**1、包的要求：**

**spacy和en\_core\_web\_sm的版本要对应，否则无法利用spacy.load()函数加载en\_core\_web\_sm模型。**

**Pip install spacy==2.2.2**

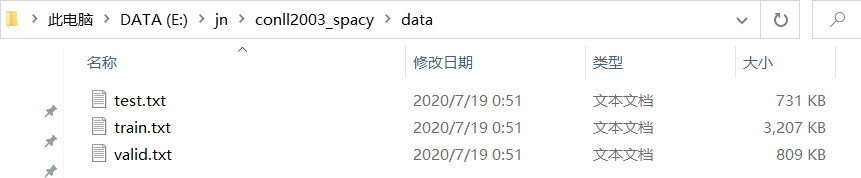
**先下载en\_core\_web\_sm模型的离线包，再pip install。**

**https://blog.csdn.net/hjzgj263446/article/details/103527952**

**Pip install E:/jn/conll2003\_spacy/en\_core\_web\_sm-2.2.5.tar.gz**

**2、数据获取：**

[**https://github.com/davidsbatista/NER-datasets/tree/master/CONLL2003/**](https://github.com/davidsbatista/NER-datasets/tree/master/CONLL2003/)

****

**训练集：train.txt**

**测试集：test.txt**

**验证集：valid.txt**

**3、加载en\_core\_web\_md模型并生成新链接。**

**En\_spacy= spacy.load('en\_core\_web\_md')**

**!python -m spacy link en\_core\_web\_sm en**

**注意：在执行spacy.load('en\_core\_web\_md')代码时报错。**

**OSError: [E053] Could not read config.cfg from C:\Users\pc\AppData\Local\Programs\Python\Python38\Lib\site-packages\en\_core\_web\_md\en\_core\_web\_md-2.2.5\config.cfg。**

**问题的原因是：下载的en\_core\_web\_md与spacy的版本不兼容，我使用的en\_core\_web\_md版本是2.2.5，但是安装spacy最新的版本是3.0.0。**

**因此我把spacy卸载后，重装spacy2.x的版本。**

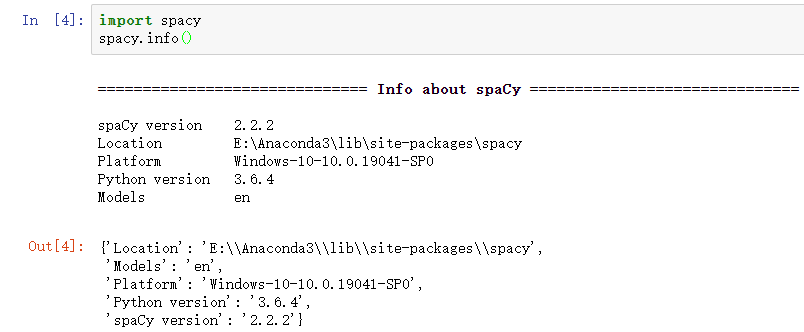
**pip uninstall spacy**

**pip install spacy==2.2.2**

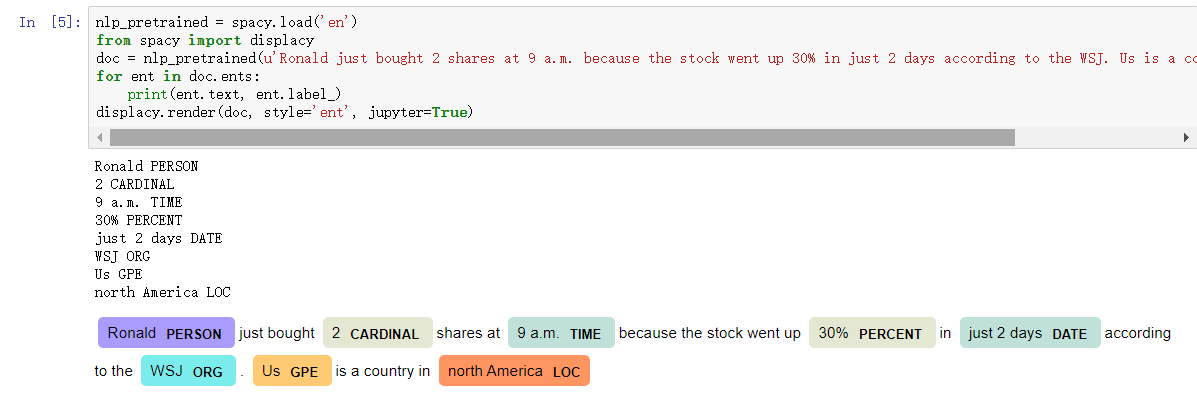
**4、导入spacy包，并查看相关信息。**

**import spacy**

**spacy.info()**

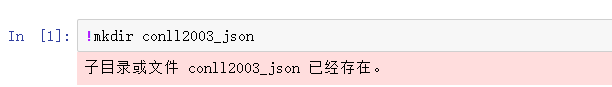


**5、加载en模型，并实现信息抽取小例子。**

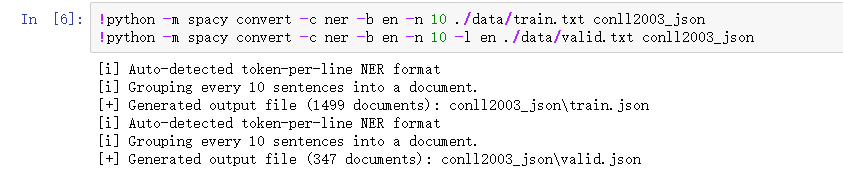


**6、训练和验证数据集的处理**

**生成conll2003\_json文件夹，用于存放conll2003数据集的json格式。**

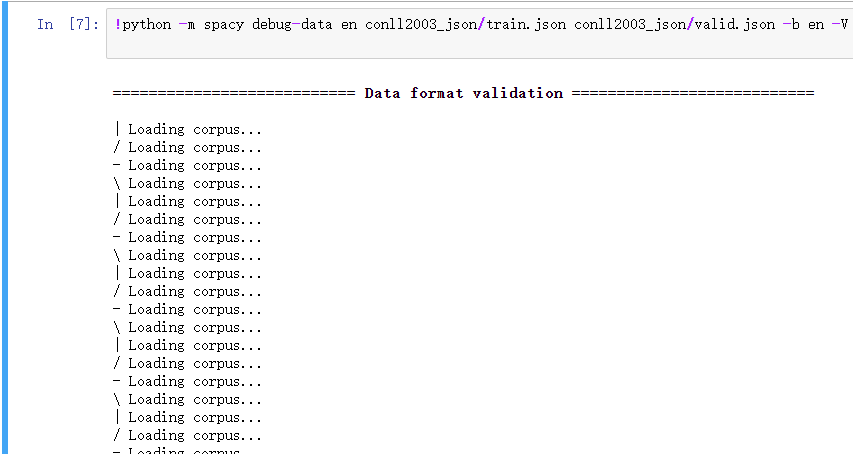
****

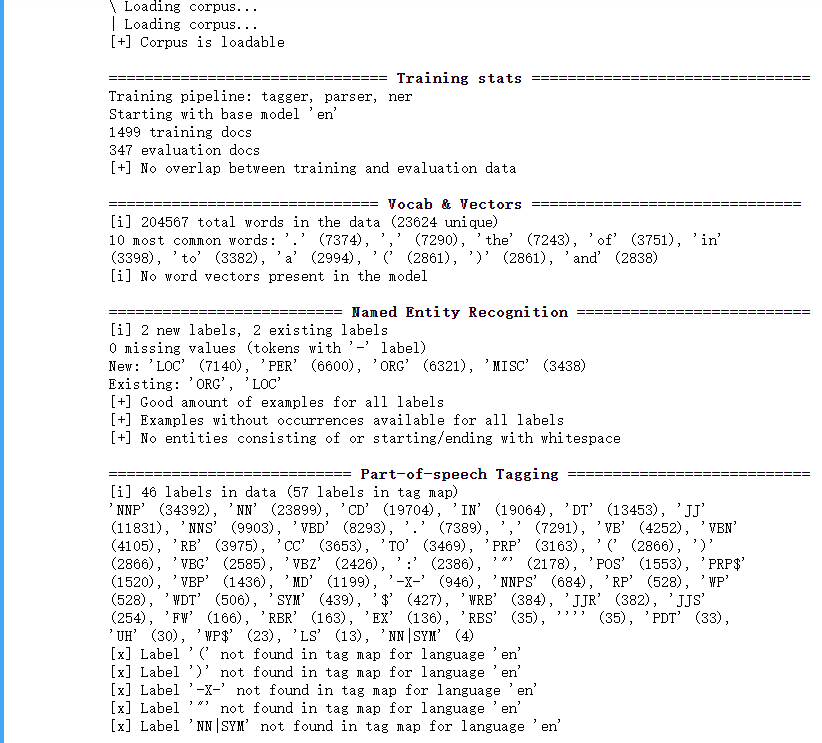
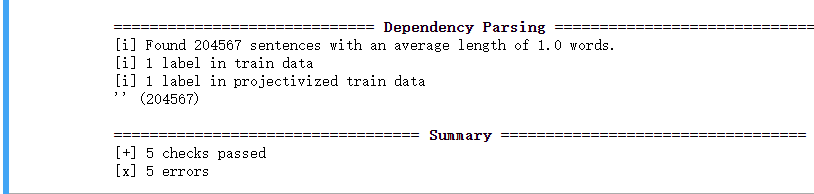
**转换训练和验证数据**



****

**7、调试生成的 json 训练和验证数据**



**8、在 Coll2003 上训练 Spacy NER 模型**

**E:\jn\conll2003\_spacy>python -m spacy train en ns\_models2 conll2003\_json/train.json conll2003\_json/valid.json -p ner**

**⚠ Output directory is not empty**

**This can lead to unintended side effects when saving the model. Please use an**

**empty directory or a different path instead. If the specified output path**

**doesn't exist, the directory will be created for you.**

**Training pipeline: ['ner']**

**Starting with blank model 'en'**

**Counting training words (limit=0)**

**Itn NER Loss NER P NER R NER F Token % CPU WPS**

**--- --------- ------ ------ ------ ------- -------**

**1 21614.232 79.307 78.559 78.931 100.000 25144**

**2 10503.866 84.781 83.911 84.344 100.000 27258**

**3 7182.972 86.149 85.729 85.938 100.000 26709**

**4 5415.420 87.356 86.856 87.105 100.000 27392**

**5 4292.472 87.485 87.176 87.330 100.000 27854**

**6 3633.963 87.840 87.529 87.684 100.000 24526**

**7 3117.598 88.196 87.647 87.921 100.000 24000**

**8 2856.826 88.284 87.883 88.083 100.000 28538**

**9 2600.745 88.121 87.765 87.943 100.000 27510**

**10 2307.530 88.260 87.933 88.096 100.000 27532**

**11 2149.664 88.236 87.984 88.110 100.000 24902**

**12 1951.191 88.519 88.236 88.378 100.000 27082**

**13 1759.552 88.295 88.102 88.198 100.000 27893**

**14 1742.362 88.289 88.051 88.170 100.000 23972**

**15 1692.663 88.203 87.832 88.018 100.000 27683**

**16 1752.455 88.266 87.984 88.125 100.000 22724**

**17 1547.677 88.579 88.236 88.407 100.000 25318**

**18 1418.538 88.656 88.388 88.522 100.000 27759**

**19 1393.818 88.379 88.186 88.282 100.000 27660**

**20 1337.872 88.160 87.967 88.063 100.000 27808**

**21 1303.301 88.200 88.051 88.125 100.000 25435**

**22 1254.123 88.162 87.984 88.073 100.000 29117**

**23 1160.237 88.276 88.068 88.172 100.000 22845**

**24 1197.359 88.164 88.001 88.082 100.000 28177**

**25 1029.577 87.978 87.816 87.897 100.000 27064**

**26 1086.861 88.068 87.816 87.941 100.000 24037**

**27 1074.834 88.036 87.799 87.917 100.000 26151**

**28 950.825 88.100 87.715 87.907 100.000 25705**

**29 971.487 88.061 87.765 87.913 100.000 27319**

**30 957.332 88.112 87.816 87.964 100.000 22391**

**✔ Saved model to output directory**

**ns\_models2\model-final**

**✔ Created best model**

**ns\_models2\model-best**

**9、评估**

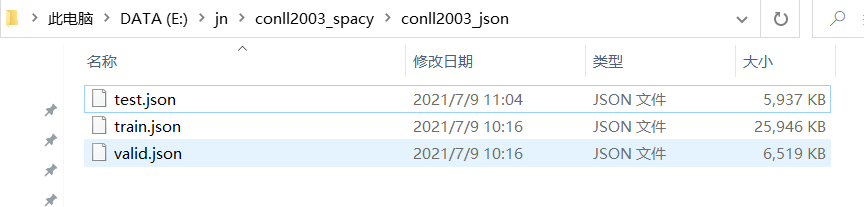
**在这个阶段，将使用测试数据进行模型评估，因此将test.txt也转换为test.json格式。**

**E:\jn\conll2003\_spacy>python -m spacy convert -c ner -n 10 ./data/test.txt conll2003\_json**

**ℹ Auto-detected token-per-line NER format**

**ℹ Grouping every 10 sentences into a document.**

**✔ Generated output file (369 documents): conll2003\_json\test.json**



**10、评估**

**确保训练和测试数据示例不重叠。**

**E:\jn\conll2003\_spacy>python -m spacy debug-data en conll2003\_json/train.json conll2003\_json/test.json -b en -V**

**=========================== Data format validation ===========================**

**✔ Corpus is loadable**

**=============================== Training stats ===============================**

**Training pipeline: tagger, parser, ner**

**Starting with base model 'en'**

**1499 training docs**

**369 evaluation docs**

**✔ No overlap between training and evaluation data**

**============================== Vocab & Vectors ==============================**

**ℹ 204567 total words in the data (23624 unique)**

**10 most common words: '.' (7374), ',' (7290), 'the' (7243), 'of' (3751), 'in'**

**(3398), 'to' (3382), 'a' (2994), '(' (2861), ')' (2861), 'and' (2838)**

**ℹ No word vectors present in the model**

**========================== Named Entity Recognition ==========================**

**ℹ 2 new labels, 2 existing labels**

**0 missing values (tokens with '-' label)**

**New: 'LOC' (7140), 'PER' (6600), 'ORG' (6321), 'MISC' (3438)**

**Existing: 'ORG', 'LOC'**

**✔ Good amount of examples for all labels**

**✔ Examples without occurrences available for all labels**

**✔ No entities consisting of or starting/ending with whitespace**

**=========================== Part-of-speech Tagging ===========================**

**ℹ 46 labels in data (57 labels in tag map)**

**'NNP' (34392), 'NN' (23899), 'CD' (19704), 'IN' (19064), 'DT' (13453), 'JJ'**

**(11831), 'NNS' (9903), 'VBD' (8293), '.' (7389), ',' (7291), 'VB' (4252), 'VBN'**

**(4105), 'RB' (3975), 'CC' (3653), 'TO' (3469), 'PRP' (3163), '(' (2866), ')'**

**(2866), 'VBG' (2585), 'VBZ' (2426), ':' (2386), '"' (2178), 'POS' (1553), 'PRP$'**

**(1520), 'VBP' (1436), 'MD' (1199), '-X-' (946), 'NNPS' (684), 'RP' (528), 'WP'**

**(528), 'WDT' (506), 'SYM' (439), '$' (427), 'WRB' (384), 'JJR' (382), 'JJS'**

**(254), 'FW' (166), 'RBR' (163), 'EX' (136), 'RBS' (35), '''' (35), 'PDT' (33),**

**'UH' (30), 'WP$' (23), 'LS' (13), 'NN|SYM' (4)**

**✘ Label '"' not found in tag map for language 'en'**

**✘ Label '(' not found in tag map for language 'en'**

**✘ Label ')' not found in tag map for language 'en'**

**✘ Label '-X-' not found in tag map for language 'en'**

**✘ Label 'NN|SYM' not found in tag map for language 'en'**

**============================= Dependency Parsing =============================**

**ℹ Found 204567 sentences with an average length of 1.0 words.**

**ℹ 1 label in train data**

**ℹ 1 label in projectivized train data**

**'' (204567)**

**================================== Summary ==================================**

**✔ 5 checks passed**

**✘ 5 errors**

**11、在 Coll2003 上测试 Spacy NER 模型**

**在测试集上评估最佳模型选项的精度 (NER P)、召回率 (NER R) 和 f 分数 (NER F)。**

**E:\jn\conll2003\_spacy>python -m spacy evaluate ns\_models2/model-best conll2003\_json/test.json**

**================================== Results ==================================**

**Time 1.67 s**

**Words 46666**

**Words/s 28024**

**TOK 100.00**

**POS 0.00**

**UAS 0.00**

**LAS 0.00**

**NER P 80.84**

**NER R 81.36**

**NER F 81.10**

**Textcat 0.00**

