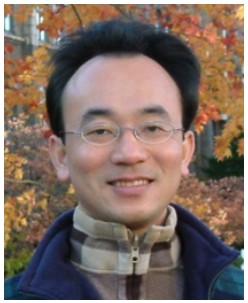
Tamkang University 淡江大學

人工智慧文本分析
(AI for Text Analytics)
文本分析的基礎：自然語言處理
(Foundations of Text Analytics: Natural Language Processing; NLP)

NTPU 國立臺北大學 National Taipei University

1091AITA02
MBA, IMTKU (M2455) (8418) (Fall 2020)
Thu 3, 4 (10:10-12:00) (B206)



Min-Yuh Day
戴敏育
Associate Professor
副教授
Institute of Information Management, National Taipei University
國立臺北大學 資訊管理研究所
https://web.ntpu.edu.tw/~myday

2020-09-24

課程大綱 (Syllabus)

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

1 2020/09/17 人工智慧文本分析課程介紹
(Course Orientation on Artificial Intelligence for Text Analytics)

2 2020/09/24 文本分析的基礎：自然語言處理
(Foundations of Text Analytics: Natural Language Processing; NLP)

3 2020/10/01 中秋節 (Mid-Autumn Festival) 放假一天 (Day off)

4 2020/10/08 Python自然語言處理
(Python for Natural Language Processing)

5 2020/10/15 處理和理解文本
(Processing and Understanding Text)

6 2020/10/22 文本表達特徵工程
(Feature Engineering for Text Representation)

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
7	2020/10/29	人工智慧文本分析個案研究 I (Case Study on Artificial Intelligence for Text Analytics I)
8	2020/11/05	文本分類 (Text Classification)
9	2020/11/12	文本摘要和主題模型 (Text Summarization and Topic Models)
10	2020/11/19	期中報告 (Midterm Project Report)
11	2020/11/26	文本相似度和分群 (Text Similarity and Clustering)
12	2020/12/03	語意分析和命名實體識別 (Semantic Analysis and Named Entity Recognition; NER)

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
13	2020/12/10	情感分析 (Sentiment Analysis)
14	2020/12/17	人工智慧文本分析個案研究 II (Case Study on Artificial Intelligence for Text Analytics II)
15	2020/12/24	深度學習和通用句子嵌入模型 (Deep Learning and Universal Sentence-Embedding Models)
16	2020/12/31	問答系統與對話系統 (Question Answering and Dialogue Systems)
17	2021/01/07	期末報告 I (Final Project Presentation I)
18	2021/01/14	期末報告 II (Final Project Presentation II)

Outline

- Text Analytics and Text Mining
- Natural Language Processing (NLP)
- Text Analytics with Python

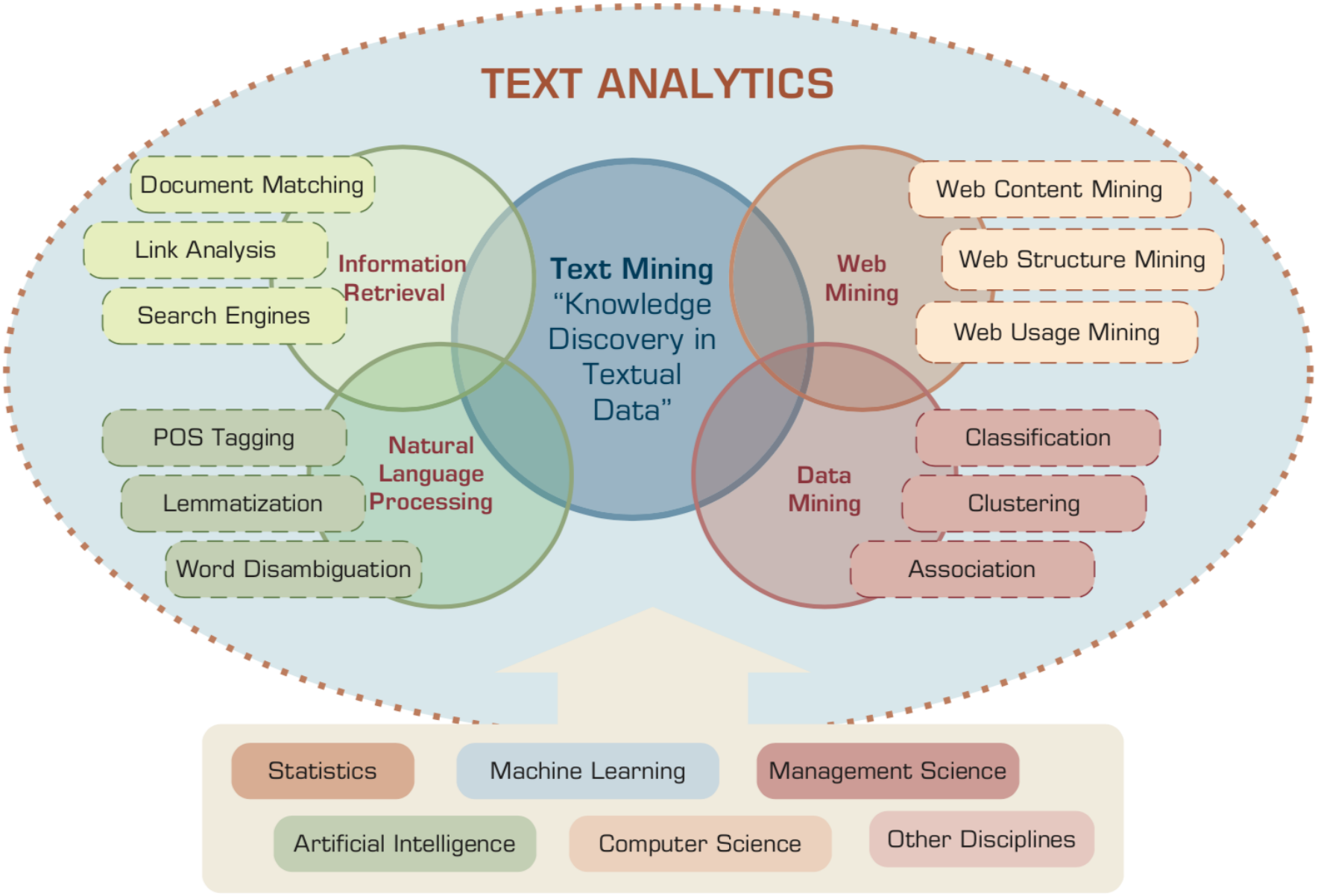
Text Analytics (TA)

Text Mining (TM)

Natural Language Processing (NLP)

Artificial Intelligence (AI)

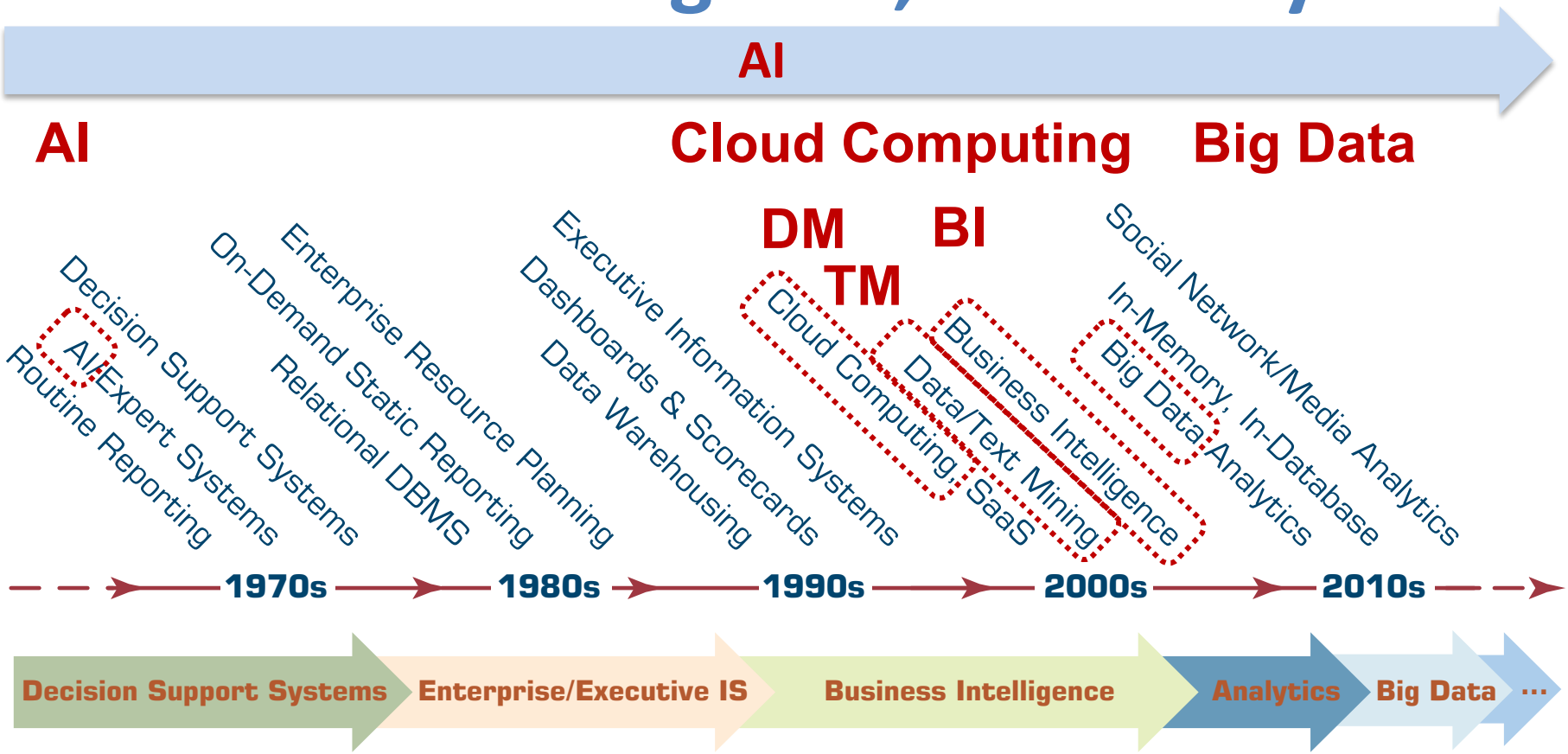
Text Analytics and Text Mining



Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

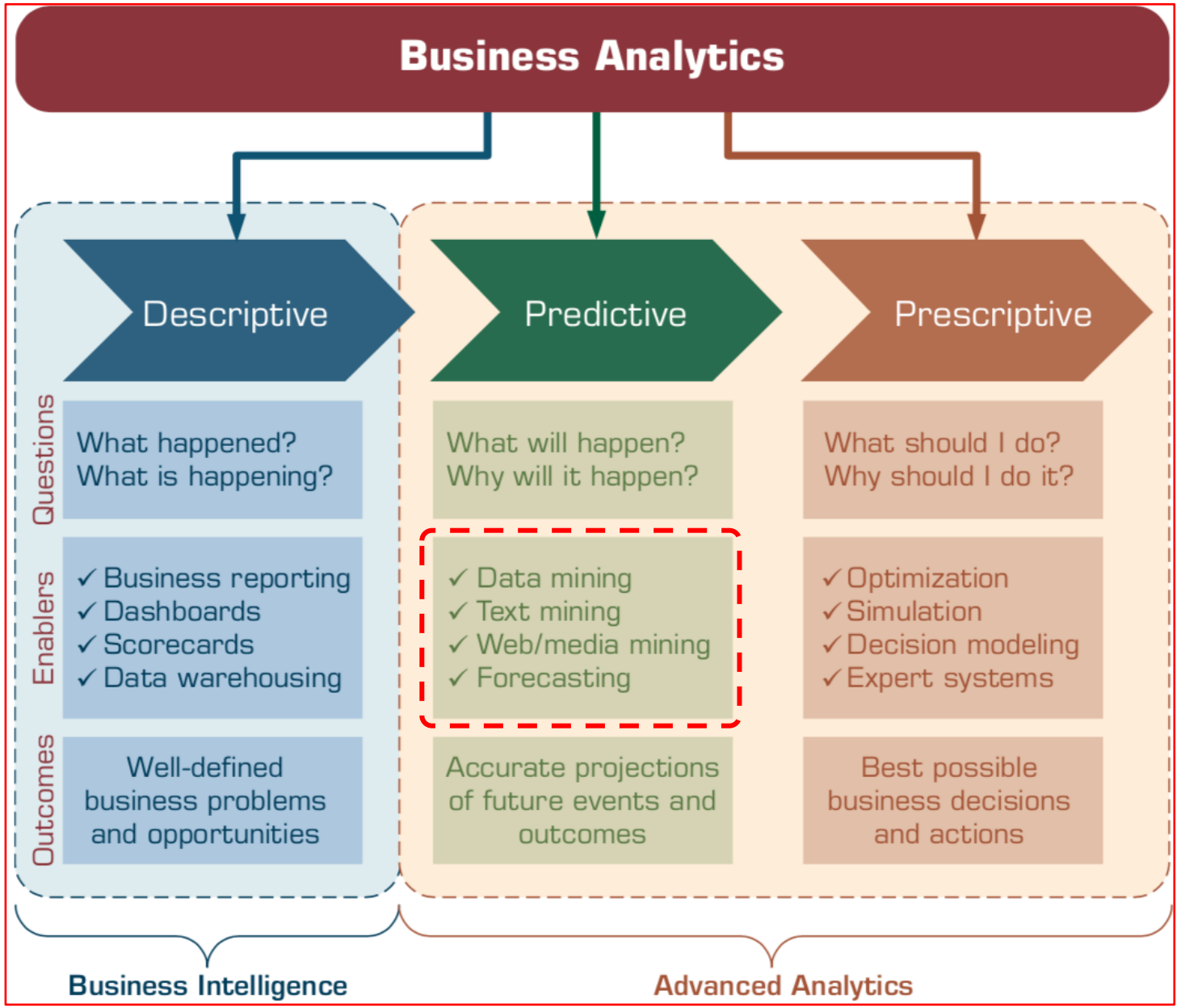
AI, Big Data, Cloud Computing

Evolution of Decision Support, Business Intelligence, and Analytics



Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017),
Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Business Analytics



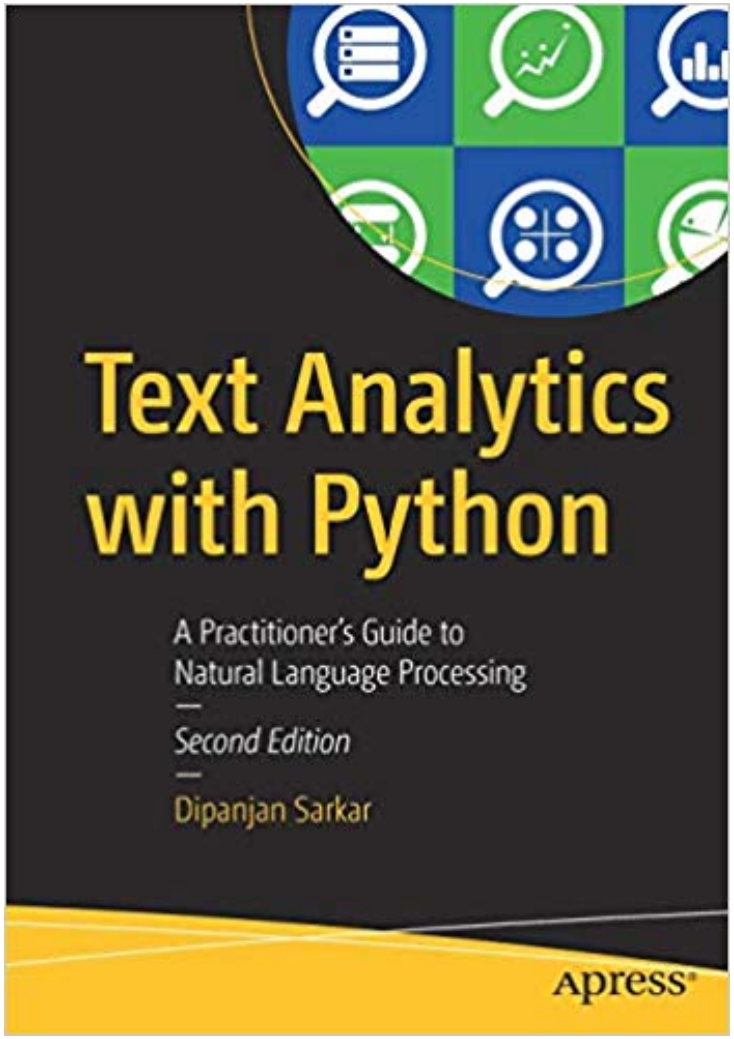
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Text Analytics and Text Mining

Dipanjan Sarkar (2019),

Text Analytics with Python:

A Practitioner's Guide to Natural Language Processing, Second Edition. APress.



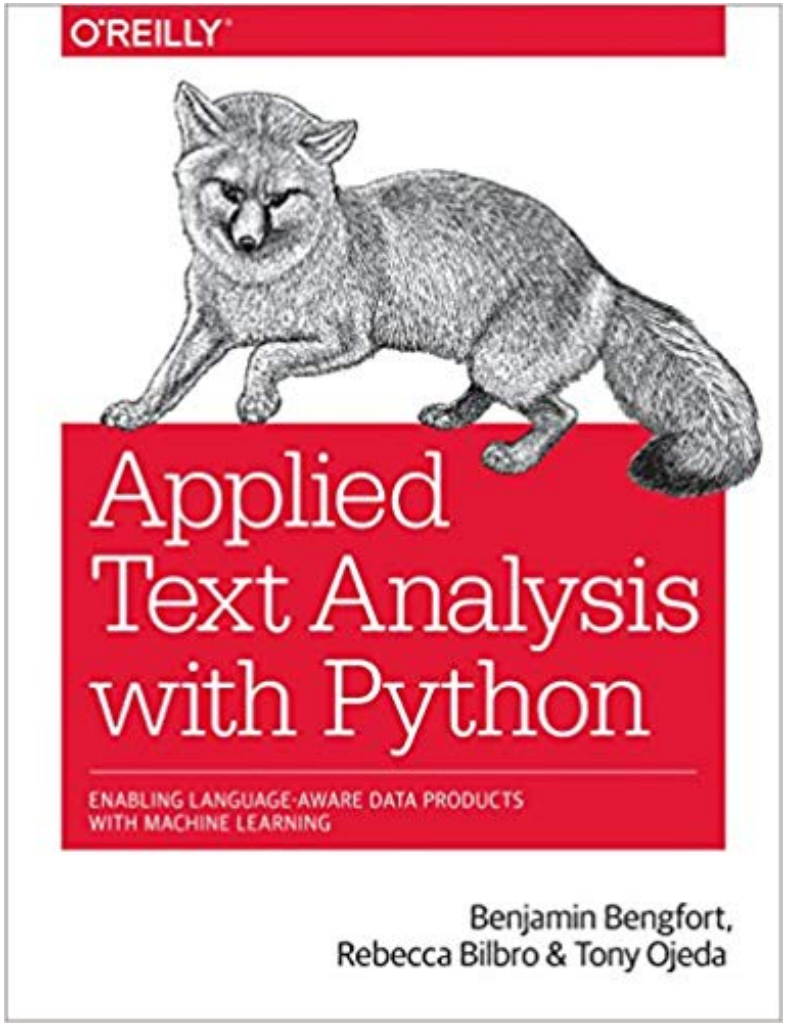
Source: https://www.amazon.com/Text-Analytics-Python-Practitioners-Processing/dp/1484243536

Benjamin Bengfort, Rebecca Bilbro, and Tony Ojeda (2018),

Applied Text Analysis with Python:

Enabling Language-Aware Data Products with Machine Learning,

O'Reilly.

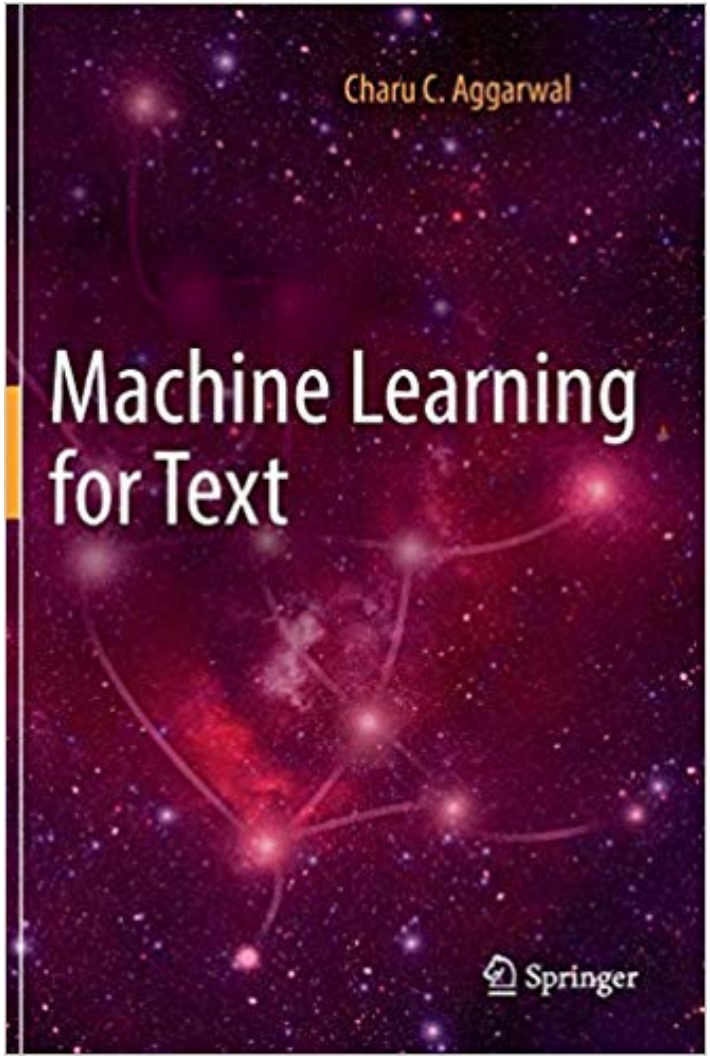


Source: https://www.amazon.com/Applied-Text-Analysis-Python-Language-Aware/dp/1491963042

Charu C. Aggarwal (2018),

Machine Learning for Text,

Springer



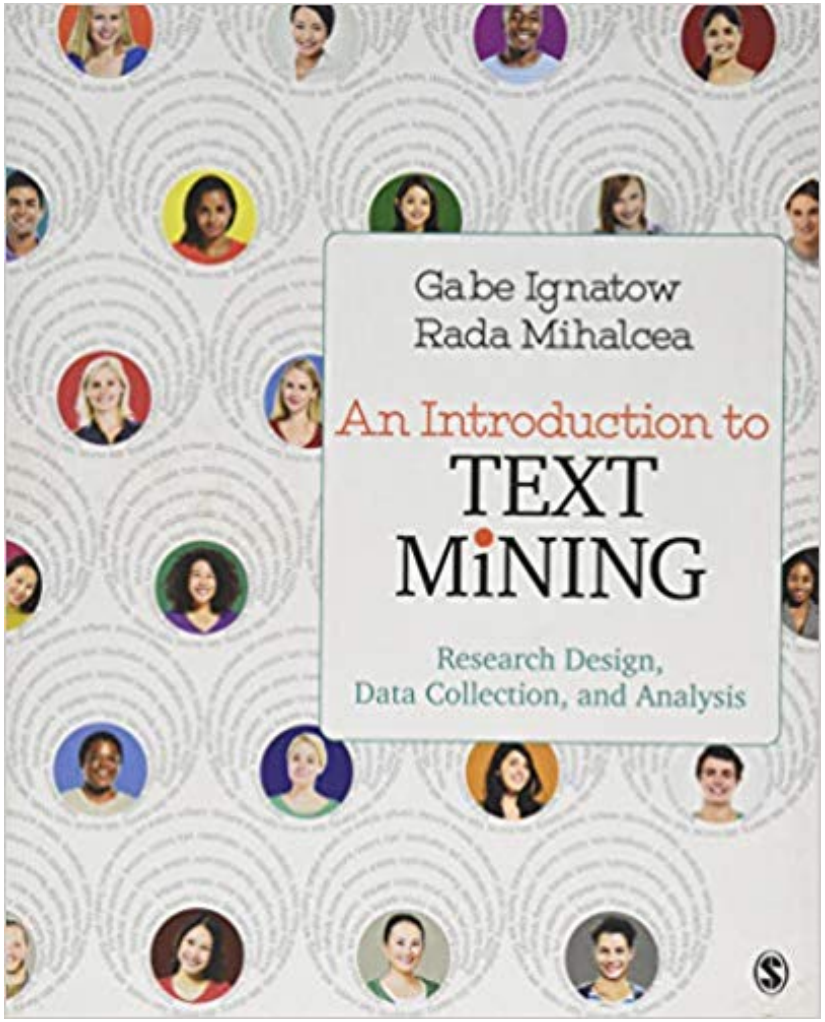
Source: https://www.amazon.com/Machine-Learning-Text-Charu-Aggarwal/dp/3319735306

Gabe Ignatow and Rada F. Mihalcea (2017),

An Introduction to Text Mining:

Research Design, Data Collection, and Analysis,

SAGE Publications.

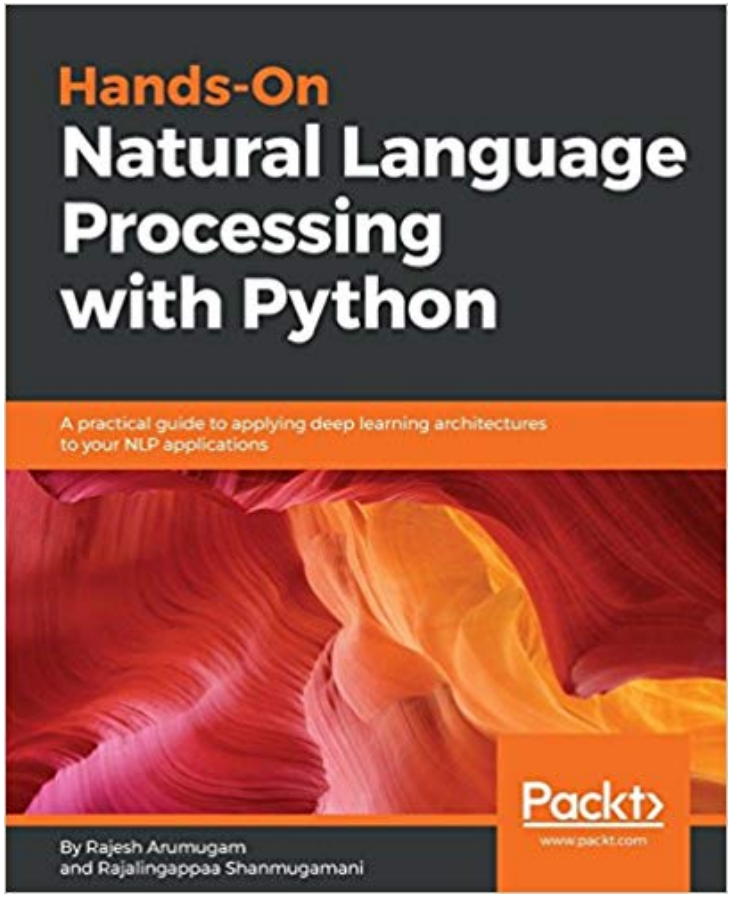


Source: https://www.amazon.com/Introduction-Text-Mining-Research-Collection/dp/1506337007

Rajesh Arumugam (2018),

Hands-On Natural Language Processing with Python:

A practical guide to applying deep learning architectures to your NLP applications, Packt



https://www.amazon.com/Hands-Natural-Language-Processing-Python/dp/178913949X

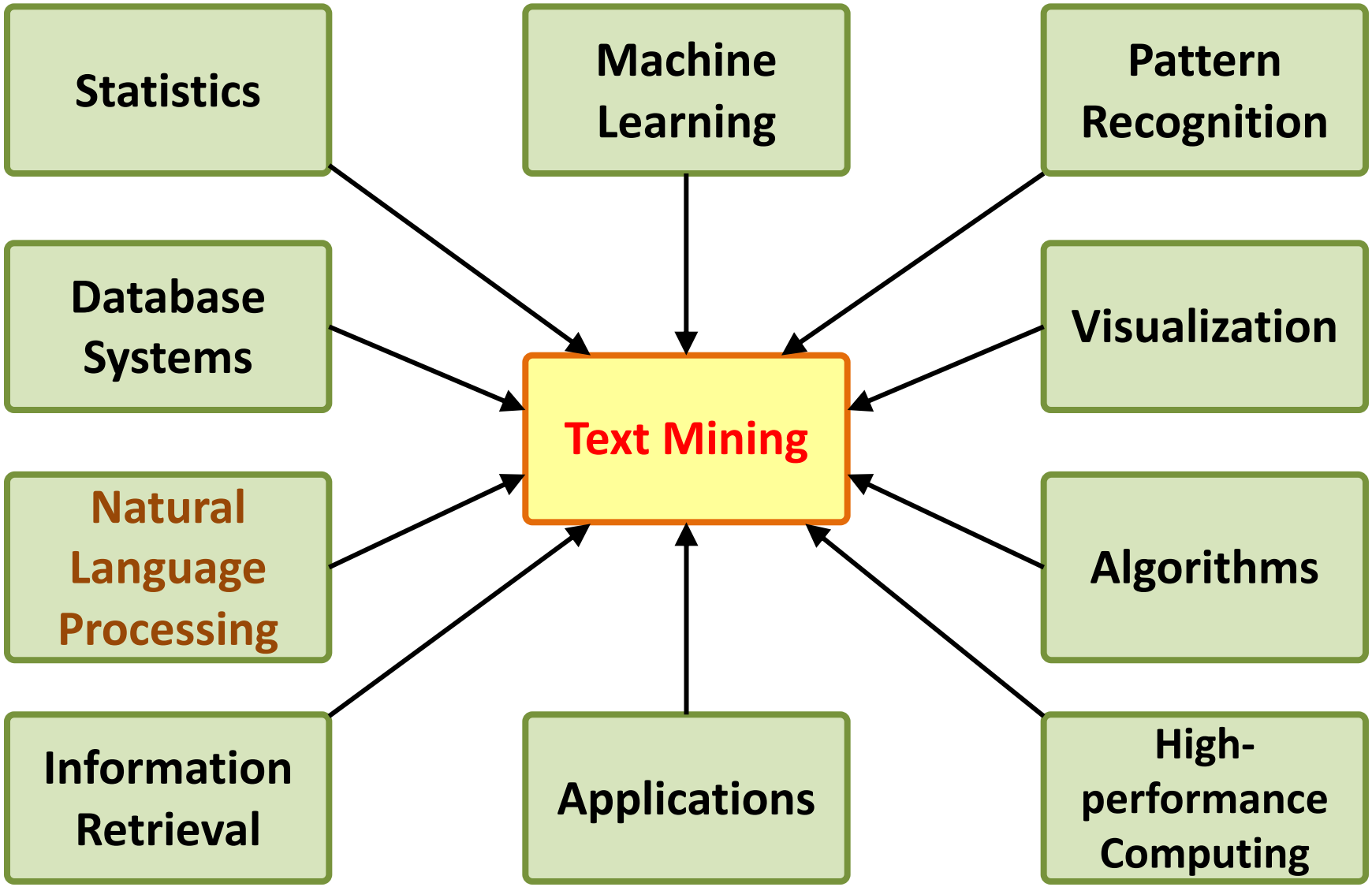
Text Analytics

- **Text Analytics** =
  Information Retrieval +
  Information Extraction +
  Data Mining +
  Web Mining
- **Text Analytics** =
  Information Retrieval +
  Text Mining

Text Mining

- Text Data Mining
- Knowledge Discovery in Textual Databases

Text Mining Technologies

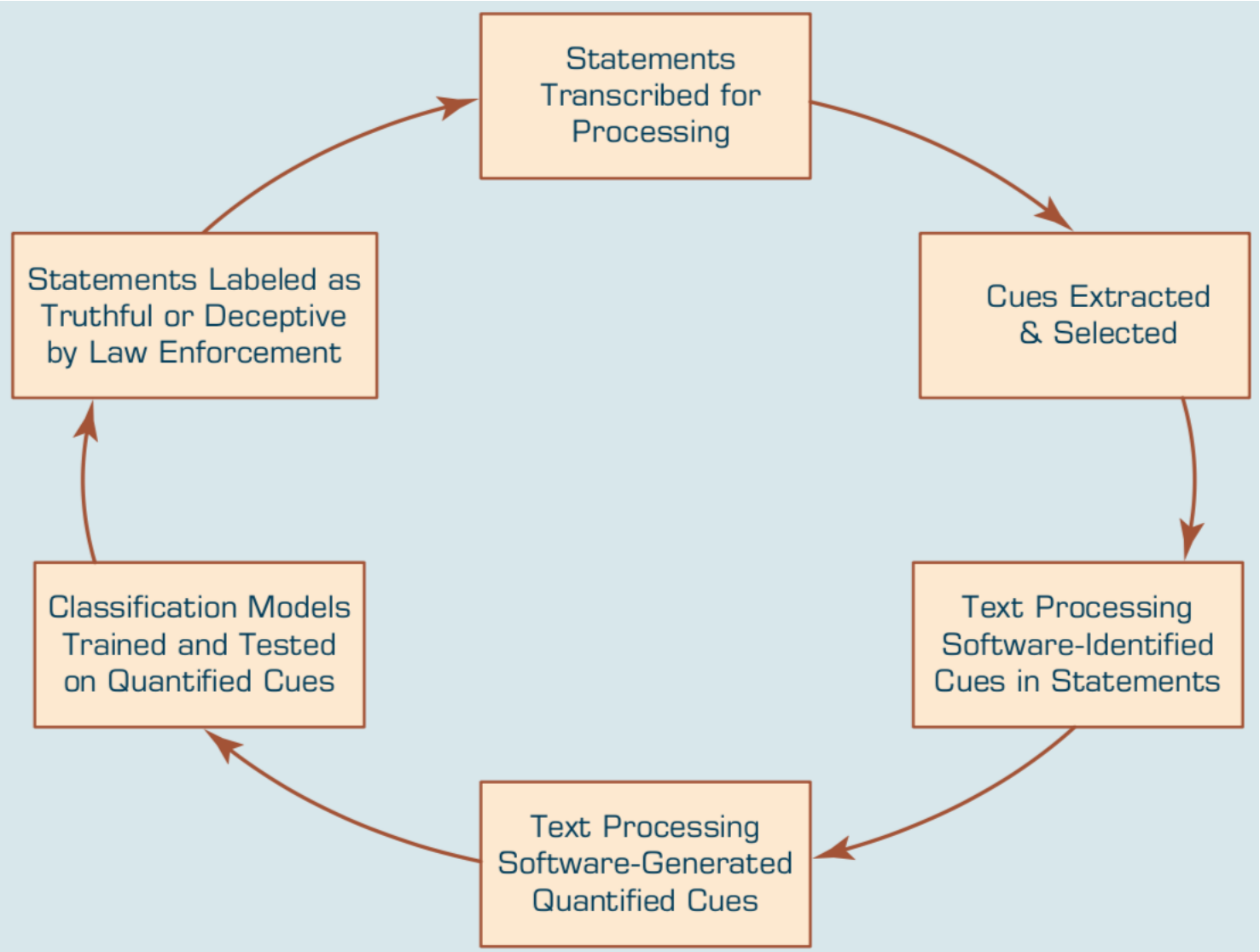


Adapted from: Jiawei Han and Micheline Kamber (2011), Data Mining: Concepts and Techniques, Third Edition, Elsevier

Application Areas of Text Mining

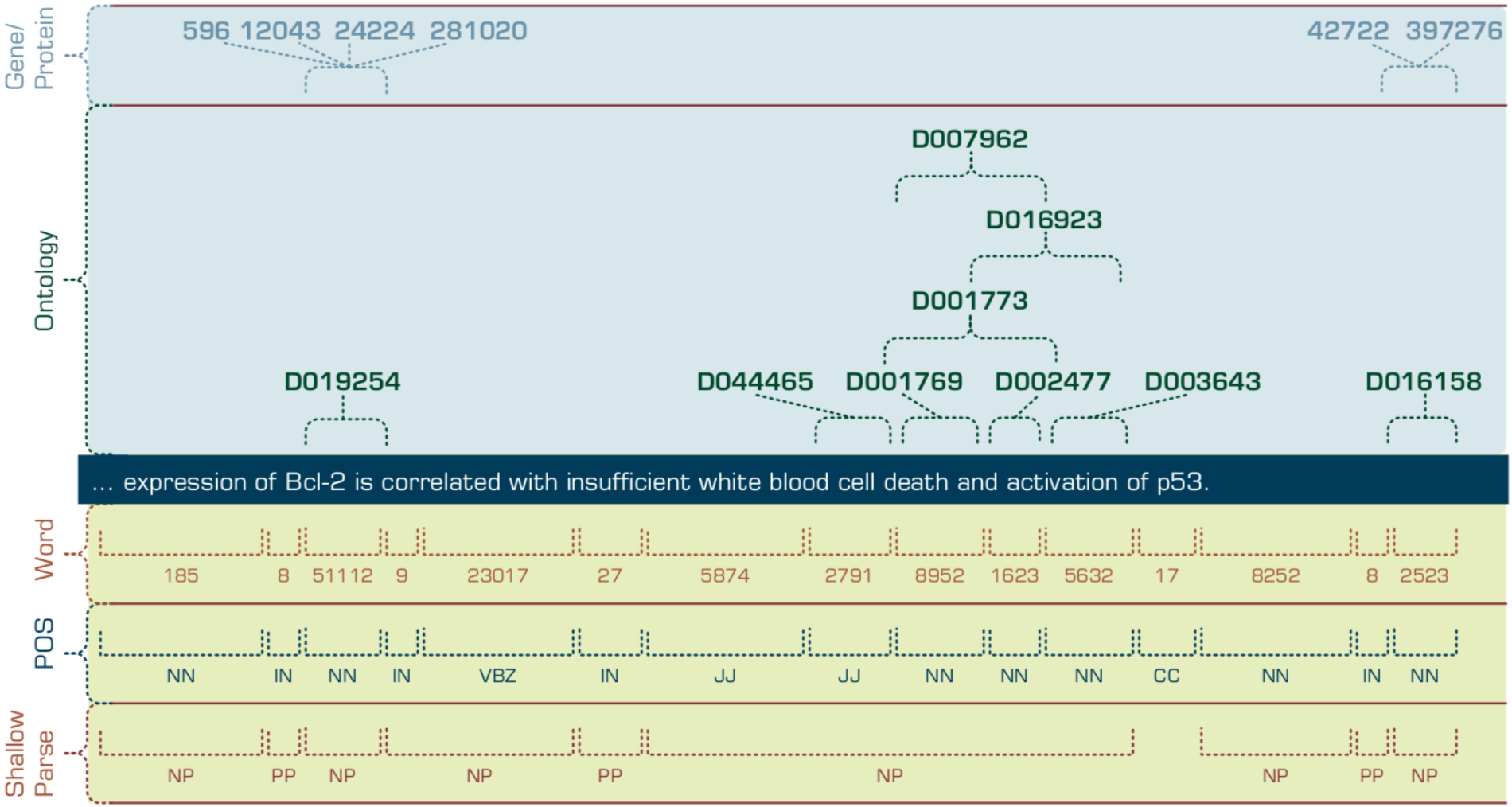
- Information extraction
- Topic tracking
- Summarization
- Categorization
- Clustering
- Concept linking
- Question answering

Text-Based Deception-Detection Process



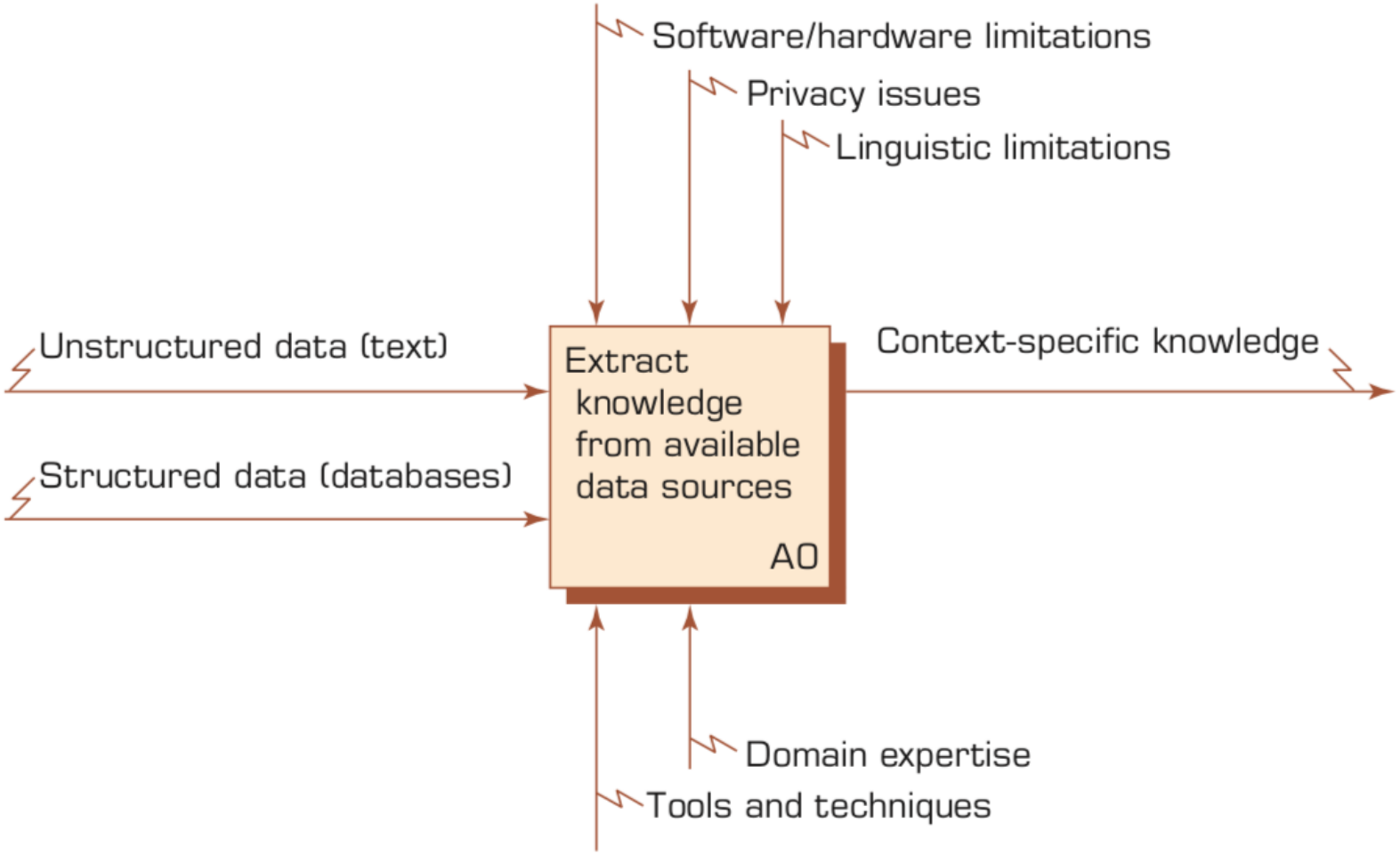
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Multilevel Analysis of Text for Gene/Protein Interaction Identification



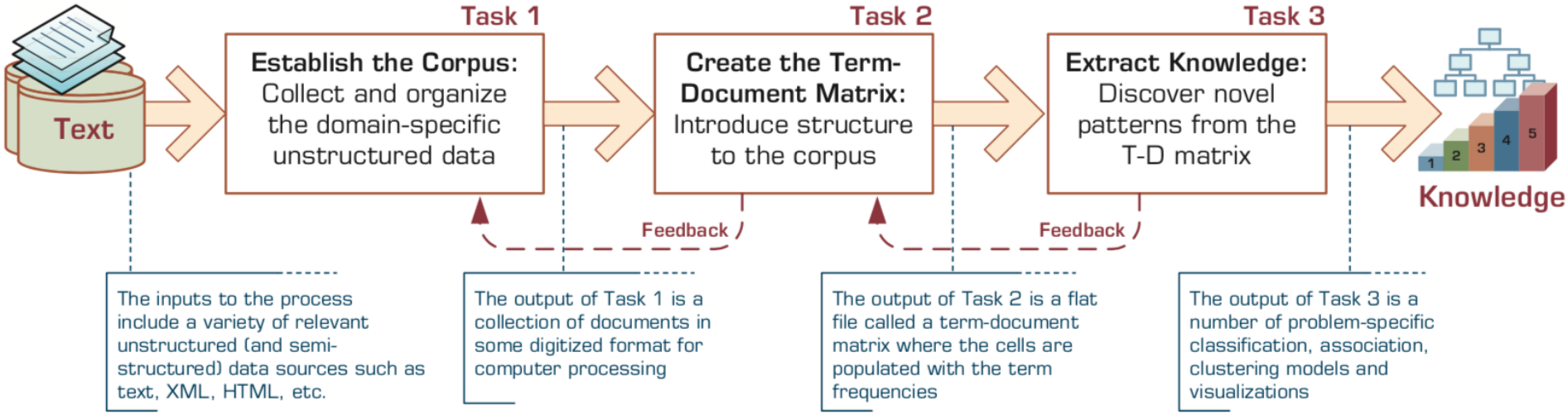
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Context Diagram for the Text Mining Process



Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

The Three-Step/Task Text Mining Process



Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Term–Document Matrix

Documents \ Terms	Investment Risk	Project Management	Software Engineering	Development	SAP	…	
Document 1	1			1			
Document 2		1					
Document 3			3		1		
Document 4		1					
Document 5			2	1			
Document 6	1			1			
…							

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Emotions


Love

Anger

Joy

Sadness

Surprise

Fear

Source: Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,” Springer, 2nd Edition,

Example of Opinion: review segment on iPhone

"I bought an iPhone a few days ago.

It was such a nice phone.

The touch screen was really cool.

The voice quality was clear too.

However, my mother was mad with me as I did not tell her before I bought it.

She also thought the phone was too expensive, and wanted me to return it to the shop. … "

Source: Bing Liu (2011) , "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data," Springer, 2nd Edition,

Example of Opinion: review segment on iPhone

"(1) I bought an iPhone a few days ago.

(2) It was such a **nice** phone.

(3) The touch screen was really **cool**.

+Positive Opinion

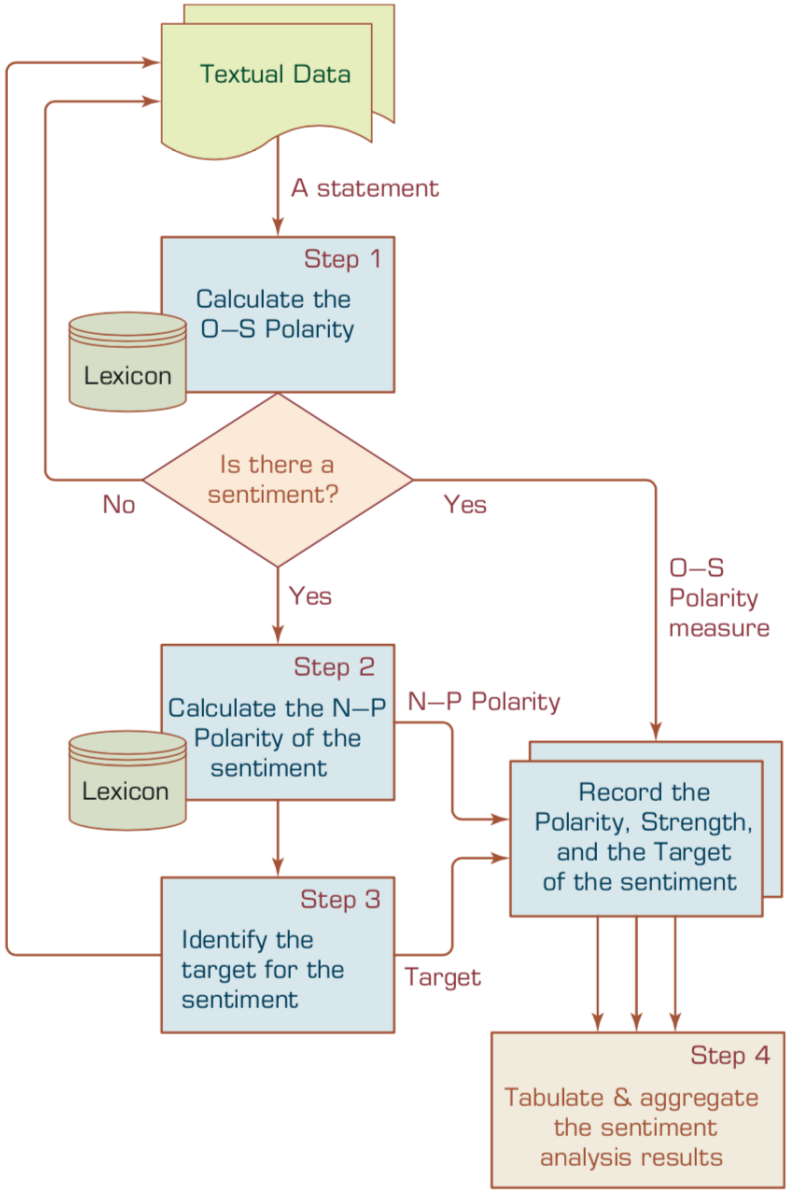
(4) The voice quality was **clear** too.

(5) However, my mother was mad with me as I did not tell her before I bought it.

(6) She also thought the phone was too **expensive**, and wanted me to return it to the shop. … "

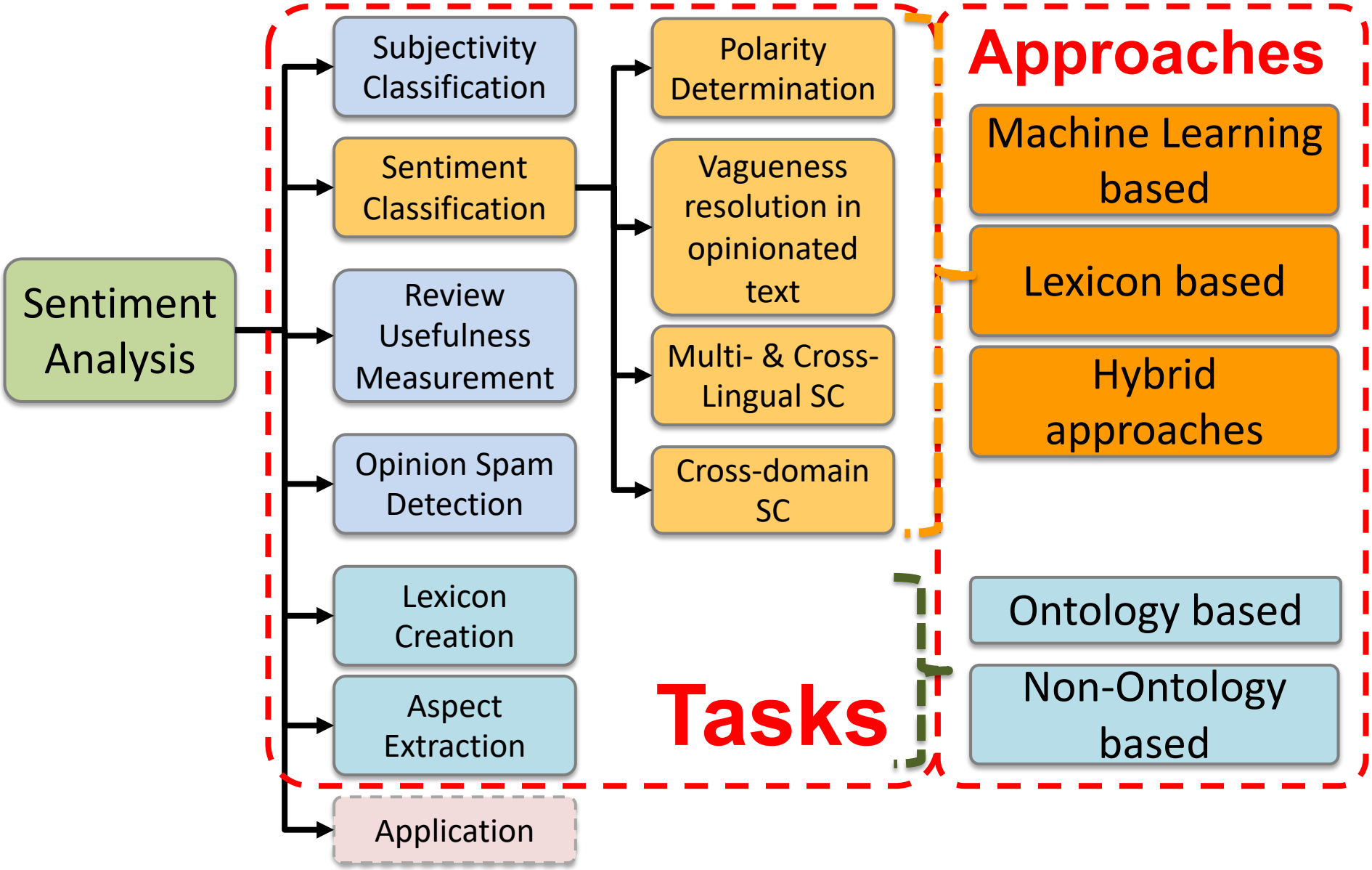
-Negative Opinion

A Multistep Process to Sentiment Analysis



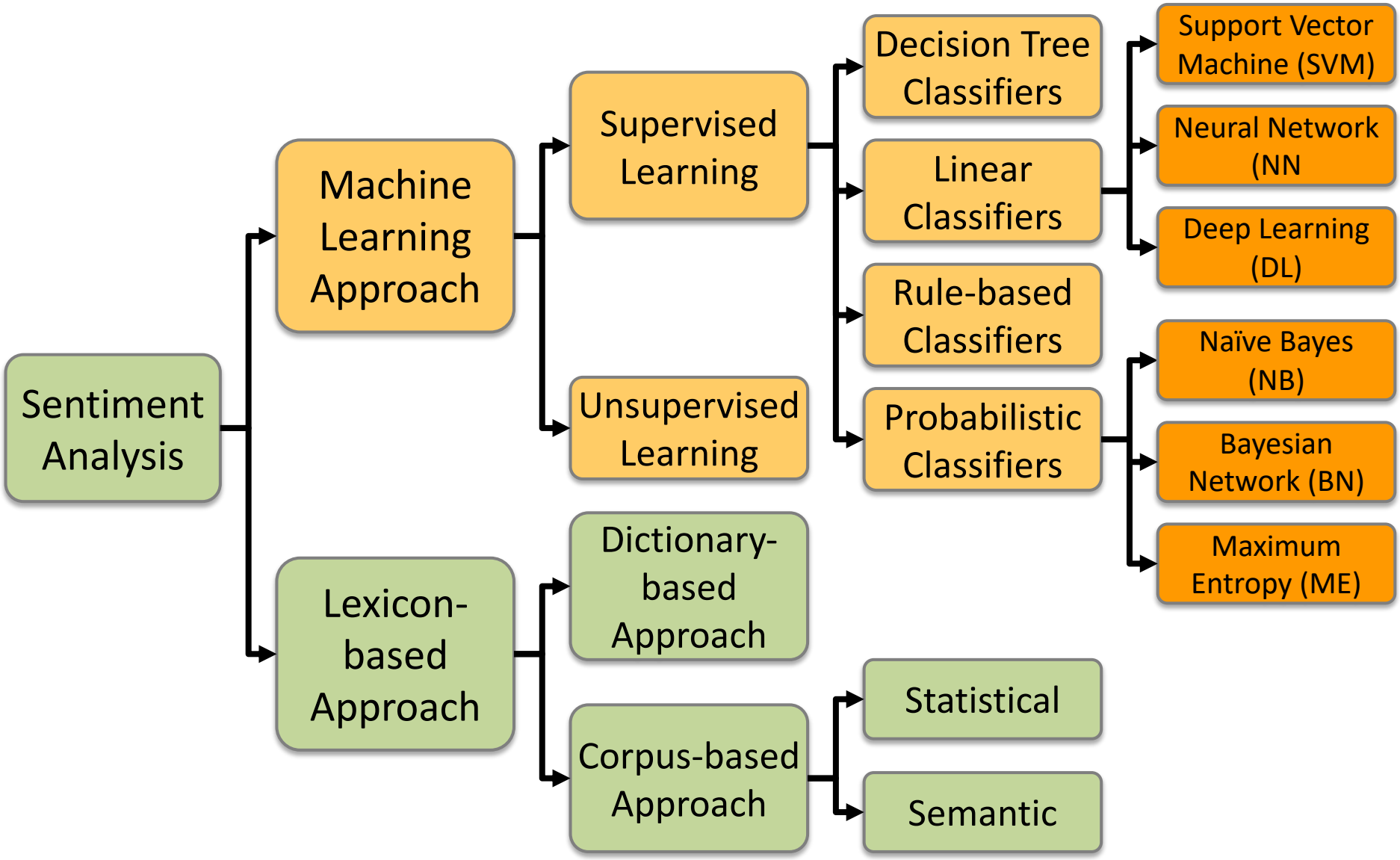
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Sentiment Analysis



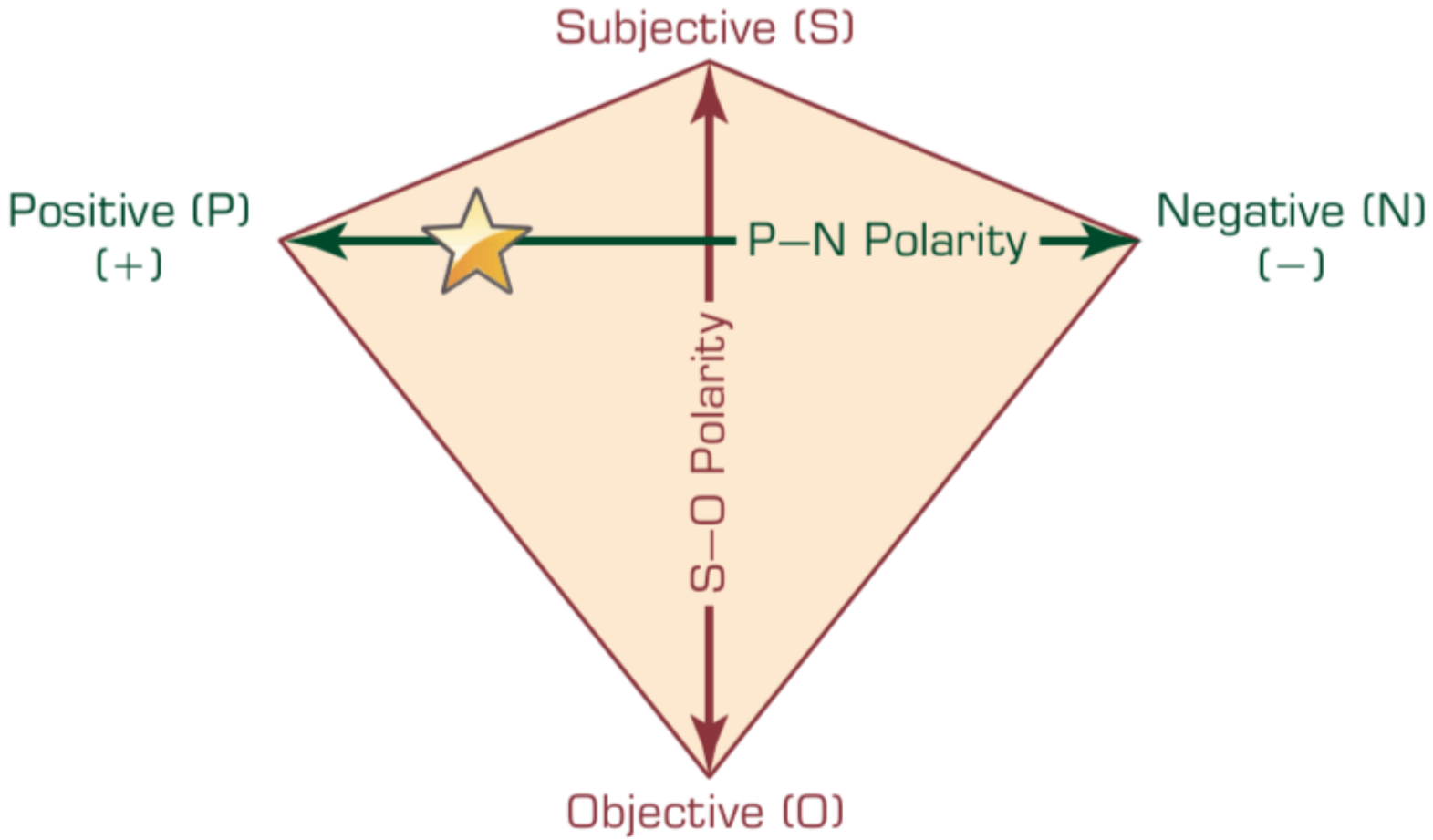
Source: Kumar Ravi and Vadlamani Ravi (2015), "A survey on opinion mining and sentiment analysis: tasks, approaches and applications." Knowledge-Based Systems, 89, pp.14-46.

Sentiment Classification Techniques



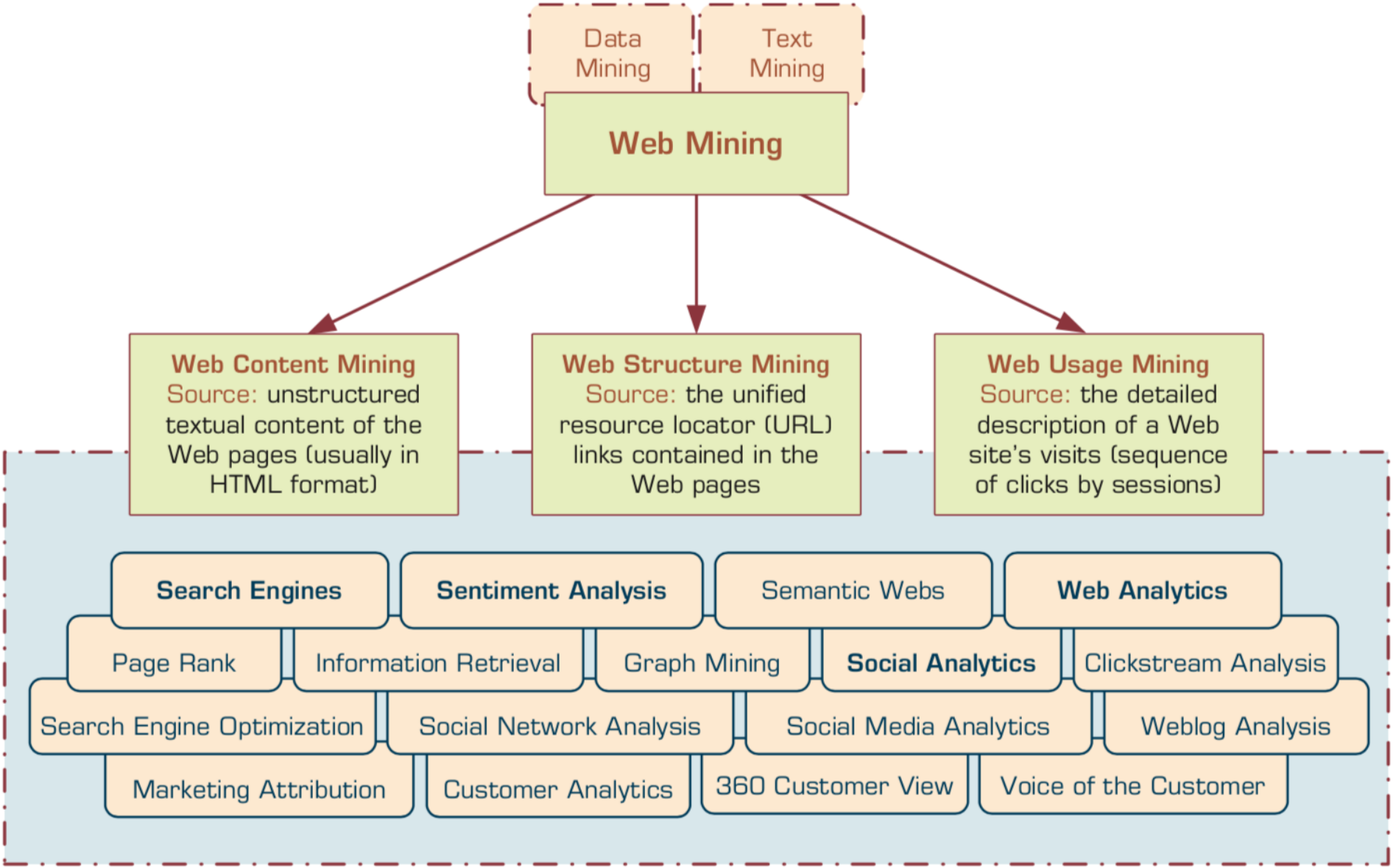
Source: Jesus Serrano-Guerrero, Jose A. Olivas, Francisco P. Romero, and Enrique Herrera-Viedma (2015), "Sentiment analysis: A review and comparative analysis of web services," Information Sciences, 311, pp. 18-38.

P–N Polarity and S–O Polarity Relationship



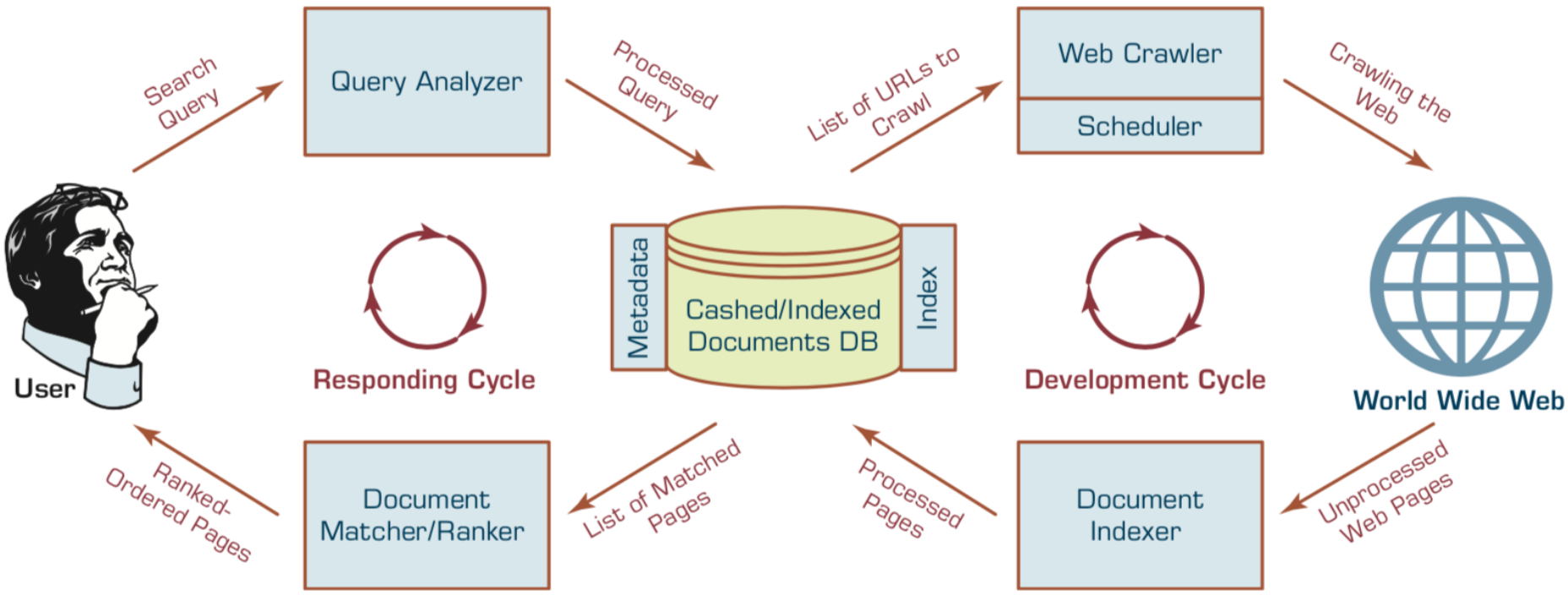
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Taxonomy of Web Mining



Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Structure of a Typical Internet Search Engine

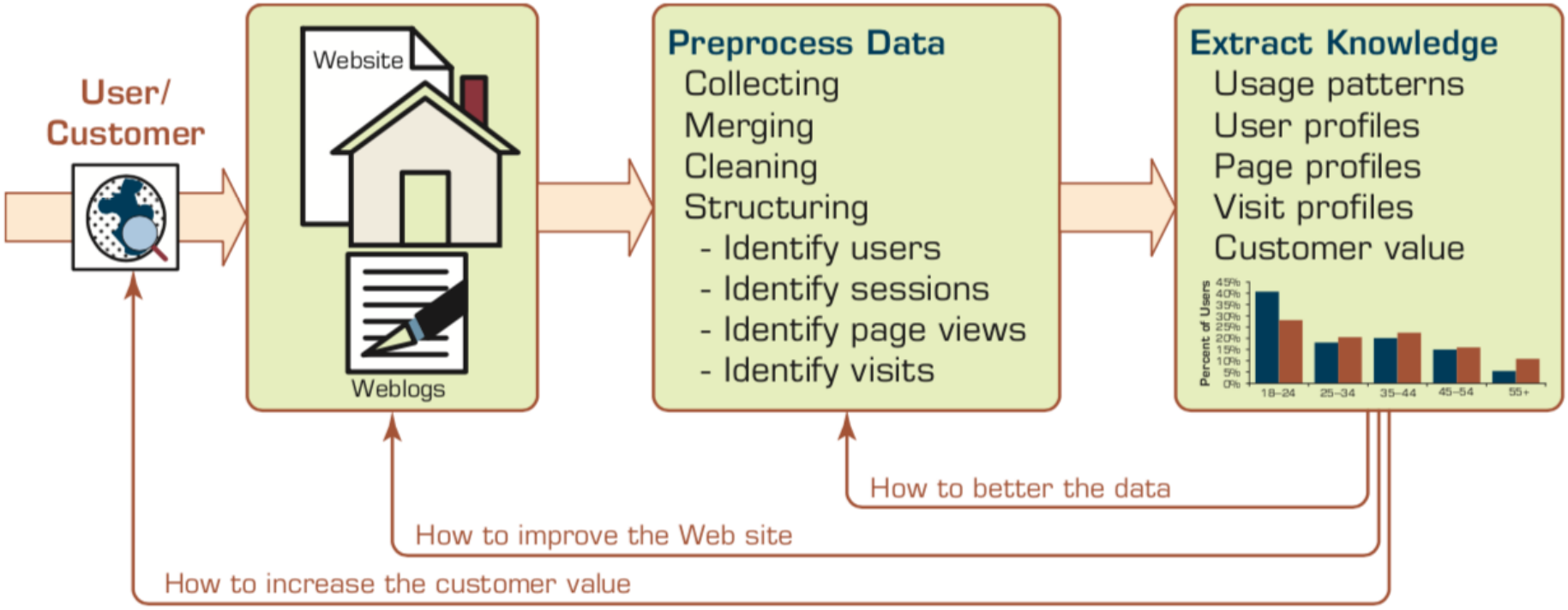


Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Web Usage Mining (Web Analytics)

- **Web usage mining** (**Web analytics**) is the extraction of useful information from data generated through Web page visits and transactions.
- **Clickstream Analysis**

Extraction of Knowledge from Web Usage Data

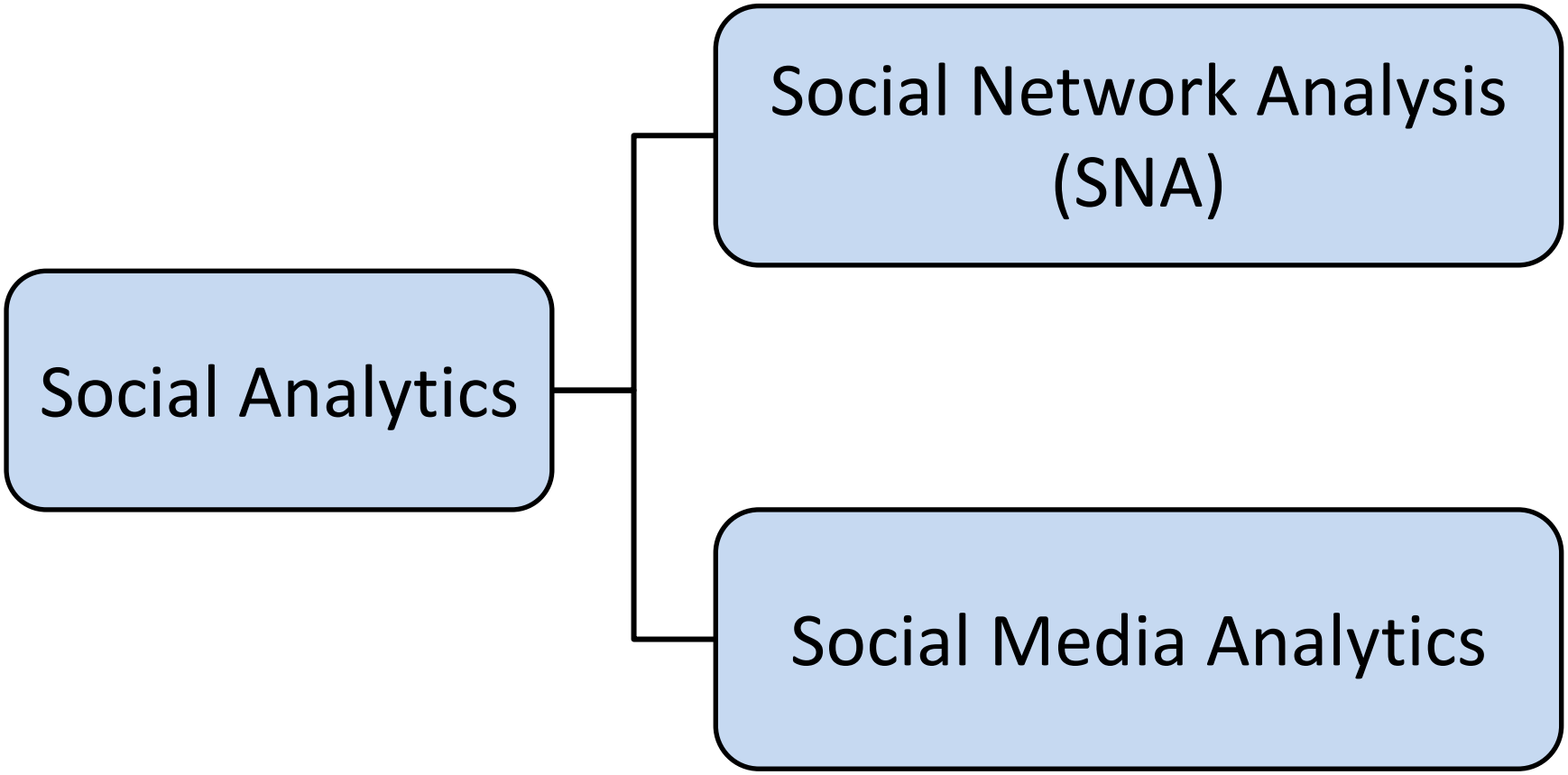


Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Social Analytics

- Social analytics is defined as monitoring, analyzing, measuring and interpreting digital interactions and relationships of people, topics, ideas and content.

Branches of Social Analytics



Text Mining Technologies

Text Mining (TM)

Natural Language Processing (NLP)

Text Mining Concepts

- 85-90 percent of all corporate data is in some kind of unstructured form (e.g., text)
- Unstructured corporate data is doubling in size every 18 months
- Tapping into these information sources is not an option, but a need to stay competitive
- Answer: text mining
  - A semi-automated process of extracting knowledge from unstructured data sources
  - a.k.a. text data mining or knowledge discovery in textual databases

Text mining

Text Data Mining

Intelligent Text Analysis

Knowledge-Discovery in Text (KDT)

Source: Vishal Gupta and Gurpreet S. Lehal (2009), "A survey of text mining techniques and applications," Journal of emerging technologies in web intelligence, vol. 1, no. 1, pp. 60-76.

Text Mining
(text data mining)

the process of deriving high-quality information from text

http://en.wikipedia.org/wiki/Text_mining

Text Mining:

the process of extracting interesting and non-trivial information and knowledge from unstructured text.

Source: Vishal Gupta and Gurpreet S. Lehal (2009), "A survey of text mining techniques and applications," Journal of emerging technologies in web intelligence, vol. 1, no. 1, pp. 60-76.

Text Mining:

discovery by computer of

new, previously

unknown information,

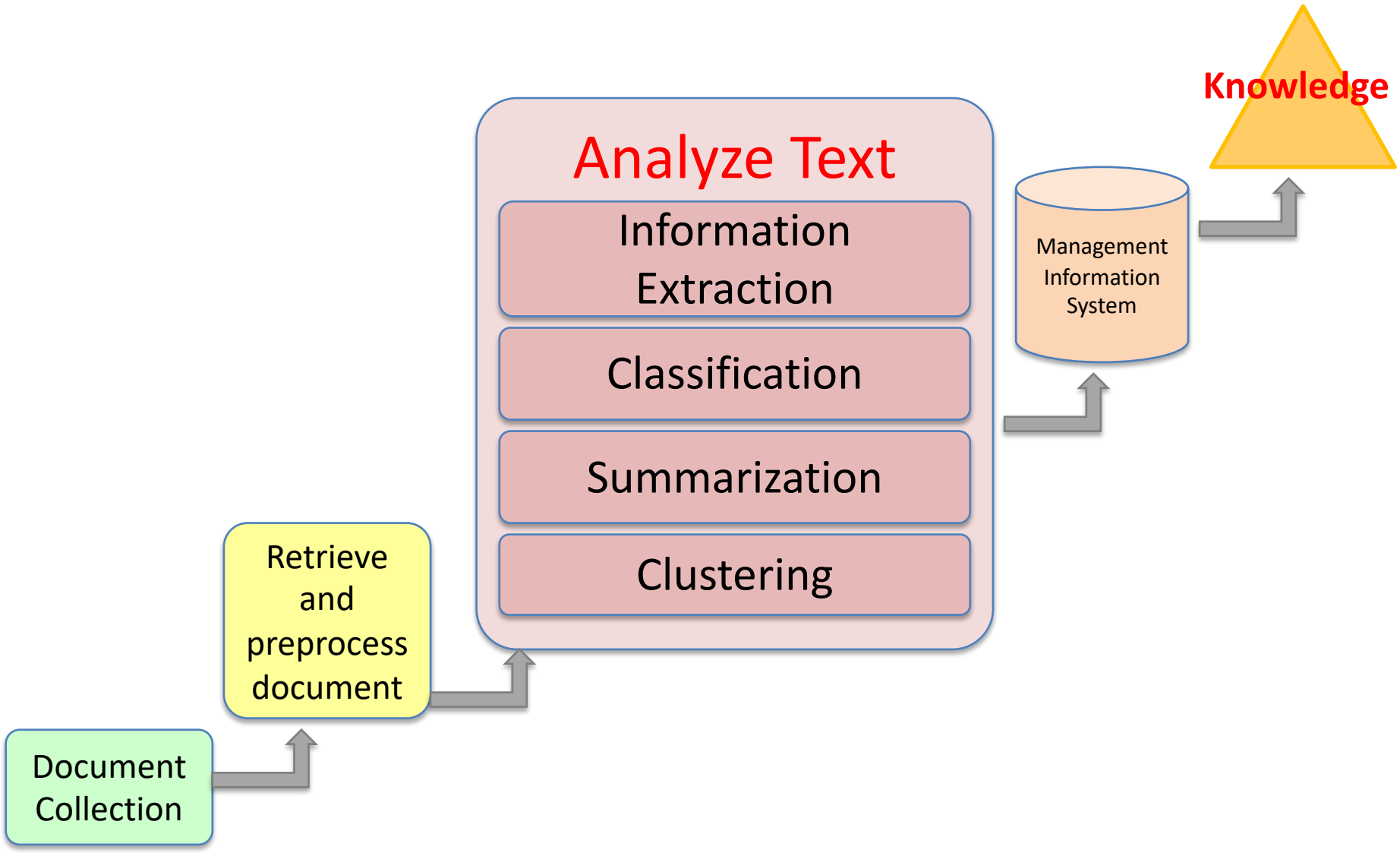
by automatically

extracting information

from different written resources.

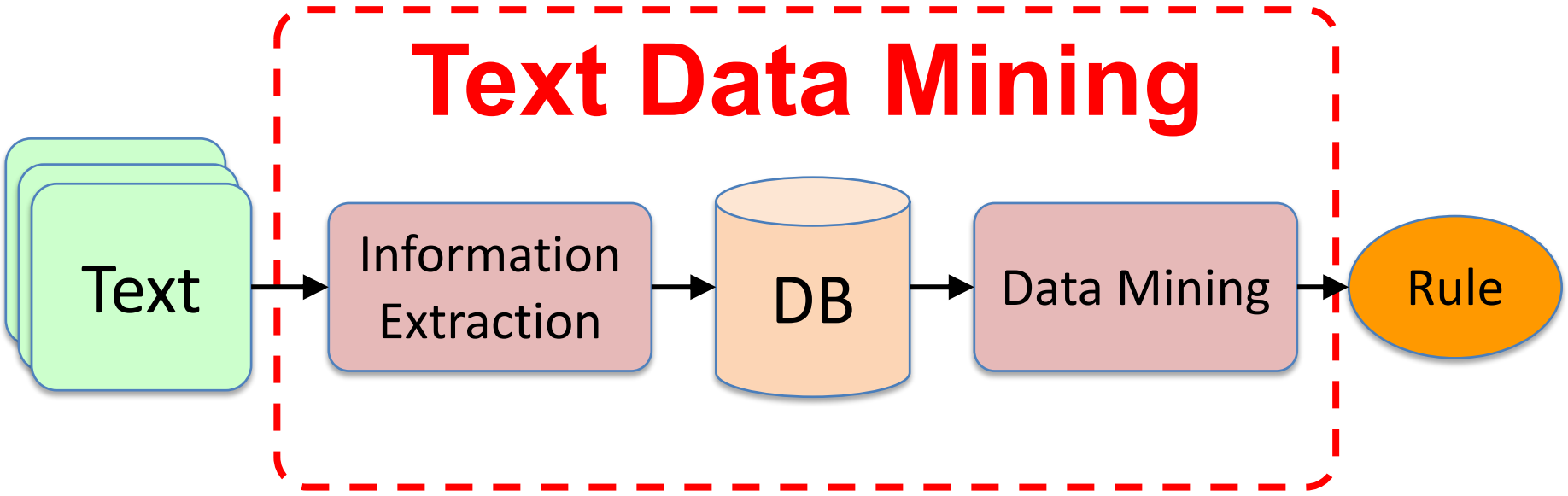
Source: Vishal Gupta and Gurpreet S. Lehal (2009), "A survey of text mining techniques and applications," Journal of emerging technologies in web intelligence, vol. 1, no. 1, pp. 60-76.

An example of Text Mining



Source: Vishal Gupta and Gurpreet S. Lehal (2009), "A survey of text mining techniques and applications," Journal of emerging technologies in web intelligence, vol. 1, no. 1, pp. 60-76.

Overview of Information Extraction based Text Mining Framework



Source: Vishal Gupta and Gurpreet S. Lehal (2009), "A survey of text mining techniques and applications," Journal of emerging technologies in web intelligence, vol. 1, no. 1, pp. 60-76.

Natural Language Processing (NLP)

- **Natural language processing (NLP)** is an important component of **text mining** and is a subfield of **artificial intelligence** and **computational linguistics**.

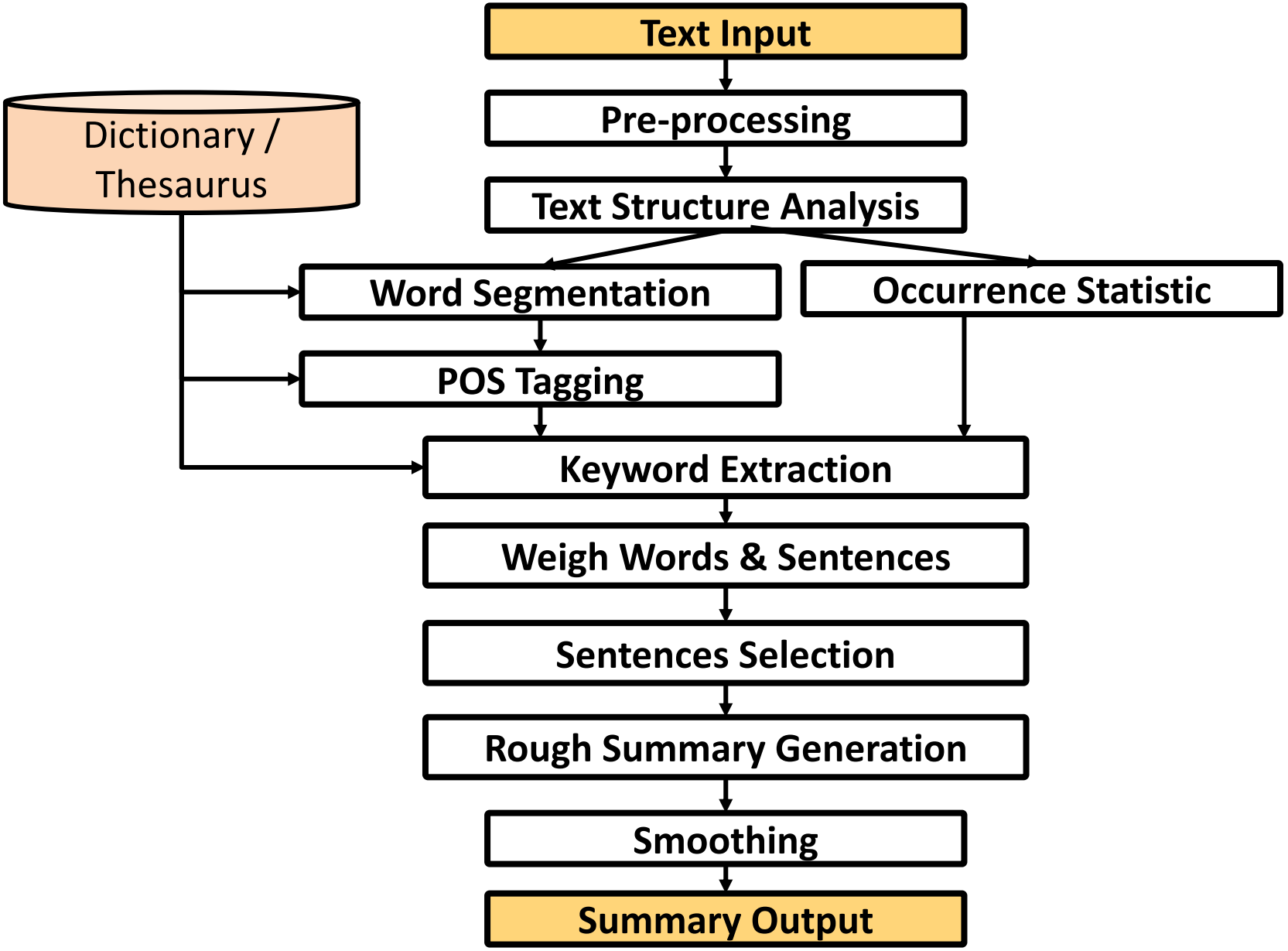
Natural Language Processing (NLP) and Text Mining

- **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

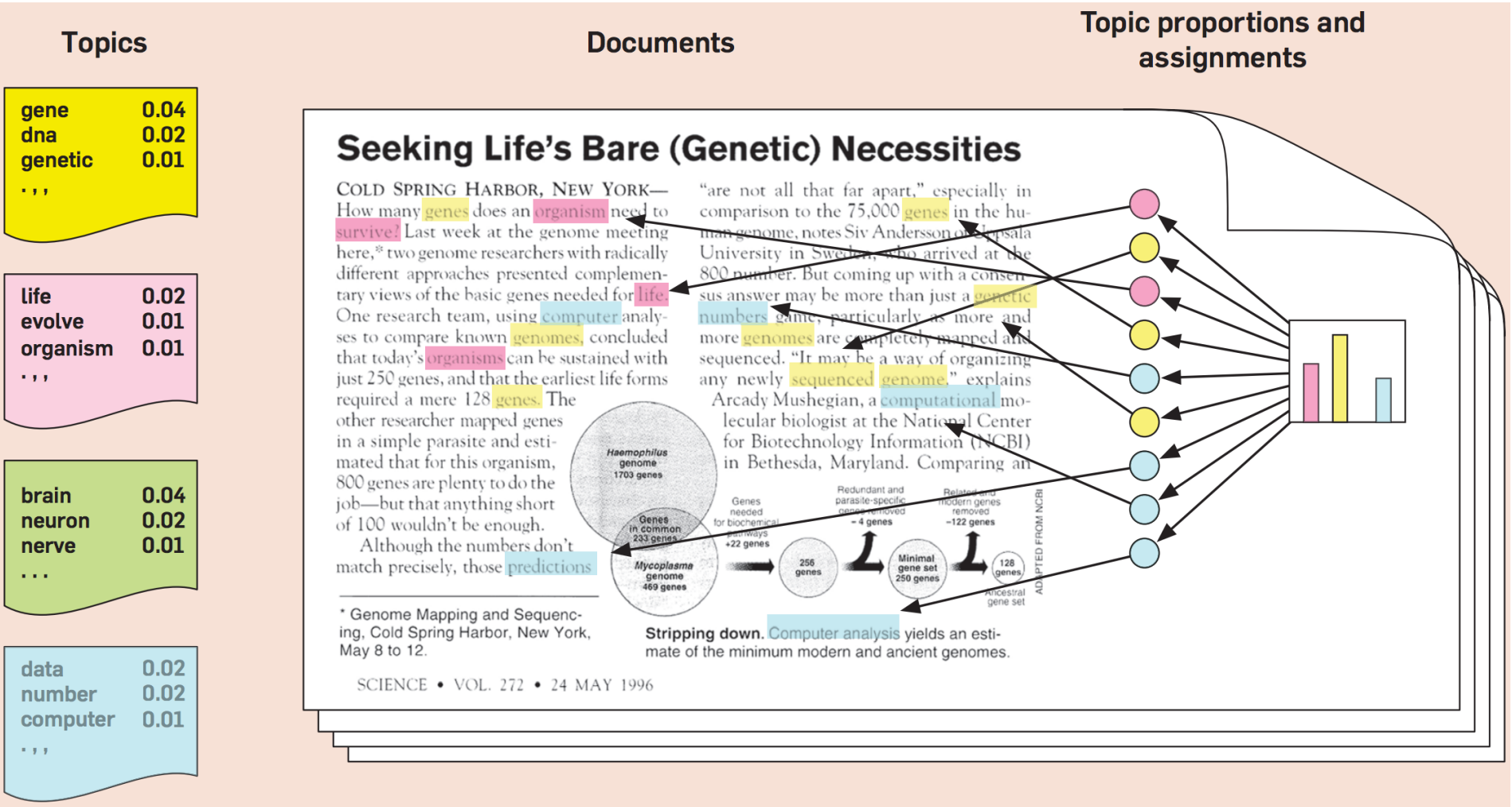
Source: Nitin Hardeniya (2015), NLTK Essentials, Packt Publishing; Florian Leitner (2015), Text mining - from Bayes rule to dependency parsing

Text Summarization



Source: Vishal Gupta and Gurpreet S. Lehal (2009), "A survey of text mining techniques and applications," Journal of emerging technologies in web intelligence, vol. 1, no. 1, pp. 60-76.

Topic Modeling



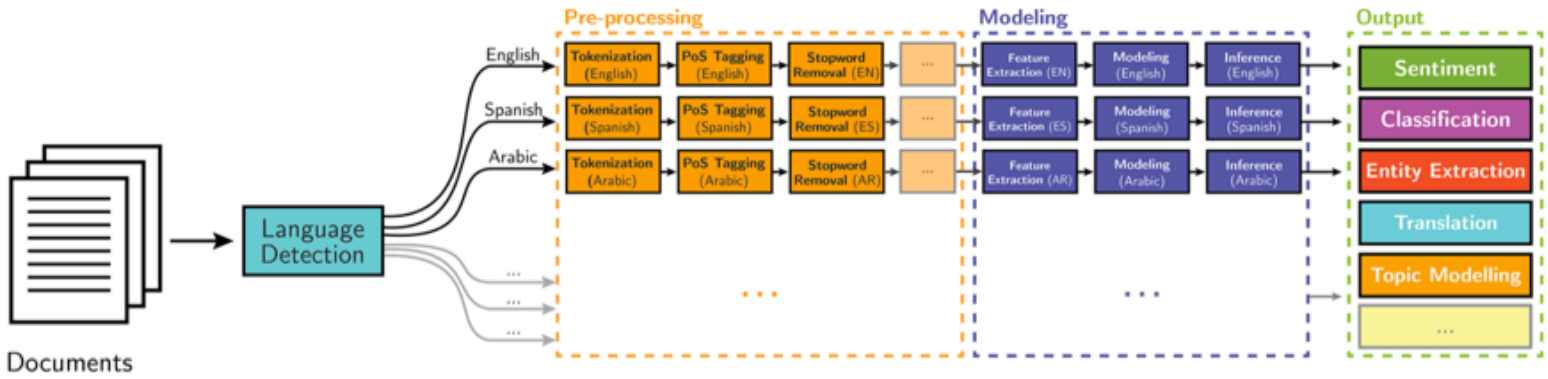
Natural Language Processing (NLP)

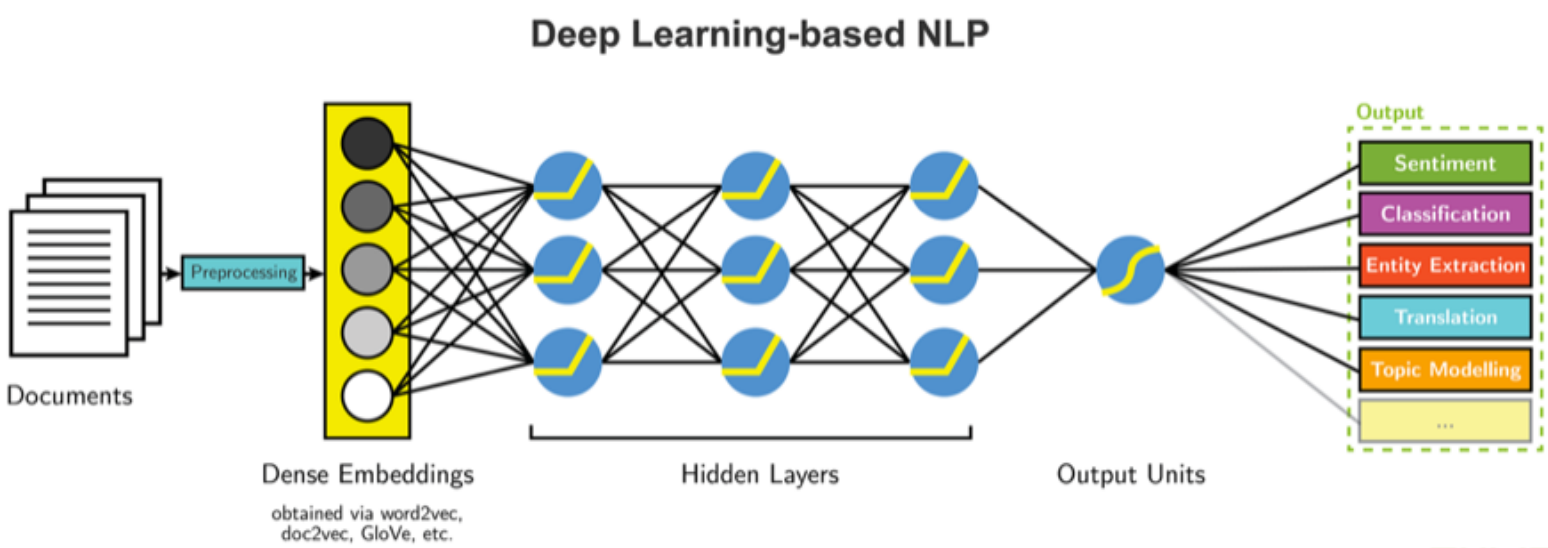
- Part-of-speech tagging
- Text segmentation
- Word sense disambiguation
- Syntactic ambiguity
- Imperfect or irregular input
- Speech acts

NLP Tasks

- Question answering
- Automatic summarization
- Natural language generation
- Natural language understanding
- Machine translation
- Foreign language reading
- Foreign language writing.
- Speech recognition
- Text-to-speech
- Text proofing
- Optical character recognition

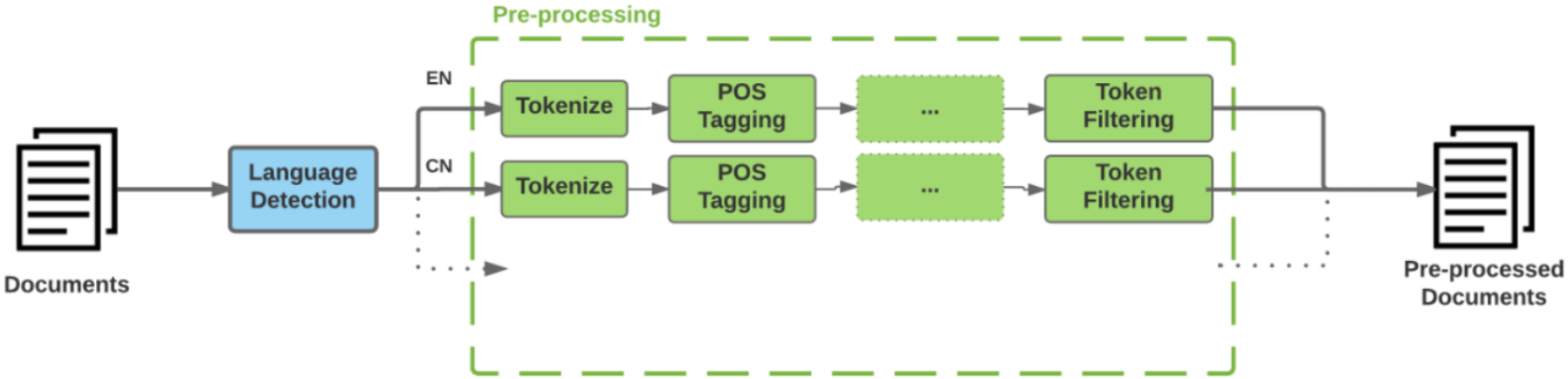
NLP

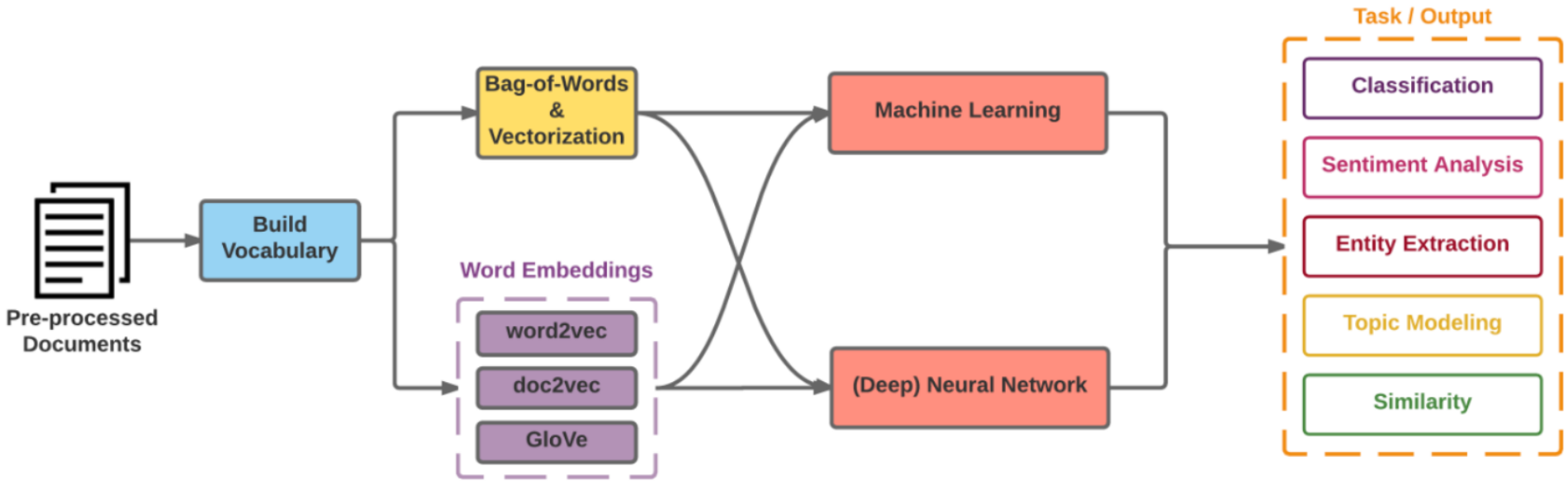




Source: http://blog.aylien.com/leveraging-deep-learning-for-multilingual/

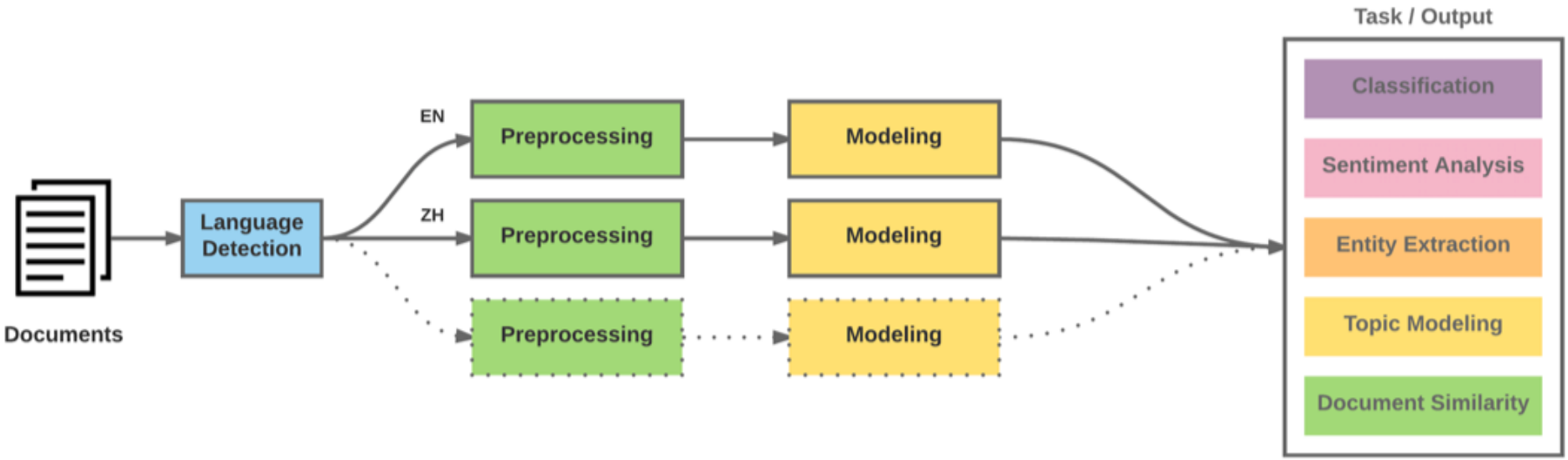
Modern NLP Pipeline





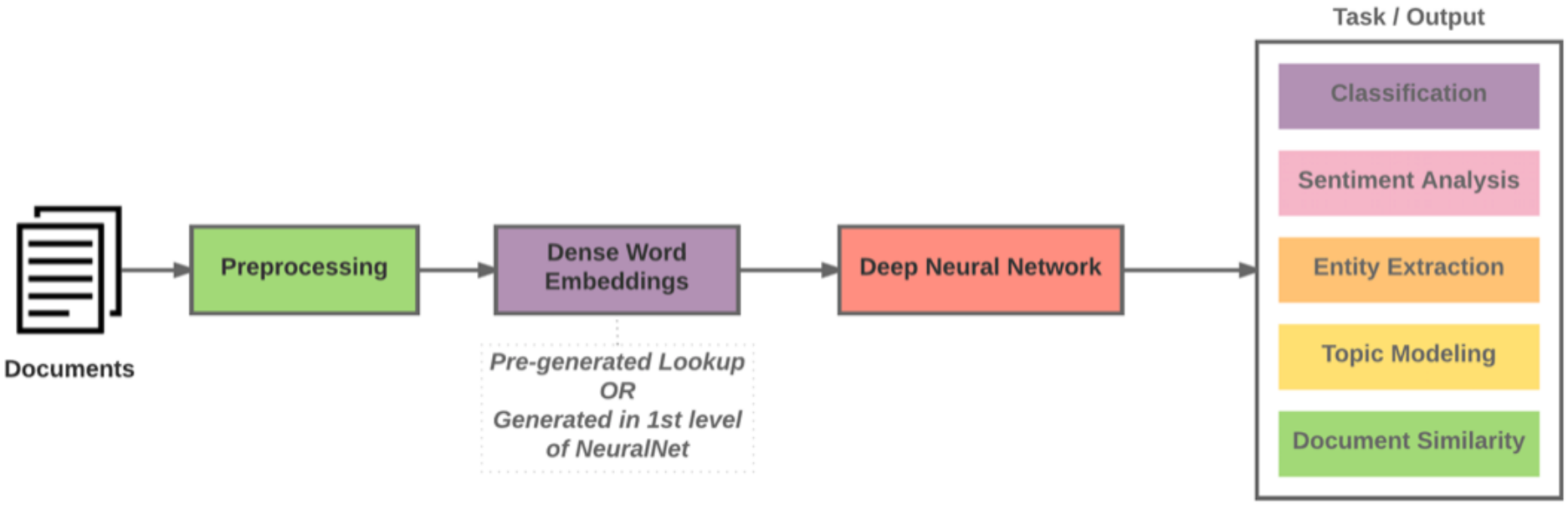
Source: https://github.com/fortiema/talks/blob/master/opendata2016sh/pragmatic-nlp-opendata2016sh.pdf

Modern NLP Pipeline



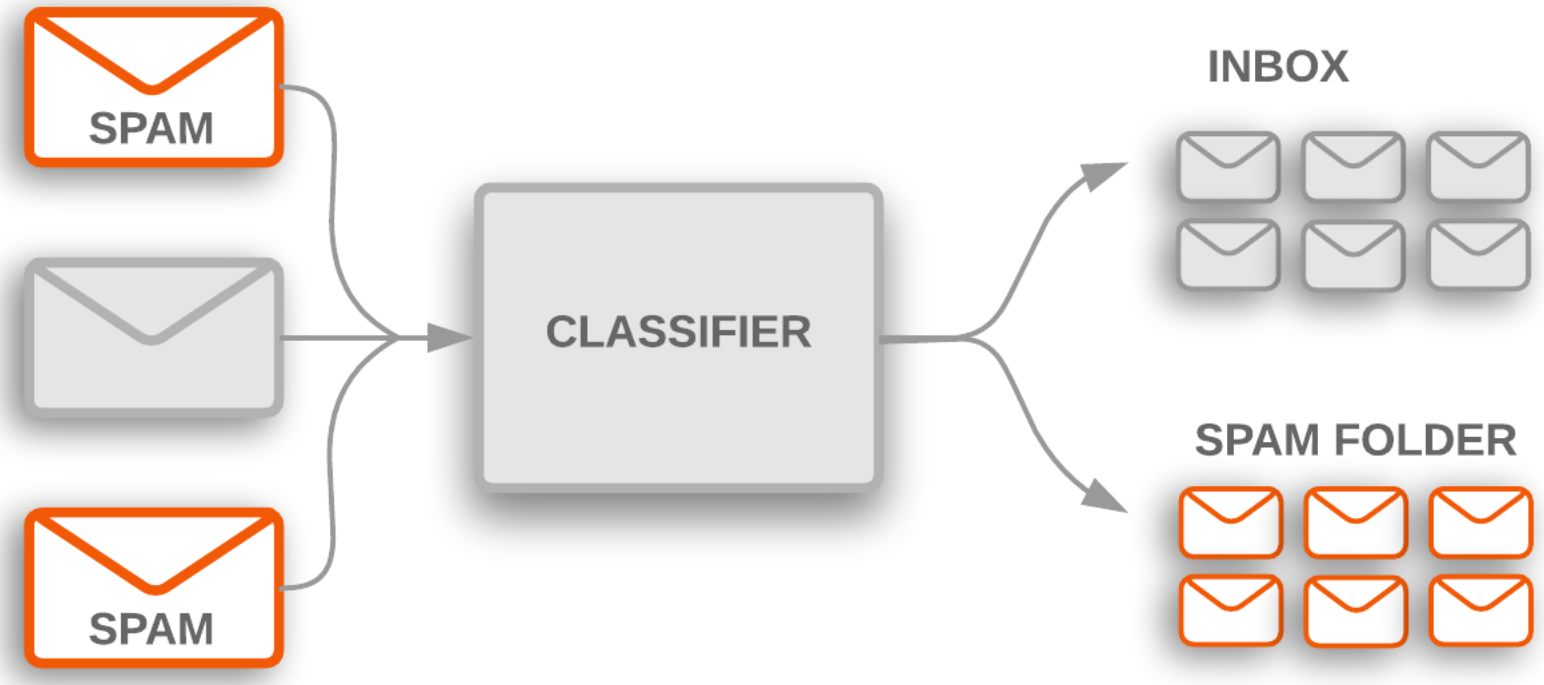
Source: http://mattfortier.me/2017/01/31/nlp-intro-pt-1-overview/

Deep Learning NLP



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

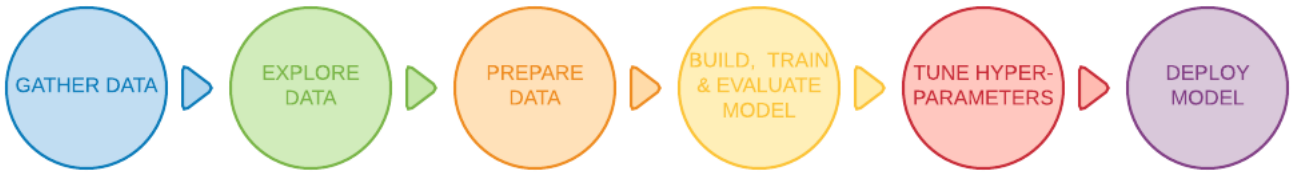
Text Classification



Source: https://developers.google.com/machine-learning/guides/text-classification/

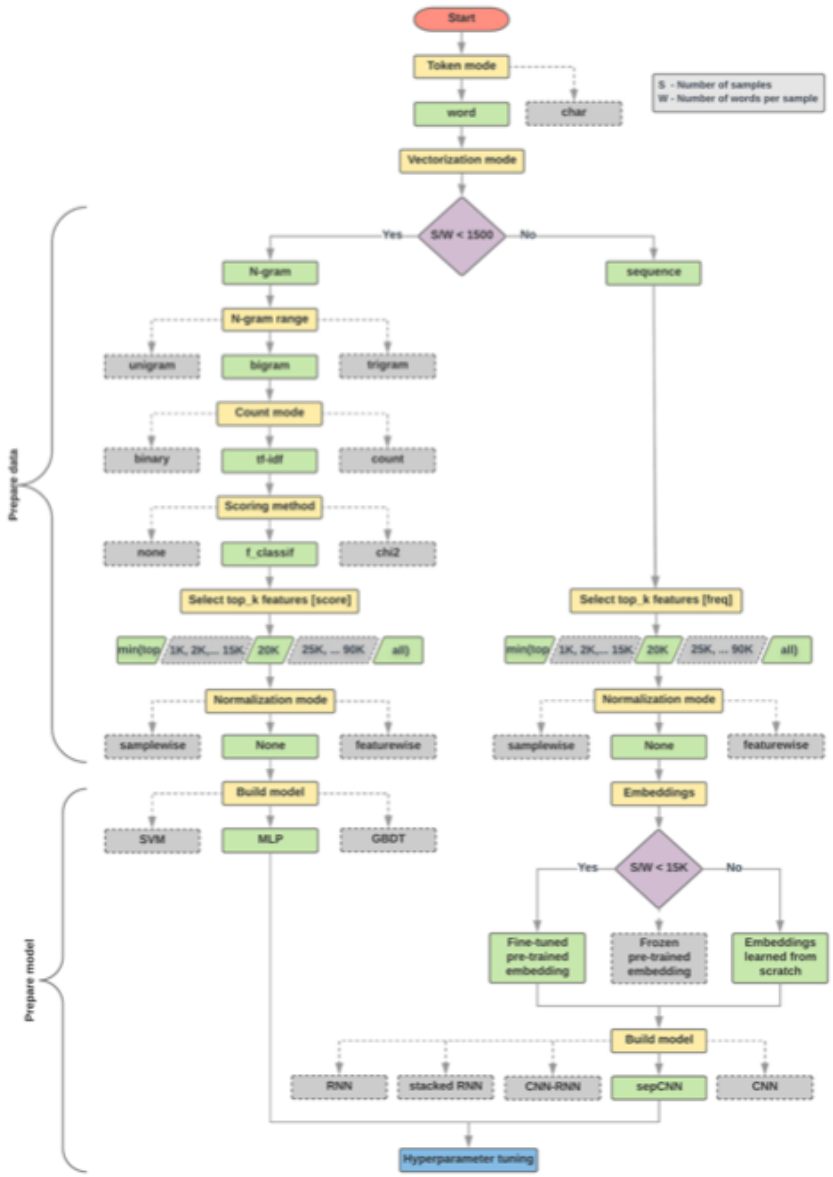
Text Classification Workflow

- Step 1: Gather Data
- Step 2: Explore Your Data
- Step 2.5: Choose a Model*
- Step 3: Prepare Your Data
- Step 4: Build, Train, and Evaluate Your Model
- Step 5: Tune Hyperparameters
- Step 6: Deploy Your Model



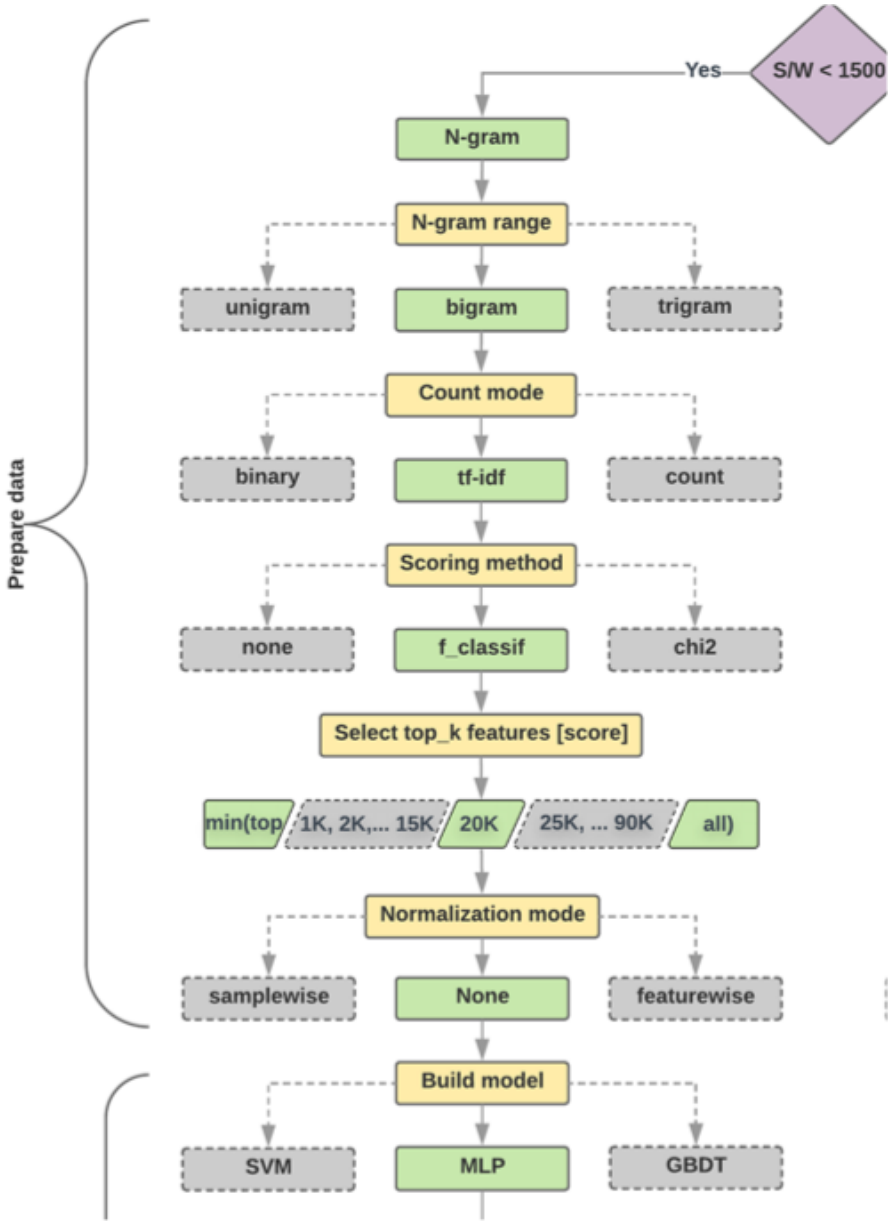
Source: https://developers.google.com/machine-learning/guides/text-classification/

Text Classification Flowchart



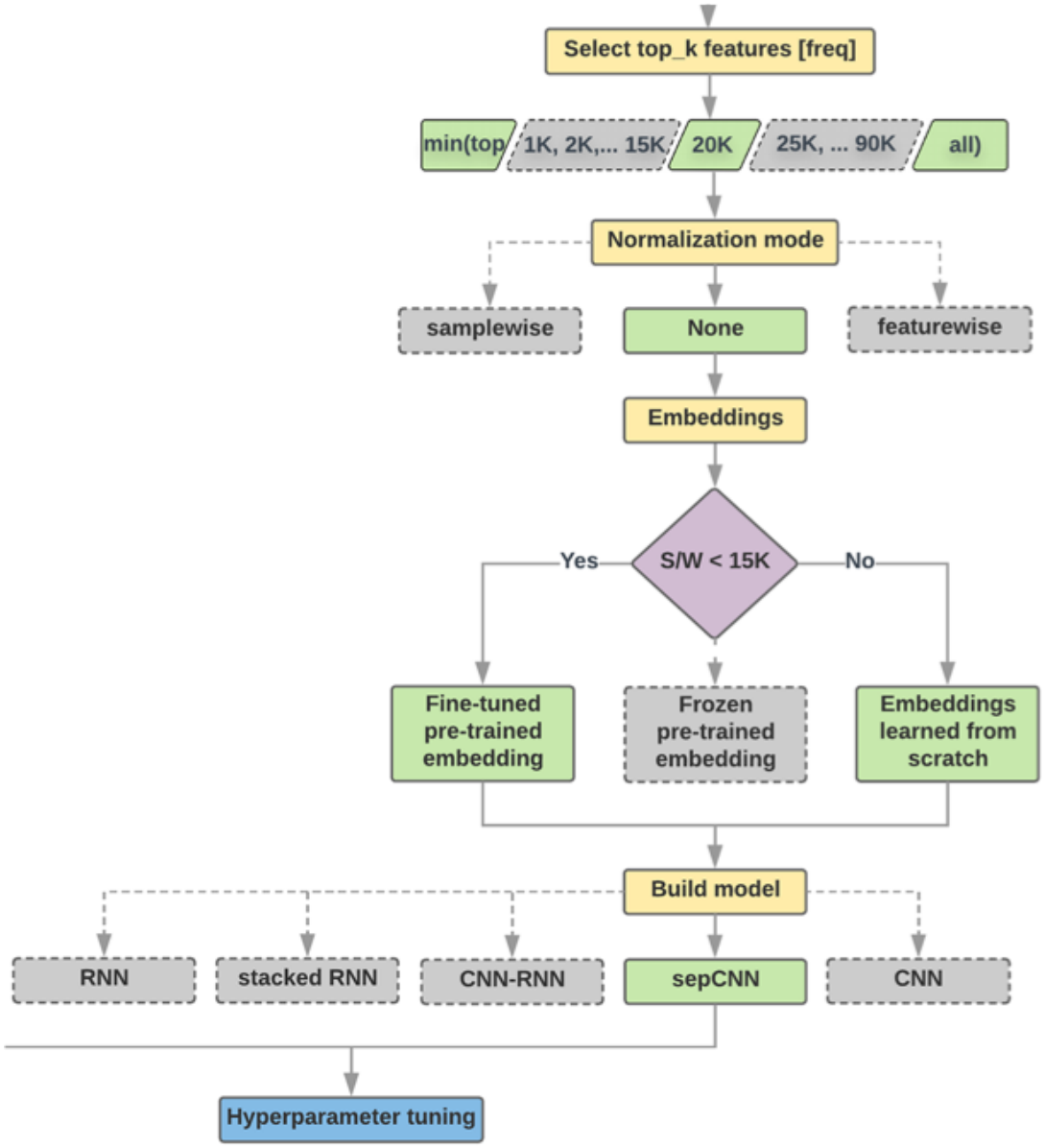
Source: https://developers.google.com/machine-learning/guides/text-classification/step-2-5

Text Classification S/W<1500: N-gram



Source: https://developers.google.com/machine-learning/guides/text-classification/step-2-5

Text Classification S/W>=1500: Sequence

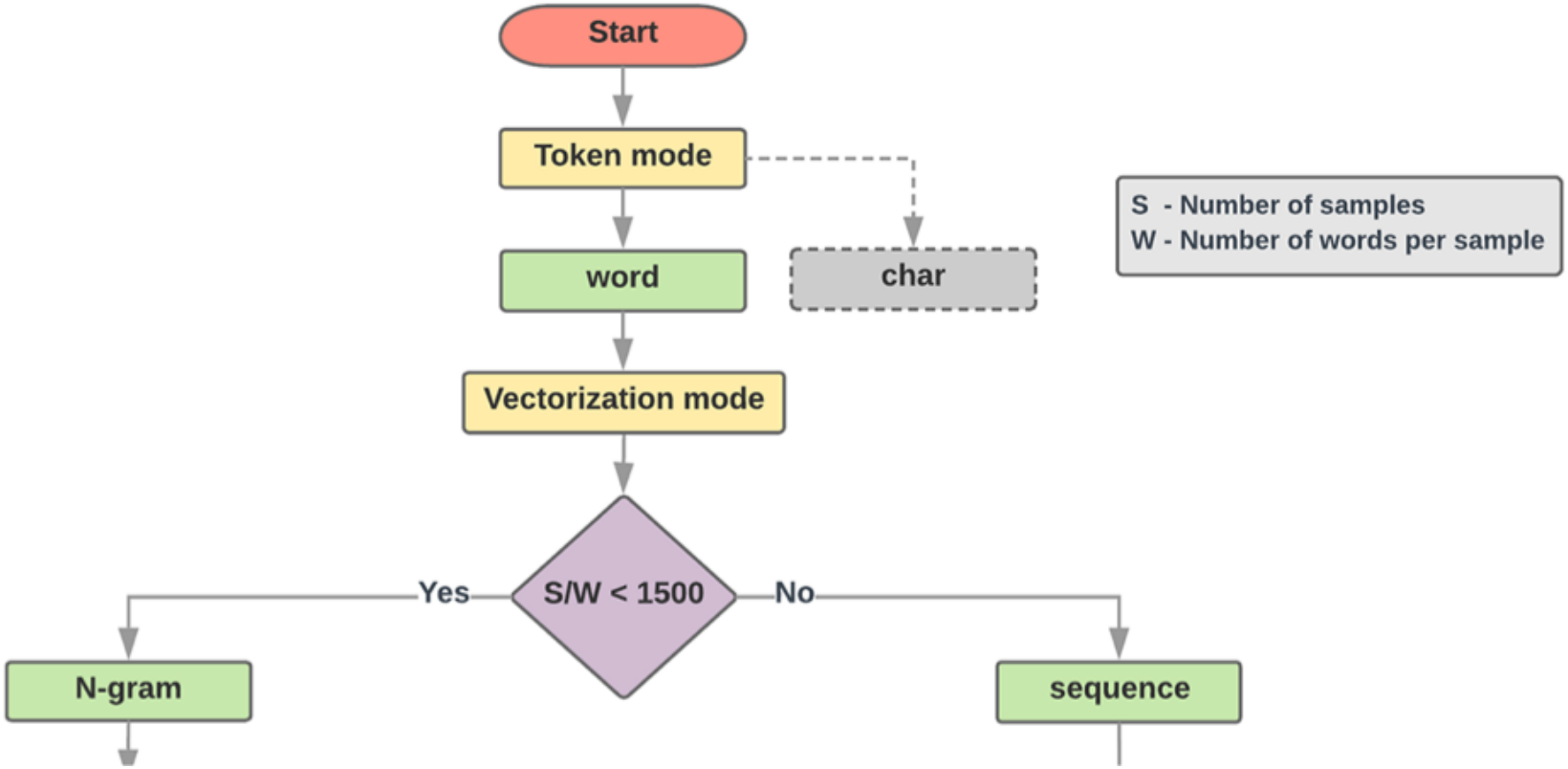


Source: https://developers.google.com/machine-learning/guides/text-classification/step-2-5

Step 2.5: Choose a Model

Samples/Words < 1500

150,000/100 = 1500



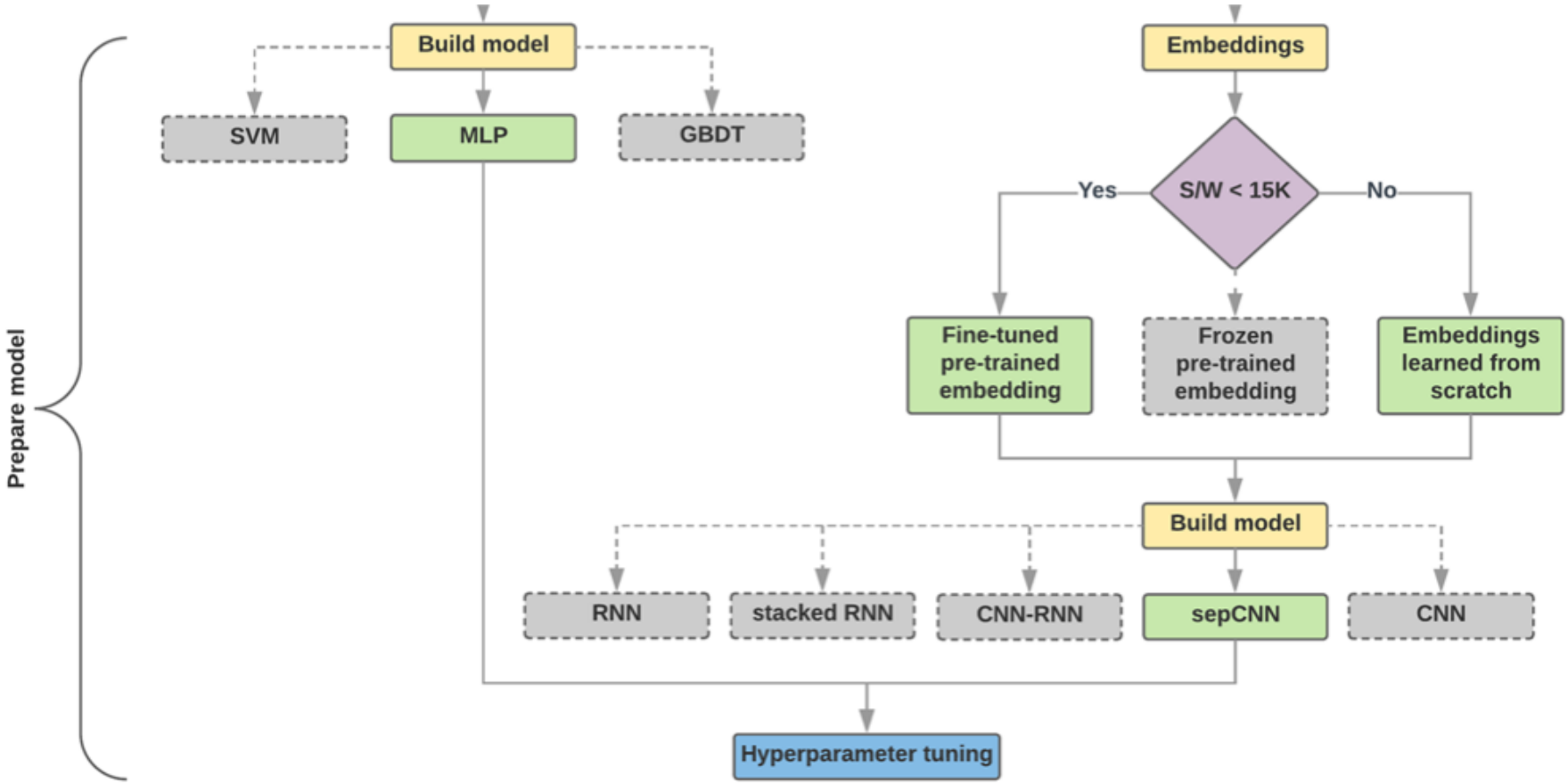
IMDb review dataset,
the samples/words-per-sample ratio is ~ 144

Source: https://developers.google.com/machine-learning/guides/text-classification/step-2-5

Step 2.5: Choose a Model

Samples/Words < 15,000

1,500,000/100 = 15,000



Source: https://developers.google.com/machine-learning/guides/text-classification/step-2-5

Step 3: Prepare Your Data

```
Texts:
T1: 'The mouse ran up the clock'
T2: 'The mouse ran down'

Token Index:
{'the': 1, 'mouse': 2, 'ran': 3, 'up': 4, 'clock': 5, 'down': 6,}.
    NOTE: 'the' occurs most frequently,
          so the index value of 1 is assigned to it.
          Some libraries reserve index 0 for unknown tokens,
          as is the case here.

Sequence of token indexes:
T1: 'The mouse ran up the clock' =
      [1,  2,   3, 4,  1,  5]
T1: 'The mouse ran down' =
      [1,   2,   3,  6]
```

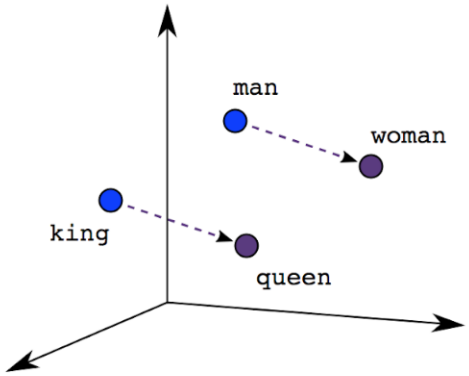
Source: https://developers.google.com/machine-learning/guides/text-classification/step-3

One-hot encoding

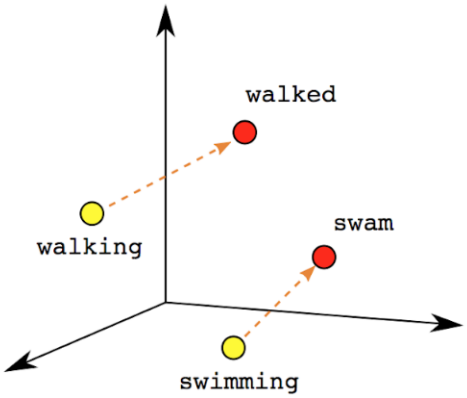
```
                'The mouse ran up the clock' =

The       1   [ [0, 1, 0, 0, 0, 0, 0],
mouse     2     [0, 0, 1, 0, 0, 0, 0],
ran       3     [0, 0, 0, 1, 0, 0, 0],
up        4     [0, 0, 0, 0, 1, 0, 0],
the       1     [0, 1, 0, 0, 0, 0, 0],
clock     5     [0, 0, 0, 0, 0, 1, 0] ]

                [0, 1, 2, 3, 4, 5, 6]
```

Source: https://developers.google.com/machine-learning/guides/text-classification/step-3

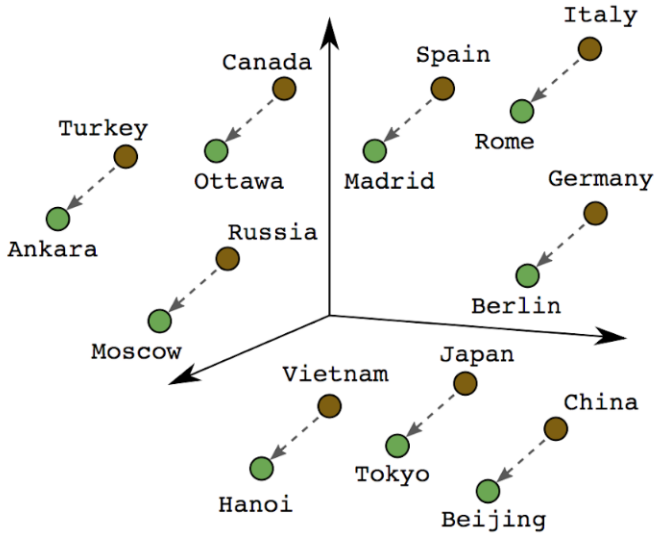
Word embeddings



Male-Female



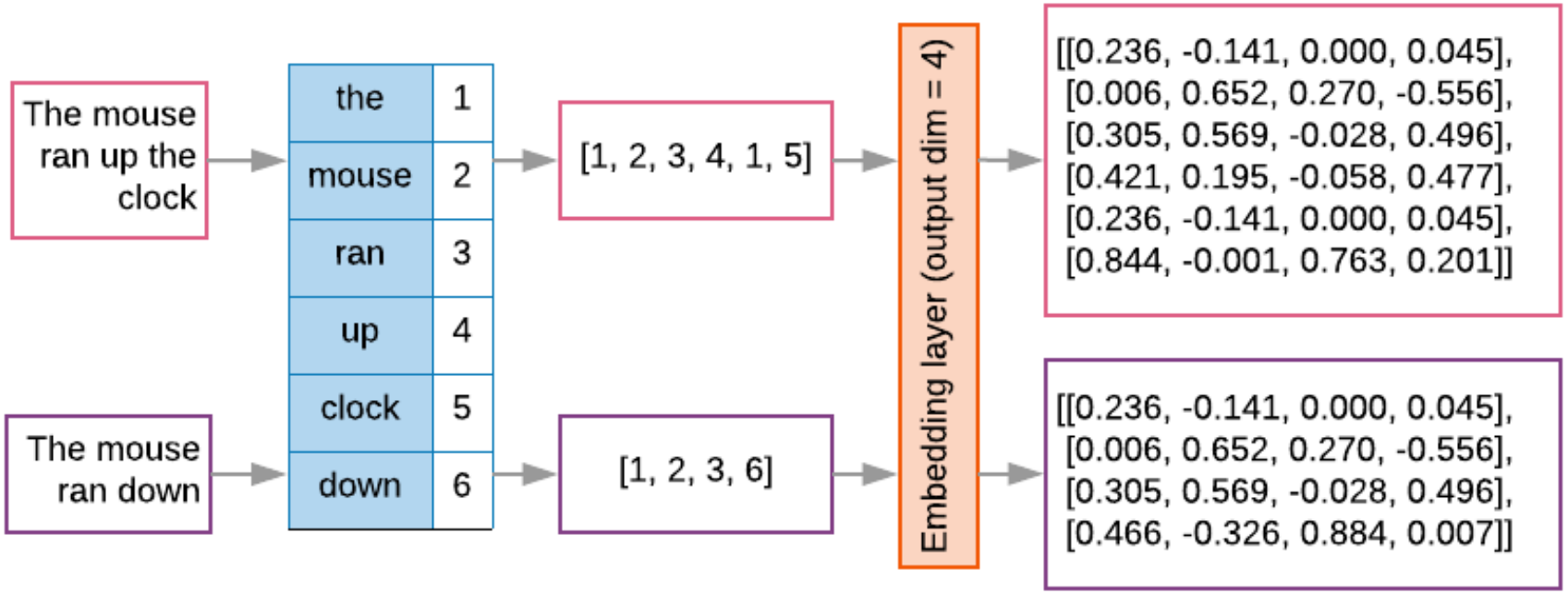
Verb Tense



Country-Capital

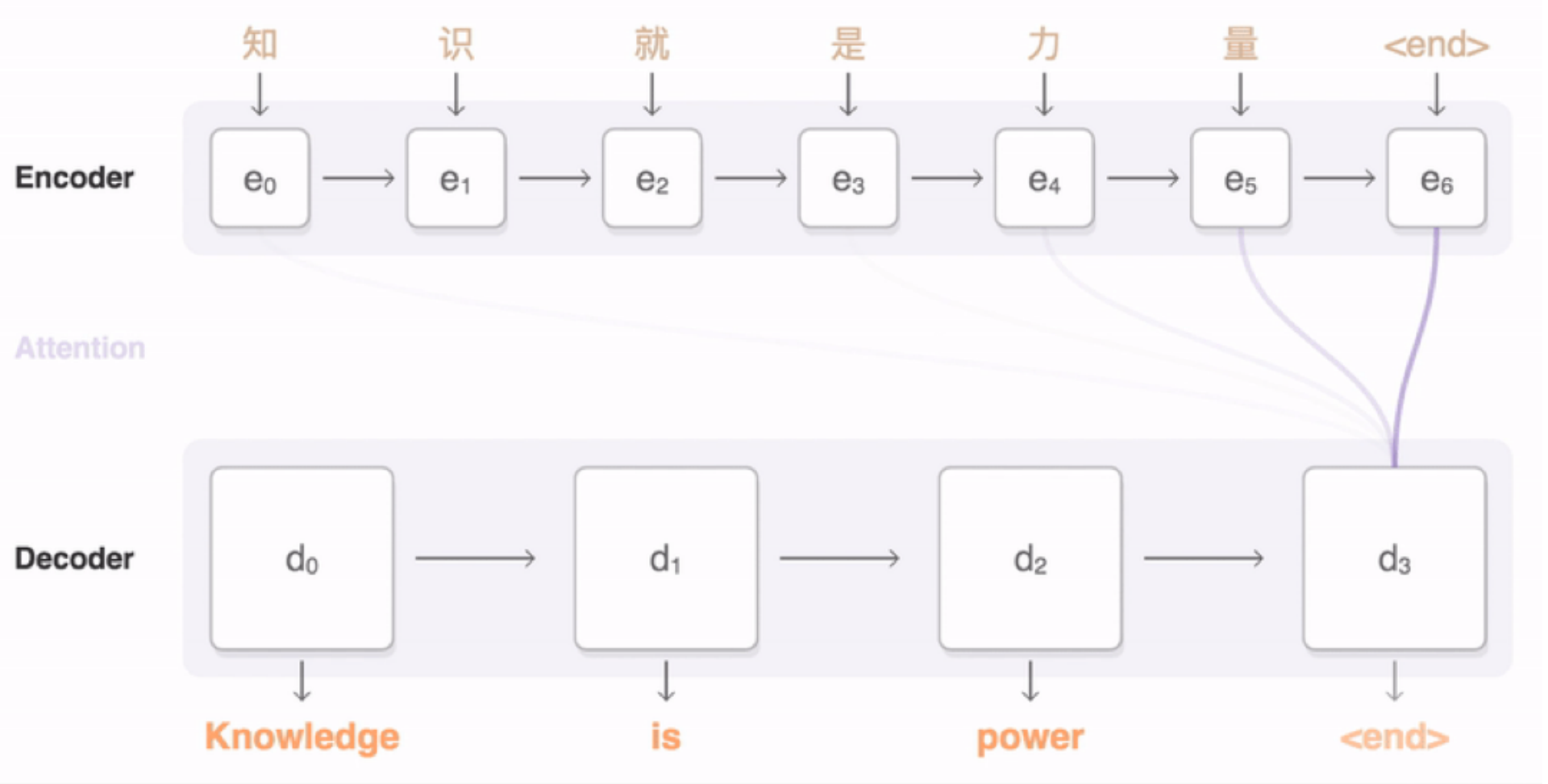
Source: https://developers.google.com/machine-learning/guides/text-classification/step-3

Word embeddings



Source: https://developers.google.com/machine-learning/guides/text-classification/step-3

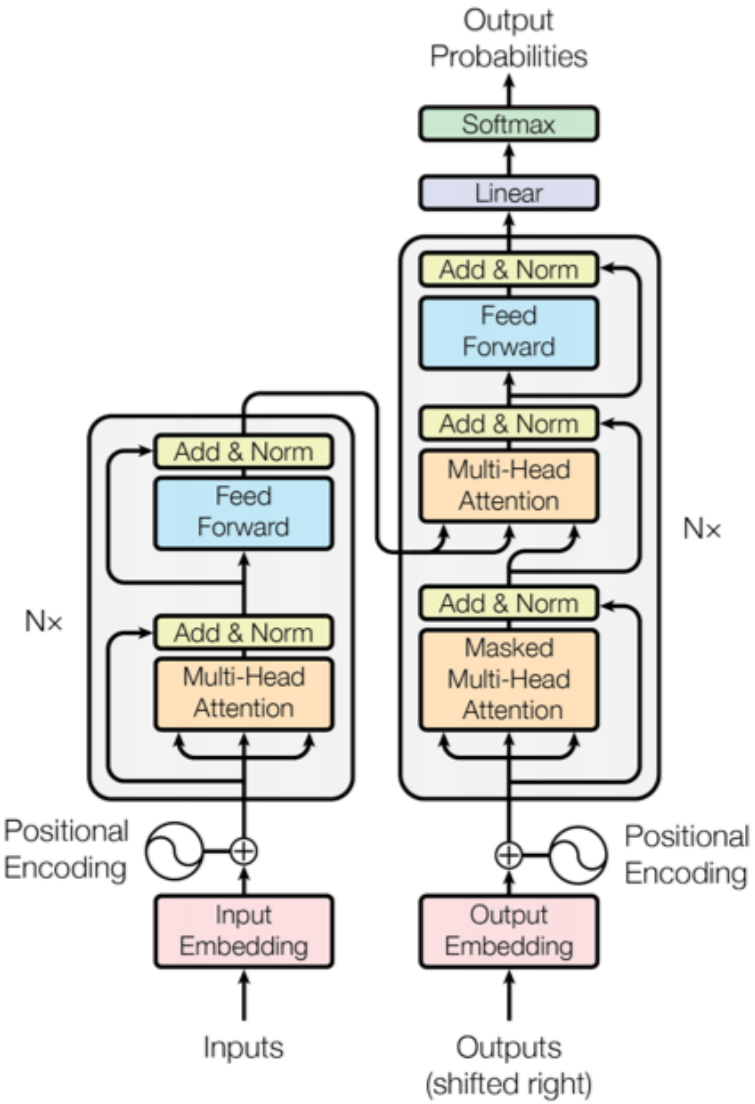
Sequence to Sequence (Seq2Seq)



Source: https://google.github.io/seq2seq/

Transformer (Attention is All You Need)

(Vaswani et al., 2017)

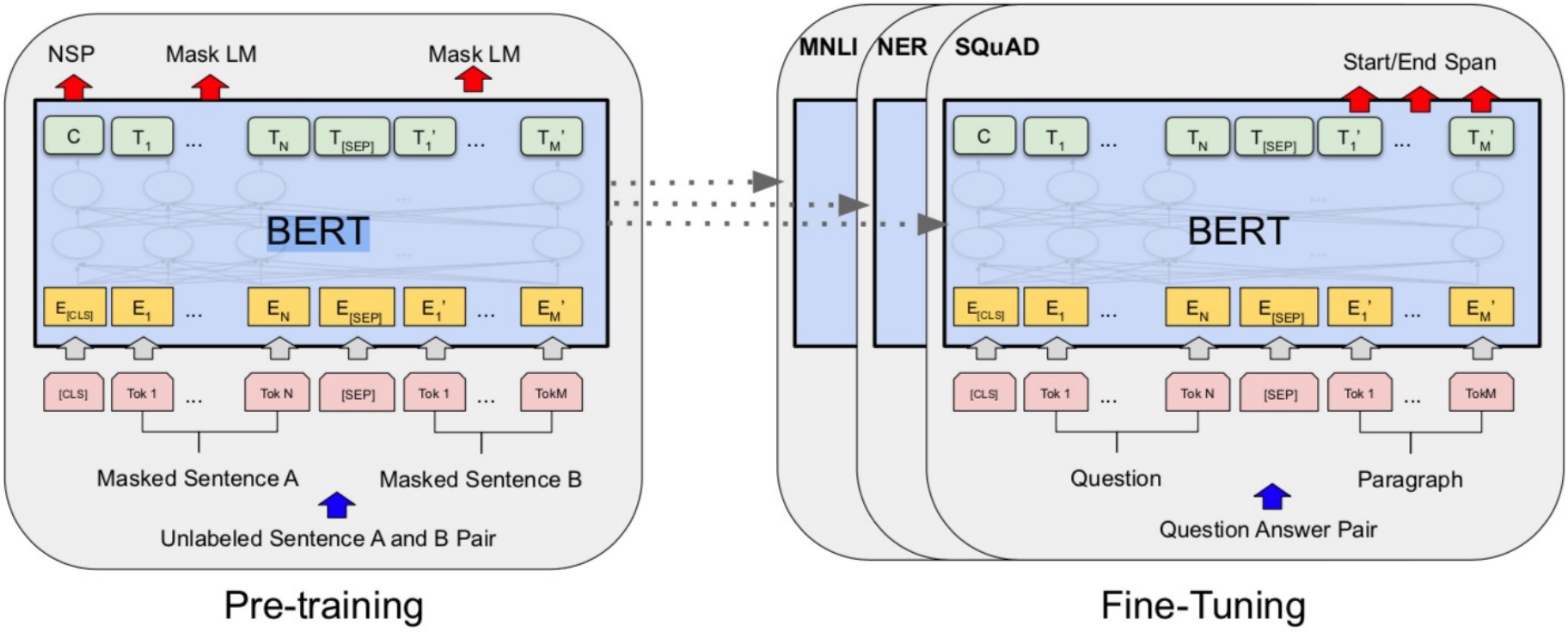


Source: Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In *Advances in neural information processing systems*, pp. 5998-6008. 2017.

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT (Bidirectional Encoder Representations from Transformers)

Overall pre-training and fine-tuning procedures for BERT



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: Pre-training of Deep Bidirectional Transformers for Language Understanding

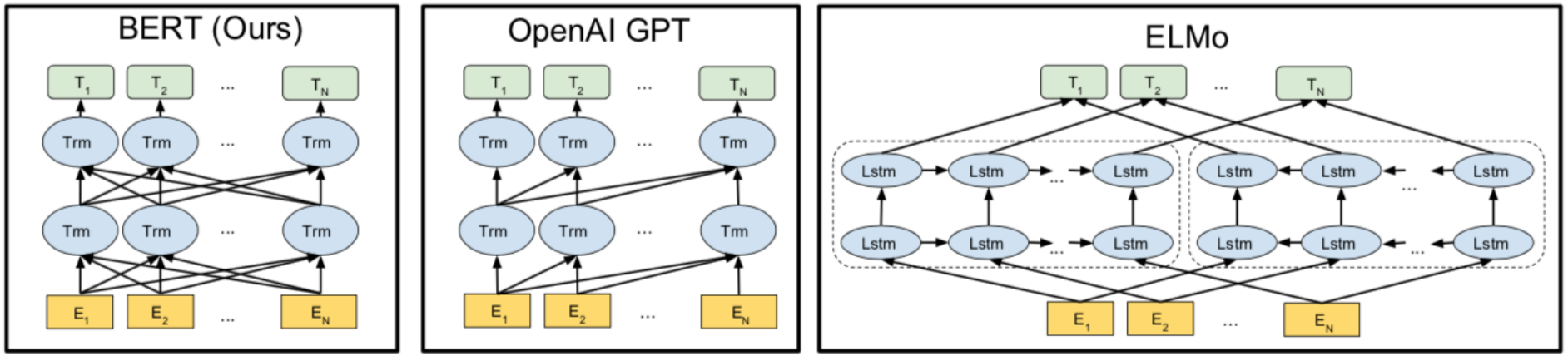
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

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

Bidirectional Encoder Representations from Transformers



Pre-training model architectures

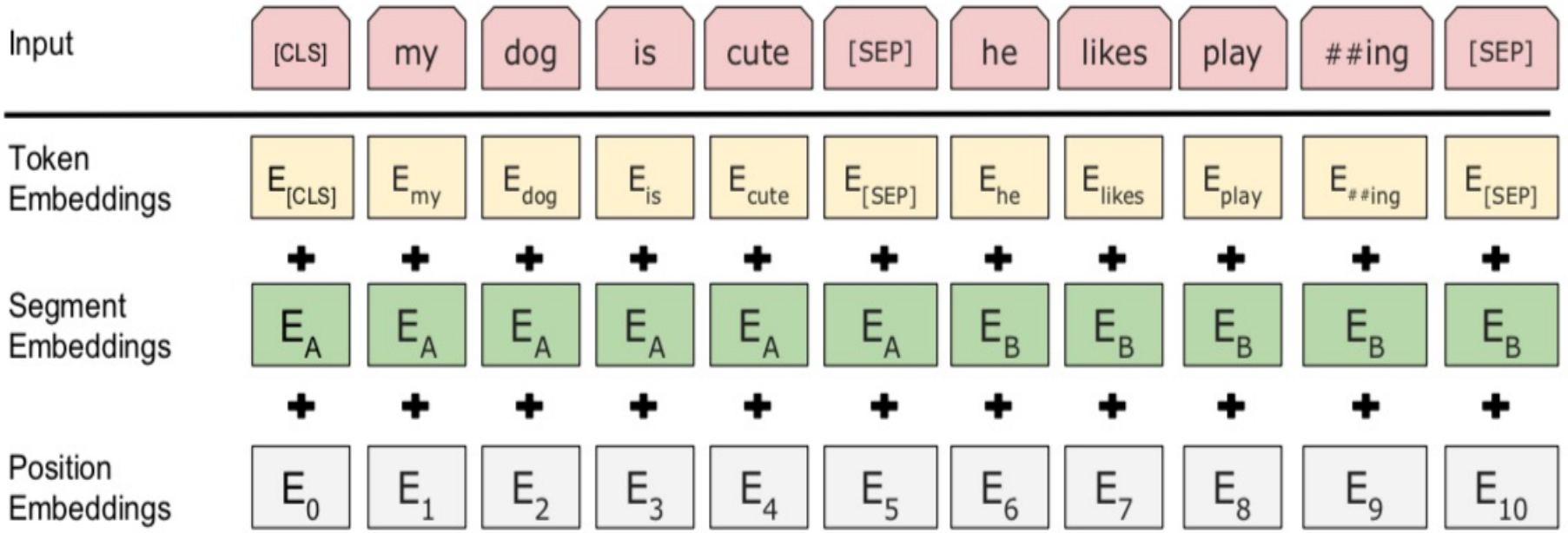
BERT uses a bidirectional Transformer.
OpenAI GPT uses a left-to-right Transformer.
ELMo uses the concatenation of independently trained left-to-right and right- to-left LSTM to generate features for downstream tasks.
Among three, only BERT representations are jointly conditioned on both left and right context in all layers.

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: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT (Bidirectional Encoder Representations from Transformers)

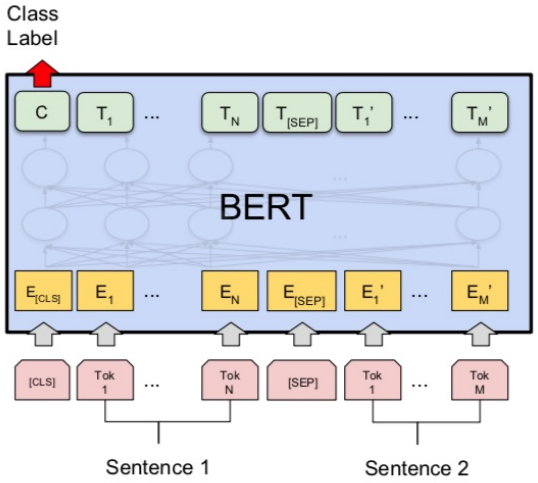
BERT input representation



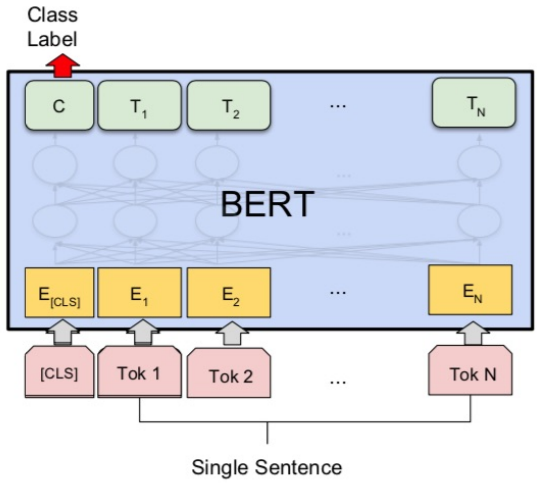
The input embeddings is the sum of the token embeddings, the segmentation embeddings and the position embeddings.

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.

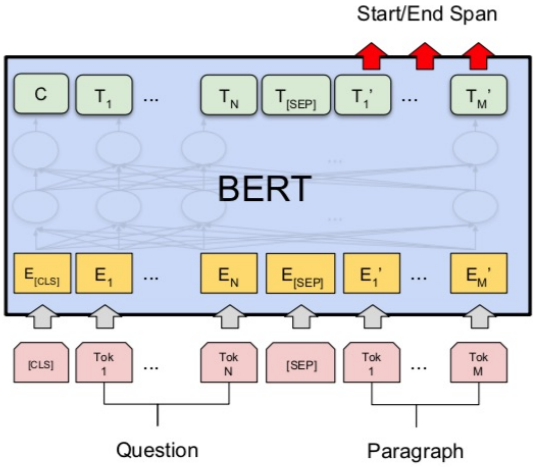
Fine-tuning BERT on NLP Tasks



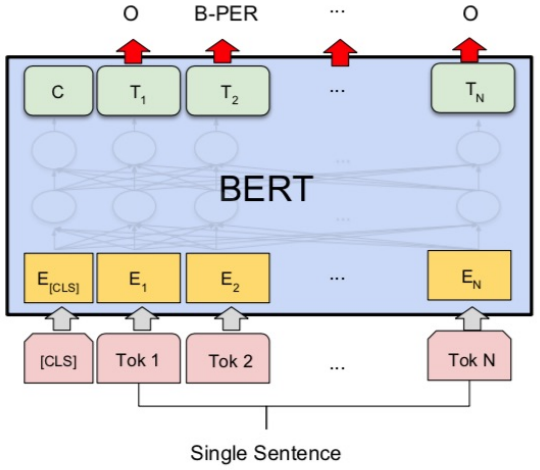
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



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



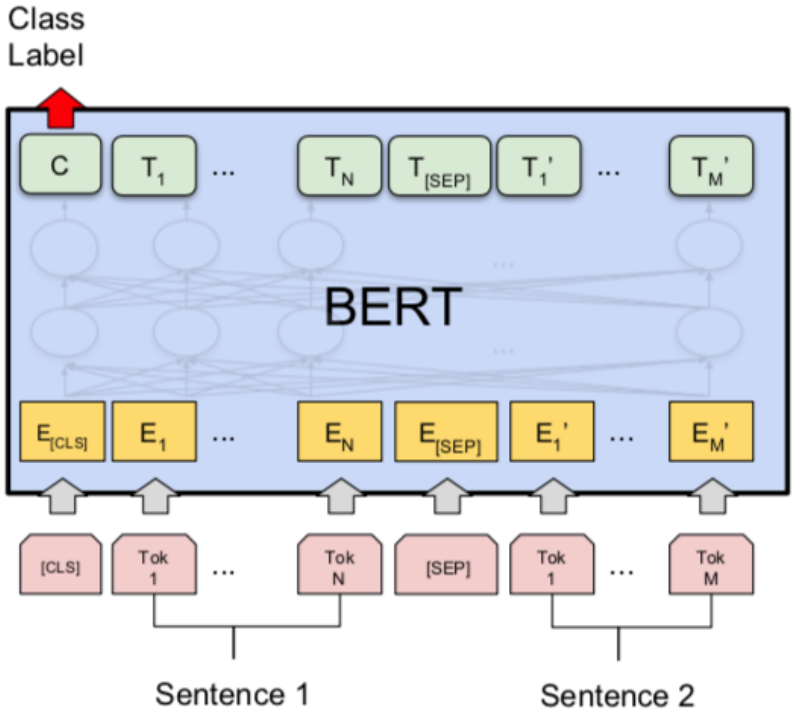
(c) Question Answering Tasks: SQuAD v1.1



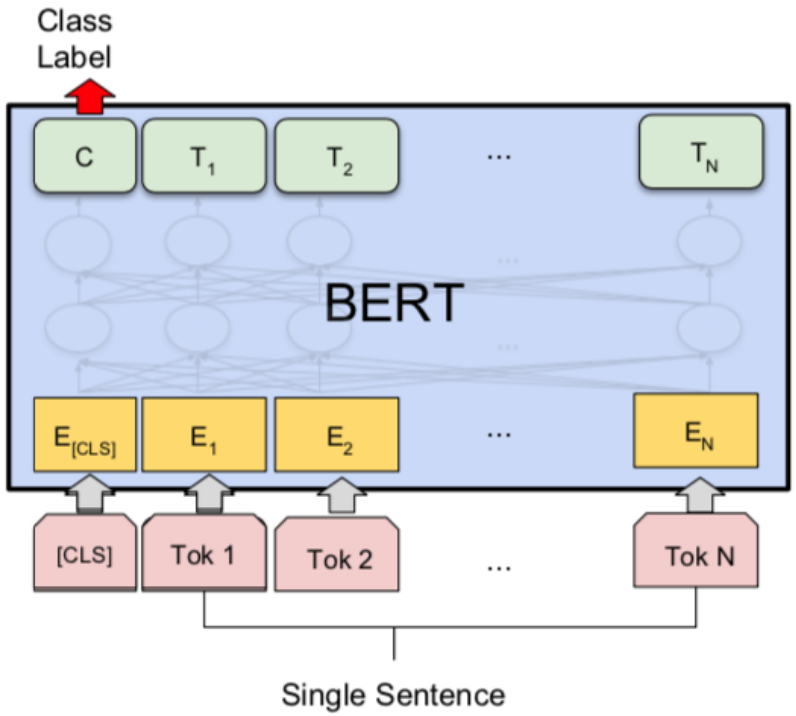
(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

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

BERT Sequence-level tasks



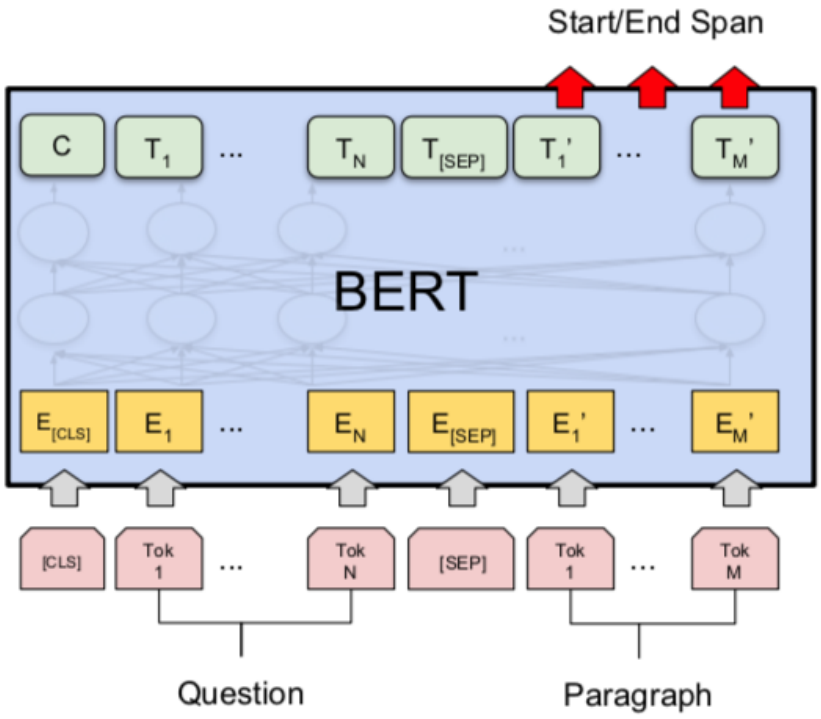
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



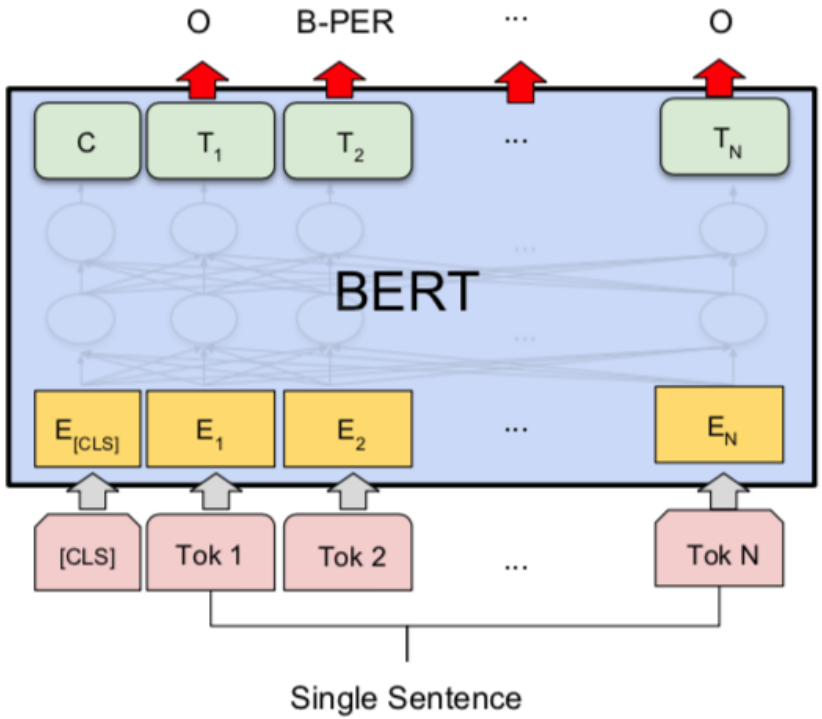
(b) Single Sentence Classification Tasks: SST-2, CoLA

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

General Language Understanding Evaluation (GLUE) benchmark

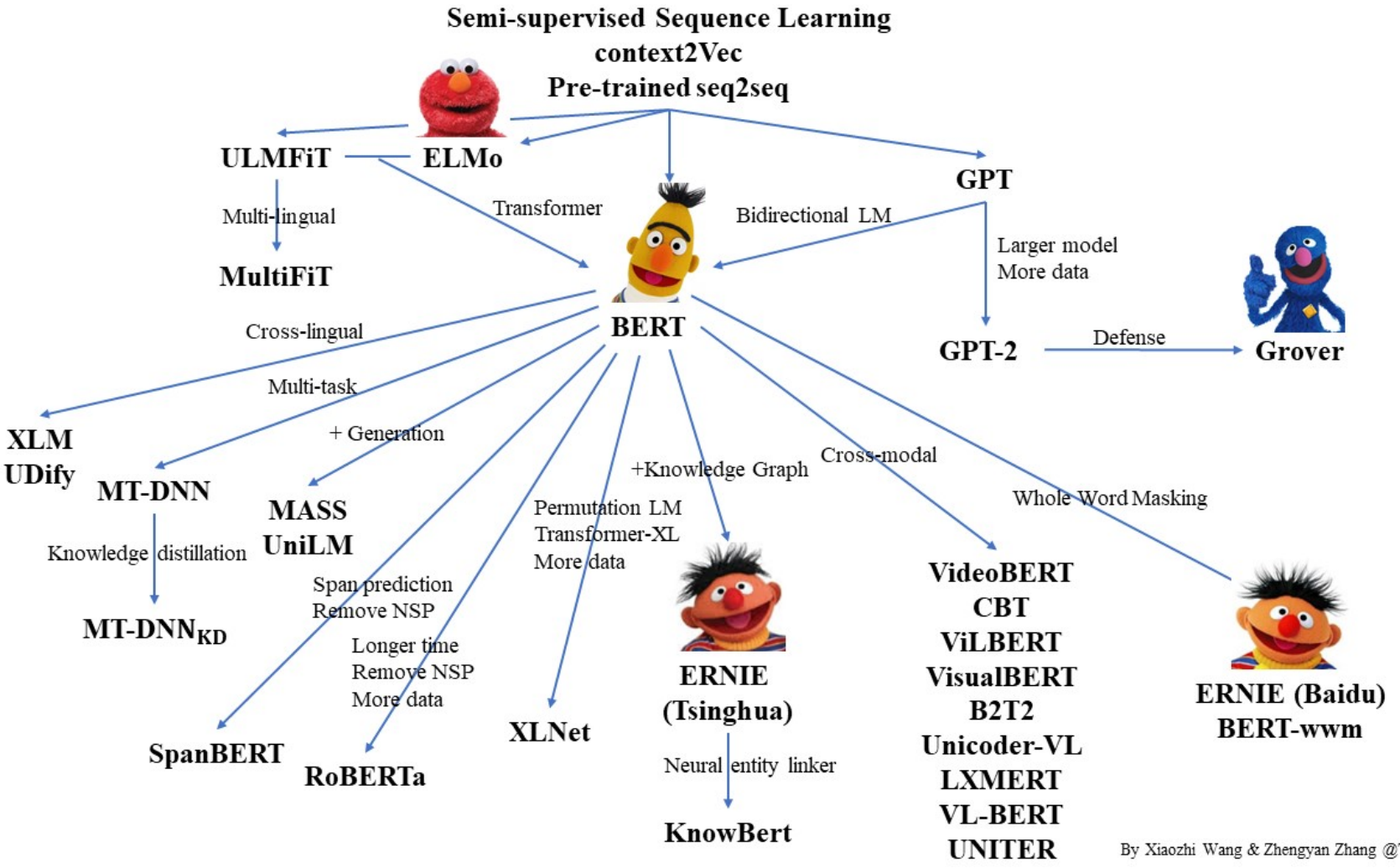
GLUE Test results

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

MNLI: Multi-Genre Natural Language Inference
QQP: Quora Question Pairs
QNLI: Question Natural Language Inference
SST-2: The Stanford Sentiment Treebank
CoLA: The Corpus of Linguistic Acceptability
STS-B:The Semantic Textual Similarity Benchmark
MRPC: Microsoft Research Paraphrase Corpus
RTE: Recognizing Textual Entailment

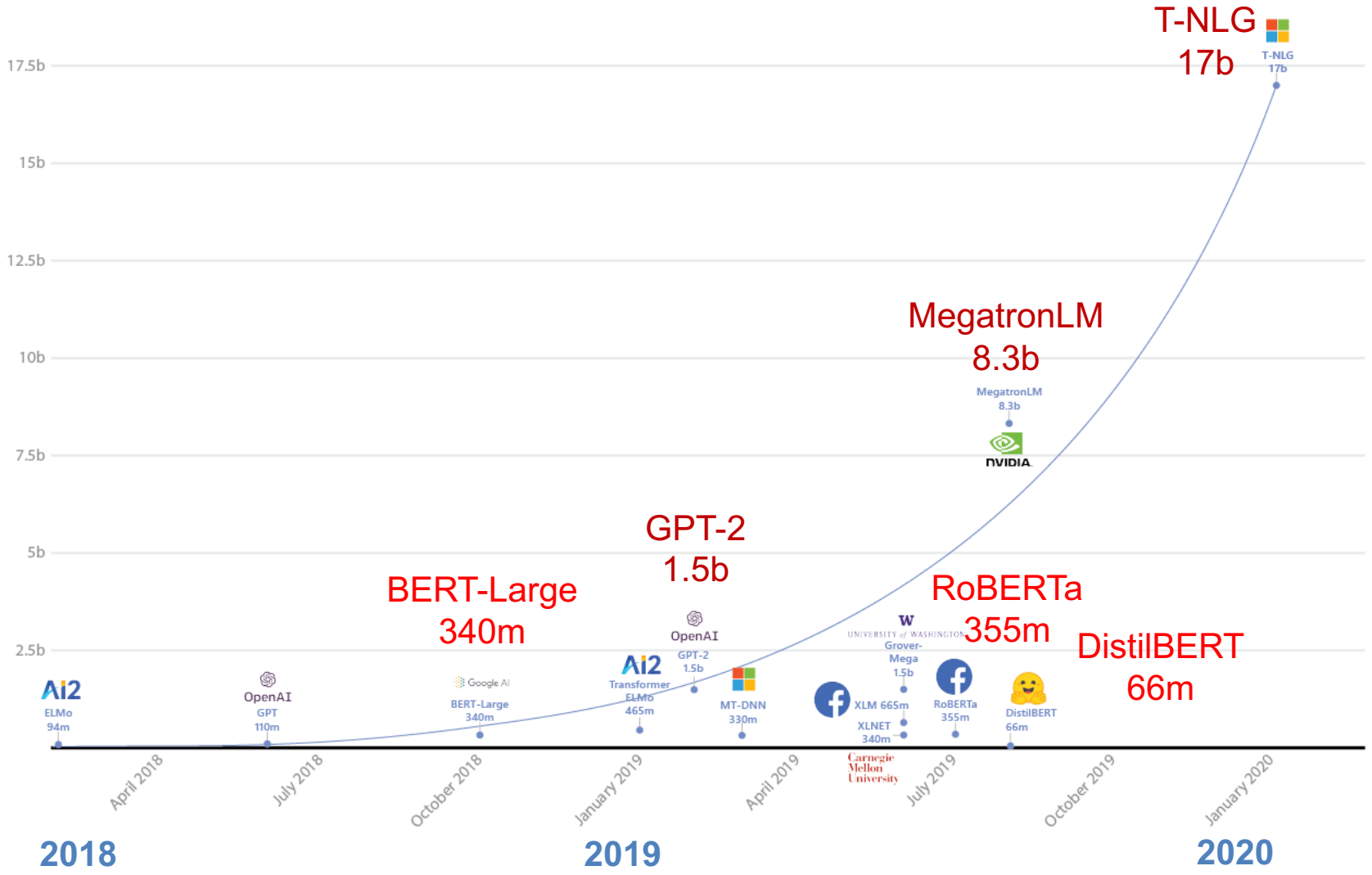
Source: Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018).
"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv preprint arXiv:1810.04805

Pre-trained Language Model (PLM)



Source: https://github.com/thunlp/PLMpapers

Turing Natural Language Generation (T-NLG)



Source: https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

Transformers

State-of-the-art Natural Language Processing for TensorFlow 2.0 and PyTorch

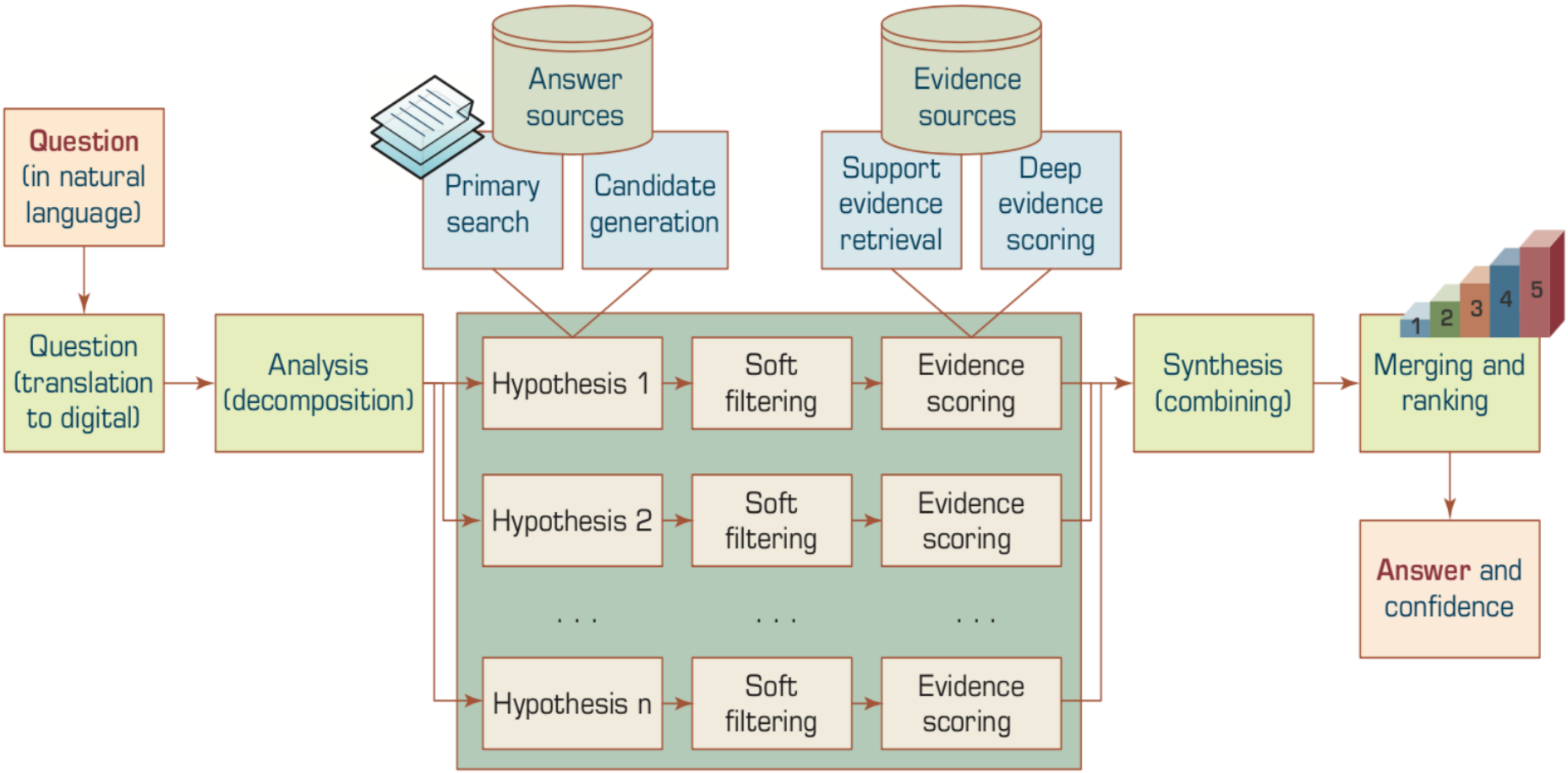
- Transformers
  - pytorch-transformers
  - pytorch-pretrained-bert
- provides state-of-the-art general-purpose architectures
  - (BERT, GPT-2, RoBERTa, XLM, DistilBert, XLNet, CTRL...)
  - for Natural Language Understanding (NLU) and Natural Language Generation (NLG) with over 32+ pretrained models in 100+ languages and deep interoperability between TensorFlow 2.0 and PyTorch.

Source: https://github.com/huggingface/transformers

Transfer Learning in Natural Language Processing

Source: Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf (2019), "Transfer learning in natural language processing." In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials, pp. 15-18.

A High-Level Depiction of DeepQA Architecture

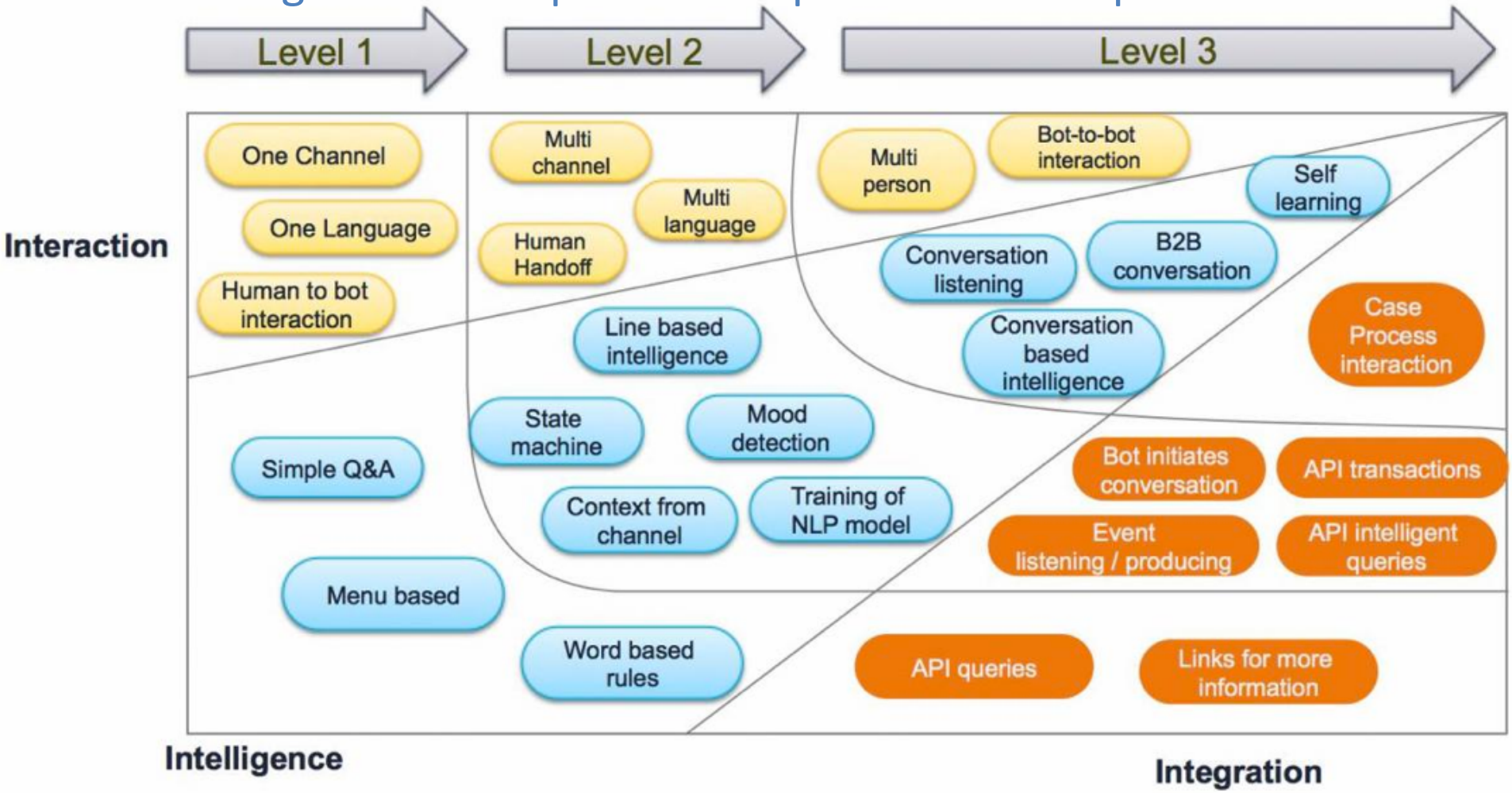


Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Chatbots

Bot Maturity Model

Customers want to have simpler means to interact with businesses and get faster response to a question or complaint.



Source: https://www.capgemini.com/2017/04/how-can-chatbots-meet-expectations-introducing-the-bot-maturity/

Dialogue on Airline Travel Information System (ATIS)

The ATIS (Airline Travel Information System) Dataset

https://www.kaggle.com/siddhadev/atis-dataset-from-ms-cntk

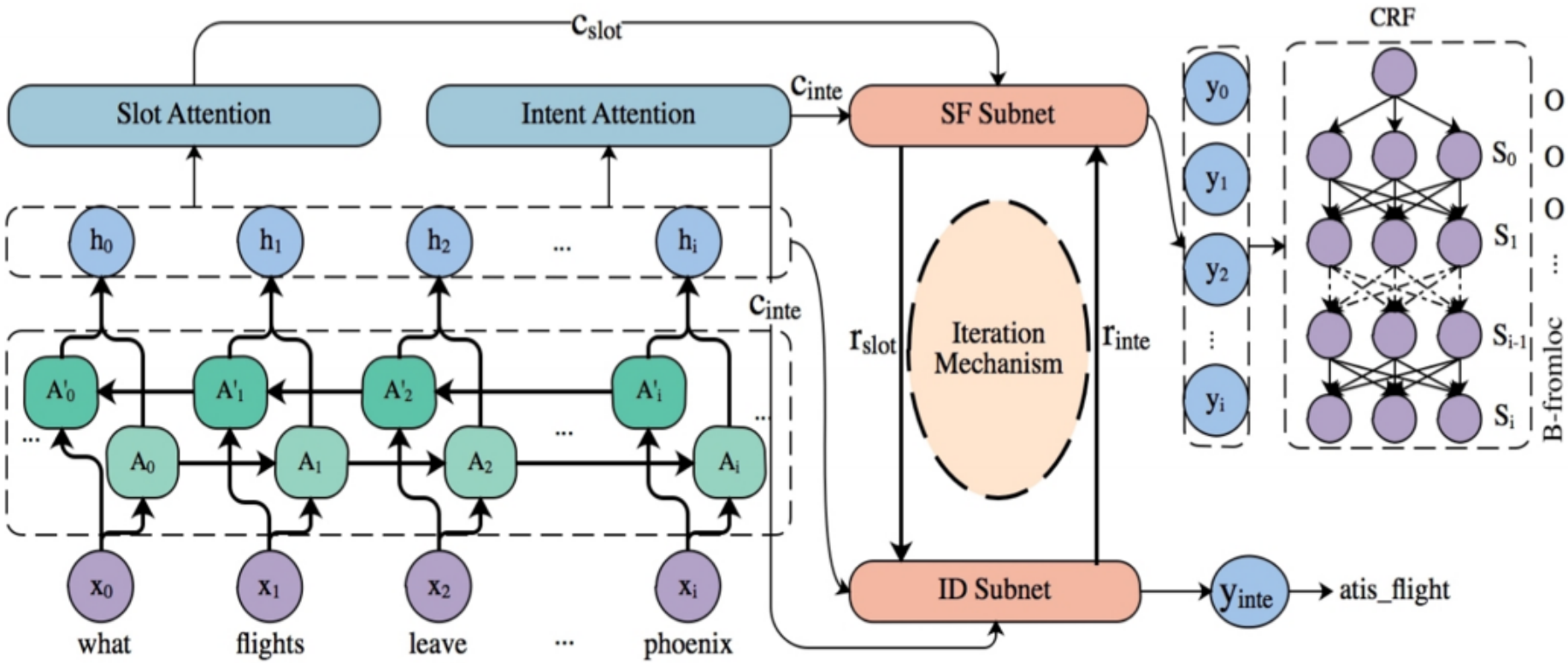
Sentence	what	flights	leave	from	phoenix
Slots	O	O	O	O	B-fromloc
Intent	atis_flight				

Training samples: 4978
Testing samples: 893
Vocab size: 943
Slot count: 129
Intent count: 26

Source: Haihong, E., Peiqing Niu, Zhongfu Chen, and Meina Song. "A novel bi-directional interrelated model for joint intent detection and slot filling." In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5467-5471. 2019.

SF-ID Network (E et al., 2019)
Slot Filling (SF)
Intent Detection (ID)

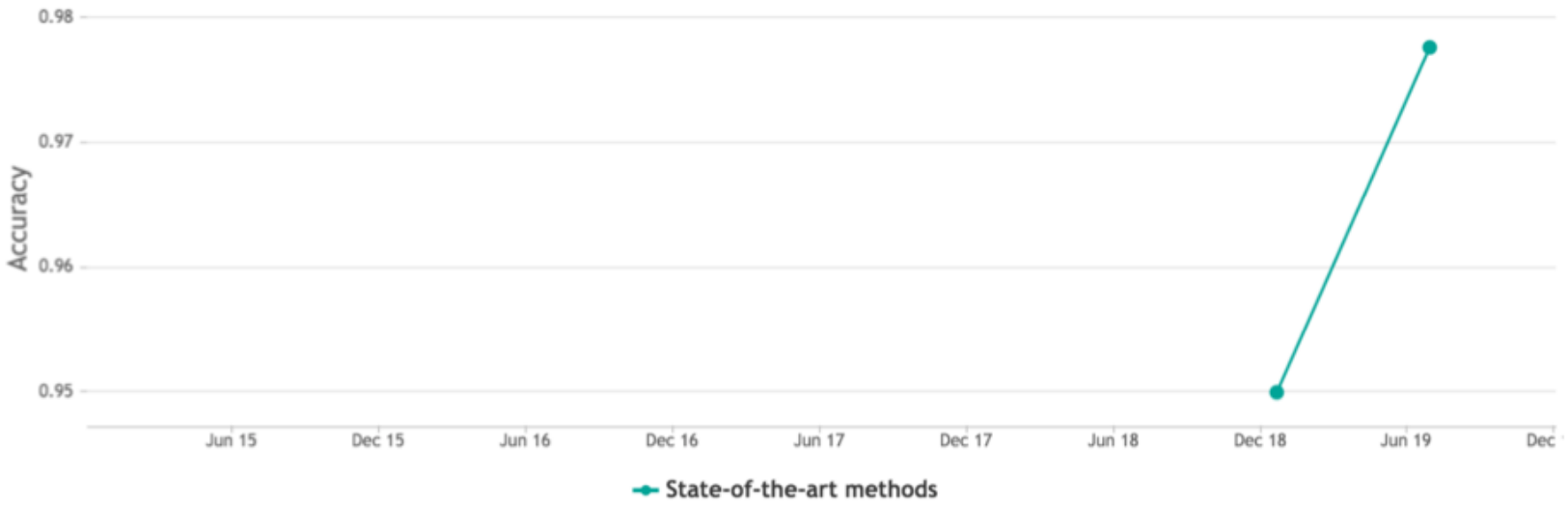
A Novel Bi-directional Interrelated Model for Joint Intent Detection and Slot Filling



Source: Haihong, E., Peiqing Niu, Zhongfu Chen, and Meina Song. "A novel bi-directional interrelated model for joint intent detection and slot filling." In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5467-5471. 2019.

Intent Detection on ATIS State-of-the-art

Intent Detection on ATIS



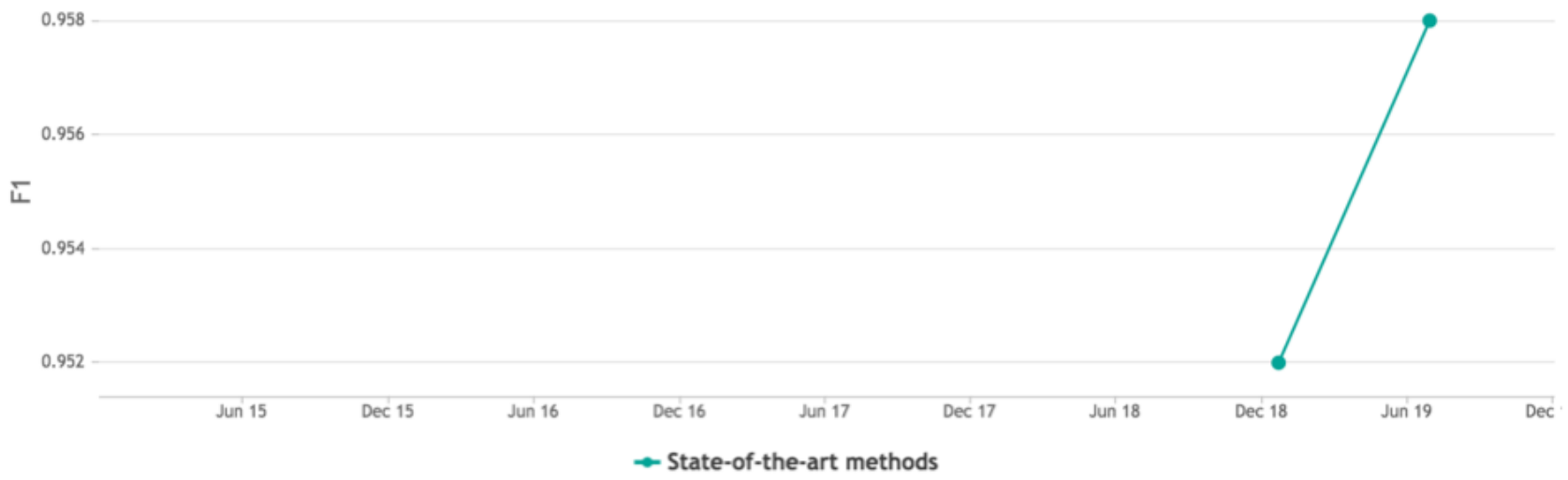
Edit

RANK	METHOD	ACCURACY	PAPER TITLE	YEAR	PAPER	CODE
1	**SF-ID**	0.9776	A Novel Bi-directional Interrelated Model for Joint Intent Detection and Slot Filling	2019		
2	**Capsule-NLU**	0.950	Joint Slot Filling and Intent Detection via Capsule Neural Networks	2018		

Source: https://paperswithcode.com/sota/intent-detection-on-atis

Slot Filling on ATIS State-of-the-art

Slot Filling on ATIS



Edit

RANK	METHOD	F1	PAPER TITLE	YEAR	PAPER	CODE
1	SF-ID	0.958	A Novel Bi-directional Interrelated Model for Joint Intent Detection and Slot Filling	2019		
2	Capsule-NLU	0.952	Joint Slot Filling and Intent Detection via Capsule Neural Networks	2018		

Source: https://paperswithcode.com/sota/slot-filling-on-atis

Restaurants Dialogue Datasets

- MIT Restaurant Corpus
  - https://groups.csail.mit.edu/sls/downloads/restaurant/
- CamRest676
  (Cambridge restaurant dialogue domain dataset)
  - https://www.repository.cam.ac.uk/handle/1810/260970
- DSTC2 (Dialog State Tracking Challenge 2 & 3)
  - http://camdial.org/~mh521/dstc/

任務型對話系統
The Evaluation of Chinese Human-Computer Dialogue Technology, SMP2019-ECDT

- 自然語言理解
  Natural Language Understanding (NLU)
- 對話管理
  Dialog Management (DM)
- 自然語言生成
  Natural Language Generation (NLG)

Source: http://conference.cipsc.org.cn/smp2019/evaluation.html
https://github.com/OnionWang/SMP2019-ECDT-NLU

NLP Libraries and Tools

Natural Language Processing with Python
– Analyzing Text with the Natural Language Toolkit

Natural Language Processing with Python

– Analyzing Text with the Natural Language Toolkit

NLTK

Steven Bird, Ewan Klein, and Edward Loper

This version of the NLTK book is updated for Python 3 and NLTK 3. The first edition of the book, published by O'Reilly, is available at http://nltk.org/book_1ed/. (There are currently no plans for a second edition of the book.)

http://www.nltk.org/book/

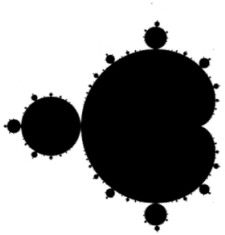
spaCy

https://spacy.io/

gensim

https://radimrehurek.com/gensim/

TextBlob



TextBlob

Star 3,777

TextBlob is a Python (2 and 3) library for processing textual data. It provides a consistent API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, and more.

Useful Links

TextBlob @ PyPI
TextBlob @ GitHub
Issue Tracker

Stay Informed

Follow @sloria

Donate

If you find TextBlob useful,

TextBlob: Simplified Text Processing

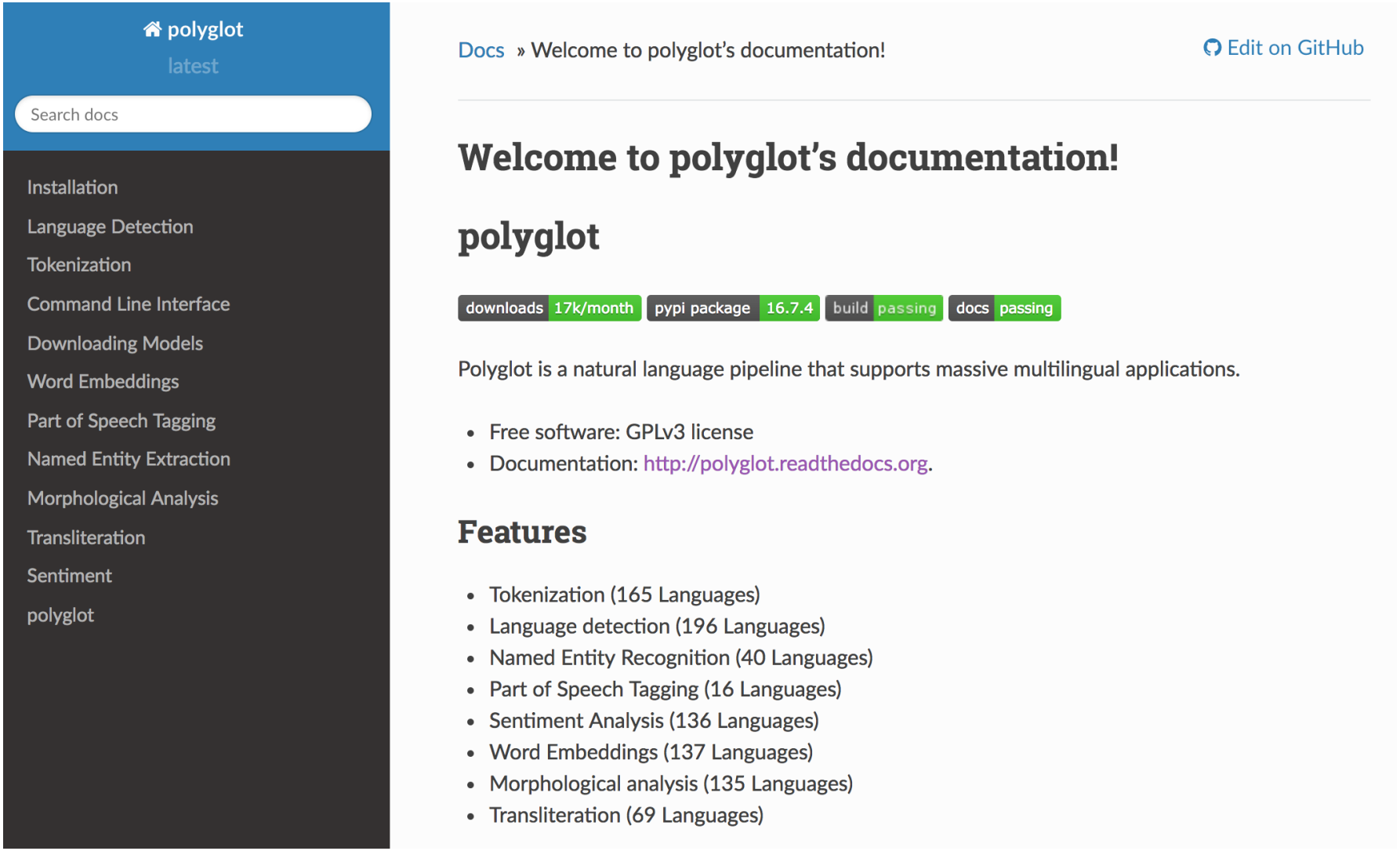
Release v0.12.0. (Changelog)

TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

```
from textblob import TextBlob

text = '''
The titular threat of The Blob has always struck me as the ultimate movie
monster: an insatiably hungry, amoeba-like mass able to penetrate
virtually any safeguard, capable of--as a doomed doctor chillingly
describes it--"assimilating flesh on contact.
Snide comparisons to gelatin be damned, it's a concept with the most
devastating of potential consequences, not unlike the grey goo scenario
proposed by technological theorists fearful of
artificial intelligence run rampant.
'''

blob = TextBlob(text)
blob.tags           # [('The', 'DT'), ('titular', 'JJ'),
                    #  ('threat', 'NN'), ('of', 'IN'), ...]

blob.noun_phrases   # WordList(['titular threat', 'blob',
                    #           'ultimate movie monster',
                    #           'amoeba-like mass', ...])

for sentence in blob.sentences:
    print(sentence.sentiment.polarity)
# 0.060
```

https://textblob.readthedocs.io

Polyglot



https://polyglot.readthedocs.io/

scikit-learn

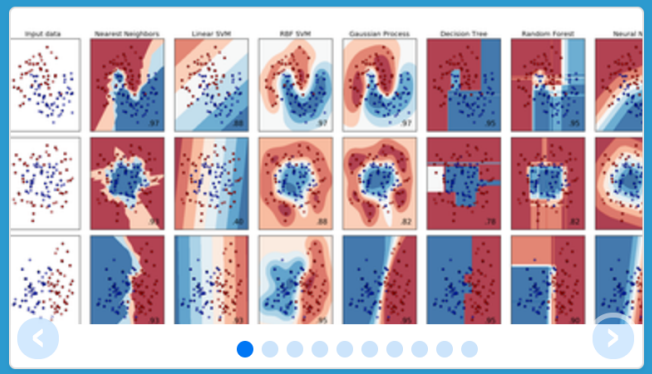
scikit learn

Home | Installation | Documentation | Examples

powered by Google

Google Custom Search | Search

Fork me on GitHub



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.
Algorithms: SVM, nearest neighbors, random forest, ... — Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, ridge regression, Lasso, ... — Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, ... — Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning

Preprocessing

Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.
Modules: preprocessing, feature extraction.

http://scikit-learn.org/

http://nlp.stanford.edu/software/index.shtml

The Stanford Natural Language Processing Group

home · people · teaching · research · publications · software · events · local

The Stanford NLP Group makes parts of our Natural Language Processing software available to everyone. These are statistical NLP toolkits for various major computational linguistics problems. They can be incorporated into applications with human language technology needs.

All the software we distribute here is written in Java. All recent distributions require Oracle Java 6+ or OpenJDK 7+. Distribution packages include components for command-line invocation, jar files, a Java API, and source code. A number of helpful people have extended our work with bindings or translations for other languages. As a result, much of this software can also easily be used from Python (or Jython), Ruby, Perl, Javascript, and F# or other .NET languages.

Supported software distributions

This code is being developed, and we try to answer questions and fix bugs on a best-effort basis.

All these software distributions are open source, **licensed under the GNU General Public License** (v2 or later). Note that this is the *full* GPL, which allows many free uses, but *does not allow* its incorporation into any type of distributed proprietary software, even in part or in translation. **Commercial licensing** is also available; please contact us if you are interested.

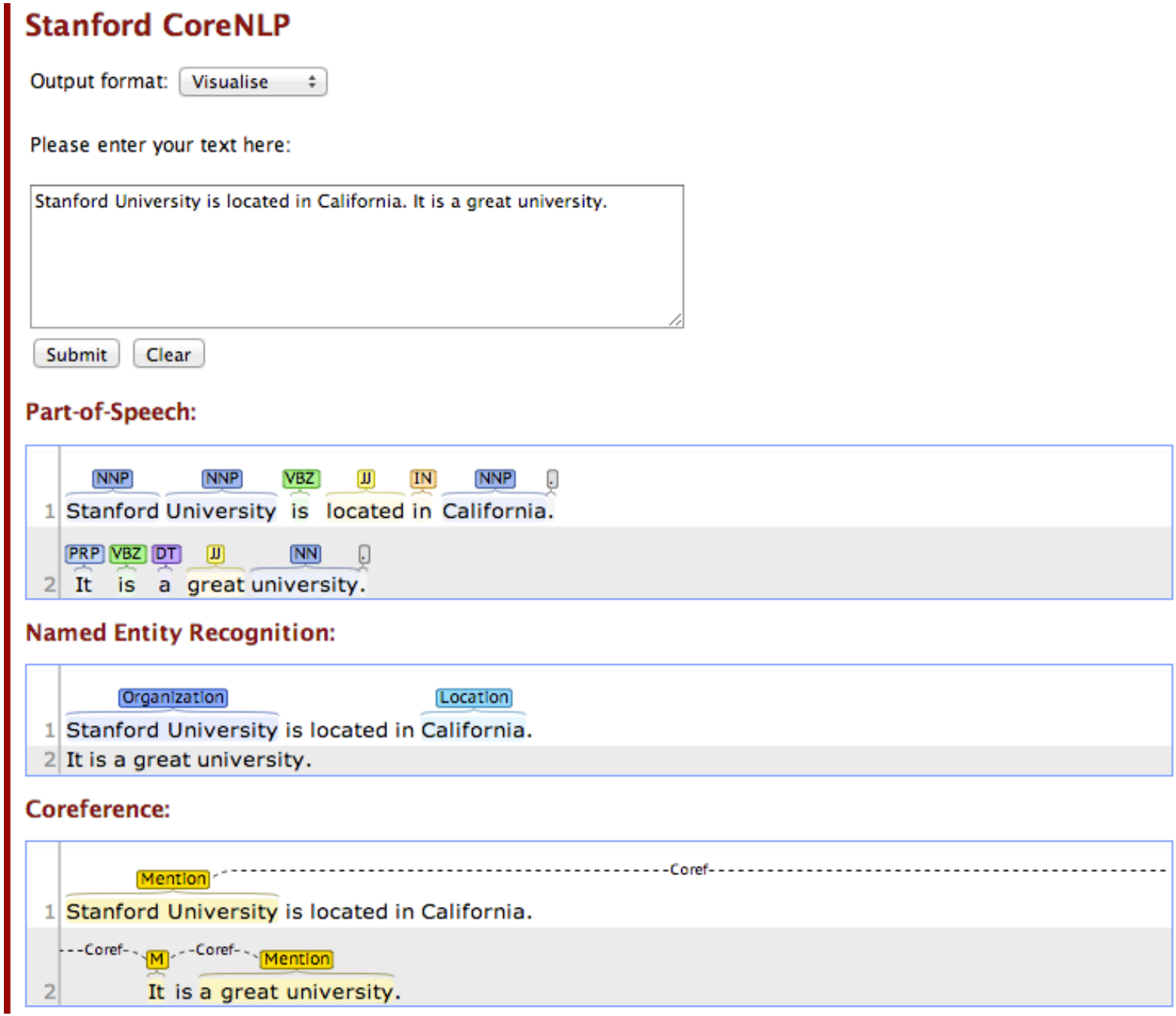
Stanford CoreNLP
: An integrated suite of natural language processing tools for English and (mainland) Chinese in Java, including tokenization, part-of-speech tagging, named entity recognition, parsing, and coreference. See also: Stanford Deterministic Coreference Resolution, and the online CoreNLP demo, and the CoreNLP FAQ.

Stanford Parser
: Implementations of probabilistic natural language parsers in Java: highly optimized PCFG and dependency parsers, a lexicalized PCFG parser, and a deep learning reranker. See also: Online parser demo, the Stanford Dependencies page, and Parser FAQ.

Stanford POS Tagger
: A maximum-entropy (CMM) part-of-speech (POS) tagger for English,

Stanford NLP Software

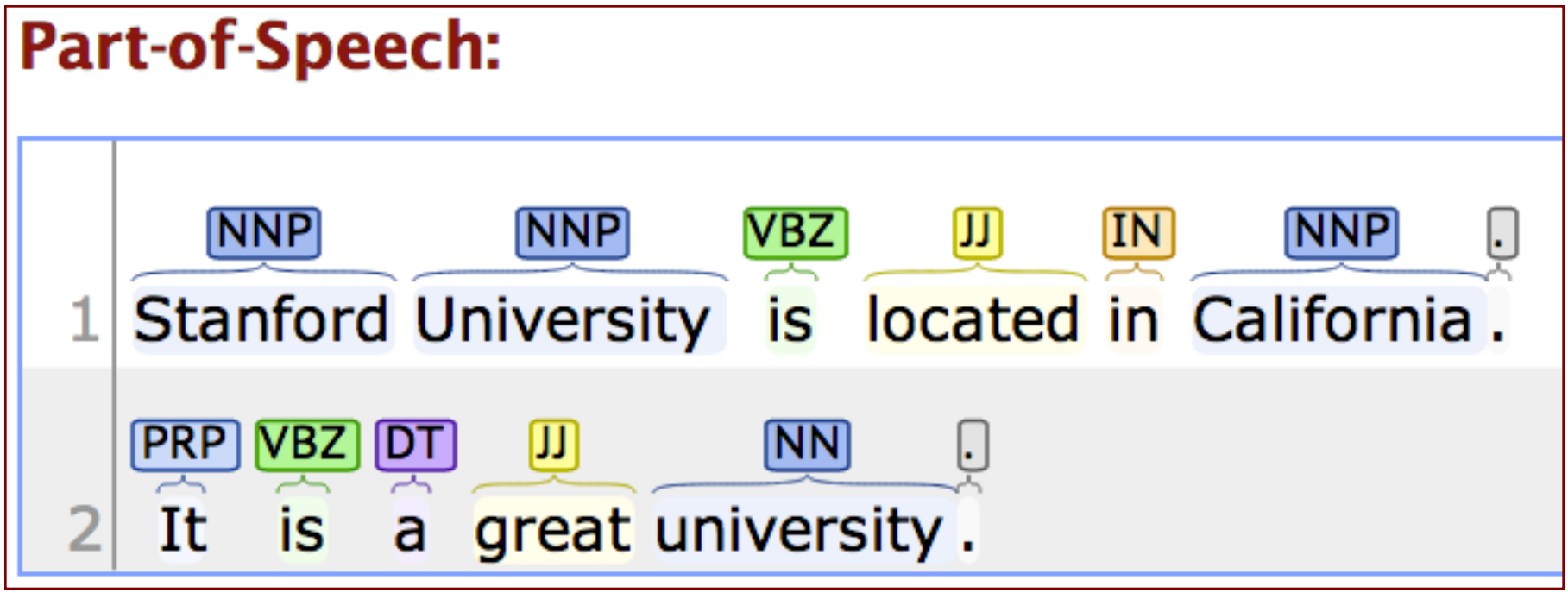
Stanford CoreNLP http://nlp.stanford.edu:8080/corenlp/process



Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

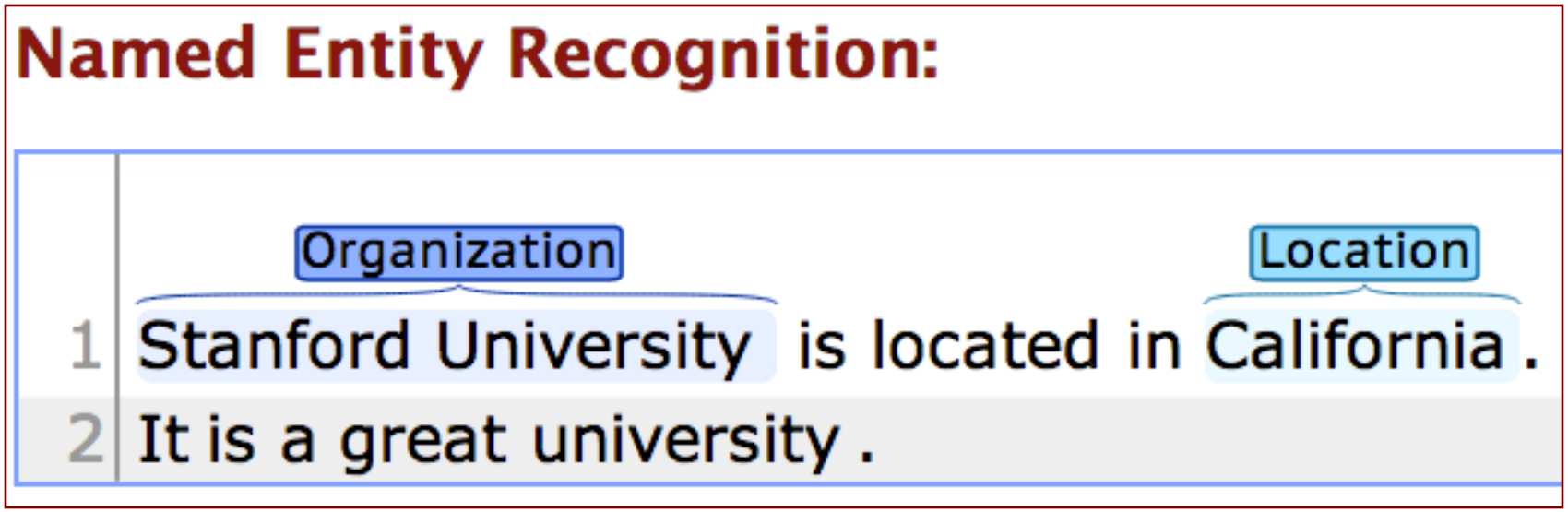
Stanford University is located in California.
It is a great university.



Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

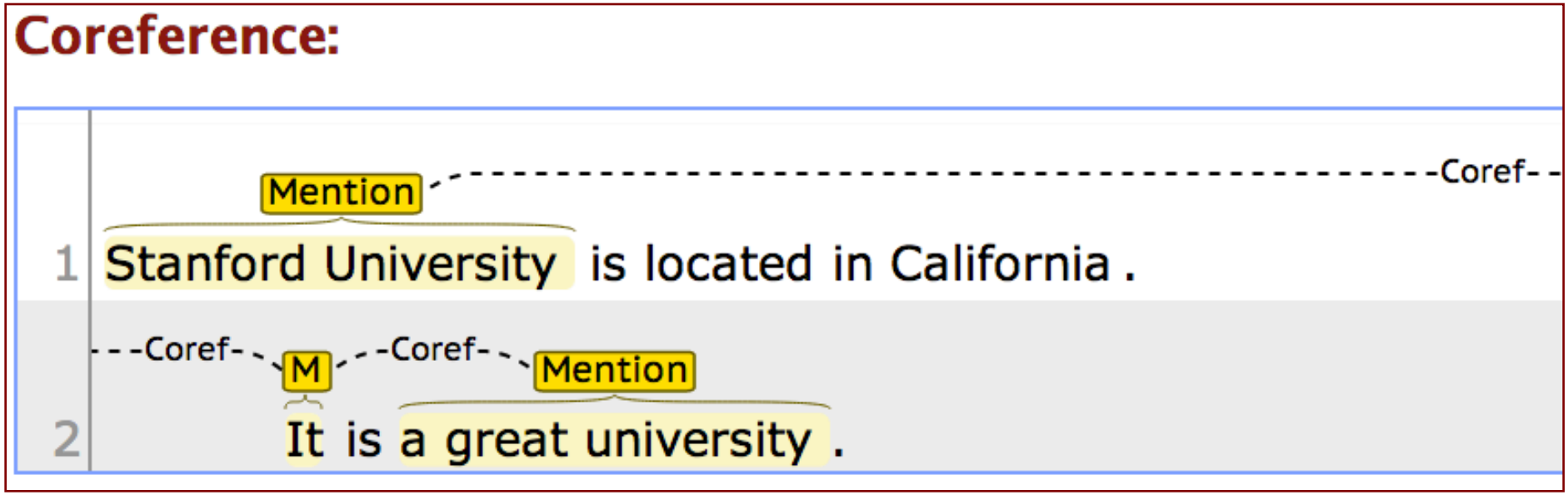
Stanford University is located in California.
It is a great university.



Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

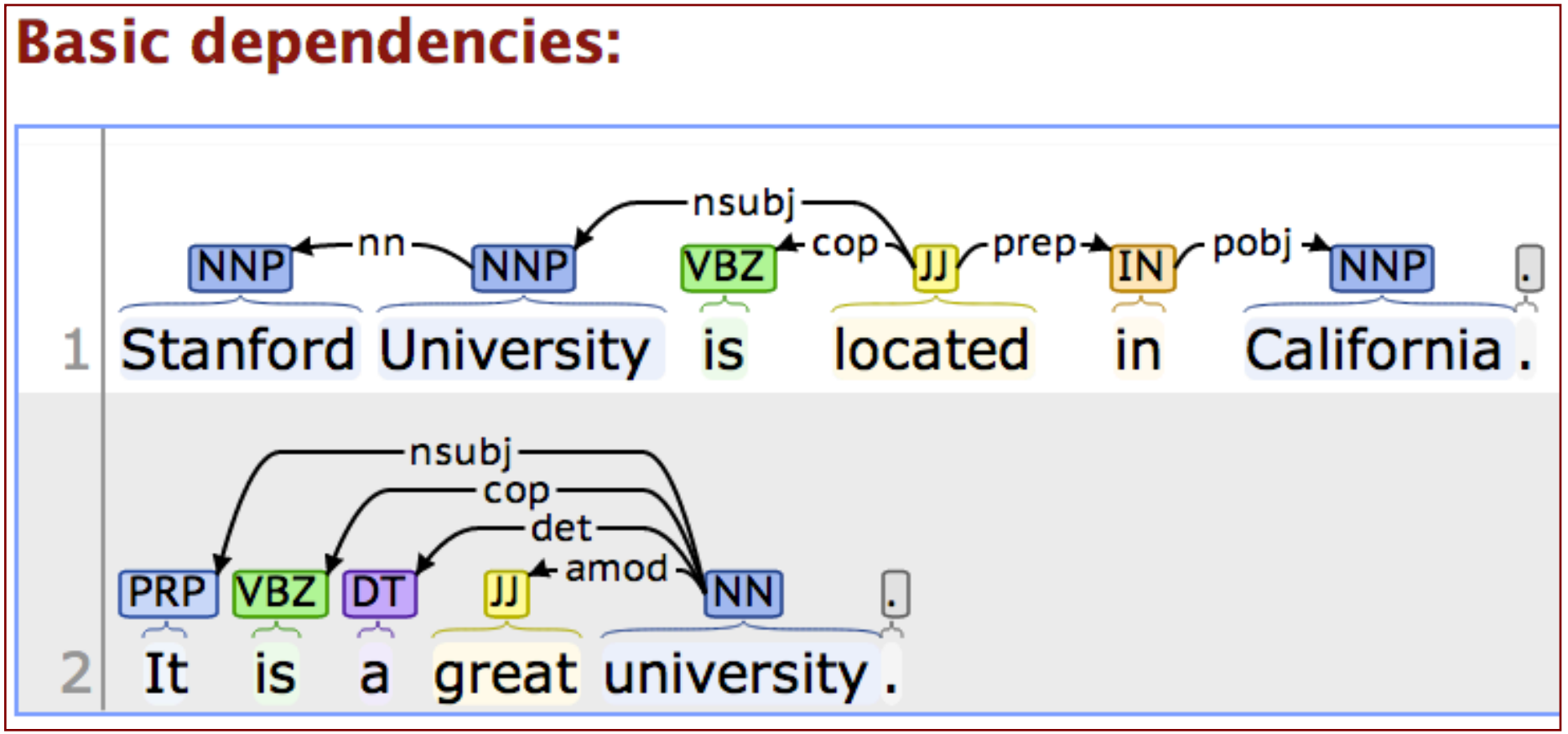
Stanford University is located in California.
It is a great university.



Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

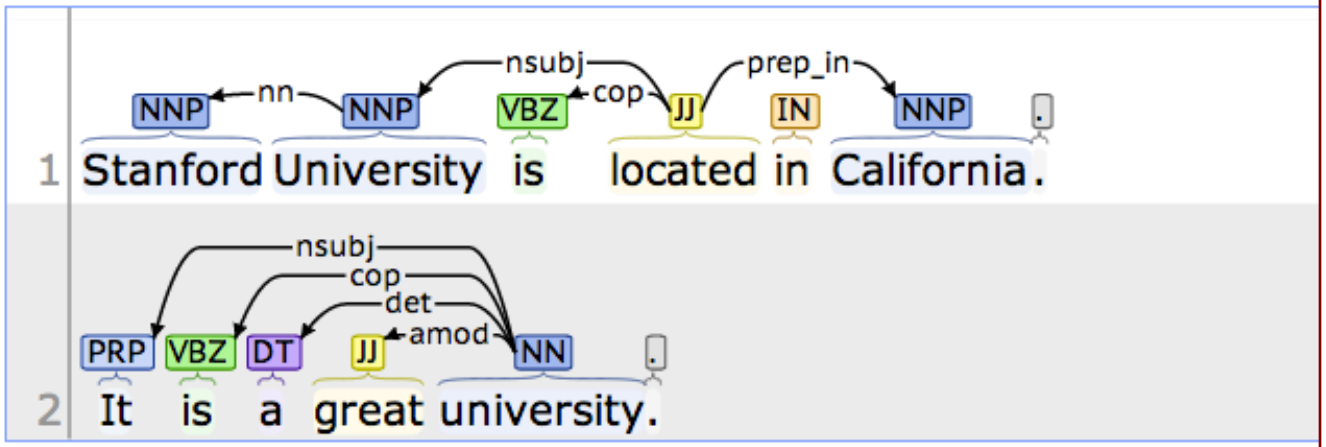
Stanford University is located in California.
It is a great university.



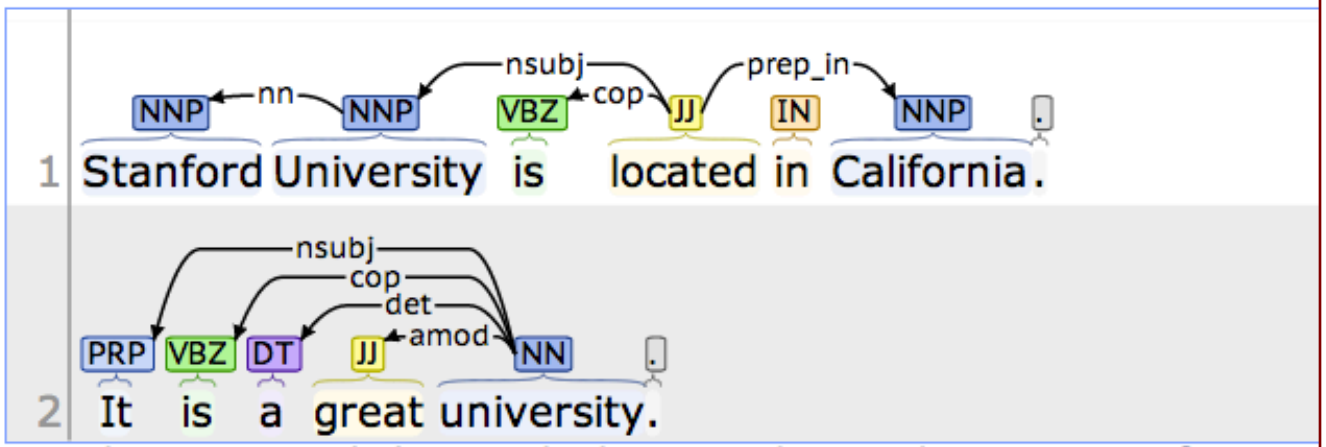
Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

Collapsed dependencies:



Collapsed CC-processed dependencies:



Visualisation provided using the brat visualisation/annotation software.

Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

Output format: Pretty print

Please enter your text here:

Stanford University is located in California. It is a great university.

Submit Clear

Stanford CoreNLP XML Output

Document

Document Info

Sentences

Sentence #1

Tokens

Id	Word	Lemma	Char begin	Char end	POS	NER	Normalized NER	Speaker
1	Stanford	Stanford	0	8	NNP	ORGANIZATION		PER0
2	University	University	9	19	NNP	ORGANIZATION		PER0
3	is	be	20	22	VBZ	O		PER0
4	located	located	23	30	JJ	O		PER0
5	in	in	31	33	IN	O		PER0
6	California	California	34	44	NNP	LOCATION		PER0
7	.	.	44	45	.	O		PER0

Parse tree

(ROOT (S (NP (NNP Stanford) (NNP University)) (VP (VBZ is) (ADJP (JJ located) (PP (IN in) (NP (NNP California))))) (. .)))

Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

Stanford University is located in California.
It is a great university.

Sentence #1

Tokens

Id	Word	Lemma	Char begin	Char end	POS	NER	Normalized NER	Speaker
1	Stanford	Stanford	0	8	NNP	ORGANIZATION		PER0
2	University	University	9	19	NNP	ORGANIZATION		PER0
3	is	be	20	22	VBZ	O		PER0
4	located	located	23	30	JJ	O		PER0
5	in	in	31	33	IN	O		PER0
6	California	California	34	44	NNP	LOCATION		PER0
7	.	.	44	45	.	O		PER0

Parse tree

(ROOT (S (NP (NNP Stanford) (NNP University)) (VP (VBZ is) (ADJP (JJ located) (PP (IN in) (NP (NNP California))))) (. .)))

Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

Stanford University is located in California.
It is a great university.

Sentence #2

Tokens

Id	Word	Lemma	Char begin	Char end	POS	NER	Normalized NER	Speaker
1	It	it	46	48	PRP	O		PER0
2	is	be	49	51	VBZ	O		PER0
3	a	a	52	53	DT	O		PER0
4	great	great	54	59	JJ	O		PER0
5	university	university	60	70	NN	O		PER0
6	.	.	70	71	.	O		PER0

Parse tree

(ROOT (S (NP (PRP It)) (VP (VBZ is) (NP (DT a) (JJ great) (NN university))) (. .)))

Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

Stanford University is located in California. It is a great university.

Coreference resolution graph

1.

Sentence	Head	Text	Context
1	2 (gov)	Stanford University	
2	1	It	
2	5	a great university	

Tokens

Id	Word	Lemma	Char begin	Char end	POS	NER	Normalized NER	Speaker
1	Stanford	Stanford	0	8	NNP	ORGANIZATION		PER0
2	University	University	9	19	NNP	ORGANIZATION		PER0
3	is	be	20	22	VBZ	O	PER0	
4	located	located	23	30	JJ	O	PER0	
5	in	in	31	33	IN	O	PER0	
6	California	California	34	44	NNP	LOCATION	PER0	
7	.	.	44	45	.	O	PER0	

Parse tree
(ROOT (S (NP (NNP Stanford) (NNP University)) (VP (VBZ is) (ADJP (JJ located) (PP (IN in) (NP (NNP California))))) (. .)))

Uncollapsed dependencies

root (ROOT-0 , located-4)
nn (University-2 , Stanford-1)
nsubj (located-4 , University-2)
cop (located-4 , is-3)
prep (located-4 , in-5)
pobj (in-5 , California-6)
Collapsed dependencies

root (ROOT-0 , located-4)
nn (University-2 , Stanford-1)
nsubj (located-4 , University-2)
cop (located-4 , is-3)
prep_in (located-4 , California-6)
Collapsed dependencies with CC processed

root (ROOT-0 , located-4)
nn (University-2 , Stanford-1)
nsubj (located-4 , University-2)
cop (located-4 , is-3)
prep_in (located-4 , California-6)

Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

Stanford University is located in California.
It is a great university.

Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

Output format: XML

Please enter your text here:

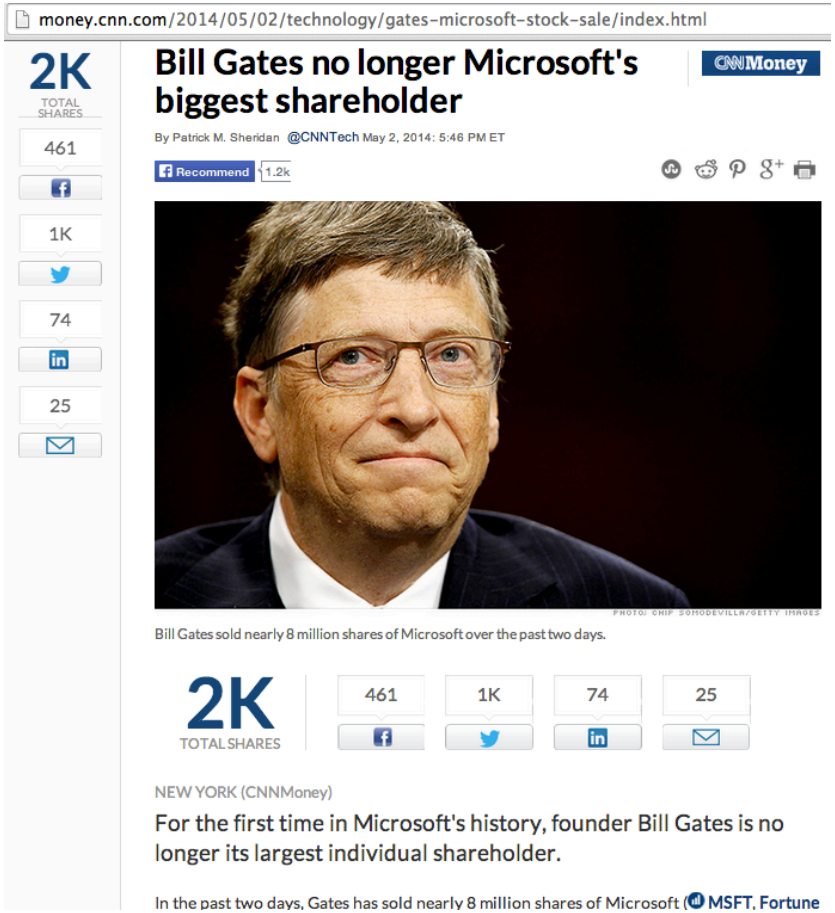
Stanford University is located in California. It is a great university.

Submit Clear

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet href="CoreNLP-to-HTML.xsl" type="text/xsl"?>
<root>
  <document>
    <sentences>
      <sentence id="1">
        <tokens>
          <token id="1">
            <word>Stanford</word>
            <lemma>Stanford</lemma>
            <CharacterOffsetBegin>0</CharacterOffsetBegin>
            <CharacterOffsetEnd>8</CharacterOffsetEnd>
            <POS>NNP</POS>
            <NER>ORGANIZATION</NER>
            <Speaker>PER0</Speaker>
          </token>
          <token id="2">
            <word>University</word>
            <lemma>University</lemma>
            <CharacterOffsetBegin>9</CharacterOffsetBegin>
            <CharacterOffsetEnd>19</CharacterOffsetEnd>
            <POS>NNP</POS>
            <NER>ORGANIZATION</NER>
            <Speaker>PER0</Speaker>
          </token>
```

NER for News Article

http://money.cnn.com/2014/05/02/technology/gates-microsoft-stock-sale/index.html



Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

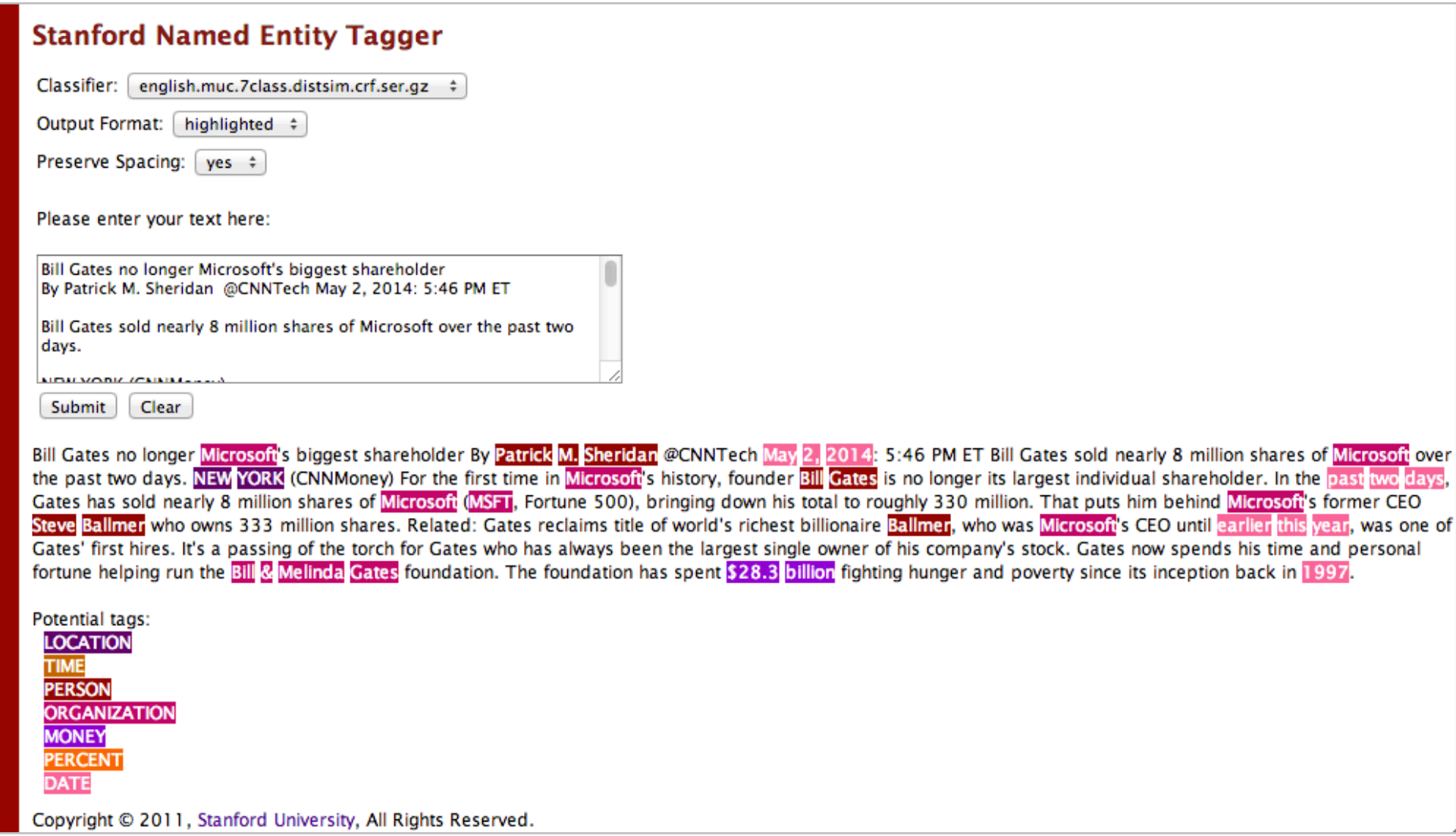
NEW YORK (CNNMoney)

For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder.
In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million.

That puts him behind Microsoft's former CEO Steve Ballmer who owns 333 million shares.
Related: Gates reclaims title of world's richest billionaire
Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires.
It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation.
The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.

Stanford Named Entity Tagger (NER)

http://nlp.stanford.edu:8080/ner/process



Stanford Named Entity Tagger (NER)

http://nlp.stanford.edu:8080/ner/process

Stanford Named Entity Tagger

Classifier: english.muc.7class.distsim.crf.ser.gz

Output Format: inlineXML

Preserve Spacing: yes

Please enter your text here:

Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

NEW YORK (CNNMoney)

Submit Clear

Bill Gates no longer <ORGANIZATION>Microsoft</ORGANIZATION>'s biggest shareholder By <PERSON>Patrick M. Sheridan</PERSON> @CNNTech <DATE>May 2, 2014</DATE>: 5:46 PM ET Bill Gates sold nearly 8 million shares of <ORGANIZATION>Microsoft</ORGANIZATION> over the past two days. <LOCATION>NEW YORK</LOCATION> (CNNMoney) For the first time in <ORGANIZATION>Microsoft</ORGANIZATION>'s history, founder <PERSON>Bill Gates</PERSON> is no longer its largest individual shareholder. In the <DATE>past two days</DATE>, Gates has sold nearly 8 million shares of <ORGANIZATION>Microsoft</ORGANIZATION> (<ORGANIZATION>MSFT</ORGANIZATION>, Fortune 500), bringing down his total to roughly 330 million. That puts him behind <ORGANIZATION>Microsoft</ORGANIZATION>'s former CEO <PERSON>Steve Ballmer</PERSON> who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire <PERSON>Ballmer</PERSON>, who was <ORGANIZATION>Microsoft</ORGANIZATION>'s CEO until <DATE>earlier this year</DATE>, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the <ORGANIZATION>Bill & Melinda Gates</ORGANIZATION> foundation. The foundation has spent <MONEY>$28.3 billion</MONEY> fighting hunger and poverty since its inception back in <DATE>1997</DATE>.

Stanford Named Entity Tagger (NER)

http://nlp.stanford.edu:8080/ner/process

Stanford Named Entity Tagger

Classifier: english.muc.7class.distsim.crf.ser.gz

Output Format: xml

Preserve Spacing: yes

Please enter your text here:

Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

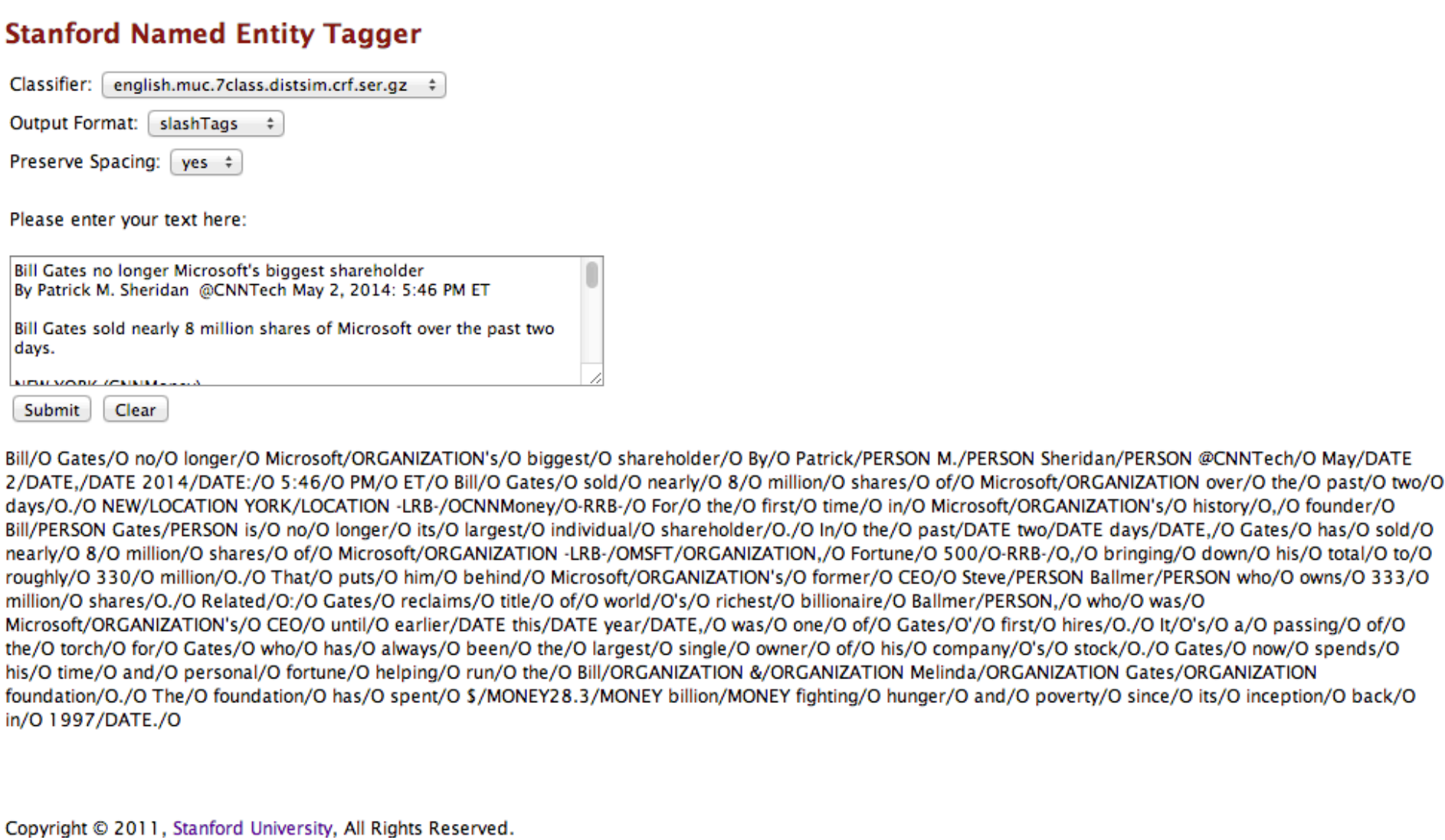
NEW YORK (CNNMoney)

Submit Clear

<wi num="0" entity="O">Bill</wi> <wi num="1" entity="O">Gates</wi> <wi num="2" entity="O">no</wi> <wi num="3" entity="O">longer</wi> <wi num="4" entity="ORGANIZATION">Microsoft</wi><wi num="5" entity="O">'s</wi> <wi num="6" entity="O">biggest</wi> <wi num="7" entity="O">shareholder</wi> <wi num="8" entity="O">By</wi> <wi num="9" entity="PERSON">Patrick</wi> <wi num="10" entity="PERSON">M.</wi> <wi num="11" entity="PERSON">Sheridan</wi> <wi num="12" entity="O">@CNNTech</wi> <wi num="13" entity="DATE">May</wi> <wi num="14" entity="DATE">2</wi><wi num="15" entity="DATE">,</wi> <wi num="16" entity="DATE">2014</wi><wi num="17" entity="O">:</wi> <wi num="18" entity="O">5:46</wi> <wi num="19" entity="O">PM</wi> <wi num="20" entity="O">ET</wi> <wi num="21" entity="O">Bill</wi> <wi num="22" entity="O">Gates</wi> <wi num="23" entity="O">sold</wi> <wi num="24" entity="O">nearly</wi> <wi num="25" entity="O">8</wi> <wi num="26" entity="O">million</wi> <wi num="27" entity="O">shares</wi> <wi num="28" entity="O">of</wi> <wi num="29" entity="ORGANIZATION">Microsoft</wi> <wi num="30" entity="O">over</wi> <wi num="31" entity="O">the</wi> <wi num="32" entity="O">past</wi> <wi num="33" entity="O">two</wi> <wi num="34" entity="O">days</wi><wi num="35" entity="O">.</wi> <wi num="0" entity="LOCATION">NEW</wi> <wi num="1" entity="LOCATION">YORK</wi> <wi num="2" entity="O">-LRB-</wi><wi num="3" entity="O">CNNMoney</wi><wi num="4" entity="O">-RRB-</wi> <wi num="5" entity="O">For</wi> <wi num="6" entity="O">the</wi> <wi num="7" entity="O">first</wi> <wi num="8" entity="O">time</wi> <wi num="9" entity="O">in</wi> <wi num="10" entity="ORGANIZATION">Microsoft</wi><wi num="11" entity="O">'s</wi> <wi num="12" entity="O">history</wi><wi num="13" entity="O">,</wi> <wi num="14" entity="O">founder</wi> <wi num="15" entity="PERSON">Bill</wi> <wi num="16" entity="PERSON">Gates</wi> <wi num="17" entity="O">is</wi> <wi num="18" entity="O">no</wi> <wi num="19" entity="O">longer</wi> <wi num="20" entity="O">its</wi> <wi num="21" entity="O">largest</wi> <wi num="22" entity="O">individual</wi> <wi num="23" entity="O">shareholder</wi><wi num="24" entity="O">.</wi> <wi num="0" entity="O">In</wi> <wi num="1" entity="O">the</wi> <wi num="2" entity="DATE">past</wi> <wi num="3" entity="DATE">two</wi> <wi num="4" entity="DATE">days</wi><wi num="5" entity="O">,</wi> <wi num="6" entity="O">Gates</wi> <wi num="7" entity="O">has</wi> <wi num="8" entity="O">sold</wi>

Stanford Named Entity Tagger (NER)

http://nlp.stanford.edu:8080/ner/process



Stanford Named Entity Tagger

Classifier: english.muc.7class.distsim.crf.ser.gz

Output Format: slashTags

Preserve Spacing: yes

Please enter your text here:

Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

NEW YORK (CNNMoney)

Submit Clear

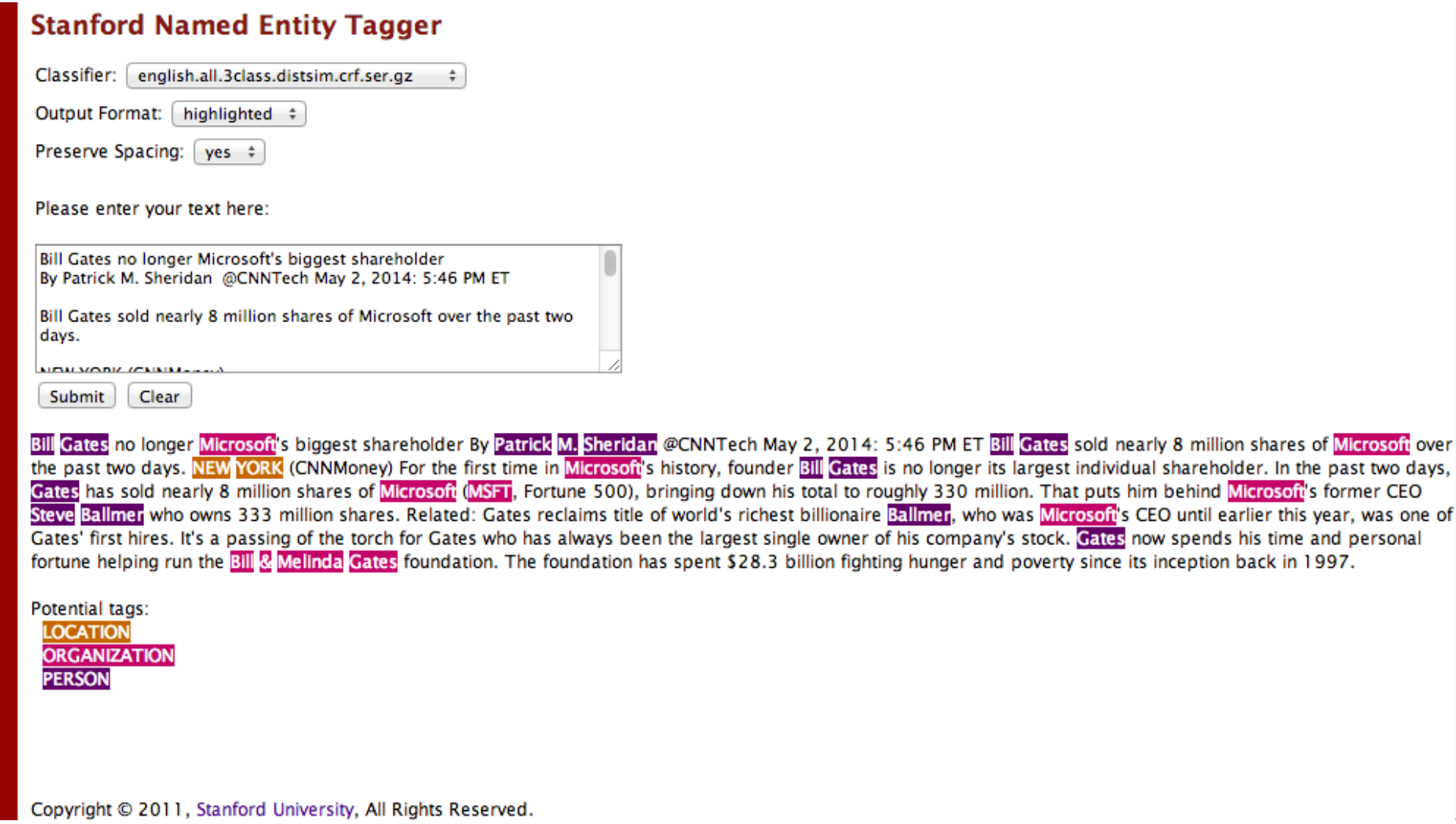
Bill/O Gates/O no/O longer/O Microsoft/ORGANIZATION's/O biggest/O shareholder/O By/O Patrick/PERSON M./PERSON Sheridan/PERSON @CNNTech/O May/DATE 2/DATE,/DATE 2014/DATE:/O 5:46/O PM/O ET/O Bill/O Gates/O sold/O nearly/O 8/O million/O shares/O of/O Microsoft/ORGANIZATION over/O the/O past/O two/O days/O./O NEW/LOCATION YORK/LOCATION -LRB-/OCNNMoney/O-RRB-/O For/O the/O first/O time/O in/O Microsoft/ORGANIZATION's/O history/O,/O founder/O Bill/PERSON Gates/PERSON is/O no/O longer/O its/O largest/O individual/O shareholder/O./O In/O the/O past/DATE two/DATE days/DATE,/O Gates/O has/O sold/O nearly/O 8/O million/O shares/O of/O Microsoft/ORGANIZATION -LRB-/OMSFT/ORGANIZATION,/O Fortune/O 500/O-RRB-/O,/O bringing/O down/O his/O total/O to/O roughly/O 330/O million/O./O That/O puts/O him/O behind/O Microsoft/ORGANIZATION's/O former/O CEO/O Steve/PERSON Ballmer/PERSON who/O owns/O 333/O million/O shares/O./O Related/O:/O Gates/O reclaims/O title/O of/O world/O's/O richest/O billionaire/O Ballmer/PERSON,/O who/O was/O Microsoft/ORGANIZATION's/O CEO/O until/O earlier/DATE this/DATE year/DATE,/O was/O one/O of/O Gates/O'/O first/O hires/O./O It/O's/O a/O passing/O of/O the/O torch/O for/O Gates/O who/O has/O always/O been/O the/O largest/O single/O owner/O of/O his/O company/O's/O stock/O./O Gates/O now/O spends/O his/O time/O and/O personal/O fortune/O helping/O run/O the/O Bill/ORGANIZATION &/ORGANIZATION Melinda/ORGANIZATION Gates/ORGANIZATION foundation/O./O The/O foundation/O has/O spent/O $/MONEY28.3/MONEY billion/MONEY fighting/O hunger/O and/O poverty/O since/O its/O inception/O back/O in/O 1997/DATE./O

Stanford Named Entity Tagger (NER)

http://nlp.stanford.edu:8080/ner/process

Stanford Named Entity Tagger (NER)

http://nlp.stanford.edu:8080/ner/process



Classifier: english.muc.**7class**.distsim.crf.ser.gz

Bill Gates no longer Microsoft's biggest shareholder By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET Bill Gates sold nearly 8 million shares of Microsoft over the past two days. NEW YORK (CNNMoney) For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.

Potential tags:
- LOCATION
- TIME
- PERSON
- ORGANIZATION
- MONEY
- PERCENT
- DATE

Classifier: english.all.**3class**.distsim.crf.ser.gz

Bill Gates no longer Microsoft's biggest shareholder By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET Bill Gates sold nearly 8 million shares of Microsoft over the past two days. NEW YORK (CNNMoney) For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.

Potential tags:
- LOCATION
- ORGANIZATION
- PERSON

Stanford Named Entity Tagger (NER)

http://nlp.stanford.edu:8080/ner/process

Stanford NER Output Format: inlineXML

Bill Gates no longer <ORGANIZATION>Microsoft</ORGANIZATION>'s biggest shareholder By <PERSON>Patrick M. Sheridan</PERSON> @CNNTech <DATE>May 2, 2014</DATE>: 5:46 PM ET Bill Gates sold nearly 8 million shares of <ORGANIZATION>Microsoft</ORGANIZATION> over the past two days. <LOCATION>NEW YORK</LOCATION> (CNNMoney) For the first time in <ORGANIZATION>Microsoft</ORGANIZATION>'s history, founder <PERSON>Bill Gates</PERSON> is no longer its largest individual shareholder. In the <DATE>past two days</DATE>, Gates has sold nearly 8 million shares of <ORGANIZATION>Microsoft</ORGANIZATION> (<ORGANIZATION>MSFT</ORGANIZATION>, Fortune 500), bringing down his total to roughly 330 million. That puts him behind <ORGANIZATION>Microsoft</ORGANIZATION>'s former CEO <PERSON>Steve Ballmer</PERSON> who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire <PERSON>Ballmer</PERSON>, who was <ORGANIZATION>Microsoft</ORGANIZATION>'s CEO until <DATE>earlier this year</DATE>, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the <ORGANIZATION>Bill & Melinda Gates</ORGANIZATION> foundation. The foundation has spent <MONEY>$28.3 billion</MONEY> fighting hunger and poverty since its inception back in <DATE>1997</DATE>.

Stanford Named Entity Tagger (NER)

http://nlp.stanford.edu:8080/ner/process

Stanford NER Output Format: slashTags

Bill/O Gates/O no/O longer/O Microsoft/ORGANIZATION's/O biggest/O shareholder/O By/O Patrick/PERSON M./PERSON Sheridan/PERSON @CNNTech/O May/DATE 2/DATE,/DATE 2014/DATE:/O 5:46/O PM/O ET/O Bill/O Gates/O sold/O nearly/O 8/O million/O shares/O of/O Microsoft/ORGANIZATION over/O the/O past/O two/O days/O./O NEW/LOCATION YORK/LOCATION -LRB-/OCNNMoney/O-RRB-/O For/O the/O first/O time/O in/O Microsoft/ORGANIZATION's/O history/O,/O founder/O Bill/PERSON Gates/PERSON is/O no/O longer/O its/O largest/O individual/O shareholder/O./O In/O the/O past/DATE two/DATE days/DATE,/O Gates/O has/O sold/O nearly/O 8/O million/O shares/O of/O Microsoft/ORGANIZATION -LRB-/OMSFT/ORGANIZATION,/O Fortune/O 500/O-RRB-/O,/O bringing/O down/O his/O total/O to/O roughly/O 330/O million/O./O That/O puts/O him/O behind/O Microsoft/ORGANIZATION's/O former/O CEO/O Steve/PERSON Ballmer/PERSON who/O owns/O 333/O million/O shares/O./O Related/O:/O Gates/O reclaims/O title/O of/O world/O's/O richest/O billionaire/O Ballmer/PERSON,/O who/O was/O Microsoft/ORGANIZATION's/O CEO/O until/O earlier/DATE this/DATE year/DATE,/O was/O one/O of/O Gates/O'/O first/O hires/O./O It/O's/O a/O passing/O of/O the/O torch/O for/O Gates/O who/O has/O always/O been/O the/O largest/O single/O owner/O of/O his/O company/O's/O stock/O./O Gates/O now/O spends/O his/O time/O and/O personal/O fortune/O helping/O run/O the/O Bill/ORGANIZATION &/ORGANIZATION Melinda/ORGANIZATION Gates/ORGANIZATION foundation/O./O The/O foundation/O has/O spent/O $/MONEY28.3/MONEY billion/MONEY fighting/O hunger/O and/O poverty/O since/O its/O inception/O back/O in/O 1997/DATE./O

CKIP 中研院中文斷詞系統

http://ckipsvr.iis.sinica.edu.tw/

歐巴馬(Nb) 是(SHI) 美國(Nc) 的(DE) 一(Neu) 位(Nf) 總統(Na)

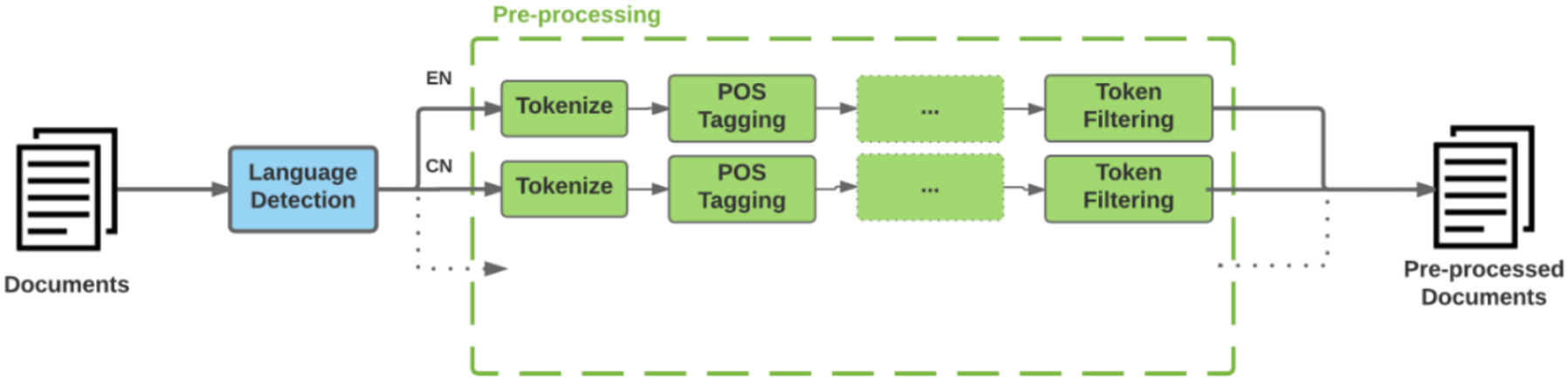
Vector Representations of Words

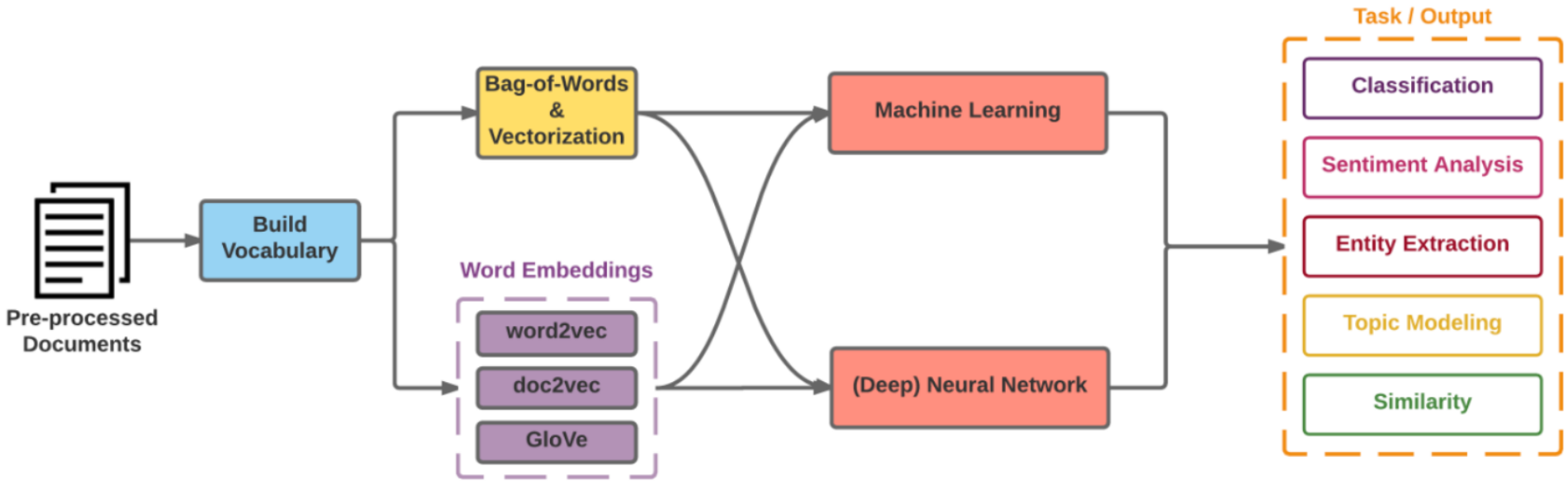
Word Embeddings

Word2Vec

GloVe

Modern NLP Pipeline





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

Facebook Research FastText

Pre-trained word vectors
Word2Vec
wiki.zh.vec (861MB)
332647 word
300 vec

Pre-trained word vectors for 90 languages, trained on Wikipedia using fastText.

These vectors in dimension 300 were obtained using the skip-gram model with default parameters.

https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

Source: Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. "Enriching word vectors with subword information." *arXiv preprint arXiv:1607.04606* (2016).

Facebook Research FastText Word2Vec: wiki.zh.vec

(861MB) (332647 word 300 vec)

wiki.zh.vec

```
31845 yg -0.3978 0.49084 -0.54621 0.078991 0.8584 -0.26163 -0.45787 0.060828 0.36513 -0.03771 0.80791 0.16613 1.4828 -0.89862 0.085965
31846 迴圈 -0.034834 0.71651 -0.4377 0.48344 0.31117 -0.51783 -0.40156 -0.057097 0.31535 -0.088301 0.23436 0.30884 1.2932 -0.6704 0.21
31847 ぶっ -0.23267 0.39349 -0.90806 -0.53805 0.59308 -0.31819 -0.64229 0.16871 0.10086 0.09342 1.0914 -0.16019 1.6954 -0.70604 -0.218
31848 三公 0.54129 0.55641 -0.4348 0.25094 0.1631 -0.10326 -0.54099 0.064742 0.13175 0.10217 0.84938 -0.10287 1.312 -0.74969 0.24025 -
31849 水貨 -0.14451 0.80455 -0.6145 0.55905 0.58307 -0.02559 -0.41088 -0.19056 -0.09178 0.33935 1.1927
31850 刚才 0.19347 0.553 -0.64736 0.26358 0.83816 -0.24098 -0.83997 -0.16232 -0.024786 -0.2483 0.69732
31851 無知 -0.0089777 0.90866 -0.25306 0.72983 0.67791 -0.3285 -0.63835 0.075295 0.4774 -0.04134 0.721
31852 好轉 -0.026068 0.92676 -0.47469 0.50129 0.67343 -0.32509 -0.32917 0.066499 0.3875 0.0011722 0.66
31853 紀事 0.40541 0.67654 -0.5351 0.30329 0.43042 -0.24675 -0.19287 0.34207 0.35516 -0.076331 0.85916
31854 變回 -0.089933 0.88136 -0.43524 0.59963 0.6403 -0.70981 -0.56788 -0.074018 0.16905 -0.086594 0.6
31855 牟尼 -0.26578 0.6434 0.028982 -0.044001 0.88297 -0.17646 -0.64672 0.040483 0.43653 0.084908 0.74
31856 埋藏 -0.0985 0.85082 -0.33363 0.24784 0.71518 -0.59054 -0.73731 0.050949 0.36726 -0.076886 0.817
31857 正大 0.21069 0.27605 -0.83862 -0.099698 0.47894 -0.32196 -0.38288 -0.01892 0.40548 -0.029619 0.7
31858 kis -0.30595 0.18482 -0.71287 -0.314 0.44776 -0.44245 -0.36447 -0.23723 0.00098801 -0.2528 0.608
31859 合奏 0.1841 0.60874 -0.51376 -0.48002 0.21506 -0.55515 -0.71746 0.030735 0.39508 -0.40856 0.6226
31860 精兵 0.25619 0.77186 -0.48847 0.23118 0.27254 0.21305 -0.3517 0.47305 0.24882 -0.34756 1.025 0.1
31861 疲勞 -0.072521 1.0381 -0.51933 0.19421 0.67573 -0.45204 -0.20126 0.22704 0.44196 0.018401 0.3473
31862 襯 -0.11771 1.4272 -1.0849 0.77532 0.87026 -0.6892 -0.3521 0.036517 0.42727 -0.1871 0.82789 -0.0
31863 小貓 -0.21554 0.73988 -0.39628 0.044656 1.0602 -0.67047 -0.54102 0.11888 0.1693 0.19343 1.0841 0.
31864 lai -0.25451 0.31596 -0.29228 -0.19144 0.99059 -0.24459 -0.66342 0.063093 -0.061142 -0.22749 0.6
31865 偏東 -0.50835 1.0943 0.043918 0.29173 1.0161 -0.32493 -0.27305 0.026946 0.46811 -0.3874 1.4049 0.
31866 大约是 -0.35726 -0.03476 -0.28672 0.075447 0.18175 -0.39421 -0.32088 0.025225 0.34808 0.074744 0.
31867 franch -0.6046 -0.3235 0.024041 -0.2756 0.74761 -0.14654 0.0082566 -0.10071 0.53593 -0.17374 0.2
31868 brazilian -0.54029 -0.63905 -0.094006 -0.68768 0.33263 -0.1583 -0.060424 0.20644 0.46234 -0.0764
31869 夾竹桃 -0.4361 0.011429 -0.078896 -0.078186 0.37747 -0.052101 -0.096683 0.10769 0.62661 -0.37252
31870 continent -0.37761 -0.72151 -0.42248 -0.81768 0.5016 -0.48569 0.13464 0.12644 0.32292 0.18099 0.
31871 我还是 0.097443 0.28929 -0.14202 0.034027 0.50621 -0.1647 -0.45849 -0.16198 0.13965 -0.33451 0.61
31872 vienna -0.25827 -0.050966 0.050502 -0.63466 0.4949 -0.17448 -0.59978 0.20269 0.37532 0.059419 0.
31873 固态 -0.12678 0.4556 -0.27108 0.12506 0.52106 -0.058477 -0.69296 0.12162 0.26508 -0.089028 0.752
31874 吉普 -0.33693 0.48335 -0.58455 0.13722 0.74856 -0.24529 -0.41125 -0.13832 0.33871 -0.12051 0.864
31875 實物 0.030096 0.65756 -0.67982 0.2203 0.38492 -0.19001 -0.53136 -0.10322 0.24523 0.15287 0.92591
31876 教职 0.11559 0.67087 -0.5111 0.14955 0.61417 -0.51571 -0.47901 0.29445 0.37629 -0.24232 0.4608 -
31877 惕 0.50469 1.5357 -0.64393 0.48668 0.69479 -0.23443 -0.47863 0.16288 0.3347 -0.51673 0.86777 0.0
31878 岸上 0.088323 0.85815 -0.485 0.30383 0.75965 -0.25031 -0.76678 0.12805 0.37641 -0.088752 0.65012
31879 议和 0.26835 0.94854 -0.27972 0.097623 0.43305 -0.031361 -0.57406 0.21608 0.3324 -0.36823 0.6987
31880 aka -0.21332 0.11216 -0.48872 -0.18531 0.79093 -0.34221 -0.51122 0.10067 0.29963 -0.075253 0.642
31881 滑鐵盧 -0.28726 0.88014 -0.39751 -0.056992 0.37408 -0.16967 -0.20673 -0.048533 -0.1978 -0.13107 0
```

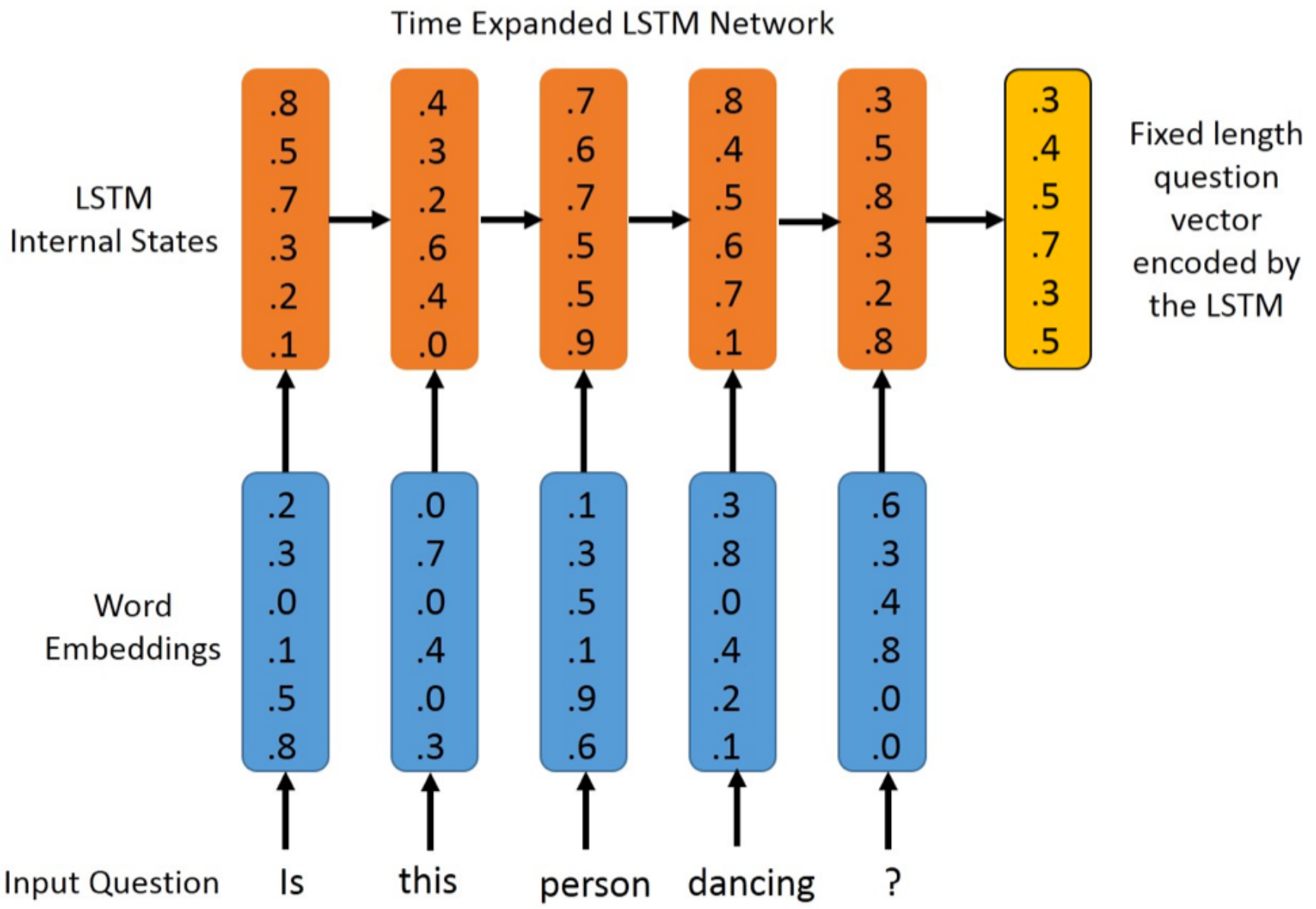
Models

The models can be downloaded from:

- Afrikaans: *bin+text*, *text*
- Albanian: *bin+text*, *text*
- Arabic: *bin+text*, *text*
- Armenian: *bin+text*, *text*
- Asturian: *bin+text*, *text*
- Azerbaijani: *bin+text*, *text*
- Bashkir: *bin+text*, *text*
- Basque: *bin+text*, *text*
- Belarusian: *bin+text*, *text*
- Bengali: *bin+text*, *text*
- Bosnian: *bin+text*, *text*
- Breton: *bin+text*, *text*
- Bulgarian: *bin+text*, *text*
- Burmese: *bin+text*, *text*
- Catalan: *bin+text*, *text*
- Cebuano: *bin+text*, *text*
- Chechen: *bin+text*, *text*
- Chinese: *bin+text*, *text*
- Chuvash: *bin+text*, *text*
- Croatian: *bin+text*, *text*
- Czech: *bin+text*, *text*

https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

Word Embeddings in LSTM RNN



Source: https://avisingh599.github.io/deeplearning/visual-qa/

NLP Tools: spaCy vs. NLTK

	SPACY	SYNTAXNET	NLTK	CORENLP
Easy installation	✓	✗	✓	✓
Python API	✓	✗	✓	✗
Multi-language support	○	✓	✓	✓
Tokenization	✓	✓	✓	✓
Part-of-speech tagging	✓	✓	✓	✓
Sentence segmentation	✓	✓	✓	✓
Dependency parsing	✓	✓	✗	✓
Entity Recognition	✓	✗	✓	✓
Integrated word vectors	✓	✗	✗	✗
Sentiment analysis	✓	✗	✓	✓
Coreference resolution	✗	✗	✗	✓

Source: https://spacy.io/docs/api/

Natural Language Processing (NLP) spaCy

1. Tokenization
2. Part-of-speech tagging
3. Sentence segmentation
4. Dependency parsing
5. Entity Recognition
6. Integrated word vectors
7. Sentiment analysis
8. Coreference resolution

Source: https://spacy.io/docs/api/

spaCy:
Fastest Syntactic Parser

SYSTEM	LANGUAGE	ACCURACY	SPEED (WPS)
spaCy	**Cython**	**91.8**	**13,963**
ClearNLP	Java	91.7	10,271
CoreNLP	Java	89.6	8,602
MATE	Java	**92.5**	550
Turbo	C++	92.4	349

Source: https://spacy.io/docs/api/

Processing Speed of NLP libraries

	ABSOLUTE (MS PER DOC)			RELATIVE (TO SPACY)		
SYSTEM	TOKENIZE	TAG	PARSE	TOKENIZE	TAG	PARSE
spaCy	**0.2ms**	**1ms**	**19ms**	**1x**	**1x**	**1x**
CoreNLP	2ms	10ms	49ms	10x	10x	2.6x
ZPar	1ms	8ms	850ms	5x	8x	44.7x
NLTK	4ms	443ms	n/a	20x	443x	n/a

Source: https://spacy.io/docs/api/

Google SyntaxNet (2016): Best Syntactic Dependency Parsing Accuracy

SYSTEM	NEWS	WEB	QUESTIONS
spaCy	92.8	n/a	n/a
Parsey McParseface	94.15	89.08	94.77
Martins et al. (2013)	93.10	88.23	94.21
Zhang and McDonald (2014)	93.32	88.65	93.37
Weiss et al. (2015)	93.91	89.29	94.17
Andor et al. (2016)	**94.44**	**90.17**	**95.40**

Source: https://spacy.io/docs/api/

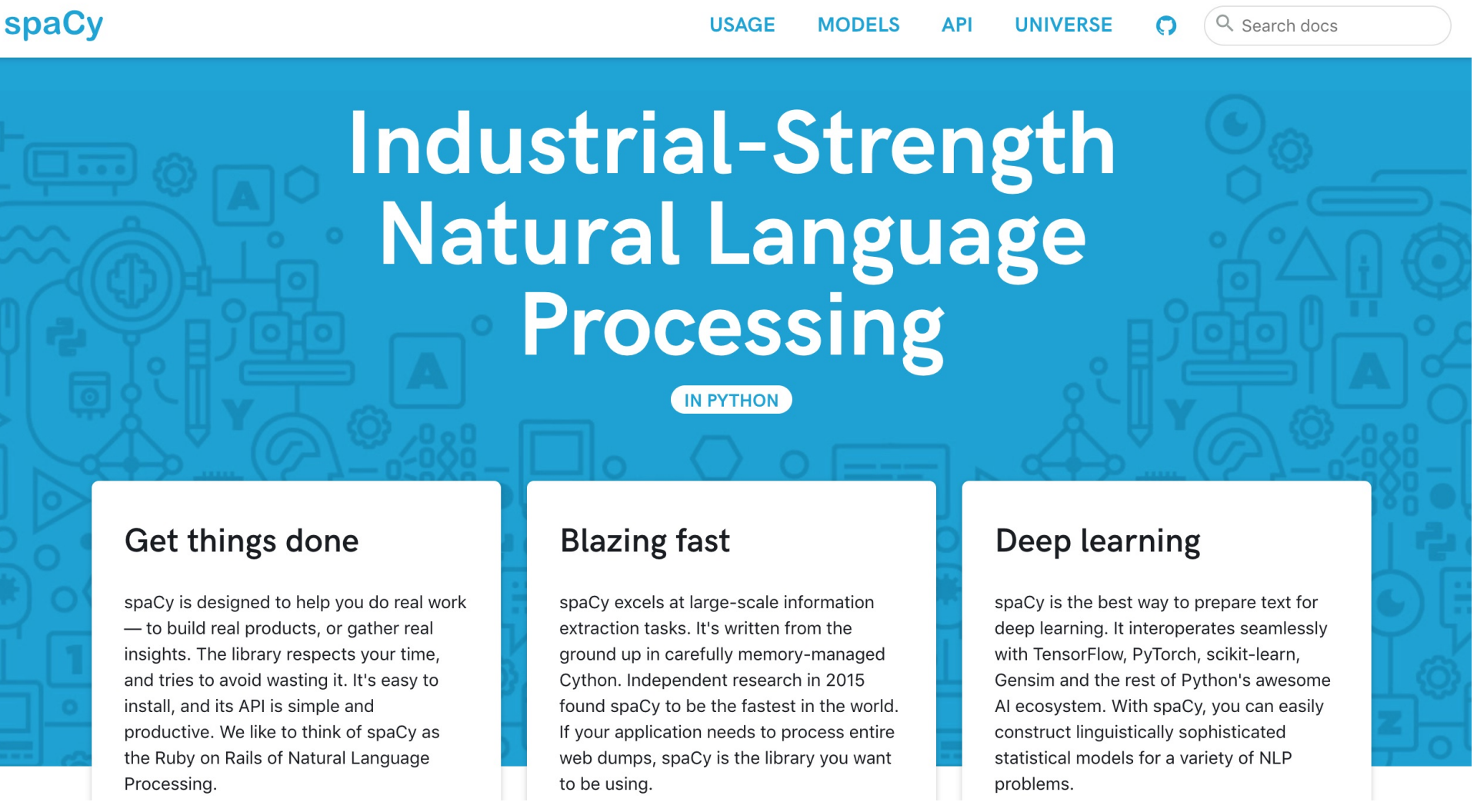
Named Entity Recognition (NER)

SYSTEM	PRECISION	RECALL	F-MEASURE
spaCy	0.7240	0.6514	0.6858
CoreNLP	**0.7914**	**0.7327**	**0.7609**
NLTK	0.5136	0.6532	0.5750
LingPipe	0.5412	0.5357	0.5384

Text Analytics with Python

Source: https://www.python.org/community/logos/

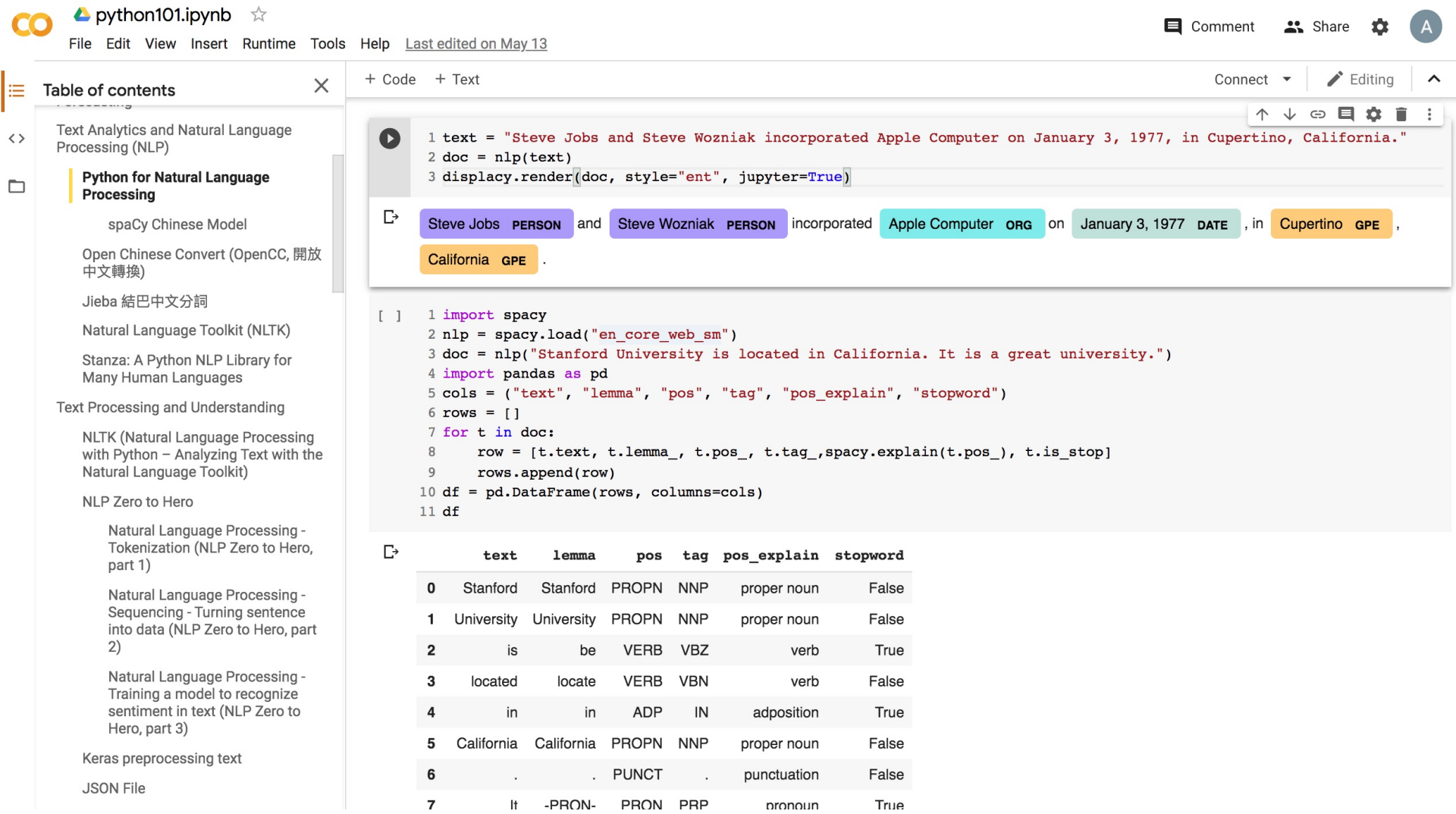
spaCy:
Natural Language Processing



https://spacy.io/

Python in Google Colab (Python101)

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

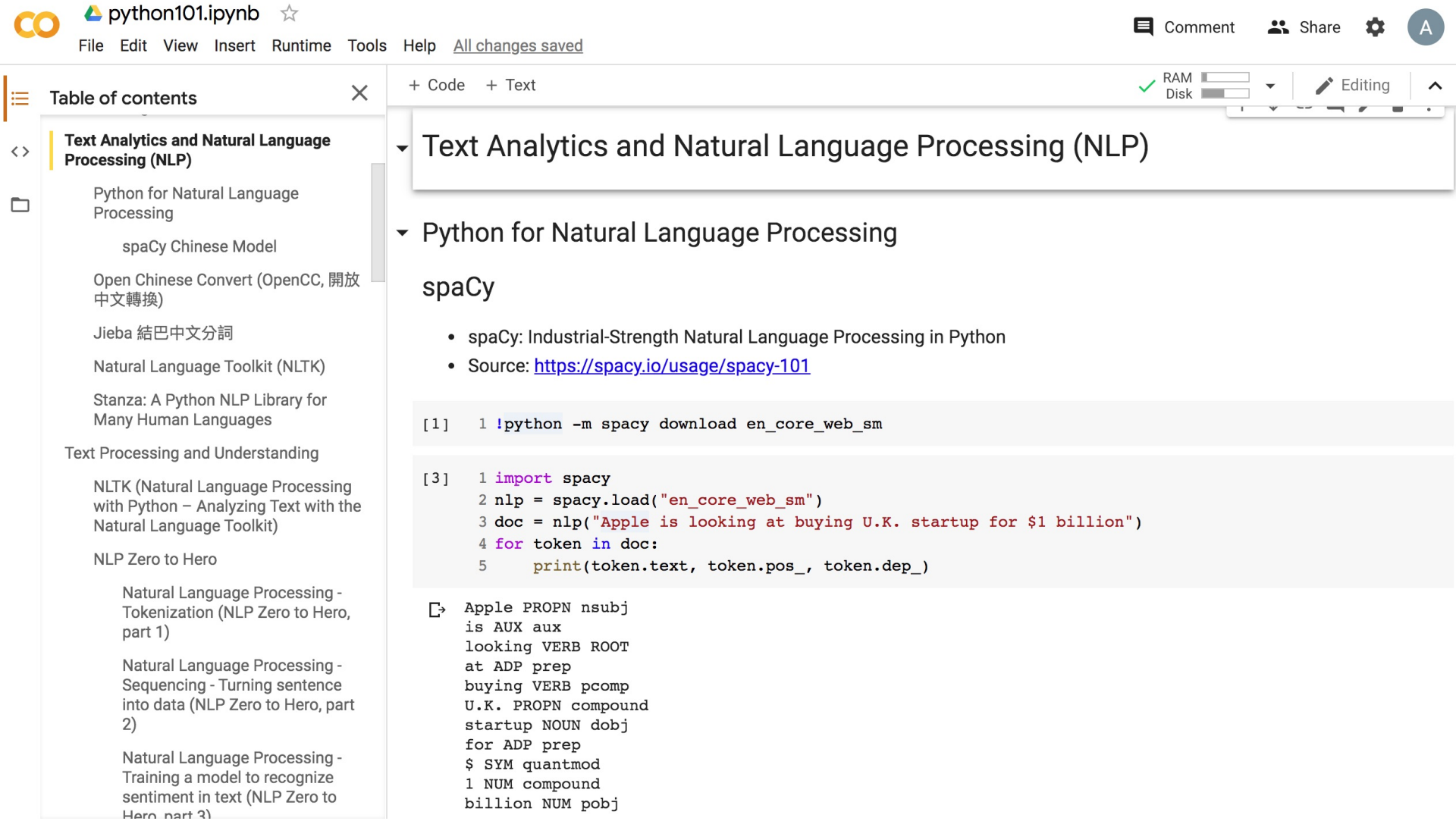


	text	lemma	pos	tag	pos_explain	stopword
0	Stanford	Stanford	PROPN	NNP	proper noun	False
1	University	University	PROPN	NNP	proper noun	False
2	is	be	VERB	VBZ	verb	True
3	located	locate	VERB	VBN	verb	False
4	in	in	ADP	IN	adposition	True
5	California	California	PROPN	NNP	proper noun	False
6	.	.	PUNCT	.	punctuation	False
7	It	-PRON-	PRON	PRP	pronoun	True

https://tinyurl.com/aintpupython101

Python in Google Colab (Python101)

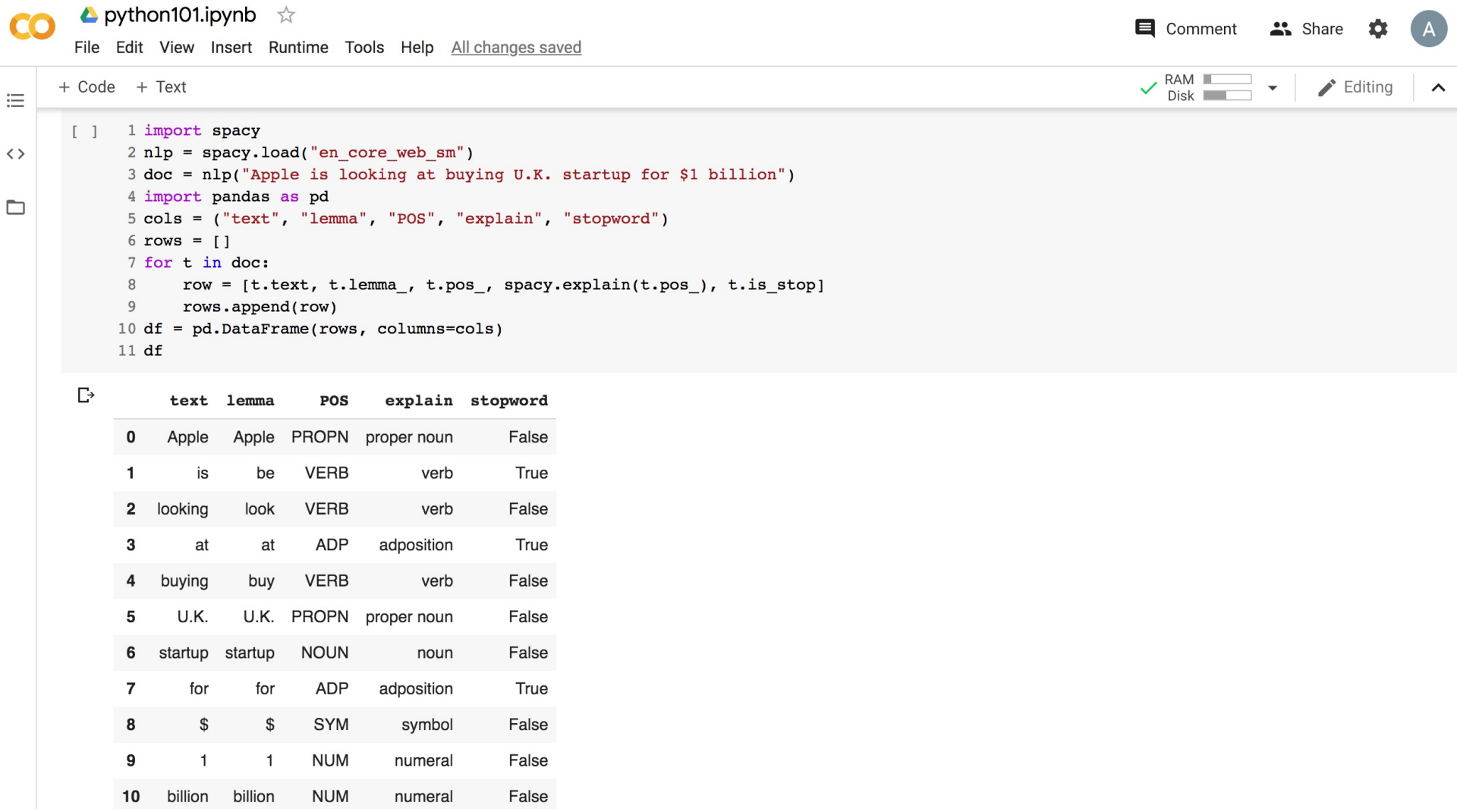
https://colab.research.google.com/drive/1FEG6DnGvwfUbeo4zJ1zTunjMqf2RkCrT



https://tinyurl.com/aintpupython101

Python in Google Colab (Python101)

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

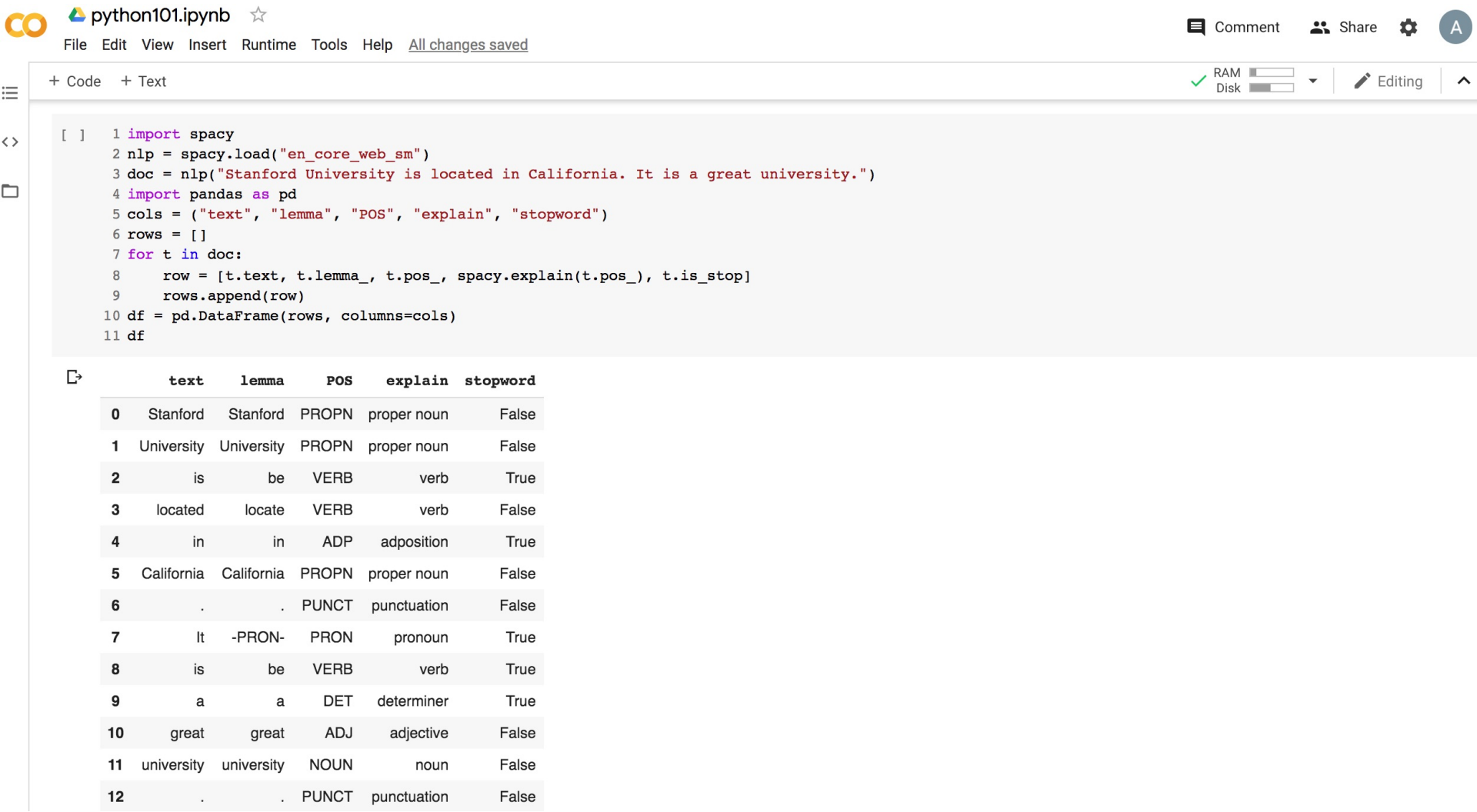


	text	lemma	POS	explain	stopword
0	Apple	Apple	PROPN	proper noun	False
1	is	be	VERB	verb	True
2	looking	look	VERB	verb	False
3	at	at	ADP	adposition	True
4	buying	buy	VERB	verb	False
5	U.K.	U.K.	PROPN	proper noun	False
6	startup	startup	NOUN	noun	False
7	for	for	ADP	adposition	True
8	$	$	SYM	symbol	False
9	1	1	NUM	numeral	False
10	billion	billion	NUM	numeral	False

https://tinyurl.com/aintpupython101

Python in Google Colab (Python101)

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



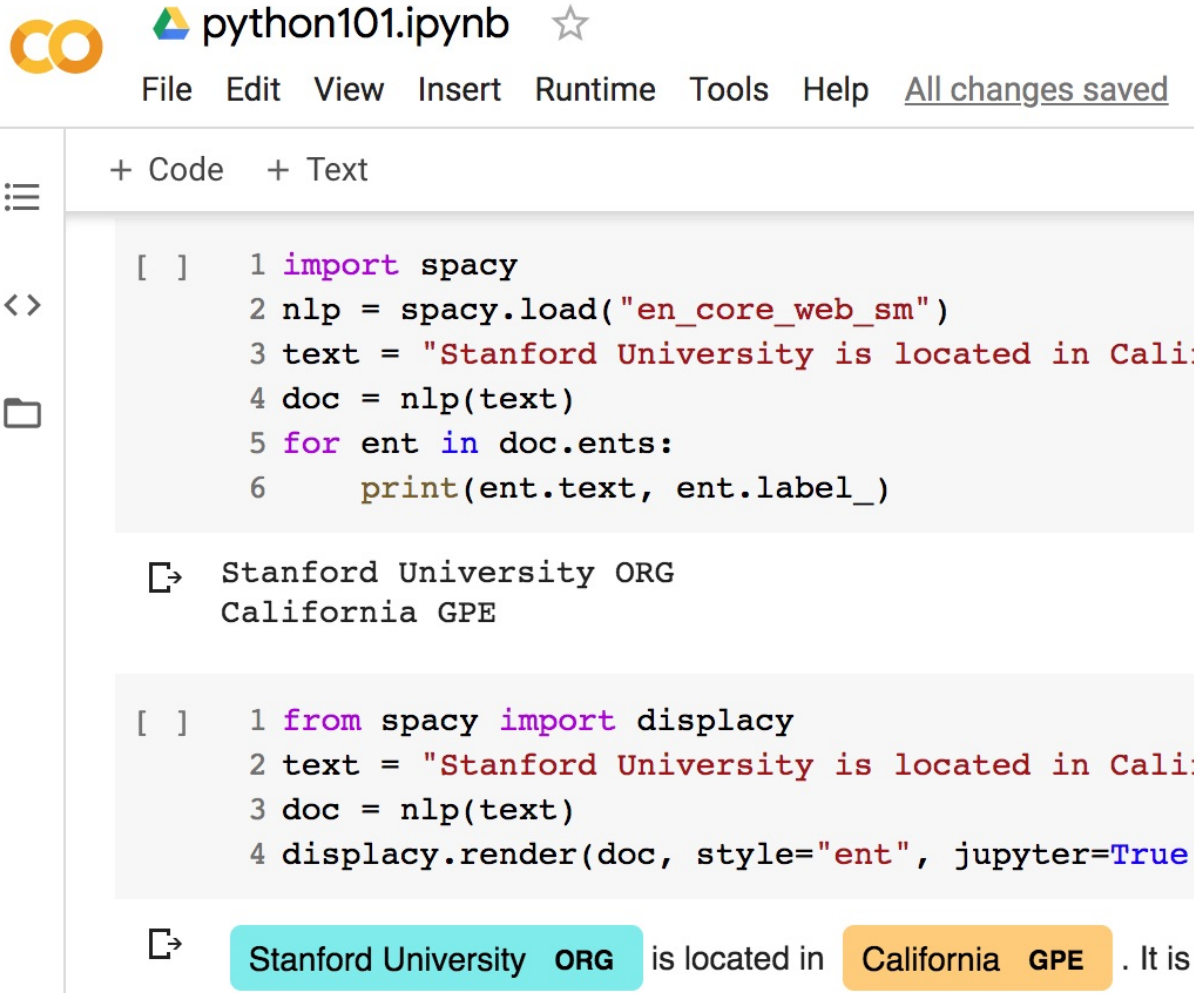

```
[ ]  1 import spacy
     2 nlp = spacy.load("en_core_web_sm")
     3 doc = nlp("Stanford University is located in California. It is a great university.")
     4 import pandas as pd
     5 cols = ("text", "lemma", "POS", "explain", "stopword")
     6 rows = []
     7 for t in doc:
     8     row = [t.text, t.lemma_, t.pos_, spacy.explain(t.pos_), t.is_stop]
     9     rows.append(row)
    10 df = pd.DataFrame(rows, columns=cols)
    11 df
```

	text	lemma	POS	explain	stopword
0	Stanford	Stanford	PROPN	proper noun	False
1	University	University	PROPN	proper noun	False
2	is	be	VERB	verb	True
3	located	locate	VERB	verb	False
4	in	in	ADP	adposition	True
5	California	California	PROPN	proper noun	False
6	.	.	PUNCT	punctuation	False
7	It	-PRON-	PRON	pronoun	True
8	is	be	VERB	verb	True
9	a	a	DET	determiner	True
10	great	great	ADJ	adjective	False
11	university	university	NOUN	noun	False
12	.	.	PUNCT	punctuation	False

https://tinyurl.com/aintpupython101

Python in Google Colab (Python101)

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



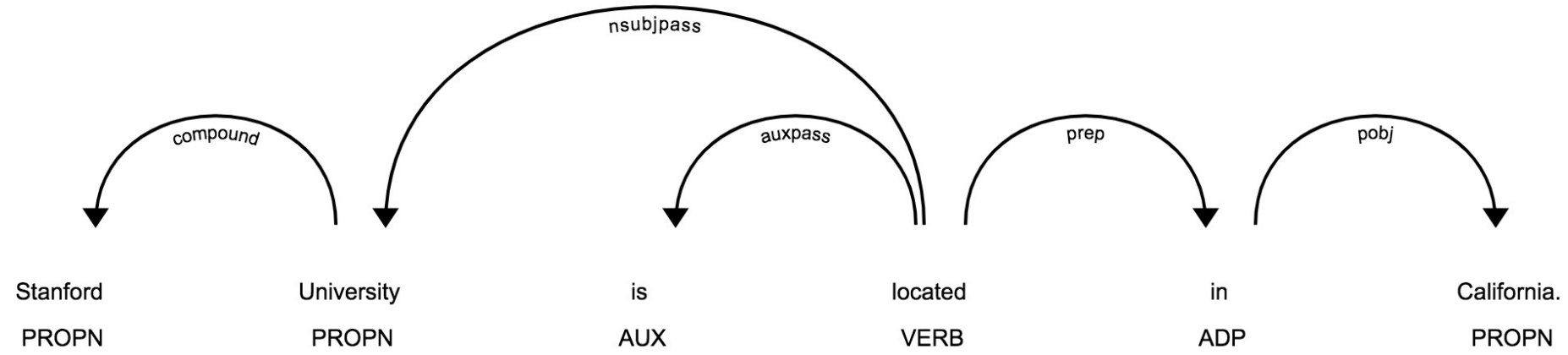
https://tinyurl.com/aintpupython101

Python in Google Colab (Python101)

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

```
1 from spacy import displacy
2 text = "Stanford University is located in California. It is a great university."
3 doc = nlp(text)
4 displacy.render(doc, style="ent", jupyter=True)
5 displacy.render(doc, style="dep", jupyter=True)
```

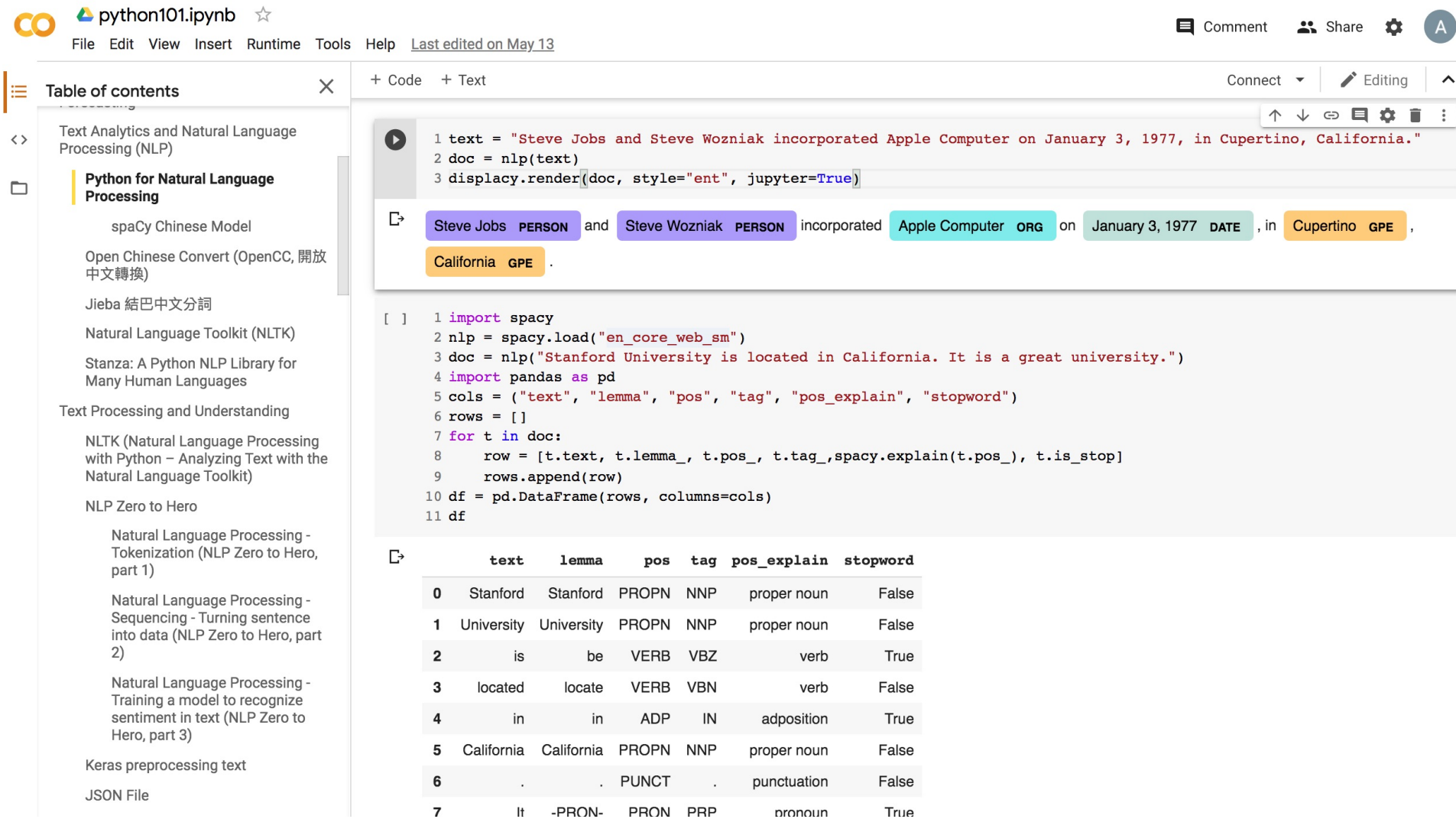
Stanford University ORG is located in California GPE . It is a great university.



https://tinyurl.com/aintpupython101

Python in Google Colab (Python101)

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



	text	lemma	pos	tag	pos_explain	stopword
0	Stanford	Stanford	PROPN	NNP	proper noun	False
1	University	University	PROPN	NNP	proper noun	False
2	is	be	VERB	VBZ	verb	True
3	located	locate	VERB	VBN	verb	False
4	in	in	ADP	IN	adposition	True
5	California	California	PROPN	NNP	proper noun	False
6	.	.	PUNCT	.	punctuation	False
7	It	-PRON-	PRON	PRP	pronoun	True

https://tinyurl.com/aintpupython101

MONPA 罔拍：

正體中文斷詞、詞性標註以及命名實體辨識的多任務模型

```
# MONPA 罔拍: 正體中文斷詞、詞性標註以及命名實體辨識的多任務模型
# Source: https://github.com/monpa-team/monpa
!pip install monpa
```

```
import monpa
sentence = "銀行產業正在改變，金融機構欲挖角科技人才"
words = monpa.cut(sentence)
print(sentence)
print(" ".join(words))
result_pseg = monpa.pseg(sentence)
for item in result_pseg:
    print(item)
```

```
銀行產業正在改變，金融機構欲挖角科技人才
銀行 產業 正在 改變 ， 金融 機構 欲 挖角 科技 人才
('銀行', 'ORG')
('產業', 'Na')
('正在', 'D')
('改變', 'VC')
('，', 'COMMACATEGORY')
('金融', 'Na')
('機構', 'Nc')
('欲', 'VK')
('挖角', 'VA')
('科技', 'Na')
('人才', 'Na')
```

https://tinyurl.com/aintpupython101

jieba
words = jieba.cut(sentence)

```
 1 import jieba
 2 import jieba.posseg as pseg
 3 sentence = "銀行產業正在改變，金融機構欲挖角科技人才"
 4 words = jieba.cut(sentence)
 5 print(sentence)
 6 print(" ".join(words))
 7 wordspos = pseg.cut(sentence)
 8 result = ''
 9 for word, pos in wordspos:
10     print(word + ' (' + pos + ')')
11     result = result + ' ' + word + '(' + pos + ')'
12 print(result.strip())
```

```
銀行產業正在改變，金融機構欲挖角科技人才
銀行 產業 正在 改變 ， 金融 機構 欲 挖角 科技人才
銀行 (n)
產業 (n)
正在 (t)
改變 (v)
， (x)
金融 (n)
機構 (n)
欲 (d)
挖角 (n)
科技人才 (n)
銀行(n) 產業(n) 正在(t) 改變(v) ，(x) 金融(n) 機構(n) 欲(d) 挖角(n) 科技人才(n)
```

https://tinyurl.com/aintpupython101

NLP Benchmark Datasets

Task	Dataset	Link
Machine Translation	WMT 2014 EN-DE WMT 2014 EN-FR	http://www-lium.univ-lemans.fr/~schwenk/cslm_joint_paper/
Text Summarization	CNN/DM Newsroom DUC Gigaword	https://cs.nyu.edu/~kcho/DMQA/ https://summari.es/ https://www-nlpir.nist.gov/projects/duc/data.html https://catalog.ldc.upenn.edu/LDC2012T21
Reading Comprehension Question Answering Question Generation	ARC CliCR CNN/DM NewsQA RACE SQuAD Story Cloze Test NarativeQA Quasar SearchQA	http://data.allenai.org/arc/ http://aclweb.org/anthology/N18-1140 https://cs.nyu.edu/~kcho/DMQA/ https://datasets.maluuba.com/NewsQA http://www.qizhexie.com/data/RACE_leaderboard https://rajpurkar.github.io/SQuAD-explorer/ http://aclweb.org/anthology/W17-0906.pdf https://github.com/deepmind/narrativeqa https://github.com/bdhingra/quasar https://github.com/nyu-dl/SearchQA
Semantic Parsing	AMR parsing ATIS (SQL Parsing) WikiSQL (SQL Parsing)	https://amr.isi.edu/index.html https://github.com/jkkummerfeld/text2sql-data/tree/master/data https://github.com/salesforce/WikiSQL
Sentiment Analysis	IMDB Reviews SST Yelp Reviews Subjectivity Dataset	http://ai.stanford.edu/~amaas/data/sentiment/ https://nlp.stanford.edu/sentiment/index.html https://www.yelp.com/dataset/challenge http://www.cs.cornell.edu/people/pabo/movie-review-data/
Text Classification	AG News DBpedia TREC 20 NewsGroup	http://www.di.unipi.it/~gulli/AG_corpus_of_news_articles.html https://wiki.dbpedia.org/Datasets https://trec.nist.gov/data.html http://qwone.com/~jason/20Newsgroups/
Natural Language Inference	SNLI Corpus MultiNLI SciTail	https://nlp.stanford.edu/projects/snli/ https://www.nyu.edu/projects/bowman/multinli/ http://data.allenai.org/scitail/
Semantic Role Labeling	Proposition Bank OneNotes	http://propbank.github.io/ https://catalog.ldc.upenn.edu/LDC2013T19

Source: Amirsina Torfi, Rouzbeh A. Shirvani, Yaser Keneshloo, Nader Tavvaf, and Edward A. Fox (2020). "Natural Language Processing Advancements By Deep Learning: A Survey." arXiv preprint arXiv:2003.01200.

Summary

- Text Analytics and Text Mining
- Natural Language Processing (NLP)
- Text Analytics with Python

References

- Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson.
- Dipanjan Sarkar (2019), Text Analytics with Python: A Practitioner's Guide to Natural Language Processing, Second Edition. APress.
- Benjamin Bengfort, Rebecca Bilbro, and Tony Ojeda (2018), Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning, O'Reilly.
- Charu C. Aggarwal (2018), Machine Learning for Text, Springer.
- Gabe Ignatow and Rada F. Mihalcea (2017), An Introduction to Text Mining: Research Design, Data Collection, and Analysis, SAGE Publications.
- Rajesh Arumugam (2018), Hands-On Natural Language Processing with Python: A practical guide to applying deep learning architectures to your NLP applications, Packt.
- Jake VanderPlas (2016), Python Data Science Handbook: Essential Tools for Working with Data, O'Reilly Media.
- 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.
- Christopher D. Manning and Hinrich Schütze (1999), Foundations of Statistical Natural Language Processing, The MIT Press.
- Bruce Croft, Donald Metzler, and Trevor Strohman (2008), Search Engines: Information Retrieval in Practice, Addison Wesley, http://www.search-engines-book.com/
- Steven Bird, Ewan Klein and Edward Loper (2009), Natural Language Processing with Python, O'Reilly Media, http://www.nltk.org/book/ , http://www.nltk.org/book_1ed/
- Bing Liu (2009), Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, Springer.
- The Super Duper NLP Repo, https://notebooks.quantumstat.com/
- Min-Yuh Day (2020), Python 101, https://tinyurl.com/aintpupython101