[spaCy 101: Everything you need to know · spaCy Usage Documentation](https://spacy.io/usage/spacy-101" \l "training)

import spacy

**NLP常见任务**

分词 Tokenization

词性标注 Part-of-speech (POS) Tagging

依存语义分析 Dependency Parsing

词形还原 Lemmatization

句子切分 Sentence Boundary Detection (SBD)

命名体识别 Named Entity Recognition (NER)

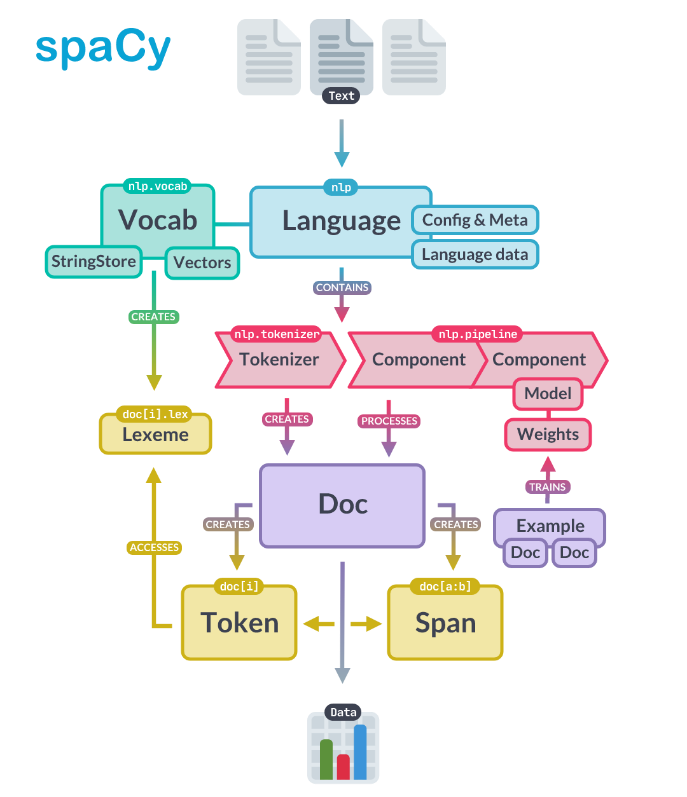
消歧 Entity Linking (EL)

文本相似度计算 Similarity

文本分类 Text Classification

序列化(将文本转化为比特串等形式进行存储)Serialization

**Spacy架构**



nlp类是处理器，将文本传入后返回一个Doc类

Doc类包含了文本分词后的数据及词性标注等信息

## Pipeline

**新建处理器**

pipeline = spacy.load('en\_core\_web\_sm') #en\_core\_web\_lg是更大的语料库

**查看处理器架构**

pipeline.analyze\_pipes(pretty=True)

**处理单条文本**

doc = pipeline(text)

**向处理器中添加处理函数**

from spacy.language import Language

@Language.component("processor")

def funct (doc): #以doc对象作为输入

函数体

return processed\_doc\_text

pipeline.add\_pipe("processor ");

## Token Class

#经pipeline处理过获得的doc，其中每个词word都是一个token类

<https://spacy.io/api/token>

for word in doc:

word.text, #原始文本

word.lemma\_ #还原词型

word.pos\_, #粗粒度的词性

word.tag\_, #细粒度的词性

word.dep\_, #依存句法成分

word.shape\_, #词形(字母以X或x替代，字母以d替代，其余字母不变，限前4个)

word.is\_alpha, #是否是单词

word.is\_stop #是否是停用词

word.is\_punct #是否是标点符号

word.sentiment #情感色彩正负(0,1)

**依存句法图**

spacy.displacy.render(doc, style='dep')