
[2, 3, 1]
[2, 5, 1]
```

A

+

```
[3]
[7]
[5]
[0, 1, 1]
[0, 5, 1]
[1, 0, 1]
[1, 6, 1]
[2, 0, 1]
```

B

=

```
[3]
[7]
[13]
[0, 0, 1]
[0, 1, 2]
[0, 5, 2]
[0, 6, 1]
[1, 0, 1]
[1, 2, 1]
[1, 3, 1]
[1, 4, 1]
[1, 6, 1]
[2, 0, 2]
[2, 1, 1]
[2, 3, 1]
[2, 5, 1]
```

C

词汇表顺序要求

数据集

文本编号	词列表					
训练文本1	少年	救出	溺水	男童	男童	
训练文本2	老人	参加	高考			
训练文本3	男童	救出	溺水	老人	溺水	救出

不重复词向量/词汇表：按词在数据集中出现的顺序排列

少年	救出	溺水	男童	老人	参加	高考

实验任务

1、将数据集的数据表示成one-hot矩阵，TF矩阵，TF-IDF矩阵，并分别保存为“onehot”，“TF”，“TFIDF”三个文件 。

2、将数据集的one-hot矩阵表示成三元组矩阵，保存为“smatrix”文件。

3、（选做）实现系数矩阵加法运算，保存为“AplusB”文件。

如果对此次实验题目有疑问，请联系庞健辉和李祥圣。

注意事项

1、作业提交地址

发生了一些奇怪的事情所以~稍后通知大家 FTP地址

2、命名方式

- 实验报告：请按照模板写，提交为：学号_拼音名字.pdf。
- 实验代码：同一个算法请尽量写成一份代码，提交为：学号_拼音名字.xxx，后缀视使用语言而定。
- 实验结果： 统一保存为txt格式，将所有实验结果进行压缩，提交为：学号_拼音名字.zip。

3、编程语言可用c++, python, matlab, java，不能使用现成库，否则扣分

4、 提交截止时间

2016年09月18日23：59：59前提交至FTP对应文件夹，否则视为迟交