

实验一：文本数据集简单处理

PPT制作及出题人：陈欣鸿，毛润泽

每周安排（初步计划）

周	课程安排
3	数据集处理
4 ~ 5	K近邻算法 朴素贝叶斯算法
6	感知机算法
7 ~ 8	决策树
9 (~10)	逻辑回归
11~12	神经网络
13 ~ 18	Project

实验课要求

- 实验课需要一定的**编程基础**以及**数学基础**，从对公式的**推导**再到**代码的实现**，都会在实验课内容中体现。
- 实验课主要包括两项内容：指导实验内容以及验收之前一次的实验内容，有**不定时签到**，验收会包括**推导公式**，**解释代码**以及**现场跑结果**。
- 实验内容不会很难，但是**绝对不水**，如果抱着**侥幸心理**抄袭代码或者敷衍实验，后果会比较严重。
- 其余详细注意事项注意查看“**实验课须知.pdf**”。

文件读写

C++:

<http://www.cnblogs.com/ifeiyun/articles/1573134.html>

Java:

<http://www.cnblogs.com/zhuocheng/archive/2011/12/12/2285290.html>

Python:

<https://www.liaoxuefeng.com/wiki/0014316089557264a6b348958f449949df42a6d3a2e542c000/001431917715991ef1ebc19d15a4afdace1169a464eccc2000>

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	词汇表					
训练文本1	苹果	手机	好用	销售		
训练文本2	市民	买	手机	手机		
训练文本3	市民	觉得	苹果	手机	贵	好用

不重复词向量/词汇表

贵	好用	觉得	买	苹果	市民	手机	销售
---	----	----	---	----	----	----	----

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。**1** 表示 **存在** 对应的单词，**0** 表示 **不存在**。

数据集

文本编号	词汇表					
训练文本1	苹果	手机	好用	销售		
训练文本2	市民	买	手机	手机		
训练文本3	市民	觉得	苹果	手机	贵	好用

One-hot矩阵

	贵	好用	觉得	买	苹果	市民	手机	销售
训练文本1	0	1	0	0	1	0	1	1
训练文本2	0	0	0	1	0	1	1	0
训练文本3	1	1	1	0	1	1	1	0

One-hot 矩阵

One-hot矩阵

	贵	好用	觉得	买	苹果	市民	手机	销售
训练文本1	0	1	0	0	1	0	1	1
训练文本2	0	0	0	1	0	1	1	0
训练文本3	1	1	1	0	1	1	1	0

标准输出：
(不重复词向量
按照出现顺序构成)



onehot.txt - 记事本

文件(F) 编辑(E) 格式(O)

```
1 1 1 1 0 0 0 0
0 1 0 0 1 1 0 0
1 1 1 0 1 0 1 1
```


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。

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

TF矩阵

	贵	好用	觉得	买	苹果	市民	手机	销售
训练文本1	0	1/4	0	0	1/4	0	1/4	1/4
训练文本2	0	0	0	1/4	0	1/4	2/4	0
训练文本3	1/6	1/6	1/6	0	1/6	1/6	1/6	0

TF矩阵

TF矩阵

	贵	好用	觉得	买	苹果	市民	手机	销售
训练文本1	0	1/4	0	0	1/4	0	1/4	1/4
训练文本2	0	0	0	1/4	0	1/4	2/4	0
训练文本3	1/6	1/6	1/6	0	1/6	1/6	1/6	0

标准输出：
(不重复词
向量**按照出
现顺序**构成)



TF.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

```
0.25 0.25 0.25 0.25 0 0 0 0
0 0.5 0 0 0.25 0.25 0 0
0.166667 0.166667 0.166667 0 0.166667 0 0.166667 0.166667
```

TF-IDF矩阵

IDF: 逆向文件频率; 假设总共有 $|D|$ 篇文章, $|\{j: t_i \in d_j\}|$ 表示出现了该单词的文章总数, IDF值的计算公式如下:

$$\text{idf}_i = \log \frac{|D|}{|\{j: t_i \in d_j\}|} \quad \text{idf}_i = \log \frac{|D|}{1 + |\{j: t_i \in d_j\}|}$$

IDF 向量:

	贵	好用	觉得	买	苹果	市民	手机	销售
IDF	$\log(3/1)$	$\log(3/2)$	$\log(3/1)$	$\log(3/1)$	$\log(3/2)$	$\log(3/2)$	$\log(3/3)$	$\log(3/1)$

思考: IDF 的第二个计算公式中分母多了个 **1** 是为什么?

TF-IDF矩阵

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

TF-IDF矩阵

	贵	好用	觉得	买	苹果	市民	手机	销售
训练文本1	0	$(1/4) \times \log(3/2)$	0	0	$(1/4) \times \log(3/2)$	0	$(1/4) \times \log(3/3)$	$(1/4) \times \log(3/1)$
训练文本2	0	0	0	$(1/4) \times \log(3/1)$	0	$(1/4) \times \log(3/2)$	$(2/4) \times \log(3/3)$	0
训练文本3	$(1/6) \times \log(3/1)$	$(1/6) \times \log(3/2)$	$(1/6) \times \log(3/1)$	0	$(1/6) \times \log(3/2)$	$(1/6) \times \log(3/2)$	$(1/6) \times \log(3/3)$	0

标准输出：

（不重复词向量
按照出现顺序构成）



TFIDF.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

```
0 -0.0719205 0 0.101366 0 0 0 0
0 -0.143841 0 0 0 0.101366 0 0
0 -0.047947 0 0 0 0 0.0675775 0.0675775
```

思考：IDF数值有什么含义？TF-IDF数值有什么含义？

稀疏矩阵三元顺序表

One-hot矩阵

	贵	好用	觉得	买	苹果	市民	手机	销售
训练文本1	0	1	0	0	1	0	1	1
训练文本2	0	0	0	1	0	1	1	0
训练文本3	1	1	1	0	1	1	1	0

三元顺序表

	3	行数	
	8	列数	
	13	数值个数	
0	0	1	1
1	0	4	1
2	0	6	1
3	0	7	1
4	1	3	1
5	1	5	1
6	1	6	1
7	2	0	1
8	2	1	1
9	2	2	1
10	2	4	1
11	2	5	1
12	2	6	1
	行号 i	列号 j	数值 k

稀疏矩阵三元顺序表

三元顺序表标准输出：

（不重复词向量**按照出现顺序**构成）

思考：为什么要用
三元顺序表表达稀
疏矩阵？

 smatrix.txt - 记事本

文件(F) 编辑(E) 格式(O)

```
[3]
[8]
[13]
[0, 0, 1]
[0, 1, 1]
[0, 2, 1]
[0, 3, 1]
[1, 1, 1]
[1, 4, 1]
[1, 5, 1]
[2, 0, 1]
[2, 1, 1]
[2, 2, 1]
[2, 4, 1]
[2, 6, 1]
[2, 7, 1]
```

矩阵加法运算

例子:

```
[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	词汇表					
训练文本1	苹果	手机	好用	销售		
训练文本2	市民	买	手机	手机		
训练文本3	市民	觉得	苹果	手机	贵	好用

不重复词向量/词汇表：按词在数据集中**出现的顺序排列**

苹果	手机	好用	销售	市民	买	觉得	贵
----	----	----	----	----	---	----	---

实验任务

- 1、将数据集“semeval”的数据表示成 One-hot 矩阵，TF 矩阵，TF-IDF 矩阵，并分别保存为“onehot.txt”，“TF.txt”，“TFIDF.txt”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.txt”文件。
- 3、实现系数矩阵加法运算，保存为“AplusB.xx”文件，xx视编程语言而定，如 c / cpp / java / py 等。

实验任务

综上：

1. 总共 **四个结果文件**：onehot.txt, tf.txt, tfidf.txt, smatrix.txt，打包，正确命名后上交ftp。
2. 代码文件 **尽量** 是写在一个代码文件里，直接正确命名后上交代码文件即可，如果有多个代码文件，打包，正确命名后上交ftp。
3. 报告中要有 **所有任务** 的结果展示，报告提交PDF版本，请勿提交word文件，避免排版混乱。

如果对此次实验题目有疑问，请联系陈欣鸿和毛润泽。

注意事项

1、作业提交地址

FTP地址：<ftp://39.108.233.34>

登录用户名与密码均为 student

提交文件夹的名字是 labx_yyyyddmmend, x为第几次实验, yyyyddmmend是指截止日期, 比如 20170927end

2、命名方式

查询“实验课须知”, 实验报告, 所有代码文件以及结果文件都需要上交。

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

4、提交截止时间

2017年09月27日23: 59: 59前