山东大学 计算机科学与技术 学院

自然语言处理 课程实验报告

学号：201600181073	姓名：唐超	班级： 智能 16
实验题目：中英文文本分词		
实验学时：	实验日期： 2019.3.18	

实验目的：

分别利用 jieba, SnowNLP, THULAC, PyNLPIR, StanfordCoreNLP, NLTK, SpaCy 等分词工具对给定的中英文文本序列进行分词并对不同分词工具产生的结果进行简要对比分析。

软件环境：

Jupyter Notebook&python3.6

实验步骤与内容：

中文分词：

1. Jieba

jieba 支持三种分词模式：

- 精确模式，试图将句子最精确地切开，适合文本分析；
- 全模式，把句子中所有的可以成词的词语都扫描出来，速度非常快，但是不能解决歧义；
- 搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。

在中文实验文本上的测试结果分别如下：

全模式：

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/ 涉嫌/ 虚假/ 宣传/ / / 中国/ 最大/ 的/ 蛋品/ 企业/ / / 而/ 被/ 罚/ 6/ 万元/ /

精确模式：

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/ 。

搜索引擎模式：

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, 涉嫌, 因涉嫌, 虚假, 宣传, “, 中国, 最大, 的, 蛋品, 企业, ”, 而, 被, 罚, 6, 万元, 。

另外，jieba 还支持用户自定义词典，以便包含 jieba 词库里没有的词，保证更高的正确率。在添加了含有“神丹牌”、“新京报”、“土鸡蛋”等词后，这些词都被正确划分。添加自定义词典后精确模式下的分词结果如下：

```
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/ 。
```

2. SnowNLP

SnowNLP 分词结果如下：

```
['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', '。']
```

不出意外地，“神丹牌”、“莲田牌”、“土鸡蛋”、“新京报”等词也没有被正确分出，另外从“暂”、“未”、“6”、“万”、“元”、“最”、“大”等分词结果可以看出，该工具的分词粒度过细。

3. THULAC

THULAC 是由清华大学自然语言处理与社会人文计算实验室研制推出的一套中文词法分析工具包，具有中文分词和词性标注功能。利用目前世界上规模最大的人工分词和词性标注中文语料库（约含 5800 万字）训练而成，模型标注能力强大。在实验文本上的测试结果如下，可以看出由于语料库的强大，之前工具分词存在问题专有名词的“神丹牌”、“莲田牌”、“土鸡蛋”、“新京报”等词都被正确划分，但是，也存在“了解”、“国家”等常用词划分错误的情况。

```
央视_v 315_m 晚会_n 曝光_v 湖北省_ns 知名_a 的_u 神丹牌_nz 、_w 莲田牌_nz “_w 土_a 鸡蛋_n ”_w 实_a 为_v 普通_a 鸡蛋_n 冒充_v ，_w 同时_c 在_p 商标_n 上_f 玩_v 猫腻_n ，_w 分别_d 注册_v “_w 鲜土_n ”_w 、_w 注册_v “_w 好_a 土_n ”_w 商标_n ，_w 让_v 消费者_n 误_d 以为_v 是_v “_w 土鸡蛋_n ”_w 。_w 3月_t 15日_t 晚间_t ，_w 新京报_nz 记者_n 就_p 此事_r 致电_v 湖北_ns 神丹_nz 健康_a 食品_n 有限公司_n 方面_n ，_w 其_r 工作_v 人员_n 表示_v 不_d 知情_v ，_w 需要_v 了_u 解_v 清楚_a 情况_n ，_w 截至_v 发稿_v 暂_d 未_d 取得_v 最新_a 回应_v 。_w 新_a 京报_n 记者_n 还_d 查询_v 发现_v ，_w 湖北_ns 神丹_nz 健康_a 食品_n 有限公司_n 为_p 农业_n 产业化_v 国_m 家_q 重点_n 龙头_n 企业_n 、_w 高新技术_n 企业_n ，_w 此前_t 曾_d 因_p 涉嫌_v 虚假_a 宣传_v “_w 中国_ns 最_d 大_a 的_u 蛋品_n 企业_n ”_w 而_c 被_p 罚_v 6万_m 元_q 。_w
```

4. PyNLPIR

PyNLPIR 在实验文档上的分词结果如下，可以看出除了在“神丹牌”、“莲田牌”、“土鸡蛋”等专有名词上的分词结果出现问题外，“央视”这样的常用词汇也被错分为两个动词，说明该工具的训练语料库不够强大，但除此之外，在测试文本上的分词结果基本令人满意。

[('央', 'verb'), ('视', 'verb'), ('315', 'numeral'), ('晚会', 'noun'), ('曝光', 'verb'), ('湖北省', 'noun'), ('知名', 'adjective'), ('的', 'particle'), ('神', 'noun'), ('丹', 'distinguishing word'), ('牌', 'noun'), ('、', 'punctuation mark'), ('莲', 'noun'), ('田', 'noun'), ('牌', 'noun'), ('“', 'punctuation mark'), ('土', 'noun'), ('鸡蛋', 'noun'), ('”', 'punctuation mark'), ('实', 'adjective'), ('为', 'verb'), ('普通', 'adjective'), ('鸡蛋', 'noun'), ('冒充', 'verb'), ('，', 'punctuation mark'), ('同时', 'conjunction'), ('在', 'preposition'), ('商标', 'noun'), ('上', 'noun of locality'), ('玩', 'verb'), ('猫腻', 'noun'), ('，', 'punctuation mark'), ('分别', 'adverb'), ('注册', 'verb'), ('“', 'punctuation mark'), ('鲜', 'adjective'), ('土', 'noun'), ('”', 'punctuation mark'), ('、', 'punctuation mark'), ('注册', 'verb'), ('“', 'punctuation mark'), ('好', 'adjective'), ('土', 'noun'), ('”', 'punctuation mark'), ('商标', 'noun'), ('，', 'punctuation mark'), ('让', 'verb'), ('消费者', 'noun'), ('误', 'adverb'), ('以为', 'verb'), ('是', 'verb'), ('“', 'punctuation mark'), ('土', 'noun'), ('鸡蛋', 'noun'), ('”', 'punctuation mark'), ('。', 'punctuation mark'), ('3月', 'time word'), ('15日', 'time word'), ('晚间', 'time word'), ('，', 'punctuation mark'), ('新京报', None), ('记者', 'noun'), ('就', 'adverb'), ('此事', 'pronoun'), ('致电', 'verb'), ('湖北', 'noun'), ('神', 'noun'), ('丹', 'distinguishing word'), ('健康', 'adjective'), ('食品', 'noun'), ('有限公司', 'noun'), ('方面', 'noun'), ('，', 'punctuation mark'), ('其', 'pronoun'), ('工作', 'verb'), ('人员', 'noun'), ('表示', 'verb'), ('不', 'adverb'), ('知', 'verb'), ('情', 'noun'), ('，', 'punctuation mark'), ('需要', 'verb'), ('了解', 'verb'), ('清楚', 'adjective'), ('情况', 'noun'), ('，', 'punctuation mark'), ('截至', 'verb'), ('发稿', 'verb'), ('暂', 'adverb'), ('未', 'adverb'), ('取得', 'verb'), ('最新', 'adjective'), ('回应', 'verb'), ('。', 'punctuation mark'), ('新京报', None), ('记者', 'noun'), ('还', 'adverb'), ('查询', 'verb'), ('发现', 'verb'), ('，', 'punctuation mark'), ('湖北', 'noun'), ('神', 'noun'), ('丹', 'distinguishing word'), ('健康', 'adjective'), ('食品', 'noun'), ('有限公司', 'noun'), ('为', 'preposition'), ('农业', 'noun'), ('产业化', 'verb'), ('国家', 'noun'), ('重点', 'noun'), ('龙头', 'noun'), ('企业', 'noun'), ('、', 'punctuation mark'), ('高新技术', 'noun'), ('企业', 'noun'), ('，', 'punctuation mark'), ('此前', 'time word'), ('曾', 'adverb'), ('因', 'preposition'), ('涉嫌', 'verb'), ('虚假', 'adjective'), ('宣传', 'verb'), ('“', 'punctuation mark'), ('中国', 'noun'), ('最', 'adverb'), ('大', 'adjective'), ('的', 'particle'), ('蛋品', 'noun'), ('企业', 'noun'), ('”', 'punctuation mark'), ('而', 'conjunction'), ('被', 'preposition'), ('罚', 'verb'), ('6万', 'numeral'), ('元', 'classifier'), ('。', 'punctuation mark')]

5. StanfordCoreNLP

该工具支持多种语言文本的分词，在中文实验文本上的测试结果如下，可以看出除了部分专有名词分词错误外并无其他明显错误。

['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', '。']

英文分词：

各个英文分词工具在实验文档上分词结果分别如下：

1. NLTK

['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', 'co-authored', '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', '.']

2. SpaCy（部分结果）

Trump
was
born
and
raised
in
the
New
York
City
borough
of
Queens
and
received
an
economics
degree
from

3. StanfordCoreNLP

Tokenize: ['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', 'co-authored', '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', '.']

结论分析与体会：

中文分词工具测试结果比较：

对 5 种中文分词工具在测试文本上的分词结果进行粗略分析，可以看出分词结果并无显著差异，除了 THULAC 对专有名词的划分结果较好外，其他的都不能对专有名词做正确的划分。但 SnowNLP 存在划分粒度过细，THULAC、PyNLPIR 对一些常见词汇划分错误，因此，单独从实验文档的测试结果来看，jieba 和 StanfordCoreNLP 的分词结果总体上更好，而 jieba 还可以自定义词典，在一定程度上解决了专业术语、专有名词的问题，因此效果最优。

英文分词工具测试结果比较：

由于有空格作为分词标志，英文分词的难度相对中文明显降低，3 种英文分词工具的分词结果都比较准确，差别不大。