胡绍泰华傲数据（Audaque）7-9月工作报告&程序说明

完成内容总结：

1. Scikit-learn (sklearn) 和Python英文情感分类语言模型 (SKL\_model\_ShaotaiHu.py)
2. Scikit-learn (sklearn) 和Python中文情感分类语言模型 (Chinese\_skl\_ShaotaiHu.py)
3. Fasttext和Python中文情感分类语言模型 (中文打分模型ShaotaiHu.py)
4. Fasttext和Python英文情感分类语言模型 (senti\_model\_Shaotai\_Hu.py)

———————————————————————————————————————

完成内容详细说明&使用：

* 情感分类模型的主要三个情感是（积极，中立，消极）。程序先简单过滤，完了用qwen来给投喂的原始语料做标签，因为是监督(supervised)。接下在进行深度过滤，包括stopwords。模型会把标好的语料分成训练集和测试集，用训练数据来练自己，程序会用测试数据来报告准确率。情感不能过滤太多stopwords，会改变文本本意。
* 主要总结：qwen大模型标语料情感太慢了，4000条至少1个小时，但是准。那Fasttext和sklearn要快很多，就是没有99%的准确率，可以到92%左右。那Fasttext 标4000条要8秒左右，sklearn只要0.5秒。Sklearn非常快，还相对更准。
* Sklearn能用的stopwords过滤工具更智能，有助于提高准确率。
* 整体说Sklearn的模型 > Fasttext的模型。

|  |
| --- |
| Scikit-learn和Python英文情感分类语言模型具体使用说明：   * 这个是我主要做的模型，也是整体准确率达到91%的模型。 * 需要的python包：re, pandas, nlppreprocess, matplotlib, sklearn, openai, random   使用步骤（参见代码）：   1. process\_file(“file\_path”) 把你的语料路径放到file\_path 2. 1-80行的程序是基础数据过滤和使用qwen大模型的对数据做情感标签，最后吐出来一个 skl\_EN.txt 标好的语料 3. qwen大模型不会标错情感但是偶尔会格式出错和写出一些乱七八糟的东西，80-100是再次过滤，吐出一个叫ccEN.txt的语料。 4. 因为sklearn是用pandas dataframe，也就是csv文件的格式，我在 100-140把txt文件变成csv文件, 以及用nlp对csv文件语料进行专对英文情感分析的stopwords过滤。 5. 140-170行的程序我用图和数据量来呈现数据的平均。例如: 6. 170-200行是用SKLearn的MultinomialNB分类模型得到的91%准确率。这个report要看macro avg。我随机分出训练和测试数据，训练模型1万次，做的pipeline把语料变成频率向量，用ti-idf来找到文字在语料里的重要性。”pline“就是模型。 7. [sklearn.naive\_bayes — scikit-learn 1.5.1 documentation](https://scikit-learn.org/stable/api/sklearn.naive_bayes.html)，模型原理是Bayes’ therorem。这个sklearn的网站还有更多模型，但是，我试了很多，MultinomialNB是整体最准的。可以改我的pipeline来做新的模型。 |
| Scikit-learn和Python中文情感分类语言模型具体使用说明：   * 跟上一个模型使用一样，就是可以分类中文。用的中文stopwords。 |
| Fasttext和Python中文情感分类语言模型具体使用说明：   * 为了方便使用，我把fasttext和sklearn模型的程序写的用法基本一样。   使用步骤（参见代码）：   1. process\_file(“file\_path”) 把你的语料路径放到file\_path 2. 1-80行的程序是基础数据过滤和使用qwen大模型的对数据做情感标签，最后吐出来一个 testCNcorpus.txt 标好的语料。过滤包括一些简单中文stopwords。 3. Qwen模型不会标错情感但是偶尔会格式出错和写出一些乱七八糟的东西，80-94是再次过滤，吐出一个名字一样textCNcorpus。 4. 95-110行的程序我用百分比来呈现数据的平均。 5. 110-150行是我随机分出训练和测试数据，训练模型50次，打印出来所有模型准确率和最高准确率。后面是保存模型的程序。 |
| Fasttext和Python英文情感分类语言模型具体使用说明：   * 和中文的一样使用方法，就是用的英文stopwords。 |