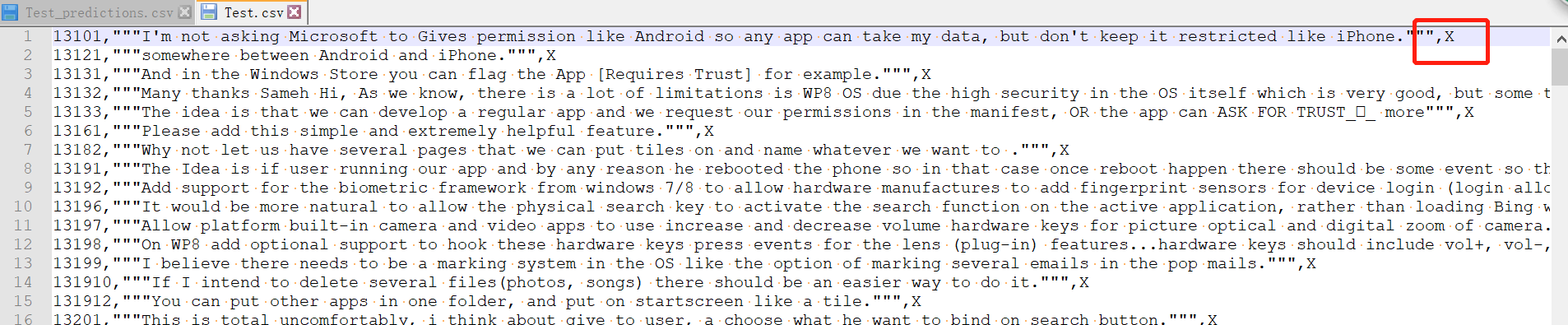
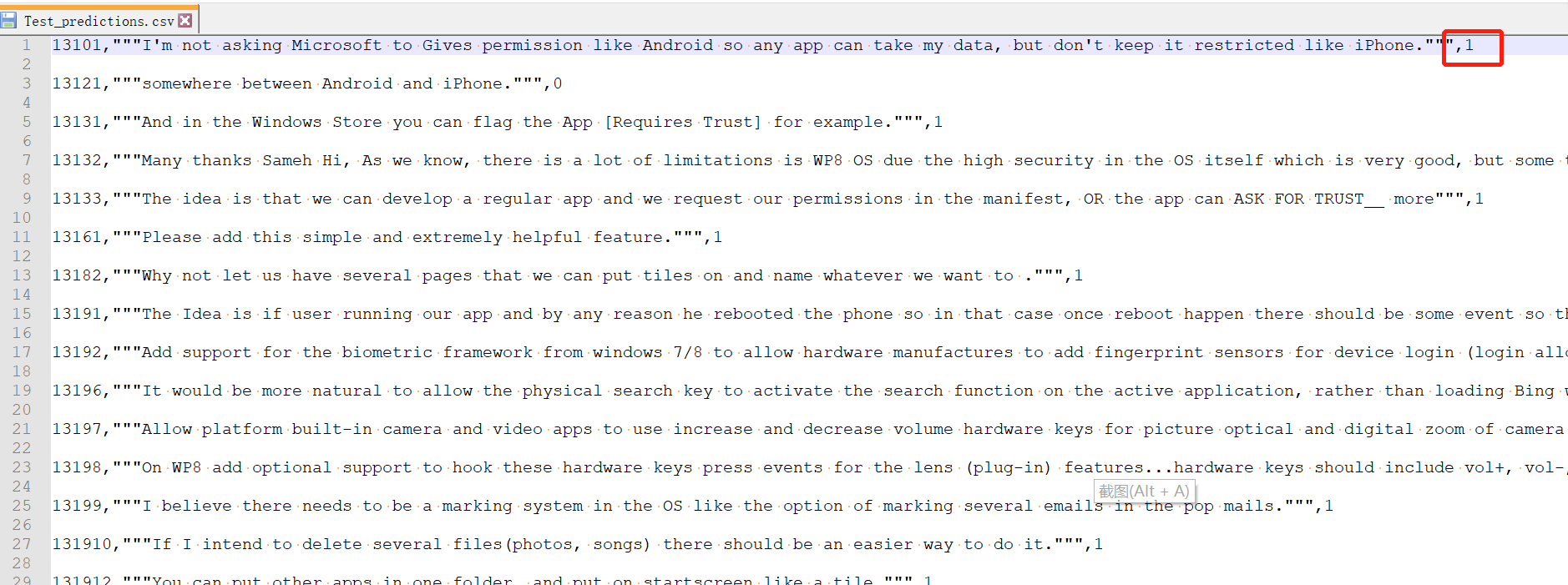
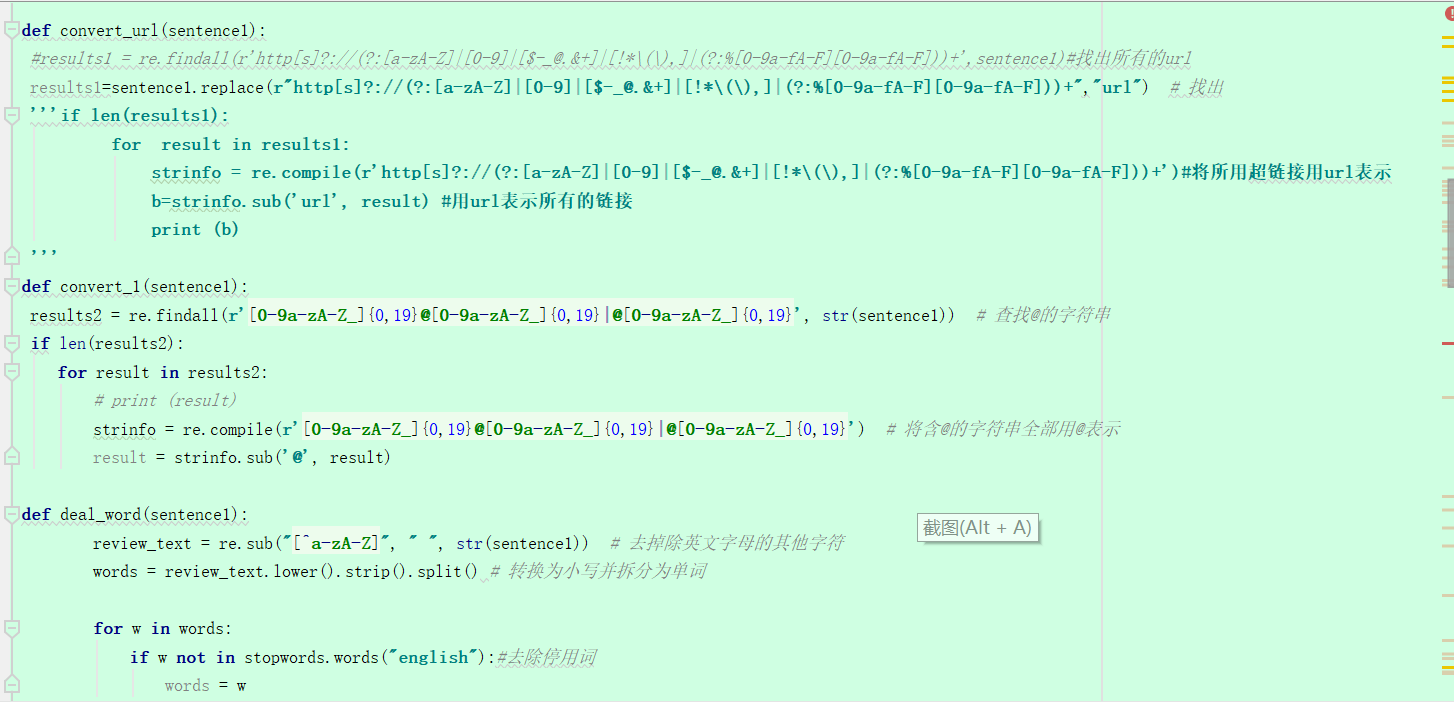
1. 运行了task9 的task A的baseline脚本，了解一下大概的结果。





1. 写脚本对相关的数据进行处理



1. 学习了相关的NLP的处理过程具体如下：

3.1.语料清洗

数据清洗，顾名思义就是在语料中找到我们感兴趣的东西，把不感兴趣的、视为噪音的内容清洗删除，常见的数据清洗方式有：人工去重、对齐、删除和标注等，或者规则提取内容、正则表达式匹配、根据词性和命名实体提取、编写脚本或者代码批处理等。

3.2 **分词**

分词而进行文本挖掘分析时，我们希望文本处理的最小单位粒度是词或者词语，所以这个时候就需要分词来将文本全部进行分词。常见的分词算法有：基于字符串匹配的分词方法、基于理解的分词方法、基于统计的分词方法和基于规则的分词方法，每种方法下面对应许多具体的算法。

3.3词性标注

词性标注，就是给每个词或者词语打词类标签，如形容词、动词、名词等。这样做可以让文本在后面的处理中融入更多有用的语言信息。

3.4**.**去停用词

停用词一般指对文本特征没有任何贡献作用的字词，比如标点符号、语气、人称等一些词。所以在一般性的文本处理中，分词之后，接下来一步就是去停用词。

3.5特征工程

做完语料预处理之后，接下来需要考虑如何把分词之后的字和词语表示成计算机能够计算的类型。有两种常用的表示模型分别是词袋模型和词向量。词袋模型（Bag of Word, BOW)， Google 团队的 Word2Vec。

3.6特征选择

使用特征选择能够找出一个特征子集，其仍然可以保留语义信息；但通过特征提取找到的特征子空间，将会丢失部分语义信息。目前，常见的特征选择方法主要有 DF、 MI、 IG、 CHI、WLLR、WFO 六种。

3.7模型训练

在特征向量选择好之后，接下来要做的事情当然就是训练模型，对于不同的应用需求，我们使用不同的模型，传统的有监督和无监督等机器学习模型， 如 KNN、SVM、Naive Bayes、决策树、GBDT、K-means 等模型；深度学习模型比如 CNN、RNN、LSTM、 Seq2Seq、FastText、TextCNN 等。