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：词性标注&命名实体识别		
实验学时：	实验日期： 2019.3.25	

实验目的：

分别利用 jieba，THULAC，StanfordCoreNLP，NLTK，SpaCy 等分词工具对给定的中英文文本序列进行词性标注&命名实体识别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：

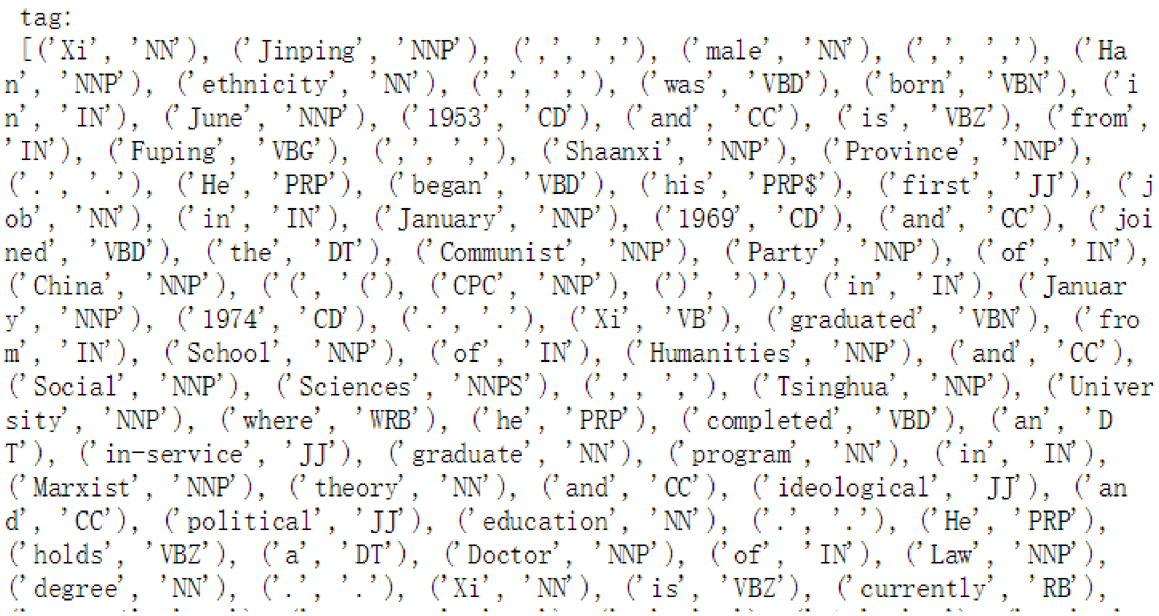
Jupyter Notebook&python3.6

实验步骤与内容：

英文文本：

1. NLTK（部分结果）

词性标注：



```
tag:
[('Xi', 'NN'), ('Jinping', 'NNP'), (',', ','), ('male', 'NN'), (',', ','), ('Han', 'NNP'), ('ethnicity', 'NN'), (',', ','), ('was', 'VBD'), ('born', 'VBN'), ('in', 'IN'), ('June', 'NNP'), ('1953', 'CD'), ('and', 'CC'), ('is', 'VBZ'), ('from', 'IN'), ('Fuping', 'VBG'), (',', ','), ('Shaanxi', 'NNP'), ('Province', 'NNP'), ('.', '.'), ('He', 'PRP'), ('began', 'VBD'), ('his', 'PRP$'), ('first', 'JJ'), ('job', 'NN'), ('in', 'IN'), ('January', 'NNP'), ('1969', 'CD'), ('and', 'CC'), ('joined', 'VBD'), ('the', 'DT'), ('Communist', 'NNP'), ('Party', 'NNP'), ('of', 'IN'), ('China', 'NNP'), ('(', '('), ('CPC', 'NNP'), (')', ')'), ('in', 'IN'), ('January', 'NNP'), ('1974', 'CD'), ('.', '.'), ('Xi', 'VB'), ('graduated', 'VBN'), ('from', 'IN'), ('School', 'NNP'), ('of', 'IN'), ('Humanities', 'NNP'), ('and', 'CC'), ('Social', 'NNP'), ('Sciences', 'NNPS'), (',', ','), ('Tsinghua', 'NNP'), ('University', 'NNP'), ('where', 'WRB'), ('he', 'PRP'), ('completed', 'VBD'), ('an', 'DT'), ('in-service', 'JJ'), ('graduate', 'NN'), ('program', 'NN'), ('in', 'IN'), ('Marxist', 'NNP'), ('theory', 'NN'), ('and', 'CC'), ('ideological', 'JJ'), ('and', 'CC'), ('political', 'JJ'), ('education', 'NN'), ('.', '.'), ('He', 'PRP'), ('holds', 'VBZ'), ('a', 'DT'), ('Doctor', 'NNP'), ('of', 'IN'), ('Law', 'NNP'), ('degree', 'NN'), ('.', '.'), ('Xi', 'NN'), ('is', 'VBZ'), ('currently', 'RB'),
```

实体识别：

```
(PERSON Shaanxi/NNP Province/NNP)
./.
He/PRP
began/VBD
his/PRP$
first/JJ
job/NN
in/IN
January/NNP
1969/CD
and/CC
joined/VBD
the/DT
(ORGANIZATION Communist/NNP Party/NNP)
of/IN
(GPE China/NNP)
(/(
(ORGANIZATION CPC/NNP)
```

2. SpaCy（部分结果）

词性标注：

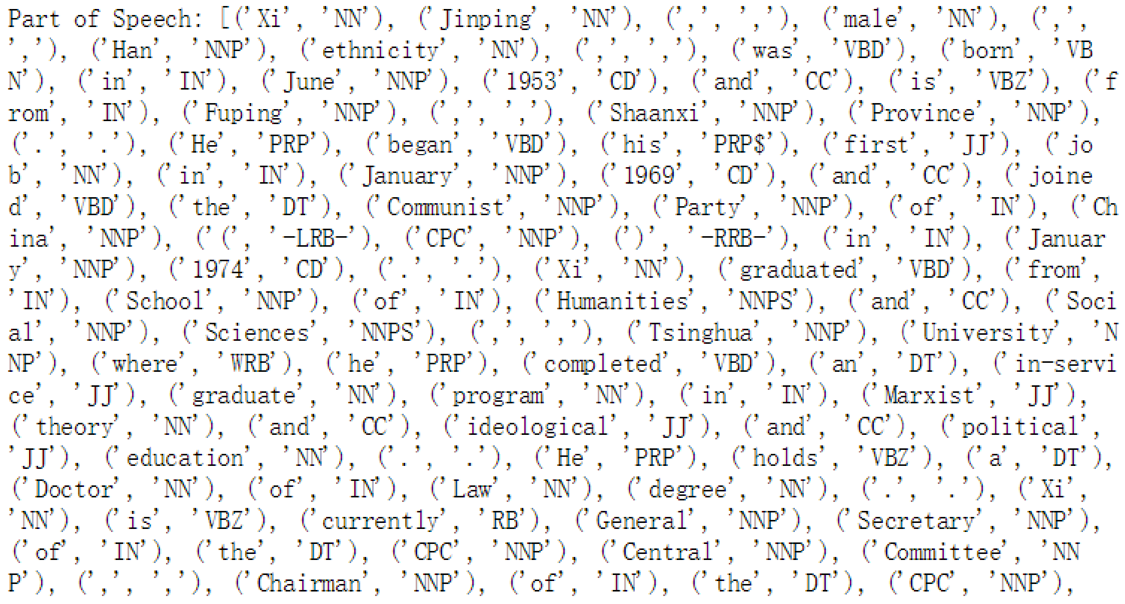
```
Xi NNP
Jinping NNP
, ,
male NN
, ,
Han NNP
ethnicity NN
, ,
was VBD
born VBN
in IN
June NNP
1953 CD
and CC
is VBZ
from IN
Fuping NNP
, ,
Shaanxi NNP
Province NNP
```

实体识别：

```
NORP ['Marxist’, 'Han’]
DATE ['January 1969’, 'June 1953’, 'January 1974’]
ORDINAL ['first’]
GPE ["the People’s Republic of China", 'PRC’, 'Fuping’, 'Shaanxi Province’]
ORG ['School of Humanities and Social Sciences’, 'the PRC Central Military Commiss
n’, 'the CPC Central Military Commission’, 'the Communist Party of China’, 'the CP
entral Committee’, 'CPC’, 'Tsinghua University’]
PERSON ['Xi Jinping’, 'Xi’]
```

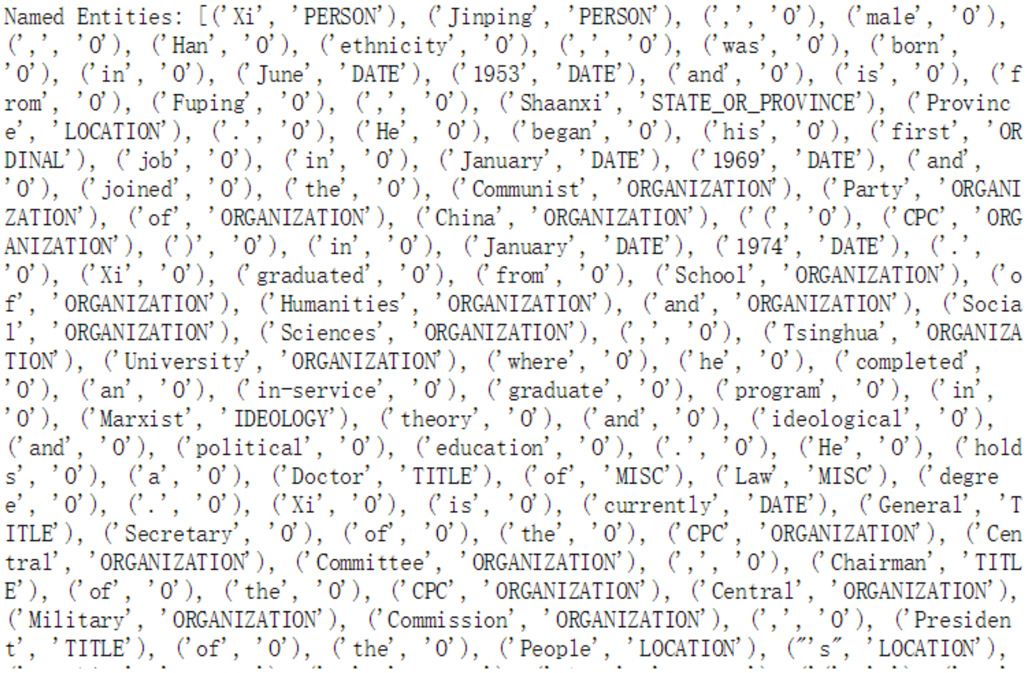
3. StanfordCoreNLP

词性标注：



```
Part of Speech: [('Xi', 'NN'), ('Jinping', 'NN'), (',', ','), ('male', 'NN'), (',',
','), ('Han', 'NNP'), ('ethnicity', 'NN'), (',', ','), ('was', 'VBD'), ('born', 'VB
N'), ('in', 'IN'), ('June', 'NNP'), ('1953', 'CD'), ('and', 'CC'), ('is', 'VBZ'), ('f
rom', 'IN'), ('Fuping', 'NNP'), (',', ','), ('Shaanxi', 'NNP'), ('Province', 'NNP'),
('.', '.'), ('He', 'PRP'), ('began', 'VBD'), ('his', 'PRP$'), ('first', 'JJ'), ('jo
b', 'NN'), ('in', 'IN'), ('January', 'NNP'), ('1969', 'CD'), ('and', 'CC'), ('joine
d', 'VBD'), ('the', 'DT'), ('Communist', 'NNP'), ('Party', 'NNP'), ('of', 'IN'), ('Ch
ina', 'NNP'), ('(', '-LRB-'), ('CPC', 'NNP'), (')', '-RRB-'), ('in', 'IN'), ('Januar
y', 'NNP'), ('1974', 'CD'), ('.', '.'), ('Xi', 'NN'), ('graduated', 'VBD'), ('from',
'IN'), ('School', 'NNP'), ('of', 'IN'), ('Humanities', 'NNPS'), ('and', 'CC'), ('Soci
al', 'NNP'), ('Sciences', 'NNPS'), (',', ','), ('Tsinghua', 'NNP'), ('University', 'N
NP'), ('where', 'WRB'), ('he', 'PRP'), ('completed', 'VBD'), ('an', 'DT'), ('in-servi
ce', 'JJ'), ('graduate', 'NN'), ('program', 'NN'), ('in', 'IN'), ('Marxist', 'JJ'),
('theory', 'NN'), ('and', 'CC'), ('ideological', 'JJ'), ('and', 'CC'), ('political',
'JJ'), ('education', 'NN'), ('.', '.'), ('He', 'PRP'), ('holds', 'VBZ'), ('a', 'DT'),
('Doctor', 'NN'), ('of', 'IN'), ('Law', 'NN'), ('degree', 'NN'), ('.', '.'), ('Xi',
'NN'), ('is', 'VBZ'), ('currently', 'RB'), ('General', 'NNP'), ('Secretary', 'NNP'),
('of', 'IN'), ('the', 'DT'), ('CPC', 'NNP'), ('Central', 'NNP'), ('Committee', 'NN
P'), (',', ','), ('Chairman', 'NNP'), ('of', 'IN'), ('the', 'DT'), ('CPC', 'NNP'),
```

实体识别：



```
Named Entities: [('Xi', 'PERSON'), ('Jinping', 'PERSON'), (',', 'O'), ('male', 'O'),
(',', 'O'), ('Han', 'O'), ('ethnicity', 'O'), (',', 'O'), ('was', 'O'), ('born',
'O'), ('in', 'O'), ('June', 'DATE'), ('1953', 'DATE'), ('and', 'O'), ('is', 'O'), ('f
rom', 'O'), ('Fuping', 'O'), (',', 'O'), ('Shaanxi', 'STATE_OR_PROVINCE'), ('Provinc
e', 'LOCATION'), ('.', 'O'), ('He', 'O'), ('began', 'O'), ('his', 'O'), ('first', 'OR
DINAL'), ('job', 'O'), ('in', 'O'), ('January', 'DATE'), ('1969', 'DATE'), ('and',
'O'), ('joined', 'O'), ('the', 'O'), ('Communist', 'ORGANIZATION'), ('Party', 'ORGANI
ZATION'), ('of', 'ORGANIZATION'), ('China', 'ORGANIZATION'), ('(', 'O'), ('CPC', 'ORG
ANIZATION'), (')', 'O'), ('in', 'O'), ('January', 'DATE'), ('1974', 'DATE'), ('.',
'O'), ('Xi', 'O'), ('graduated', 'O'), ('from', 'O'), ('School', 'ORGANIZATION'), ('o
f', 'ORGANIZATION'), ('Humanities', 'ORGANIZATION'), ('and', 'ORGANIZATION'), ('Socia
l', 'ORGANIZATION'), ('Sciences', 'ORGANIZATION'), (',', 'O'), ('Tsinghua', 'ORGANIZA
TION'), ('University', 'ORGANIZATION'), ('where', 'O'), ('he', 'O'), ('completed',
'O'), ('an', 'O'), ('in-service', 'O'), ('graduate', 'O'), ('program', 'O'), ('in',
'O'), ('Marxist', 'IDEOLOGY'), ('theory', 'O'), ('and', 'O'), ('ideological', 'O'),
('and', 'O'), ('political', 'O'), ('education', 'O'), ('.', 'O'), ('He', 'O'), ('hold
s', 'O'), ('a', 'O'), ('Doctor', 'TITLE'), ('of', 'MISC'), ('Law', 'MISC'), ('degre
e', 'O'), ('.', 'O'), ('Xi', 'O'), ('is', 'O'), ('currently', 'DATE'), ('General', 'T
ITLE'), ('Secretary', 'O'), ('of', 'O'), ('the', 'O'), ('CPC', 'ORGANIZATION'), ('Cen
tral', 'ORGANIZATION'), ('Committee', 'ORGANIZATION'), (',', 'O'), ('Chairman', 'TITL
E'), ('of', 'O'), ('the', 'O'), ('CPC', 'ORGANIZATION'), ('Central', 'ORGANIZATION'),
('Military', 'ORGANIZATION'), ('Commission', 'ORGANIZATION'), (',', 'O'), ('Presiden
t', 'TITLE'), ('of', 'O'), ('the', 'O'), ('People', 'LOCATION'), ("'s", 'LOCATION'),
```

中文文本：

1. Jieba

词性标注：

```
“ x
青年报 n
杯赛 n
” x
U19 eng
邀请赛 n
在 p
越南 ns
芽庄 n
进行 v
， x
前 f
国脚 n
曲波 nr
挂帅 n
的 uj
```

2. StanfordCoreNLP

词性标注：

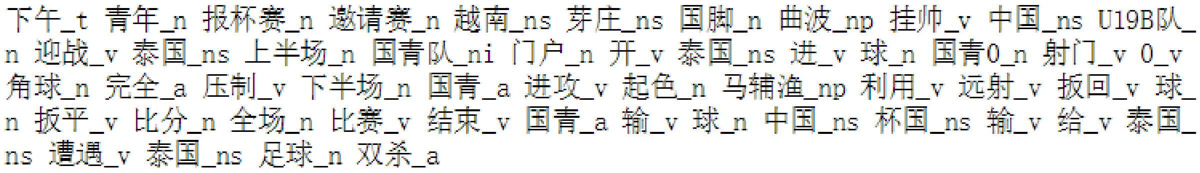
```
[（’3月’，’NT’），（’23日’，’NT’），（’下午’，’NT’），（’，’，’PU’），（’“’，’PU’），（’青年
报’，’NN’），（’杯赛’，’NN’），（’”’，’PU’），（’U19’，’NN’），（’邀请赛’，’NN’），（’在’，
’P’），（’越南’，’NR’），（’芽’，’NN’），（’庄’，’NR’），（’进行’，’VV’），（’，’，’PU’），
（’前’，’JJ’），（’国脚’，’NN’），（’曲波’，’NR’），（’挂帅’，’VV’），（’的’，’DEC’），（’中国’，
’NR’），（’U19’，’NN’），（’B’，’NN’），（’队’，’NN’），（’迎战’，’VV’），（’泰国’，’NR’），（’U1
9’，’NN’），（’，’，’PU’），（’上半场’，’NN’），（’国青队’，’NR’），（’的’，’DEG’），（’门户’，
’NN’），（’大开’，’VV’），（’，’，’PU’），（’泰国’，’NR’），（’在’，’P’），（’第11’，’OD’），
（’分钟’，’M’），（’和’，’CC’），（’第17’，’OD’），（’分钟’，’M’），（’连’，’AD’），（’进’，’V
V’），（’2’，’CD’），（’球’，’NN’），（’，’，’PU’），（’半场’，’NN’），（’国青’，’NN’），（’0’，
’CD’），（’射门’，’VV’），（’0’，’CD’），（’角球’，’NN’），（’，’，’PU’），（’几乎’，’AD’），
（’被’，’LB’），（’完全’，’AD’），（’压制’，’VV’），（’。’，’PU’），（’下半场’，’NN’），（’，’，
’PU’），（’国青’，’VA’），（’的’，’DEC’），（’进攻’，’NN’），（’一度’，’AD’），（’有所’，’VV’），
（’起色’，’NN’），（’，’，’PU’），（’并’，’AD’），（’由’，’P’），（’马辅渔’，’NR’），（’利用’，
’VV’），（’远射’，’NN’），（’扳回’，’VV’），（’一’，’CD’），（’球’，’NN’），（’，’，’PU’），
（’但’，’AD’），（’最终’，’AD’），（’未’，’AD’），（’能’，’VV’），（’扳平’，’VV’），（’比分’，’N
N’），（’。’，’PU’），（’全’，’DT’），（’场’，’NN’），（’比赛’，’NN’），（’结束’，’VV’），（’，’，
’PU’），（’国青’，’NN’），（’1-2’，’CD’），（’输球’，’VV’），（’，’，’PU’），（’继’，’P’），（’中
国’，’NR’），（’杯’，’NN’），（’国足’，’NN’），（’0-1’，’NN’），（’输给’，’VV’），（’泰国’，’N
R’），（’之后’，’LC’），（’，’，’PU’），（’3’，’CD’），（’天’，’M’），（’内’，’LC’），（’遭遇’，
’VV’），（’泰国’，’NR’），（’足球’，’NN’），（’双杀’，’NN’），（’。’，’PU’）]
```

实体识别：

```
[（’3月’，’DATE’），（’23日’，’DATE’），（’下午’，’TIME’），（’，’，’O’），（’“’，’O’），（’青
年报’，’O’），（’杯赛’，’O’），（’”’，’O’），（’U19’，’O’），（’邀请赛’，’O’），（’在’，’O’），
（’越南’，’COUNTRY’），（’芽’，’O’），（’庄’，’O’），（’进行’，’O’），（’，’，’O’），（’前’，
’O’），（’国脚’，’O’），（’曲波’，’PERSON’），（’挂帅’，’O’），（’的’，’O’），（’中国’，’COUNTR
Y’），（’U19’，’O’），（’B’，’O’），（’队’，’O’），（’迎战’，’O’），（’泰国’，’COUNTRY’），（’U1
9’，’O’），（’，’，’O’），（’上半场’，’ORGANIZATION’），（’国青队’，’ORGANIZATION’），（’的’，
’O’），（’门户’，’O’），（’大开’，’O’），（’，’，’O’），（’泰国’，’COUNTRY’），（’在’，’O’），
（’第11’，’ORDINAL’），（’分钟’，’MISC’），（’和’，’O’），（’第17’，’ORDINAL’），（’分钟’，’MI
SC’），（’连’，’O’），（’进’，’O’），（’2’，’NUMBER’），（’球’，’O’），（’，’，’O’），（’半场’，
’MISC’），（’国青’，’MISC’），（’0’，’NUMBER’），（’射门’，’O’），（’0’，’NUMBER’），（’角球’，
’O’），（’，’，’O’），（’几乎’，’O’），（’被’，’O’），（’完全’，’O’），（’压制’，’O’），（’。’，
’O’），（’下半场’，’O’），（’，’，’O’），（’国青’，’O’），（’的’，’O’），（’进攻’，’O’），（’一
度’，’O’），（’有所’，’O’），（’起色’，’O’），（’，’，’O’），（’并’，’O’），（’由’，’O’），（’马
辅渔’，’PERSON’），（’利用’，’O’），（’远射’，’O’），（’扳回’，’O’），（’一’，’NUMBER’），
（’球’，’O’），（’，’，’O’），（’但’，’O’），（’最终’，’O’），（’未’，’O’），（’能’，’O’），（’扳
平’，’O’），（’比分’，’O’），（’。’，’O’），（’全’，’O’），（’场’，’O’），（’比赛’，’O’），（’结
束’，’O’），（’，’，’O’），（’国青’，’O’），（’1-2’，’NUMBER’），（’输球’，’O’），（’，’，’O’），
（’继’，’O’），（’中国’，’COUNTRY’），（’杯’，’O’），（’国足’，’O’），（’0-1’，’MISC’），（’输
给’，’MISC’），（’泰国’，’MISC’），（’之后’，’MISC’），（’，’，’O’），（’3’，’NUMBER’），
（’天’，’MISC’），（’内’，’MISC’），（’遭遇’，’O’），（’泰国’，’COUNTRY’），（’足球’，’O’），
（’双杀’，’O’），（’。’，’O’）]
```

3. Thulac

词性标注：



```
下午_t 青年_n 报杯赛_n 邀请赛_n 越南_ns 芽庄_ns 国脚_n 曲波_np 挂帅_v 中国_ns U19B队_
n 迎战_v 泰国_ns 上半场_n 国青队_ni 门户_n 开_v 泰国_ns 进_v 球_n 国青0_n 射门_v 0_v
角球_n 完全_a 压制_v 下半场_n 国青_a 进攻_v 起色_n 马辅渔_np 利用_v 远射_v 扳回_v 球_
n 扳平_v 比分_n 全场_n 比赛_v 结束_v 国青_a 输_v 球_n 中国_ns 杯国_ns 输_v 给_v 泰国_
ns 遭遇_v 泰国_ns 足球_n 双杀_a
```

结论分析与体会：

从各个工具的测试结果可以看出，在词性标注方面各个工具的效果都差不多，而命名实体识别方面，stanfordcorenlp 的效果在中英文文档的测试结果都明显更好。