题目：多文档处理并计算 TF-IDF

姓名： 叶康 学号：MG1633093 邮箱：604379334@qq.com 联系方式：18851822212

(南京大学 计算机科学与技术系, 南京 210093)

1 实现细节

1.1 Task1

通过正则表达式 re.split(r'\W',line)以特殊符号为界进行分词，并通过 re.search(r'\w*\d+\w*',lowword)和 re.findall(r'[a-zA-Z\s]+',lowword)只选出由英文字母组成的单词；

1.2 Task2

将网上找到的英文停用词表文件读进去， 并将上一步处理后得到的单词进一步处理删除停用词；

1.3 Task3

在分词去停用词的同时，记录每篇文章中出现的单词次数，每篇文章构成一个字典，其中 key 值为单词，value 值为其在该文档中出现的次数，并与此同时返回一个 list，list 当中每个元素为即为这样一个字典。除此之外还有一个 classcount 字典，key 值为所有文档中出现过的单词，value 值为出现的文档次数；

1.4 Task4

最后一步根据得到的 list 和 classcount 计算 TF-IDF 值，并将 vocablist 和 TF-IDF 写到相应 txt 文件中去。

2 结果

2.1 实验设置

代码是在 Windows 10 系统下 Python3.5 环境中运行的，其中停用词表来源于网络详见附录，paper 数据来源于 ICML 会议。

2.2 实验结果

能够成功计算 TF-IDF 并将其写入文件中：

```
280:0.000494163172817   336:0.000243217143706   383:0.00100280074558    510:0.000170456204085   630:0.000945984632308
706:0.000328445245749   727:0.000229009089057   876:0.000183293500675   884:4.79647080489e-05   1149:0.000318957457176
1200:4.04236014695e-05  1211:0.000175980262818  1257:0.000878335281922  1347:0.000736968864214  1352:0.00115878254972
1542:0.000232449289845  1548:0.00139278036121   1568:3.33209312094e-05  1570:3.90377672016e-05  1574:0.00023594247756
```

能够成功得出词典并将其写入文件中：

```
abramowitz      abramson        abrera  abridged        abrupt  abruptly        abscisic        abscissae       absence
absences        absent  absil  absne   absolute        absolutely      absolution      absorb  absorbed        absorber
absorbers       absorbing       absorbs absorption      absoute absps   abstain abstains        abstrach        abstract
abstractbeamsearch      abstracted      abstraction     abstractions    abstractive     abstractly      abstracts
absurd  abul    abundance       abundances      abundant        abuse   abusing abusively       abwhere acad    acade
academia        academic        academic        academies       academy acar    acbpj   accel   accelartor      accelerate
```

3 附录：

停用词表见 stopwords.txt 文件