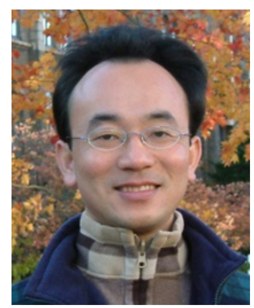
Tamkang University
淡江大學

# 文字探勘 (Text Mining)

# 語意分析和命名實體識別 (Semantic Analysis and Named Entity Recognition; NER)

1082TM09
MBA, BDABI, TKU (E3611) (8480) (Spring 2020)
Mon, 7, 8, 9 (14:10-17:00) (B206)


Chichang Jou
周清江
Associate Professor
副教授
cjou@mail.tku.edu.tw



Min-Yuh Day
戴敏育
Associate Professor
副教授
myday@mail.tku.edu.tw

Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系


2020-05-11

# 課程大綱 (Syllabus)

週次 (Week) 日期 (Date) 內容 (Subject/Topics)

1 2020/03/02 文字探勘課程介紹
(Course Orientation on Text Mining)

2 2020/03/09 文字探勘基礎：自然語言處理
(Foundations of Text Mining:
Natural Language Processing; NLP)

3 2020/03/16 Python自然語言處理
(Python for Natural Language Processing)

4 2020/03/23 處理和理解文本 (Processing and Understanding Text)

5 2020/03/30 文本表達特徵工程
(Feature Engineering for Text Representation)

6 2020/04/06 人工智慧文本分析個案研究 I
(Case Study on Artificial Intelligence for Text Analytics I)

# 課程大綱 (Syllabus)

週次 (Week)　日期 (Date)　內容 (Subject/Topics)

7 2020/04/13 文本分類
(Text Classification)

8 2020/04/20 文本摘要和主題模型
(Text Summarization and Topic Models)

9 2020/04/27 期中報告 (Midterm Project Report)

10 2020/05/04 文本相似度和分群
(Text Similarity and Clustering)

11 2020/05/11 語意分析和命名實體識別
(Semantic Analysis and Named Entity Recognition; NER)

12 2020/05/18 情感分析
(Sentiment Analysis)

# 課程大綱 (Syllabus)

週次 (Week) 日期 (Date) 內容 (Subject/Topics)

13 2020/05/25 人工智慧文本分析個案研究 II
(Case Study on Artificial Intelligence for Text Analytics II)

14 2020/06/01 深度學習和通用句子嵌入模型
(Deep Learning and Universal Sentence-Embedding Models)

15 2020/06/08 問答系統與對話系統
(Question Answering and Dialogue Systems)

16 2020/06/15 期末報告 I (Final Project Presentation I)

17 2020/06/22 期末報告 II (Final Project Presentation II)

18 2020/06/29 教師彈性補充教學

# Semantic Analysis and Named Entity Recognition (NER)

# Outline

- Semantic Analysis
  - WordNet
  - Word sense disambiguation
- Named Entity Recognition (NER)

# Semantic Analysis

- **Semantics**
  - the study of meaning
- **Linguistic semantics**
  - the study of meaning in natural language.

# Semantic Analysis and NER

- WordNet and synsets
  - Analyzing lexical semantic relations
  - Word sense disambiguation
- Named entity recognition
- Analyzing semantic representations

Source: Dipanjan Sarkar (2019), Text Analytics with Python: A Practitioner's Guide to Natural Language Processing, Second Edition. APress.

# WordNet
# A Lexical Database for English

PRINCETON UNIVERSITY

WordNet

A Lexical Database for English

Search...

- What is WordNet
- People
- News
- Use Wordnet Online
- Download
- Citing WordNet
- License and Commercial Use
- Related Projects
- Documentation
- Publications
- Frequently Asked

## What is WordNet?

*Any opinions, findings, and conclusions or recommendations expressed in this material are those of the creators of WordNet and do not necessarily reflect the views of any funding agency or Princeton University.*

When writing a paper or producing a software application, tool, or interface based on WordNet, it is necessary to properly cite the source. Citation figures are critical to WordNet funding.

## About WordNet

WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser. WordNet is also freely and publicly available for download. WordNet's structure makes it a useful tool for computational linguistics and natural language processing.

WordNet superficially resembles a thesaurus, in that it groups words together based on their

## Note

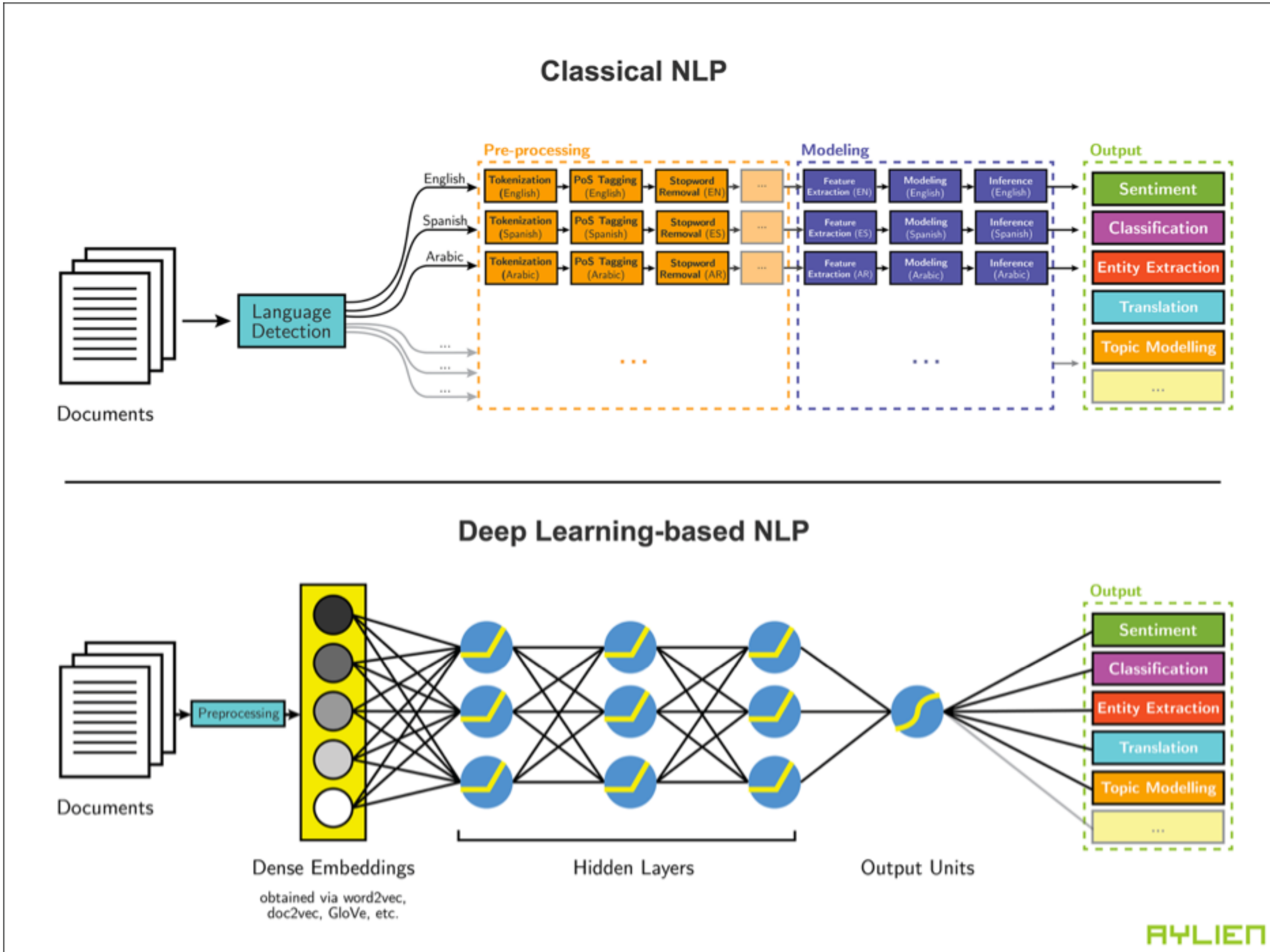
Due to funding and staffing issues, we are no longer able to accept comment and suggestions.

We get numerous questions regarding topics that are addressed on our FAQ page. If you have a problem or question regarding something you downloaded from the "Related projects" page, you must contact the developer directly.

Please note that any changes

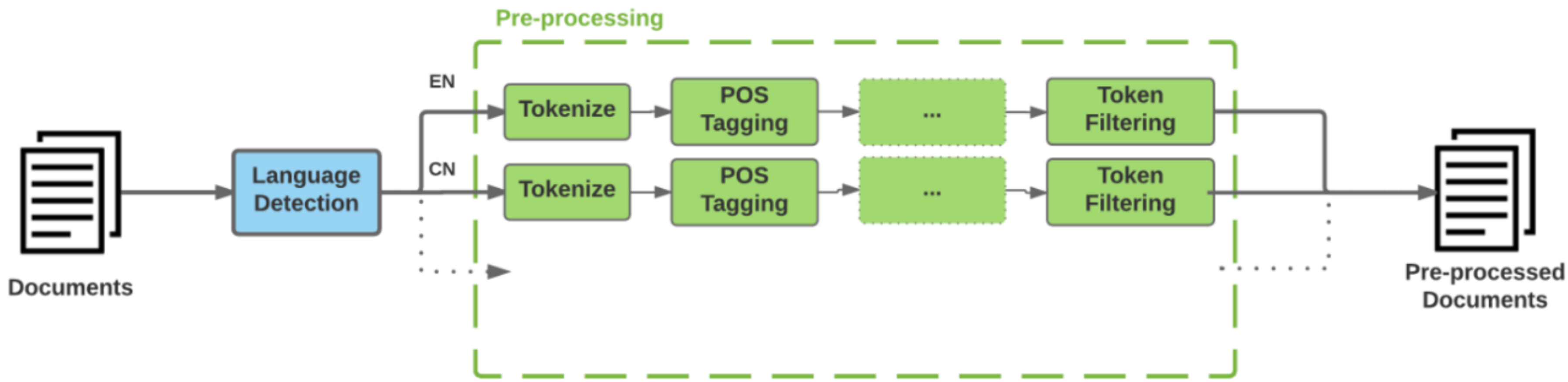
https://wordnet.princeton.edu/

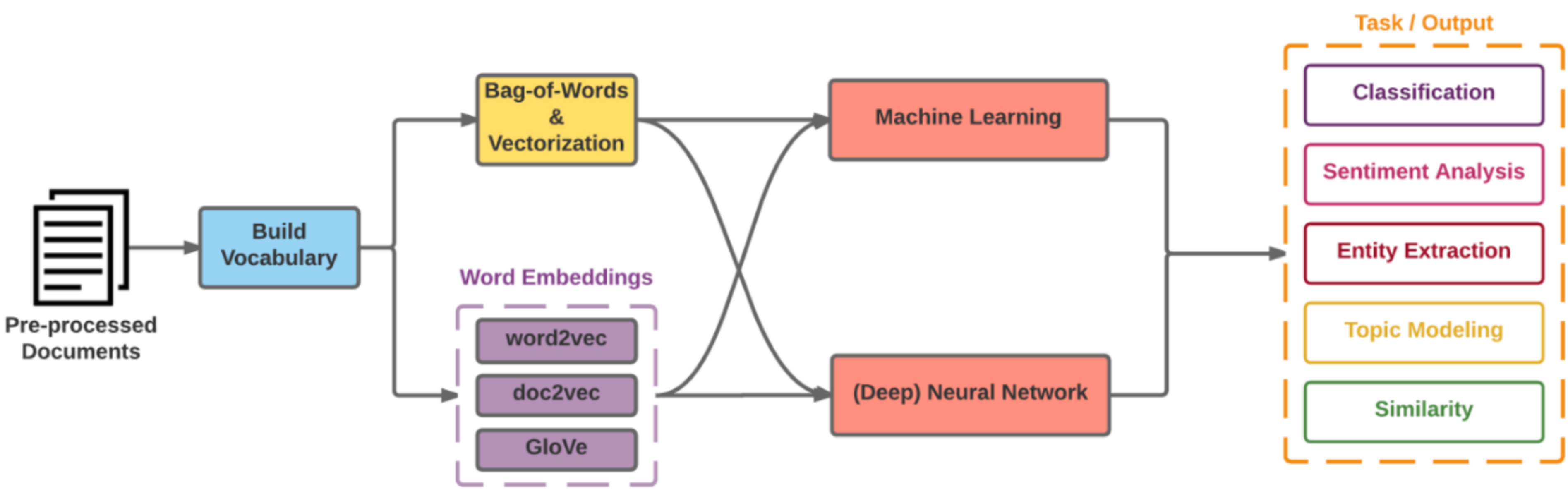
# NLP




Source: http://blog.aylien.com/leveraging-deep-learning-for-multilingual/

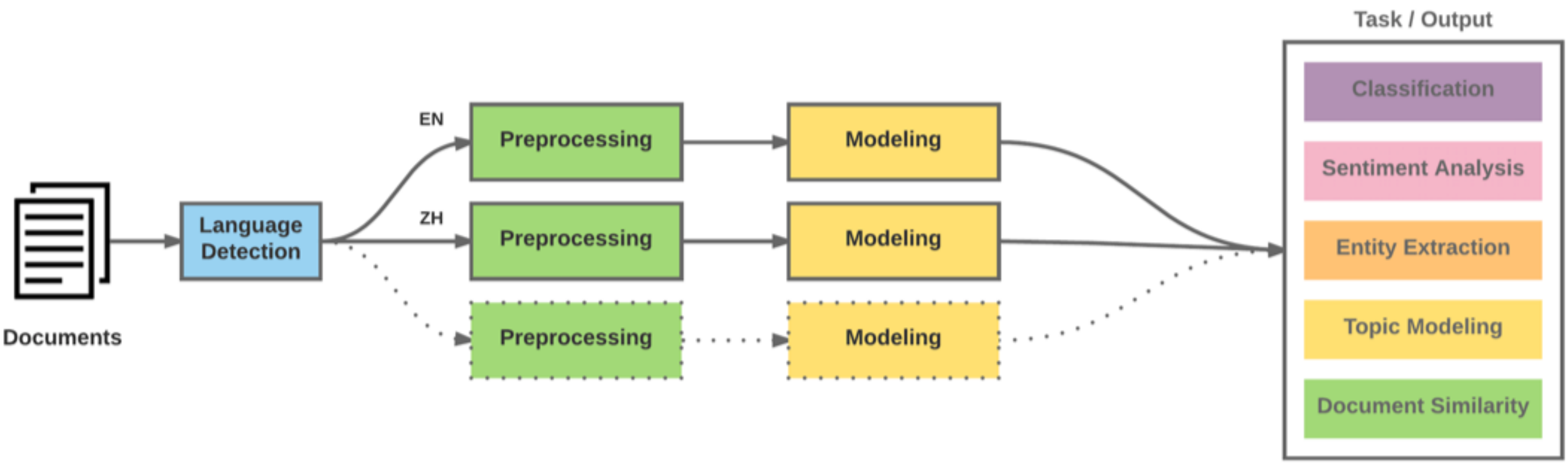
# Modern NLP Pipeline



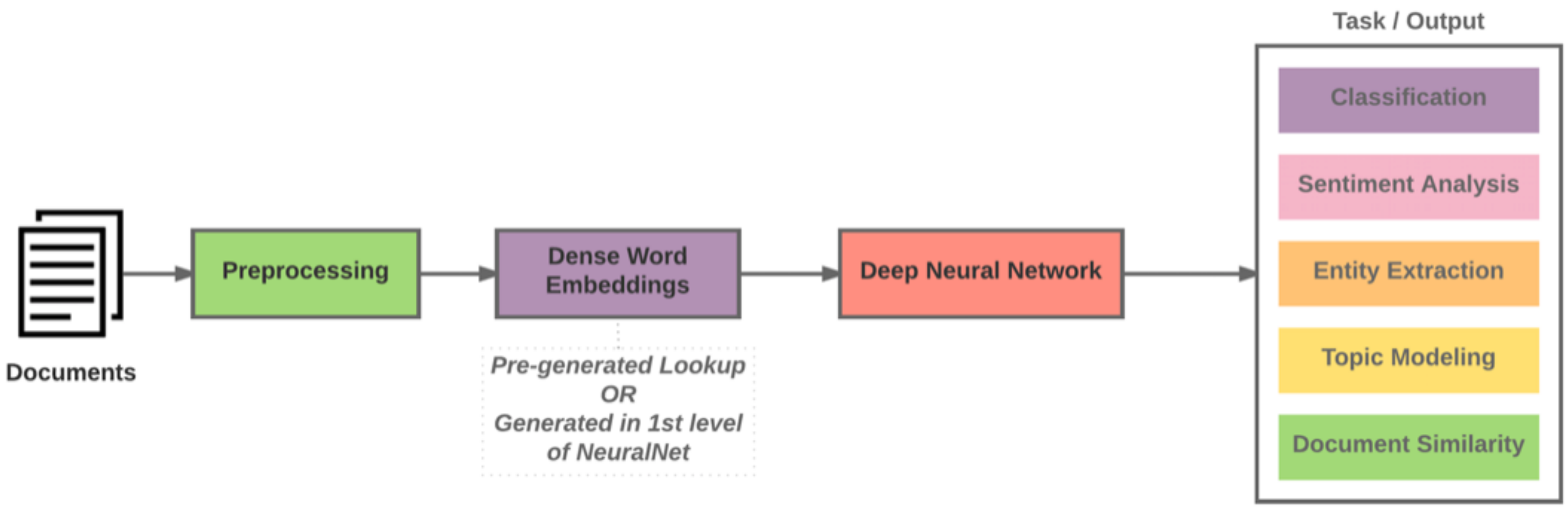

Source: https://github.com/fortiema/talks/blob/master/opendata2016sh/pragmatic-nlp-opendata2016sh.pdf

# Modern NLP Pipeline




Source: http://mattfortier.me/2017/01/31/nlp-intro-pt-1-overview/

# Deep Learning NLP




Source: http://mattfortier.me/2017/01/31/nlp-intro-pt-1-overview/

# Natural Language Processing (NLP) and Text Mining

- Raw text
- Sentence Segmentation
- Tokenization
- Part-of-Speech (POS)
- Stop word removal
- Stemming / Lemmatization
- Dependency Parser
- String Metrics & Matching

| word's stem | word's lemma |
|---|---|
| am → am | am → be |
| having → hav | having → have |

Source: Nitin Hardeniya (2015), NLTK Essentials, Packt Publishing; Florian Leitner (2015), Text mining - from Bayes rule to dependency parsing

# Analyzing Lexical Semantic Relationships

- Entailments
- Homonyms and Homographs
- Synonyms and Antonyms
- Hyponyms and Hypernyms
- Holonyms and Meronyms
- Semantic Relationships and Similarity

Source: Dipanjan Sarkar (2019), Text Analytics with Python: A Practitioner's Guide to Natural Language Processing, Second Edition. APress.

# Word Sense Disambiguation

- Lesk algorithm (Lesk, 1986)
  - leverage dictionary or vocabulary definitions for a word we want to disambiguate in a body of text and compare the words in these definitions with a section of text surrounding our word of interest.
  - The main objective is to return the synset with the maximum number of overlapping words or terms between the context sentence and the different definitions from each synset for the word we target for disambiguation.

Source: Dipanjan Sarkar (2019), Text Analytics with Python: A Practitioner's Guide to Natural Language Processing, Second Edition. APress.

# Named Entity Recognition (NER)

- **Named entities**
  - represent real-world objects
  - people, places, organizations
  - proper names
- **Named entity recognition**
  - Entity chunking
  - Entity extraction

Source: Dipanjan Sarkar (2019), Text Analytics with Python: A Practitioner's Guide to Natural Language Processing, Second Edition. APress.

# NER: OntoNotes 5 Named Entities

| SID | TYPE | DESCRIPTION |
|---|---|---|
| 1 | PERSON | People, including fictional. |
| 2 | NORP | Nationalities or religious or political groups. |
| 3 | FAC | Buildings, airports, highways, bridges, etc. |
| 4 | ORG | Companies, agencies, institutions, etc. |
| 5 | GPE | Countries, cities, states. |
| 6 | LOC | Non-GPE locations, mountain ranges, bodies of water. |
| 7 | PRODUCT | Objects, vehicles, foods, etc. (Not services.) |
| 8 | EVENT | Named hurricanes, battles, wars, sports events, etc. |
| 9 | WORK_OF_ART | Titles of books, songs, etc. |
| 10 | LAW | Named documents made into laws. |
| 11 | LANGUAGE | Any named language. |
| 12 | DATE | Absolute or relative dates or periods. |
| 13 | TIME | Times smaller than a day. |
| 14 | PERCENT | Percentage, including ”%“. |
| 15 | MONEY | Monetary values, including unit. |
| 16 | QUANTITY | Measurements, as of weight or distance. |
| 17 | ORDINAL | “first”, “second”, etc. |
| 18 | CARDINAL | Numerals that do not fall under another type. |

Source: https://spacy.io/api/annotation#named-entities

# NER: Wikipedia Named Entities

| SID | TYPE | DESCRIPTION |
|---|---|---|
| 1 | PER | Named person or family. |
| 2 | LOC | Name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains). |
| 3 | ORG | Named corporate, governmental, or other organizational entity. |
| 4 | MISC | Miscellaneous entities, e.g. events, nationalities, products or works of art. |

Source: [https://spacy.io/api/annotation#named-entities](https://spacy.io/api/annotation#named-entities)

# NER IOB Scheme

| TAG | ID | DESCRIPTION |
|---|---|---|
| "I" | 1 | Token is inside an entity. |
| "O" | 2 | Token is outside an entity. |
| "B" | 3 | Token begins an entity. |
| "" | 0 | No entity tag is set (missing value). |

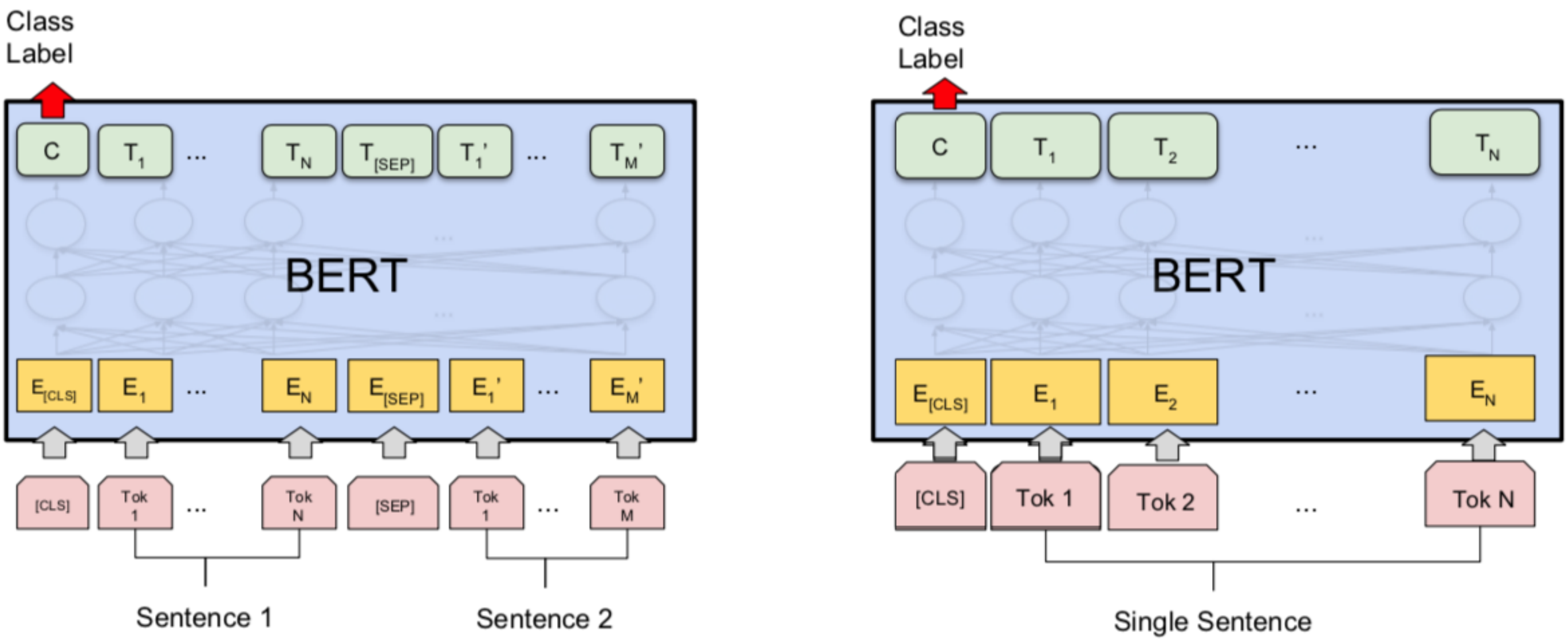
Source: https://spacy.io/api/annotation#named-entities

# NER BILUO Scheme

| TAG | DESCRIPTION |
| --- | --- |
| BEGIN | The first token of a multi-token entity. |
| IN | An inner token of a multi-token entity. |
| LAST | The final token of a multi-token entity. |
| UNIT | A single-token entity. |
| OUT | A non-entity token. |

Source: https://spacy.io/api/annotation#named-entities

# BERT Sequence-level tasks

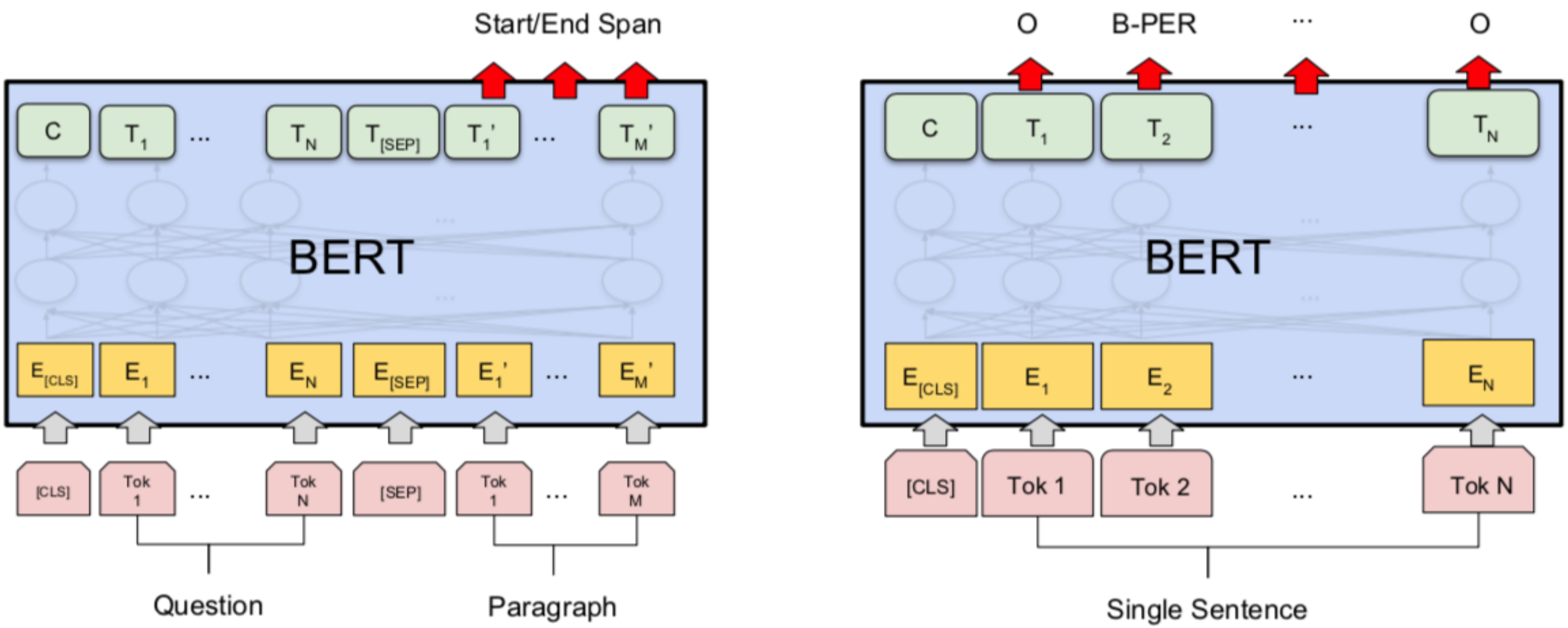



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

Source: Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv preprint arXiv:1810.04805

# BERT Token-level tasks

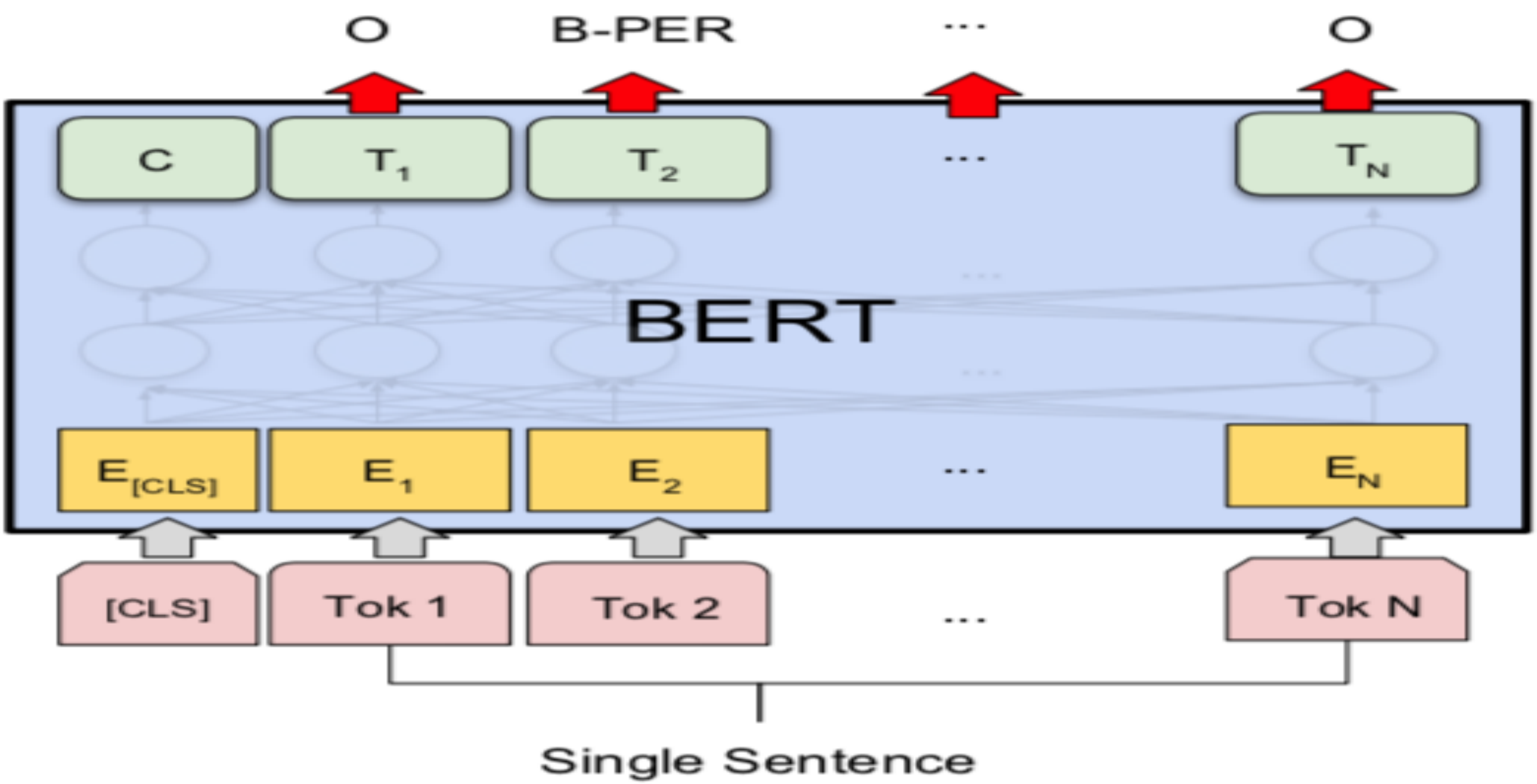



(c) Question Answering Tasks:
SQuAD v1.1

(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER

Source: Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018).
"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv preprint arXiv:1810.04805

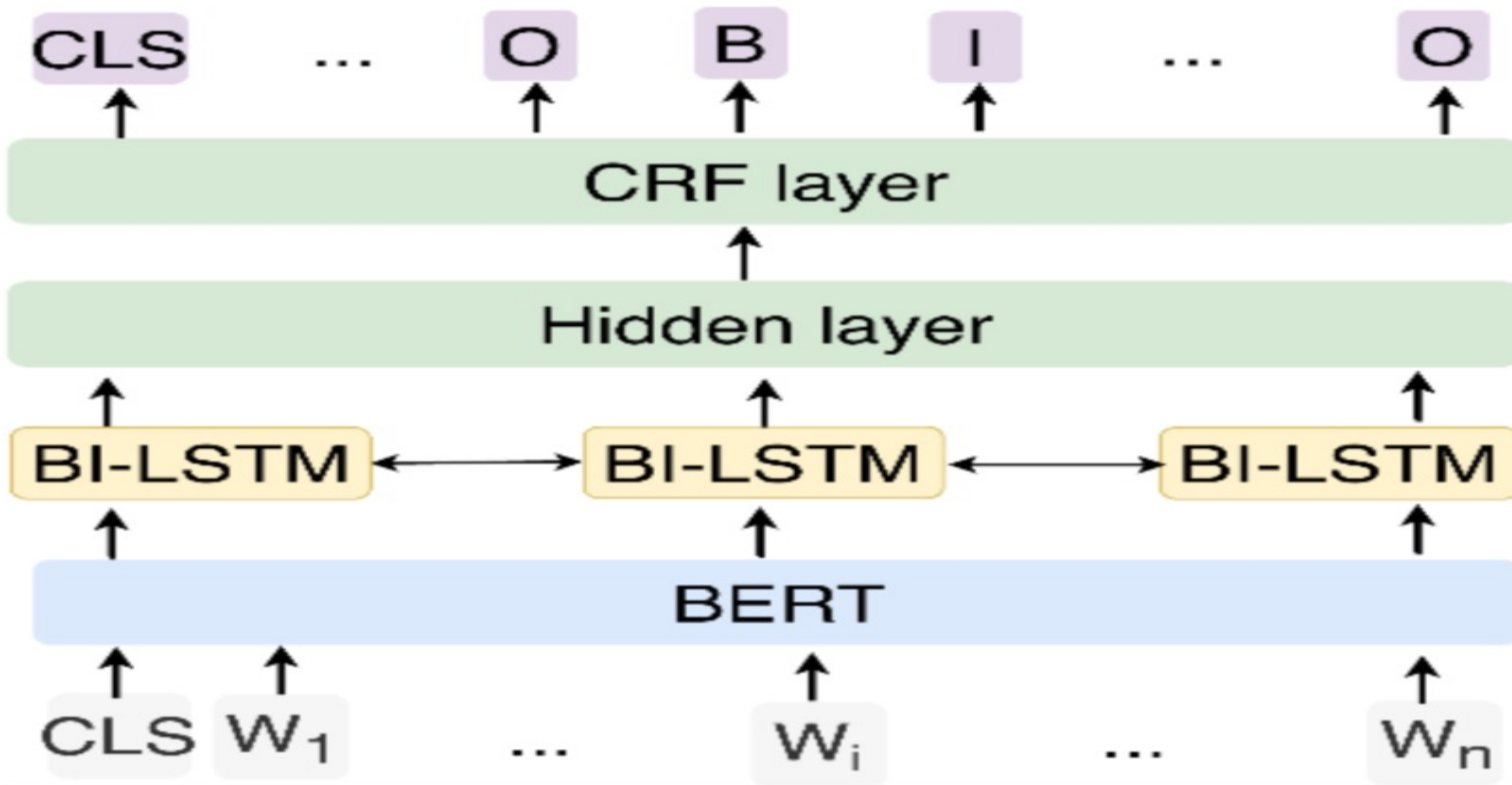
# NER: Single Sentence Tagging




(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Source: Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv preprint arXiv:1810.04805

# NER: Fine-tuning BERT with Bi-LSTM CRF




Source: Zhang, Xiaohui, Yaoyun Zhang, Qin Zhang, Yuankai Ren, Tinglin Qiu, Jianhui Ma, and Qiang Sun. "Extracting comprehensive clinical information for breast cancer using deep learning methods." International Journal of Medical Informatics 132 (2019): 103985.

# Python in Google Colab (Python101)

https://colab.research.google.com/drive/1FEG6DnGvwfUbeo4zJ1zTunjMqf2RkCrT

python101.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

Table of contents



## Semantic Analysis and Named Entity Recognition (NER)

- Source: Dipanjan Sarkar (2019), Text Analytics with Python: A Practitioner's Guide to Natural Language Processing, Second Edition. APress. https://github.com/Apress/text-analytics-w-python-2e

## Semantic Analysis

[1]

```
import nltk
from nltk.corpus import wordnet as wn
import pandas as pd
nltk.download('wordnet')
# WordNet Synsets
word = 'fruit'
synsets = wn.synsets(word)
print('Word:', word)
print('Wordnet Synsets:', len(synsets))
df = pd.DataFrame([{'Synset': synset,
                    'Part of Speech': synset.lexname(),
                    'Definition': synset.definition(),
                    'Lemmas': synset.lemma_names(),
                    'Examples': synset.examples()}
                   for synset in synsets])
df
```

```
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Unzipping corpora/wordnet.zip.
Word: fruit
Wordnet Synsets: 5
```

| | Synset | Part of Speech | Definition | Lemmas | Examples |
|---|---|---|---|---|---|
| 0 | Synset('fruit.n.01') | noun.plant | the ripened reproductive body of a seed plant | [fruit] | [] |
| 1 | Synset('yield.n.03') | noun.artifact | an amount of a product | [yield, fruit] | [] |

https://tinyurl.com/imtkupython101

# Python in Google Colab (Python101)

https://colab.research.google.com/drive/1FEG6DnGvwfUbeo4zJ1zTunjMqf2RkCrT

python101.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

Table of contents



+ Code + Text RAM Disk Editing

```
November with the intention of hearing from Zuckerberg. Since the Cambridge Analytica scandal broke, the Facebook chief
has only appeared in front of two legislatures: the American Senate and House of Representatives, and the European parliament.
Facebook has consistently rebuffed attempts from others, including the UK and Canadian parliaments, to hear from Zuckerberg.
He added that an article in the New York Times on Thursday, in which the paper alleged a pattern of behaviour from Facebook
to “delay, deny and deflect” negative news stories, “raises further questions about how recent data breaches were allegedly
dealt with within Facebook.”

re.sub: Three more countries have joined an “international grand committee” of parliaments, adding to calls for Facebook’s boss,
text_nlp: Three more countries have joined an “international grand committee” of parliaments, adding to calls for Facebook’s boss
```

```python
# print named entities in article
ner_tagged = [(word.text, word.ent_type_) for word in text_nlp]
print(ner_tagged)
```

```
[('Three', 'CARDINAL'), ('more', ''), ('countries', ''), ('have', ''), ('joined', ''), ('an', ''), ('“', ''), ('international', '
```

```python
from spacy import displacy
# visualize named entities
displacy.render(text_nlp, style='ent', jupyter=True)
```

Three CARDINAL more countries have joined an “international grand committee” of parliaments, adding to calls for Facebook’s boss, Mark Zuckerberg PERSON , to give evidence on misinformation to the coalition. Brazil GPE , Latvia GPE and Singapore GPE bring the total to eight CARDINAL different parliaments across the world, with plans to send representatives to London GPE on 27 November DATE with the intention of hearing from Zuckerberg GPE . Since the Cambridge Analytica scandal broke, the Facebook ORG chief has only appeared in front of two CARDINAL legislatures: the American Senate ORG and House of Representatives ORG , and the European NORP parliament. Facebook has consistently rebuffed attempts from others, including the UK GPE and Canadian NORP parliaments, to hear from Zuckerberg GPE . He added that an article in the New York Times ORG on Thursday DATE , in which the paper alleged a pattern of behaviour from Facebook ORG to “delay, deny and deflect” negative news stories, “raises further questions about how recent data breaches were allegedly dealt with within Facebook ORG .”

https://tinyurl.com/imtkupython101

# Summary

- Semantic Analysis
  - WordNet
  - Word sense disambiguation
- Named Entity Recognition (NER)

# References


- Dipanjan Sarkar (2019),
  Text Analytics with Python: A Practitioner's Guide to Natural Language Processing, Second Edition. APress. https://github.com/Apress/text-analytics-w-python-2e
- Benjamin Bengfort, Rebecca Bilbro, and Tony Ojeda (2018),
  Applied Text Analysis with Python, O'Reilly Media.
  https://www.oreilly.com/library/view/applied-text-analysis/9781491963036/
- HuggingFace (2020), Transformers Notebook,
  https://huggingface.co/transformers/notebooks.html
- The Super Duper NLP Repo, https://notebooks.quantumstat.com/
- Min-Yuh Day (2020), Python 101, https://tinyurl.com/imtkupython101