推荐使用的库：

***NLTK：***

分词，词性标注，命名实体等等，还有各种各样的语料库和词汇资源（教育和研究，效率低但是使用简单）

***spaCy：***

同样是提供现成nlp功能，如分词，词性分析，词性标注，依存句法分析和实体识别等等（生产环境，性能高）

***PyLucene：***

用于文本检索

分数分配：

***Dataset:***

<https://www.yelp.com/dataset>

Yelp 数据集是由 Yelp 公司提供的一个大型数据集，用于学术研究和教育用途。它包含了数百万条用户对餐馆、商店、地方服务等业务的评价和评论，以及相关的商家信息。这些数据通常用于机器学习、数据挖掘、自然语言处理等领域的研究。

这个数据集主要包括以下类型的数据：

商家信息：包括商家的地理位置、类别（如餐馆、美容院等）、星级评价等。

用户评论：用户对商家的评价和评论文本。

用户信息：用户的一些基本信息，如好友关系等。

打分信息：用户给商家的星级评价。

图片数据：与商家相关的图片。

Yelp 数据集经常用于推荐系统的研究、情感分析、文本分类、社交网络分析等领域。由于数据集的多样性和复杂性，它为研究者提供了丰富的资源来测试和改进各种数据分析技术。

***Dataset Analysis(40%)：***

***分词和词干提取：***

从数据集中随机选择一个商家（b1），提取该商家的所有评论，形成一个小数据集B1。

在进行词干提取前后，展示B1中的词频分布，可以考虑使用对数尺度来展示数据。

重复以上步骤选择另一个商家（b2）。

展示并讨论两个商家的词频分布图。

列出每个商家在词干提取前后十个最频繁的词（排除停用词），并讨论发现。

***词性标注：***

从数据集中随机选择3个句子，并进行词性标注。

显示并讨论使用两种不同的标注方法得到的结果。

***写作风格：***

随机选取以下内容：两篇StackOverflow帖子、两篇Reddit帖子、两篇《海峡时报》的新闻文章、两个任意领域的专利。

讨论它们的写作风格差异，比如句首单词是否大写，句子是否符合语法规则，专有名词是否大写等。

提供所选帖子/文章的URL。

***每个评级的最频繁形容词：***

随机选择50条评级为1的评论，提取其中最常用的10个形容词。

对评级为2、3、4、5的评论重复同样的过程。

讨论结果和方法的局限性。

***Development of a Simple Search Engine(30%)：***

使用Lucene或其他信息检索特定的库来索引和搜索评论。

讨论应该索引和搜索哪些字段（如评论ID、用户ID、商家ID、评分、日期、文本、有用、有趣、酷）。

讨论你的搜索引擎的潜在用户及其可能提交的查询类型。

详细描述对字段中的词语/术语进行的解析/语言处理，比如是否执行词干提取、大小写折叠、停用词移除。

讨论索引数据集中每10%的文档所需时间。

搜索引擎至少支持对“文本”字段进行自由文本关键词查询，包括单关键词查询和短语查询。讨论查询结果是否符合预期，记录处理查询所需的时间。

如果能使用Yelp数据集的所有文件，讨论你希望为搜索引擎提供哪些额外功能，以及为什么用户会喜欢这些功能。

***Extraction of Review Aspects(20%)：***

假设所有评论都针对餐馆，找到评论者通常想要评论的关键方面，如地点、价格、食物、服务等。

使用200条评论样本来展示你提出的方法的结果。

***Application(10%)：***

定义并开发一个简单的基于数据集的NLP/IR应用。

例如，检测包含否定表达的句子，否定通常通过如no、not、never等词表达。

可以自定义一个难度相当的应用，应用只需是一个小工具或代码片段，不需要是基于Web或移动的应用，命令行界面足够。

报告和源码的提交：

***最终报告的硬拷贝***

***提交截止日期：***

2024年5月9日（星期四），需在晚间课堂提交。

***报告格式：***

报告必须使用提供的封面页，主要内容应遵循ACM "sigconf"会议论文模板（可选用MS Word或Latex格式）。报告的主要内容不得超过10页，不包括封面页和附录。

***代码和结果：***

不要在报告中包括所有源代码和完整的结果集，但必须包含对主要功能重要的代码片段。所有使用的第三方库应予以引用。

***打印格式：***

报告应尽可能双面打印。使用塑料封面或装订环将导致2%的成绩扣分。

***审核：***

提交前，请仔细阅读自己的报告硬拷贝，确保报告中的文字和图片清晰可读。

***最终报告的软拷贝、源代码和文档***

***提交内容：***

需要提交名为 MH6301-Gxx.zip 的文件，其中 xx 是你的小组编号。这个压缩文件包括：

***Report-Gxx.PDF：***

与硬拷贝报告相同的PDF文件。

***Readme.txt：***

包括以下内容：

如果使用了第三方库，提供一个下载链接。

安装指南，说明如何设置系统以及如何使用系统（例如，命令行、输入格式、参数等）。

从系统得到的样本输出的解释。

***SourceCode 文件夹：***

仅包含你自己的源代码。数据集和库文件不应包含在软拷贝提交中，以减小文件大小。

***软拷贝提交截止日期：***

2024年5月9日晚上11:59。允许迟交，但每过一个日历日将扣除5%的成绩，直至扣完为止。软拷贝最多可提交三次，仅最后一次提交将被评分并加上时间戳。