关于数据清洗的步骤

对于英文数据的清洗，可以按照以下步骤进行：

1. 去除重复数据：通过去重操作，可以减少数据量，提高处理效率。
2. 文本处理：对于包含文本的数据，需要进行文本处理，如去除标点符号、转换大小写、分词、去除停用词等。

主要的代码操作步骤则是在数据预处理的那个py文件里面

困惑度、一致性相关解释

LDA（Latent Dirichlet Allocation）是一种文本主题建模算法，其目的是在给定文档集合的情况下，推断出这些文档中隐藏的主题。困惑度和一致性是LDA模型的两个评估指标。

1. 困惑度（Perplexity）

困惑度是LDA模型的一项内部评估指标，它用于衡量LDA模型对新文本的预测效果。它的计算方式如下：

$$Perplexity(D_{test})=exp\left(-\frac{\sum_{d\in D_{test}}log\ p(d)}{\sum_{d\in D_{test}}N_d}\right)$$

其中，D_{test}是测试集合，$p(d)$是模型对文档d的预测概率，N_d是文档d的长度（单词数）。困惑度越小，说明模型对新文本的预测效果越好。

1. 一致性（Coherence）

一致性是LDA模型的另一个评估指标，它用于衡量LDA模型推断出的主题是否具有连贯性和可解释性。一致性的计算方式如下：

首先，对于每个主题，选取其前K个最相关的词汇，然后计算这些词汇之间的相似度得分，最后取平均值。常见的相似度得分计算方法有以下几种：

- 点互信息（PMI）
- 余弦相似度（Cosine）
- Jaccard相似度（Jaccard）

通过对所有主题的一致性得分取平均值，可以得到模型的总体一致性得分。一致性得分越高，说明模型推断出的主题具有更好的连贯性和可解释性。

这里一致性我们采用的是余弦相似度

LDA的一些相关性描述

LDA (Latent Dirichlet Allocation) 是一种用于主题建模的机器学习算法。在使用LDA生成主题模型时，常常需要将主题模型以可视化的方式呈现出来。这时常常使用词云、主题关系图等方式呈现，其中主题关系图通常使用节点和边来表示主题及其之间的关系，而节点的大小和距离则分别代表不同的含义。

在LDA的主题关系图中，圆圈的大小可以代表主题的重要性或流行度，通常使用主题在语料库中出现的频率或使用度来表示。比如，一个主题在语料库中出现频率越高，它的圆圈大小就越大，反之圆圈就越小。这可以帮助人们更好地理解每个主题在整个主题模型中的重要性。

而圆圈之间的距离则可以代表主题之间的相关性或距离。主题之间的相关性可以根据它们之间的词语共现情况来计算，如果两个主题的词语重叠度越大，则它们之间的距离就越近。反之，如果两个主题的词语没有共现，则它们之间的距离就越远。这可以帮助人们更好地了解不同主题之间的关系和联系。