山东大学 计算机科学与技术 学院

人工智能导论 课程实验报告

|  |  |  |
| --- | --- | --- |
| 学号：202300130183 | 姓名： 宋浩宇 | 邮箱：202300130183＠mail.sdu.edu.cn |
| 实验题目：九、⾃然语⾔处理分词处理实验 | | |
| 实验过程：  （记录实验过程、遇到的问题和实验结果。可以适当配以关键代码辅助说明，但不要大段贴代码。）  Jieba中文分词输出结果如下：  jieba精确模式输出：  央视 / 315 / 晚会 / 曝光 / 湖北省 / 知名 / 的 / 神丹 / 牌 / 、 / 莲田牌 / “ / 土 / 鸡蛋 / ” / 实为 / 普通 / 鸡蛋 / 冒充 / ， / 同时 / 在 / 商标 / 上 / 玩 / 猫腻 / ， / 分别 / 注 /  / 册 / “ / 鲜土 / ” / 、 / 注册 / “ / 好土 / ” / 商标 / ， / 让 / 消费者 / 误以为 / 是 / “ / 土 / 鸡蛋 / ” / 。 / 3 / 月 / 15 / 日 / 晚间 / ， / 新 / 京报 / 记者 / 就 / 此事 / 致电 / 湖北 / 神 /  / 丹 / 健康 / 食品 / 有限公司 / 方面 / ， / 其 / 工作人员 / 表示 / 不知情 / ， / 需要 / 了解 / 清楚 / 情况 / ， / 截至 / 发稿 / 暂未 / 取得 / 最新 / 回应 / 。 / 新京 /  / 报 / 记者 / 还 / 查询 / 发现 / ， / 湖北 / 神丹 / 健康 / 食品 / 有限公司 / 为 / 农业 / 产业化 / 国家 / 重点 / 龙头企业 / 、 / 高新技术 / 企业 / ， / 此前 / 曾 / 因 /  / 涉嫌 / 虚假 / 宣传 / “ / 中国 / 最大 / 的 / 蛋品 / 企业 / ” / 而 / 被 / 罚 / 6 / 万元 / 。  jieba全模式输出:  央视 / 315 / 晚会 / 曝光 / 湖北 / 湖北省 / 知名 / 的 / 神丹 / 牌 / 、 / 莲 / 田 / 牌 / “ / 土鸡 / 鸡蛋 / ” / 实为 / 普通 / 鸡蛋 / 冒充 / ， / 同时 / 在 / 商标 / 标上 / 玩 / 猫腻 / ， / 分别 / 注 / /  / / 册 / “ / 鲜 / 土 / ”、 / 注册 / “ / 好 / 土 / ” / 商标 / ， / 让 / 消费 / 消费者 / 误以为 / 以为 / 是 / “ / 土鸡 / 鸡蛋 / ”。 / 3 / 月 / 15 / 日 / 晚间 / ， / 新 / 京报 / 记者 / 就此 / 此事 / 致电 / 湖北 / 神 / /  / / 丹 / 健康 / 食品 / 有限 / 有限公司 / 公司 / 方面 / ， / 其 / 工作 / 工作人员 / 作人 / 人员 / 表示 / 不知 / 不知情 / 知情 / ， / 需要 / 了解 / 清楚 / 情况 / ， / 截至 / 发稿 / 暂 / 未取 / 取得 / 最新 / 回应 / 。 / 新 / 京 / /  / / 报 / 记者 / 还 / 查询 / 发现 / ， / 湖北 / 神丹 / 健康 / 食品 / 有限 / 有限公司 / 公司 / 为 / 农业 / 农业产业 / 产业 / 产业化 / 国家 / 重点 / 龙头 / 龙头企业 / 企业 / 、 / 高新 / 高新技术 / 技术 / 企业 / ， / 此前 / 曾 / 因 / /  / / 涉嫌 / 虚假 / 宣传 / “ / 中国 / 最大 / 的 / 蛋品 / 企业 / ” / 而 / 被 / 罚 / 6 / 万元 / 。  jieba paddle模式输出:  央视 / 315 / 晚会 / 曝光 / 湖北省 / 知名 / 的 / 神丹 / 牌 / 、 / 莲田牌 / “ / 土 / 鸡蛋 / ” / 实为 / 普通 / 鸡蛋 / 冒充 / ， / 同时 / 在 / 商标 / 上 / 玩 / 猫腻 / ， / 分别 / 注 /  / 册 / “ / 鲜土 / ” / 、 / 注册 / “ / 好土 / ” / 商标 / ， / 让 / 消费者 / 误以为 / 是 / “ / 土 / 鸡蛋 / ” / 。 / 3 / 月 / 15 / 日 / 晚间 / ， / 新 / 京报 / 记者 / 就 / 此事 / 致电 / 湖北 / 神 /  / 丹 / 健康 / 食品 / 有限公司 / 方面 / ， / 其 / 工作人员 / 表示 / 不知情 / ， / 需要 / 了解 / 清楚 / 情况 / ， / 截至 / 发稿 / 暂未 / 取得 / 最新 / 回应 / 。 / 新京 /  / 报 / 记者 / 还 / 查询 / 发现 / ， / 湖北 / 神丹 / 健康 / 食品 / 有限公司 / 为 / 农业 / 产业化 / 国家 / 重点 / 龙头企业 / 、 / 高新技术 / 企业 / ， / 此前 / 曾 / 因 /  / 涉嫌 / 虚假 / 宣传 / “ / 中国 / 最大 / 的 / 蛋品 / 企业 / ” / 而 / 被 / 罚 / 6 / 万元 / 。  jieba搜索引擎模式输出:  央视 / 315 / 晚会 / 曝光 / 湖北 / 湖北省 / 知名 / 的 / 神丹 / 牌 / 、 / 莲田牌 / “ / 土 / 鸡蛋 / ” / 实为 / 普通 / 鸡蛋 / 冒充 / ， / 同时 / 在 / 商标 / 上 / 玩 / 猫腻 / ， / 分别 / 注 /  / 册 / “ / 鲜土 / ” / 、 / 注册 / “ / 好土 / ” / 商标 / ， / 让 / 消费 / 消费者 / 以为 / 误以为 / 是 / “ / 土 / 鸡蛋 / ” / 。 / 3 / 月 / 15 / 日 / 晚间 / ， / 新 / 京报 / 记者 / 就 / 此事 / 致电 / 湖北 / 神 /  / 丹 / 健康 / 食品 / 有限 / 公司 / 有限公司 / 方面 / ， / 其 / 工作 / 作人 / 人员 / 工作人员 / 表示 / 不知 / 知情 / 不知情 / ， / 需要 / 了解 / 清楚 / 情况 / ， / 截至 / 发稿 / 暂未 / 取得 / 最新 / 回应 / 。 / 新京 /  / 报 / 记者 / 还 / 查询 / 发现 / ， / 湖北 / 神丹 / 健康 / 食品 / 有限 / 公司 / 有限公司 / 为 / 农业 / 产业 / 产业化 / 国家 / 重点 / 龙头 / 企业 / 龙头企业 / 、 / 高新 / 技术 / 高新技术 / 企业 / ， / 此前 / 曾 / 因 /  / 涉嫌 / 虚假 / 宣传 / “ / 中国 / 最大 / 的 / 蛋品 / 企业 / ” / 而 / 被 / 罚 / 6 / 万元 / 。  然后很可惜因为CoreNLP的下载链接失效了，在此处只能用代码示意一下stanfordcorenlp的运行了    以上是中文分词，然后是英文分词的部分：  首先，还是非常可惜的，我下载不到CoreNLP的模型，因此还是用代码来做示意    然后是nltk英文分词，输出如下：  ['Trump', 'was', 'born', 'and', 'raised', 'in', 'the', 'New', 'York', 'City', 'borough', 'of', 'Queens', 'and', 'received', 'an', 'economics', 'degree', 'from', 'the', 'Wharton', 'School', '.']  ['He', 'was', 'appointed', 'president', 'of', 'his', 'family', "'s", 'real', 'estate', 'business', 'in', '1971', ',', 'renamed', 'it', 'The', 'Trump', 'Organization', ',', 'and', 'expanded', 'it', 'from', 'Queens', 'and', 'Brooklyn', 'into', 'Manhattan', '.']  ['The', 'company', 'built', 'or', 'renovated', 'skyscrapers', ',', 'hotels', ',', 'casinos', ',', 'and', 'golf', 'courses', '.']  ['Trump', 'later', 'started', 'various', 'side', 'ventures', ',', 'including', 'licensing', 'his', 'name', 'for', 'real', 'estate', 'and', 'consumer', 'products', '.']  ['He', 'managed', 'the', 'company', 'until', 'his', '2017', 'inauguration', '.']  ['He', 'coauthored', 'several', 'books', ',', 'including', 'The', 'Art', 'of', 'the', 'Deal', '.']  ['He', 'owned', 'the', 'Miss', 'Universe', 'and', 'Miss', 'USA', 'beauty', 'pageants', 'from', '1996', 'to', '2015', ',', 'and', 'he', 'produced', 'and', 'hosted', 'The', 'Apprentice', ',', 'a', 'reality', 'television', 'show', ',', 'from', '2003', 'to', '2015', '.']  ['Forbes', 'estimates', 'his', 'net', 'worth', 'to', 'be', '$', '3.1', 'billion', '.']  最后是SpaCy英文分词，输出如下：  ['Trump', 'was', 'born', 'and', 'raised', 'in', 'the', 'New', 'York', 'City', 'borough', 'of', 'Queens', 'and', 'received', '\n', 'an', 'economics', 'degree', 'from', 'the', 'Wharton', 'School', '.', 'He', 'was', 'appointed', 'president', 'of', '\n', 'his', 'family', "'s", 'real', 'estate', 'business', 'in', '1971', ',', 'renamed', 'it', 'The', 'Trump', 'Organization', ',', '\n', 'and', 'expanded', 'it', 'from', 'Queens', 'and', 'Brooklyn', 'into', 'Manhattan', '.', 'The', 'company', 'built', 'or', '\n', 'renovated', 'skyscrapers', ',', 'hotels', ',', 'casinos', ',', 'and', 'golf', 'courses', '.', 'Trump', 'later', 'started', '\n', 'various', 'side', 'ventures', ',', 'including', 'licensing', 'his', 'name', 'for', 'real', 'estate', 'and', '\n', 'consumer', 'products', '.', 'He', 'managed', 'the', 'company', 'until', 'his', '2017', 'inauguration', '.', 'He', 'coauthored', 'several', 'books', ',', 'including', 'The', 'Art', 'of', 'the', 'Deal', '.', 'He', 'owned', 'the', 'Miss', '\n', 'Universe', 'and', 'Miss', 'USA', 'beauty', 'pageants', 'from', '1996', 'to', '2015', ',', 'and', 'he', 'produced', 'and', '\n', 'hosted', 'The', 'Apprentice', ',', 'a', 'reality', 'television', 'show', ',', 'from', '2003', 'to', '2015', '.', 'Forbes', '\n', 'estimates', 'his', 'net', 'worth', 'to', 'be', '$', '3.1', 'billion', '.'] | | |
| 结果分析与体会：  中文分词因为用不了StanfordCoreNLP所以只能看jieba这一个库的效果，就这几种输出来说，全模式和搜索引擎模式会把所有可能的情况都输出出来，而另外的几种则都是只输出一种分词情况，首先在不考虑自定义词典的情况下，这种把所有情况都输出出来的方式可以把一些企业名等不存在于词典中的词汇但是像词汇的词划分出来，准确模式则不行，然后经过测试，jieba库只能划分中文词汇，如果输入的中文文本中含有英文或者数字，则它们不会被正确划分，而只是将他们从中文文本中摘出来而已。Jieba优点在于对中文文本处理的优化和使用便捷。与之相对，另外那几种都需要在使用之前先下载好语言模型，配置好模型路径，使用起来就没有jieba这么方便，然后nltk和spacy这两个库，经过实验一些连起来翻译和拆成两个单词分别翻译意思不同的短语也会被划分开，因此这两个库在处理英文的分词任务时应该都是进行的最简单的单词划分，不同的是nltk更像是对语言的处理，spacy更像是对文本数据的处理，理由是后者在输出结果的时候会把“\n”这种东西输出出来。 | | |