一、使用说明（重要！）

1. 由于 生成文档向量模型 与 随机梯度下降的随机性 代码运行时间不定，大概需运行10+min – 20+min。

2. 为了使运行时间加快，我们将文档ID与其对应的 文档内容 存放在sentence\_dict.txt中，请不要删除！不然你需要大概30+min来处理数据。

3. 本实验使用gensim.models中的Doc2Vec得到文档向量。正确安装依赖（或在anaconda环境下安装）可使gensim使用并行处理快速得到 文档向量model。如果无法并行处理，

这一步可能将花费几小时的漫长时间。请 在anaconda环境下安装gensim.models模块。

4. 本实验采用 Logistic Regression 学习器进行2分类学习，超参数已留在代码中，学习率为0.1，收敛梯度要求为0.01。

5. 本实验采用 不确定性采样（US） 的查询策略。我们选择在准确率到达0.75（而非0.8）时跳出，这是因为我本身的Logistic Regression的结果不好，很多queries能到达0.75但到达不了0.8。

6. 代码运行过程中输出终端的内容详细表明了程序运行流程，最后我们会输出qsize，其存储10个queries达到0.75以上正确率时,当前已标注数据集的大小。我们也会将他写在predict.txt中

7. 请将所需的数据文件：stopwords.txt、sentence\_dict.txt 、documents.txt、Hiemstra\_LM0.15\_Bo1bfree\_d\_3\_t\_10\_16.res 同 代码.py 放在同一目录下，保证代码正常运行。

8. 有问题请一定联系我，谢谢！

9. 祝每天开开心心！

二、实验过程与代码

L1 – L137 数据处理代码+Logistic regression学习器实现代码。

L138 - L149 + L161 – 276 是 使用不确定性采样查询策略 模拟主动学习的实现代码。

四、实验结果及分析

代码运行结束后即可显示结果，以下为2次实验的结果图：



