Lab 1: Simple Natural Language Processing

1. Objective

Know how to change the documents to vectors, and make predictions based on the distance.

已知 10 篇训练文本对应的公众“感动”概率值（1、0.9、0.5、0.5、0.4、0.1、0.02、0、0、0），预测 10 篇测试文本对应的公众“感动”概率值各是多少？

2. Dataset

文本编号	词列表（以空格分隔）	公众"感动"的概率
训练文本1	消防员 冲进 火场 救出 男童	1
训练文本2	公务员 患 癌症 保持 在岗	0.9
训练文本3	消防员 多次 冲进 火场 救人 不幸 身亡	0.5
训练文本4	老人 成功 进行 免费 白内障 手术	0.5
训练文本5	海豚 误 吞 排球 后 手术 成功 取出	0.4
训练文本6	6旬 老人 跳楼 自杀 身亡	0.1
训练文本7	男子 跳楼 自杀 身亡	0.02
训练文本8	疑犯 枪杀 出租车 司机	0
训练文本9	男子 枪杀 妻子 后 自杀	0
训练文本10	医师 误 把 肾脏 当 肝脏 致人 身亡	0
测试文本1	癌症 老人 成功 手术	?
测试文本2	男子 枪杀 司机 后 喝药 自杀	?
测试文本3	癌症 医师 保持 手术 清醒	?
测试文本4	男子 跳楼 自杀	?
测试文本5	男子 枪杀 老人 后 自杀	?
测试文本6	消防员 冲进 火场 将 男童 救出	?
测试文本7	出租车 司机 免费 搭载 老人	?
测试文本8	男子 误 杀 弟媳 后 自杀 身亡	?
测试文本9	医师 误 把 患者 肝脏 捅破 致人 身亡	?
测试文本10	6旬 老人 火场 救人 不幸 身亡	?

提供两种格式的数据集：Dataset_txt format.txt、Dataset_excel format.xls

第一种为 txt 格式的数据集（三列之间以 Tab 分隔），供大家方便读取文件中的数据。主要编程语言的参考代码示例如下：

C、C++读写文件：http://blog.csdn.net/kingstar158/article/details/6859379

Java 读写文件：http://blog.csdn.net/jiangxinyu/article/details/7885518

Matlab 读写文件：http://blog.csdn.net/yelbosh/article/details/8549121

Python 读写文件：http://www.cnblogs.com/allenblogs/archive/2010/09/13/1824842.html

http://sucre.iteye.com/blog/704077

第二种为 excel 格式的数据集，与上述内容一致，供大家参考。

该数据集包含 10 篇训练文本，每篇训练文本既有词列表，也有标准答案（即公众“感动”的概率值）；另有 10 篇测试文本，每篇测试文本只有词列表，其公众“感动”的概率值需要大家预测。数据集很小，供大家开始搭建你们的实验工程，建议选定一种编程语言，后续在此基础上不断完善。后期将有更大数据集。

3. Processes ~~【本部分提交的截止时间为 10 月 18 日 23:00，鼓励当场提交】~~

（1）基于 10 篇训练文本和 10 篇测试文本的词列表，生成不重复的全部词列表文件：

输入（Dataset_words.txt）：

文本编号	词列表（以空格分隔）
训练文本1	消防员 冲进 火场 救出 男童
训练文本2	公务员 患 癌症 保持 在岗
训练文本3	消防员 多次 冲进 火场 救人 不幸 身亡
训练文本4	老人 成功 进行 免费 白内障 手术
训练文本5	海豚 误 吞 排球 后 手术 成功 取出
训练文本6	6旬 老人 跳楼 自杀 身亡
训练文本7	男子 跳楼 自杀 身亡
训练文本8	疑犯 枪杀 出租车 司机
训练文本9	男子 枪杀 妻子 后 自杀
训练文本10	医师 误 把 肾脏 当 肝脏 致人 身亡
测试文本1	癌症 老人 成功 手术
测试文本2	男子 枪杀 司机 后 喝药 自杀
测试文本3	癌症 医师 保持 手术 清醒
测试文本4	男子 跳楼 自杀
测试文本5	男子 枪杀 老人 后 自杀
测试文本6	消防员 冲进 火场 将 男童 救出
测试文本7	出租车 司机 免费 搭载 老人
测试文本8	男子 误 杀 弟媳 后 自杀 身亡
测试文本9	医师 误 把 患者 肝脏 捅破 致人 身亡
测试文本10	6旬 老人 火场 救人 不幸 身亡

输出（不重复的全部词列表文件）示例如下：

```
消防员
冲进
火场
救出
男童
公务员
患
癌症
保持
在岗
消防员（这个词在第一行出现过，因此不加入不重复的全部词列表文件中）
多次
冲进（这个词在第二行出现过，因此不加入不重复的全部词列表文件中）
火场（这个词在第三行出现过，因此不加入不重复的全部词列表文件中）
救人
……
```

（2）基于上一步骤输出的不重复的全部词列表文件，将原始的 10 篇训练文本和 10 篇测试文本都转换为向量，如下所示：

文本编号	消防员	冲进	火场	救出	男童	公务员	患	癌症	保持	在岗	多次	救人	……
训练文本1	1	1	1	1	1	0	0	0	0	0	0	0	……
训练文本2	0	0	0	0	0	1	1	1	1	1	0	0	……
训练文本3	1	1	1	0	0	0	0	0	0	0	1	1	……
……													
测试文本1	0	0	0	0	0	0	0	1	0	0	0	0	……
……													

其中，第一行中的“消防员”、“冲进”等词，为上一步骤输出的不重复的全部词，以空格分隔（也可以用逗号、分号、Tab 等分隔，程序能够处理即可）；第二行中的“1”、“1”等值，获得的方式为：如果当前文本中含有“消防员”这个词，则“消防员”这一列的值即为“1”，否则为“0”。比如，训练文本 1 是“消防员 冲进 火场 救出 男童”，所以训练文本 1 的“消防员”这一列的值为“1”。训练文本 2 是“公务员 患 癌症 保持 在岗”，所以训练文本 2 的“消防员”这一列的值为“0”。据此，原始的 10 篇训练文本和 10 篇测试文本都被转换为向量，向量中各个维度上的值要么是 1，要么是 0。如下：

训练文本 1 = (1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, ……)

训练文本 2 = (0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, ……)

……

测试文本 1 = (0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, ……)

……

（3）基于上一步骤输出的向量文件，采用下述方法预测 10 篇测试文本对应的公众“感动”的概率值：

假如现在要预测测试文本 1 对应的公众“感动”的概率值，首先计算测试文本 1 的向量与 10 篇训练文本（即训练文本 1、训练文本 2、…、训练文本 10）的向量的欧式距离。这里再回顾一下欧式距离的计算公式。6 维向量(1, 1, 1, 1, 0, 0)与 6 维向量(0, 1, 0, 0, 1, 1)的欧式距离为：$\sqrt{(1-0)^2+(1-1)^2+(1-0)^2+(1-0)^2+(0-1)^2+(0-1)^2}$，即两个向量在每一个维度上的值的差的平方和，再开根号。然后，上一个步骤，我们会得到测试文本 1 的向量分别与训练文本 1 的向量、训练文本 2 的向量、…、训练文本 10 的向量的 10 个欧式距离的值，比较这 10 个欧式距离的值，得到一个最小的欧式距离的值。假如通过比较得到测试文本 1 的向量与训练文本 4 的向量的欧式距离最小，则将测试文本 1 对应的公众“感动”的概率值预测为训练文本 4 对应的公众“感动”的概率值，即 0.5。采用上述相同的方法，得到测试文本 2 至测试文本 10 对应的公众“感动”的概率值的预测值，比如分别是：0.1、0.4、0、1、0.9、0.02、1、0.5、0。

（4）运行 RunResult.bat 文件（请将 AILab.jar 与该文件放在同一目录下，若仍无法运行该文件，请在此下载 jre 后安装：http://www.java.com/en/download/manual.jsp），输入上一步骤获得的 10 篇测试文本对应的公众“感动”的概率值的预测值，得到相关系数的值，如下：

```
请依次输入10篇测试文本的预测结果，以空格或回车分隔：
0.5 0.1 0.4 0 1 0.9 0.02 1 0.5 0
你的上述预测结果与标准答案的相关系数(-1到1之间)为：
0.046830853318700745
```

（5）以自己的学号新建一个文本文件，如 99999999.txt，将“exp1、空格、以及上述相关系数的输出值（保留小数点后四位，即 0.0468）”写入该文本文件中的第一行：

```
99999999.txt
1  exp1 0.0468
```

4. More methods【本部分提交的截止时间为 10 月 28 日 23:00，鼓励当场提交】

在实现了“3. Processes”中的全部内容后，请大家继续下述扩展实验：

4.1 更改向量中值的表示方法

在“3. Processes”的（2）步骤中，每篇文本转换成的向量中，各个维度上的值要么是 1，要么是 0，如下：

文本编号	消防员	冲进	火场	救出	男童	公务员	患	癌症	保持	在岗	多次	救人	……
训练文本1	1	1	1	1	1	0	0	0	0	0	0	0	……
训练文本2	0	0	0	0	0	1	1	1	1	1	0	0	……
训练文本3	1	1	1	0	0	0	0	0	0	0	1	1	……
……													
测试文本1	0	0	0	0	0	0	0	1	0	0	0	0	……
……													

现在，请将上述向量中的值归一化，即对于每篇文本对应的向量，都将其每个维度上的值同时除以这篇文本的总词数（或者说这篇文本对应向量的所有维度上的值的和），比如，在“3. Processes”的（2）步骤中，训练文本 1 = (1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, ……)，这个向量的所有维度上的值的和为 5（因为训练文本 1 总共包含 5 个词），那么，归一化以后的训练文本 1 对应的向量 = (1/5, 1/5, 1/5, 1/5, 1/5, 0, 0, 0, 0, 0, 0, 0, ……) = (0.2, 0.2, 0.2, 0.2, 0.2, 0, 0, 0, 0, 0, 0, 0, ……)。按照上述方法，将全部的 10 篇训练文本和 10 篇测试文本对应的向量都进行归一化，然后按照“3. Processes”的后续步骤，将“exp2、空格、以及在归一化的向量文件中进行预测输出的相关系数的值（保留小数点后四位）”继续写入 99999999.txt 的第二行：

99999999.txt

```
1  exp1 0.0468
2  exp2 0.0242
```

【**注意**】上述 0.0468、0.0242 仅供举例，你们运行得到的值应该不会与之相同。另外请保存上述过程中所有的中间文件，后期实验中可能会用到。

4.2 更多的方法

在“3. Processes”的（3）步骤中，我们采用最小的欧式距离作为预测的公众“感动”的概率值，如果采用最小的 City block distance，或者最小的 Supremum distance 作为概率预测的基准，相关系数会如何变化？如果不是只采用最小的欧式距离对应的训练文本的公众“感动”的概率值作为预测值，而是将每个欧式距离的倒数等作为权重（由于欧式距离的倒数可能会是一个很大的正数，此时可能需要采用归一化的方法将其转换为(0, 1]之间）加权计算，相关系数又会如何变化？

归一化方法参考：

[1] http://baike.baidu.com/link?url=3dqGpByo1AlVQD008bNUeNYv9WtJvKa1cZhvv7rz147rhOhbCn_I84owwcgcphLHnTXAGVbti1Q4q-UiMs6j6K

[2] http://www.cnblogs.com/chaosimple/archive/2013/07/31/3227271.html

请在时间允许的前提下，开拓思路进行实验，并将所有实验当中，输出的相关系数最大的那个值，以“exp3、空格、以及最大的相关系数的值（保留小数点后四位）”的格式继续写入 99999999.txt 的第三行，比如：

99999999.txt

```
1  exp1 0.0468
2  exp2 0.0242
3  exp3 0.8000
```

以自己的学号新建一个 word（或 PDF）文件，如 99999999.doc，word（或 PDF）文件中将“4.2 更多的方法”中你的实验方法，以及关键代码（包括“3. Processes”以及“4.1 更改向量中值的表示方法”中的实现代码）简要地描述或截图出来即可。本次实验历时两周，即 2015 年 10 月 15 日、16 日，以及 2015 年 10 月 22 日、23 日，时间还算充裕。但请大家在每次实验课结束前，提交下述文件（需要提交 2 次，2015 年 10 月 15 日、16 日的本周实验课，提交 1 次；2015 年 10 月 22 日、23 日的下周实验课，提交 1 次。每次提交的进展，根据每位同学的基础会有不同，但本周实验课提交的文件中，至少应该实现了“3. Processes”中的核心代码）：

【1】将 99999999（改为你自己的学号）.txt 上传到 Lab 1 results 目录中；

【2】将 99999999（改为你自己的学号）.doc 或 99999999.pdf 上传到 Lab1 reports 目录中。

FTP 服务器：ftp://smie2.sysu.edu.cn

帐号：student0007

密码：student0007

【注意】提交上述 2 个文件前，请再三确认自己的学号无误！

附录：

（1）字符串分隔参考

C++ 分隔符（字符串）处理参考网址

http://blog.csdn.net/xw20084898/article/details/21939811

java 分隔符处理参考网址

http://www.cnblogs.com/liubiqu/archive/2008/08/14/1267867.html

python 字符串分割参考网址

http://www.jbxue.com/article/9358.html

By TA: 朱和胜、吴宏鹏

（2）Python 中文处理问题

有同学可能使用了 Python 作为完成实验的语言，这一点我们是鼓励的。一方面它比较易懂而且轻量级；另一方面它提供了很多强大的模块库，这些库可以大大简化代码。

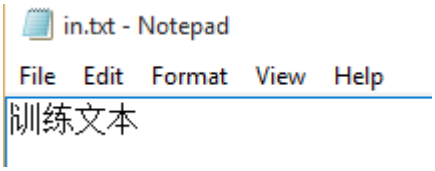
但是如果是在 Windows 的环境下使用 Python 的话就会不得不面临一个问题，那就是中文编码的处理，下面举个例子：

```
s1 = "训练文本"
fr = open('in.txt')
s2 = fr.readline()
fr.close

print "s1:", s1
print "s2:", s2

if s1 == s2:
    print "s1:", s1, "等于", "s2:", s2
else:
    print "s1:", s1, "不等于", "s2:", s2
```

这个脚本中，s1 和 s2 是两个中文字符串，其中 s1 直接由脚本给出，而 s2 的值从下面的文件读得：



将它们各自的内容输出，然后比较两个字符串，看看它们是否相等。

虽然它们的文本内容是相同的，但是运行了这个脚本之后却得到了如下结果：

```
>>>
s1: 训练文本
s2: 训练文本
s1: 训练文本 不等于 s2: 训练文本
>>>
```

这就是其编码方式不同造成的，我们可以在 Python shell 中看它们各自的编码：

```
>>> s1
'\xe8\xae\xad\xe7\xbb\x83\xe6\x96\x87\xe6\x9c\xac'

>>> s2
'\xd1\xb5\xc1\xb7\xce\xc4\xb1\xbe'
```

造成这个问题的原因是，Windows 系统的默认编码系统是 GBK，而 Python 解释器的默认编码系统是 UTF-8。

这样一来，由于 s1 是直接由代码给出，因此 s1 的编码方式是 UTF-8，而 s2 由文件读得，所以 s2 的编码方式是 GBK。

（值得一提的是，Windows 同样可以正常显示 UTF-8 编码的文件，而 GBK 编码的字符串，Python 解释器也可以将其正常的输出，但是互相比较就不行）

要解决上面的问题也非常简单：

```
>>> s3=s2.decode('gbk')

>>> s3
u'\u8bad\u7ec3\u6587\u672c'

>>> s4=s3.encode('utf8')

>>> s4
'\xe8\xae\xad\xe7\xbb\x83\xe6\x96\x87\xe6\x9c\xac'
```

上面的指令中 s3 为 s2 用 GBK 解码之后的结果得到的是“测试文本”这个字符串的标准 Unicode 表示法，而 s4 是 s3 经过 UTF-8 编码之后的结果，得到的是该 Unicode 的字节码。可以看到 s4 此时就和 s2 的值是一样的了。

所以要比较 s2 和 s1，需要进行转码：

```
print "s1:", s1
print "s2:", s2

if s1 == s2.decode('gbk').encode('utf8'):
    print "s1:", s1.decode('utf8'), "等于".decode('utf8'), "s2:", s2.decode('gbk')
else:
    print "s1:", s1, "不等于", "s2:", s2
```

这样一来，输出就是两者相等：

```
s1: 训练文本
s2: 训练文本
s1: 训练文本 等于 s2: 训练文本
```

By TA: 彭禹惟

（3）实验提交注意事项

【当次实验得分为 0 的情况】：

在当次实验提交的 deadline 之前，（1）没有提交“学号.txt”（即实验结果）文件，或者“学号.txt”文件为空；（2）只提交了“学号.txt”文件而没有提交“学号.doc”或“学号.pdf”（即实验报告）文件。

“学号.txt”（即实验结果）文件的格式请严格按照实验文档中的要求统一，最多是三行。

为防止误删除了同学们已经提交了的文件，我们决定若要重复提交实验结果或实验报告文件，请在学号后面加“_new”作为标识，比如“学号_new.txt”、“学号_new.doc”或“学号_new.pdf”。从当次实验课程结束到 deadline 之前，除了原始提交的实验结果或实验报告文件，最多只允许再提交一次更新的实验结果或实验报告文件。如有疑问，请联系 TA：朱和胜（QQ：472544864）。若还有疑问，请及时联系老师。

【注】：每次实验课结束后，TA 都会将当场提交的实验结果和实验报告文件剪切出来。因此，若发现自己当场提交的文件，后面找不到了，是正常情况。请每位当场提交了文件的同学，在 deadline 之前，再次提交这两个文件（哪怕没有任何更新），以免有遗漏。