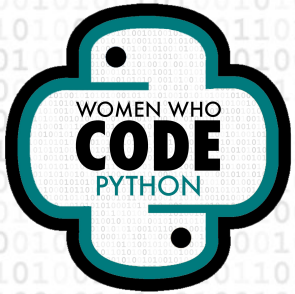
Welcome everyone!

- You can find these slides on GitHub here: https://github.com/WomenWhoCode/WWCodePython
- Please make sure your chat is set to “All panelists and attendees”.
- Some housekeeping rules:
  - Everyone will be muted throughout the webinar, but there will be opportunities for participation!
  - Please share your thoughts on the chat and/or ask questions in the Q&A.
  - The entire team is here today. Please reach out to us with any technical questions!



Discover NLP with Python

Welcome to Session #1!

THANK YOU TO OUR LEADERS!

WOMEN WHO CODE®

OUR MISSION

Inspiring women to excel in technology careers.

WOMEN WHO CODE

OUR VISION

A world where women are representative as technical executives, founders, VCs, board members and software engineers.

WOMEN WHO CODE

OUR TARGET

Engineers with two or more years of experience looking for support and resources to strengthen their influence and levelup in their careers.

WOMEN WHO CODE

CODE OF CONDUCT

WWCode is an inclusive community, dedicated to providing an empowering experience for everyone who participates in or supports our community, regardless of gender, gender identity and expression, sexual orientation, ability, physical appearance, body size, race, ethnicity, age, religion, socioeconomic status, caste, creed, political affiliation, or preferred programming language(s).

Our events are intended to inspire women to excel in technology careers, and anyone who is there for this purpose is welcome. We do not tolerate harassment of members in any form. Our **Code of Conduct** applies to all WWCode events and online communities.

Read the full version and access our incident report form at **womenwhocode.com/codeofconduct**

WOMEN WHO CODE

230,000
Members

70 networks in 20 countries
Members in 97+ countries
10K+ events
$1025 daily Conference tickets
$2M Scholarships
Access to jobs + resources
Infinite connections

WOMEN WHO CODE

OUR MOVEMENT

As the world changes, we can be a connecting force that creates a sense of belonging while the world is being asked to isolate.

WOMEN WHO CODE

NLP FOUNDATIONS

Fundamental concepts of NLP, general applications, and pre-processing

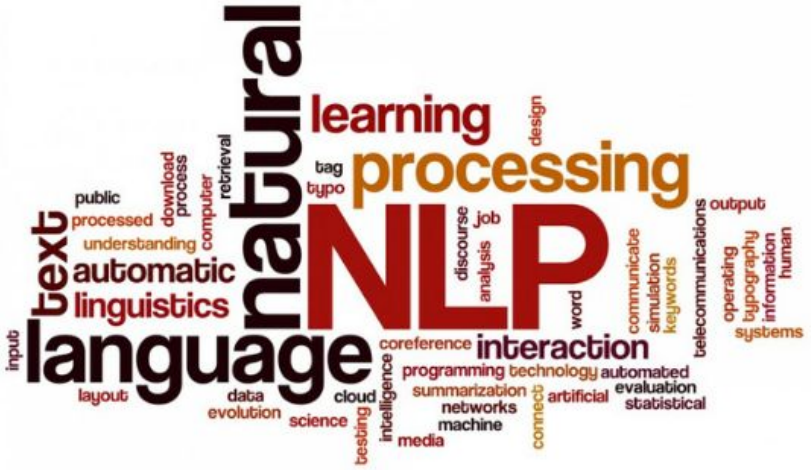
Source: https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html

AGENDA

1. What is NLP?
2. Where is NLP used?
3. Challenges in understanding natural language text
4. NLP Workflow Description
5. EDA & Pre-processing
6. Google Colab Coding!
7. Recap & Next Steps

1

What is NLP?



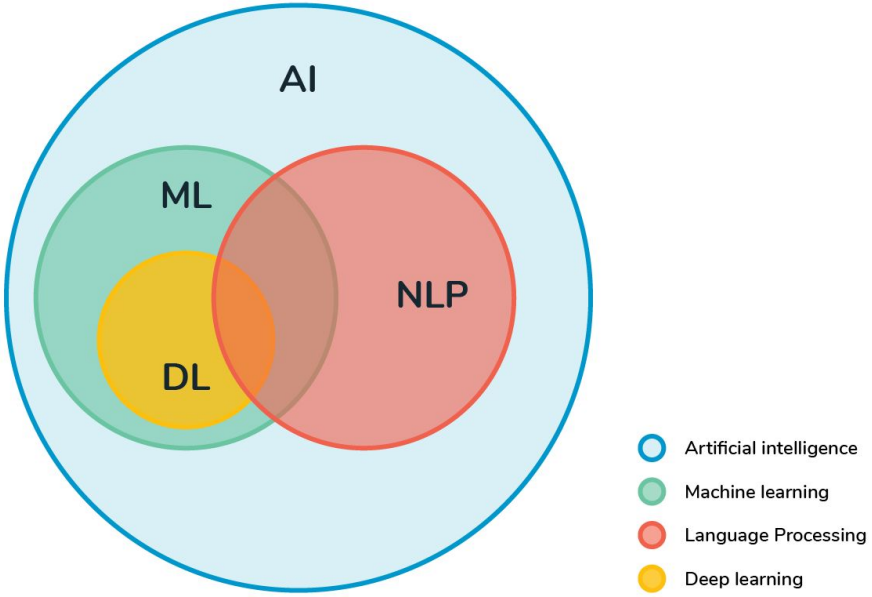
WOMEN WHO CODE PYTHON

NLP Overview

- Subfield of computer science and artificial intelligence
- Allows humans to bypass programming languages to speak to computers and use normal human speech instead
- Applications: text classification, machine translation, sentiment analysis
- Our devices nowadays: Apple's Siri, Amazon's Alexa, and Gmail's spam filter

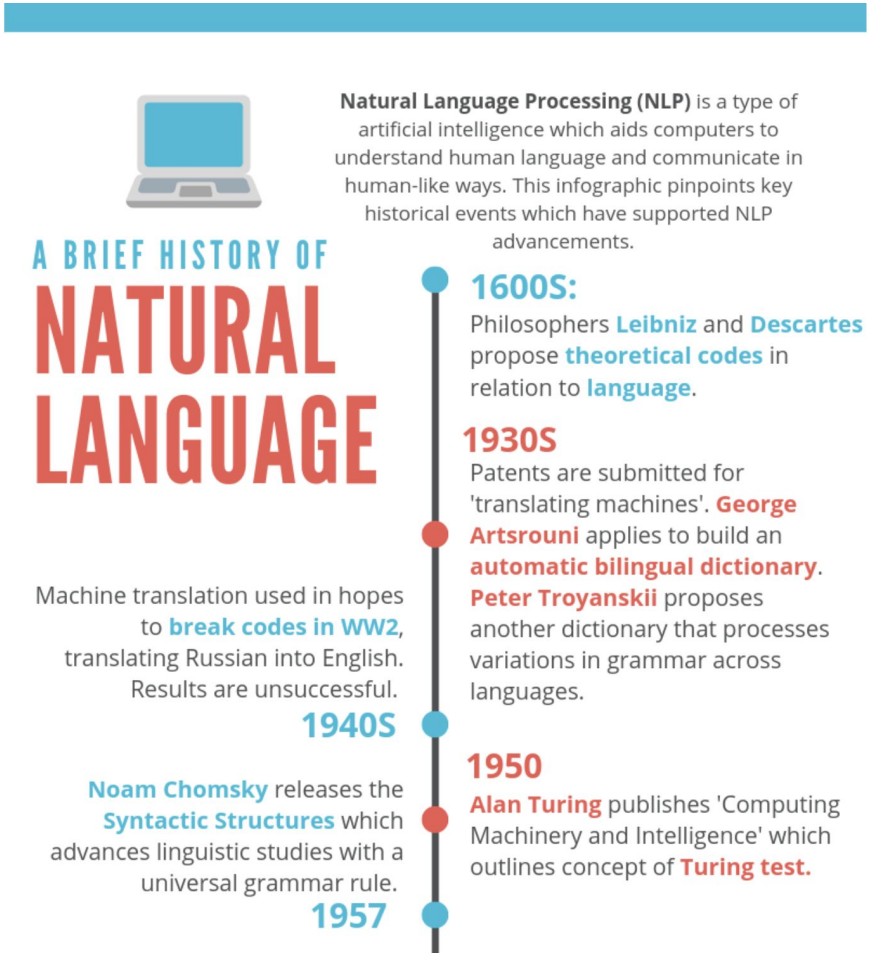
Source: raindrop74 / Shutterstock

Relationship with DS/AI/ML



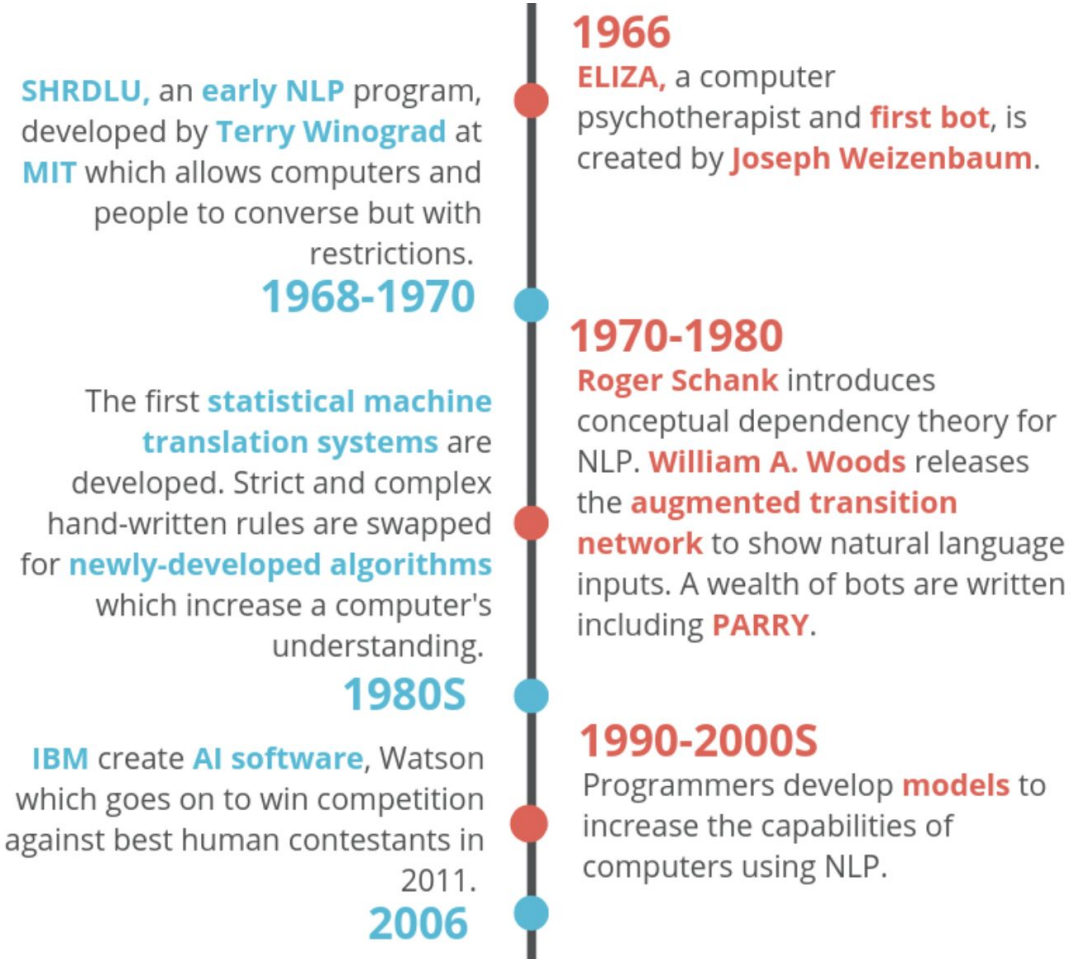
- Machine learning can help NLP powered systems adjust actions according to the historical context and patterns it picks up in a conversation
- NLP technology is human-like in the sense that more conversation can lead to better comprehension on repetitive tasks

Source: raindrop74 / Shutterstock



NLP Timeline

1600 - 1957
Part 1 of 3



NLP Timeline

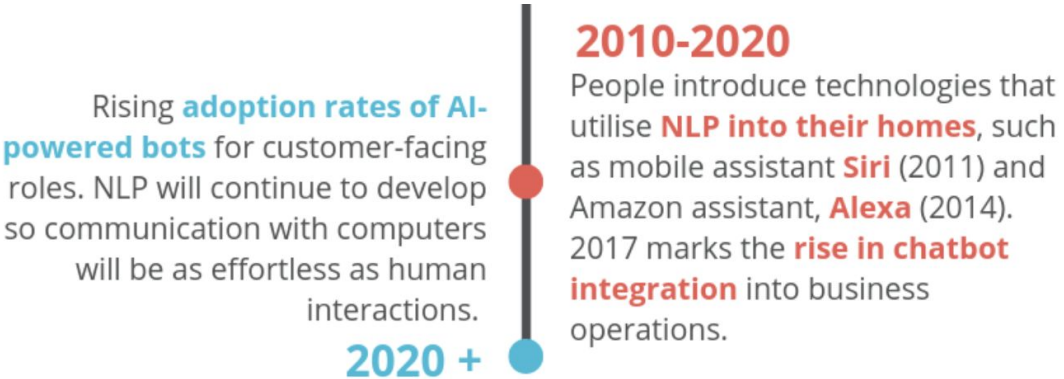
1966 - 2006

Part 2 of 3

NLP Timeline

2010 - now

Part 3 of 3



2

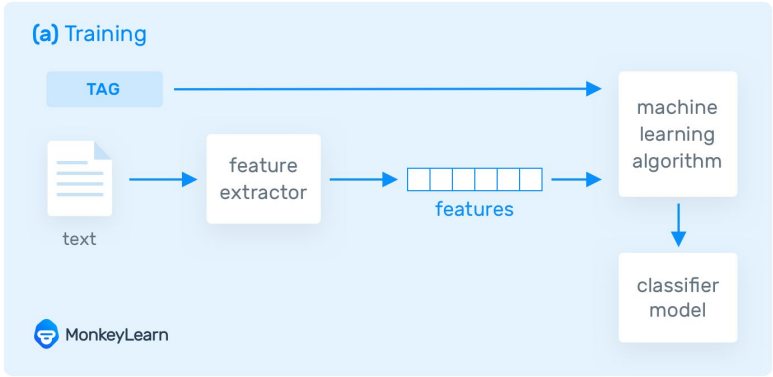
Where is NLP used?

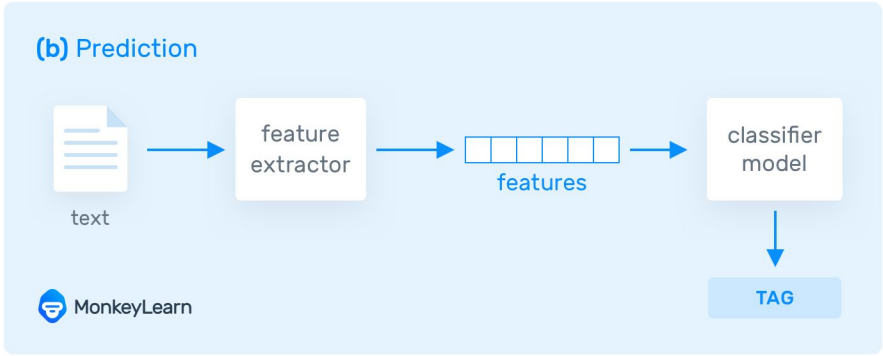
Machine Translation

- Task of automatically converting one natural language into another, preserving the meaning of the input text, and producing fluent text in the output language
- Challenging aspects:
  - the large variety of languages, alphabets and grammars
  - the task to translate a sequence of words/characters to another sequence is harder for a computer (than working with numbers alone)
  - there is no one correct answer

Text Classification

- Process of assigning tags or categories to text according to its content
- One of the fundamental tasks in NLP with broad applications
- Can be done in two different ways: manual and automatic classification





Source: https://monkeylearn.com/text-classification/

Sentiment Analysis

- Contextual mining of text which identifies and extracts subjective information in source material
- Focus on polarity (positive, negative, neutral), feelings and emotions (angry, happy, sad, etc), and intentions (e.g. interested v. not interested)

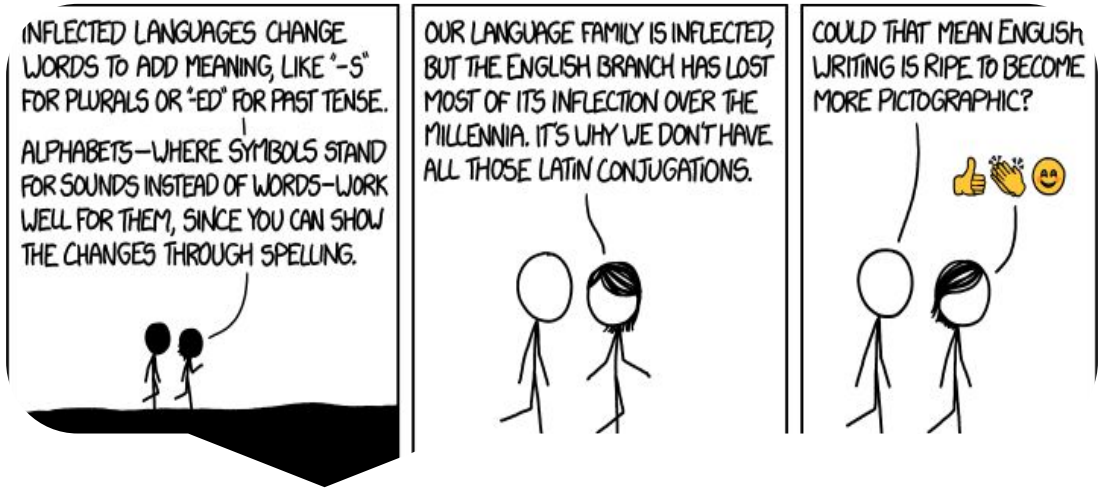
Source: https://monkeylearn.com/sentiment-analysis/

NLP in Our Everyday Lives

- Email assistant
- Ask Siri
- Answering questions
- 5 Amazing Applications:
  - Livox app
  - SignAll
  - Google Translate
  - Aircraft maintenance
  - Predictive police work

Source: Forbes.com

Language in Today's World



Source: TowardsDataScience

3

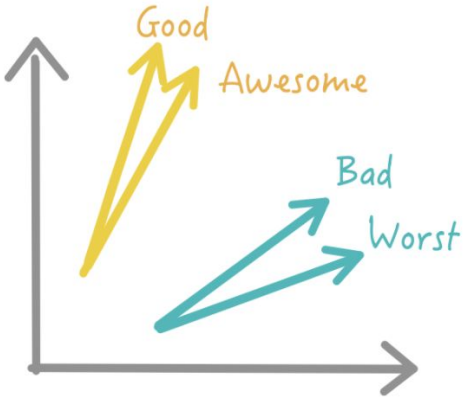
Challenges in understanding Natural Language Text

Ambiguity

- An intrinsic characteristic of human conversations, particularly challenging in NLU scenarios
- Different forms that are relevant to natural language and artificial intelligence systems
- In AI theory, the process of handling ambiguity is called disambiguation

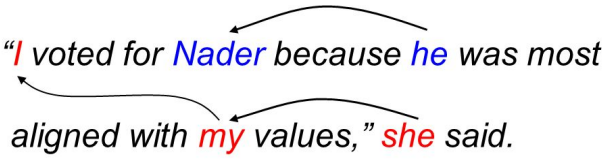
Synonymity

- We can express the same idea with different terms (which are also dependent on the specific context)
- Examples: “big” vs. “large”
- Necessary to incorporate the knowledge of synonyms and different ways to name the same object or phenomenon



Co-Reference

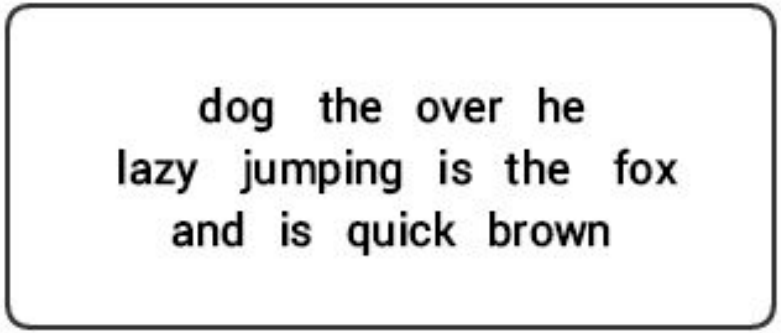
- Process of finding all expressions that refer to the same entity in a text
- Important step for a lot of higher-level NLP tasks that involve natural language understanding
- Notoriously difficult for NLP researchers, revived recently with the advent of cutting-edge techniques of deep learning and reinforcement learning.
- Coreference resolution may be instrumental in improving the performances of NLP neural architectures like RNN and LSTM



Source: https://nlp.stanford.edu/projects/coref.shtml

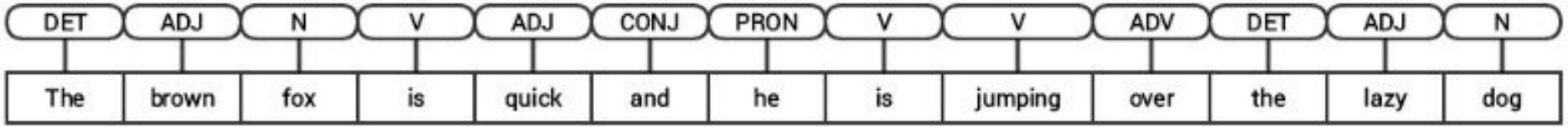
Syntactic Rules

- Knowledge about the structure and syntax of a language is helpful in many areas
- Typical parsing techniques for understanding text syntax include the following:
  - Parts of Speech (POS) Tagging
  - Shallow Parsing or Chunking
  - Constituency Parsing
  - Dependency Parsing



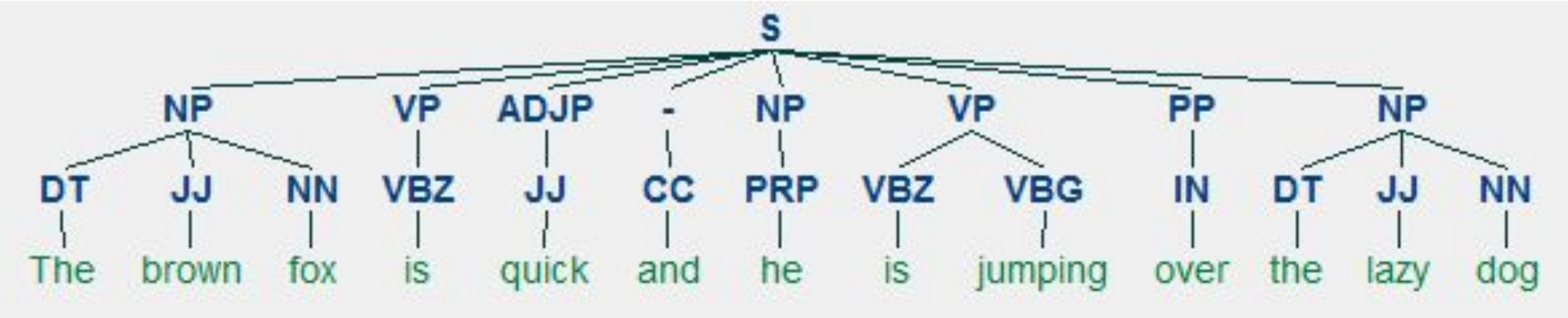
Source: https://www.kdnuggets.com/2018/08/understanding-language-syntax-and-structure-practitioners-guide-nlp-3.html

Syntactic Rules - Parts of Speech Tagging

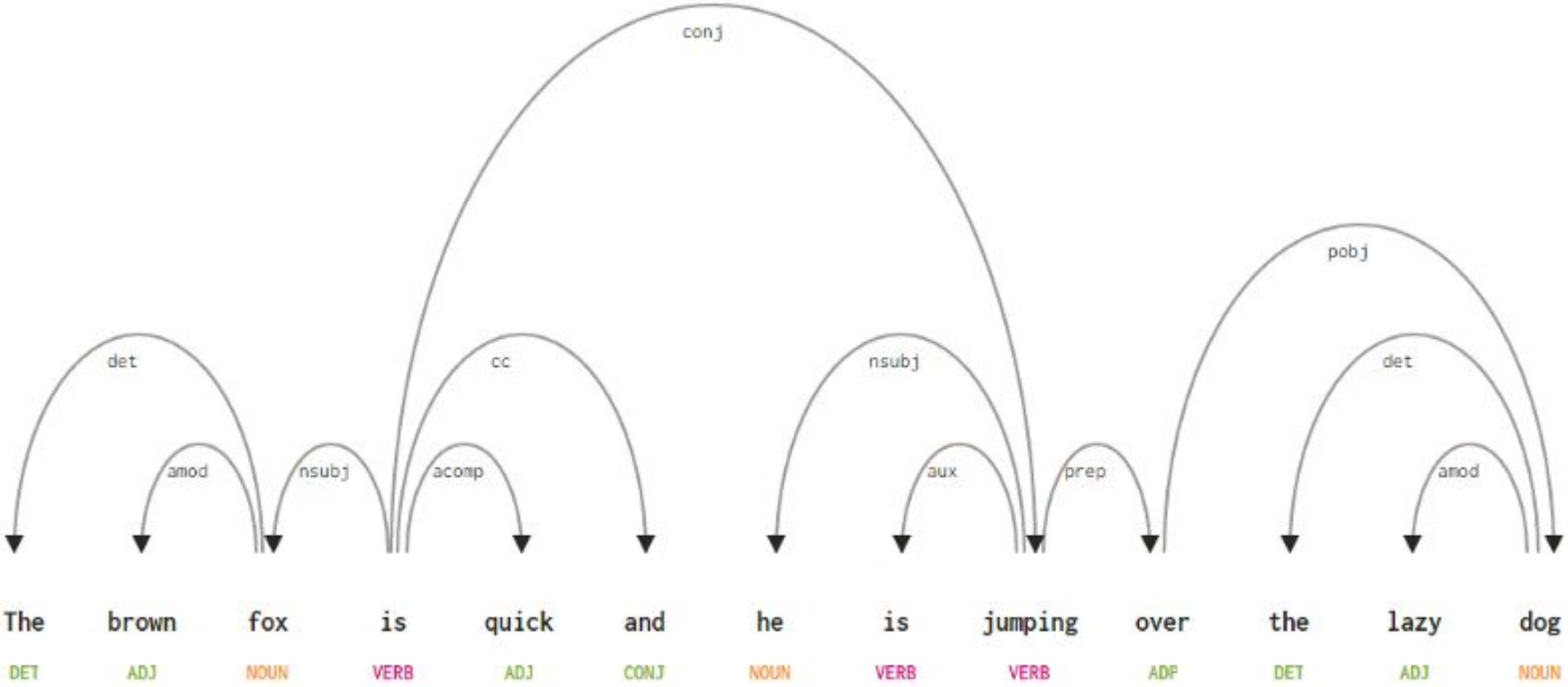


DET: Dependency tag
ADJ: Adjective
N: Noun
V: Verb
CONJ: Conjunction (coordinating)
PRON: Pronoun
ADV: Adverb

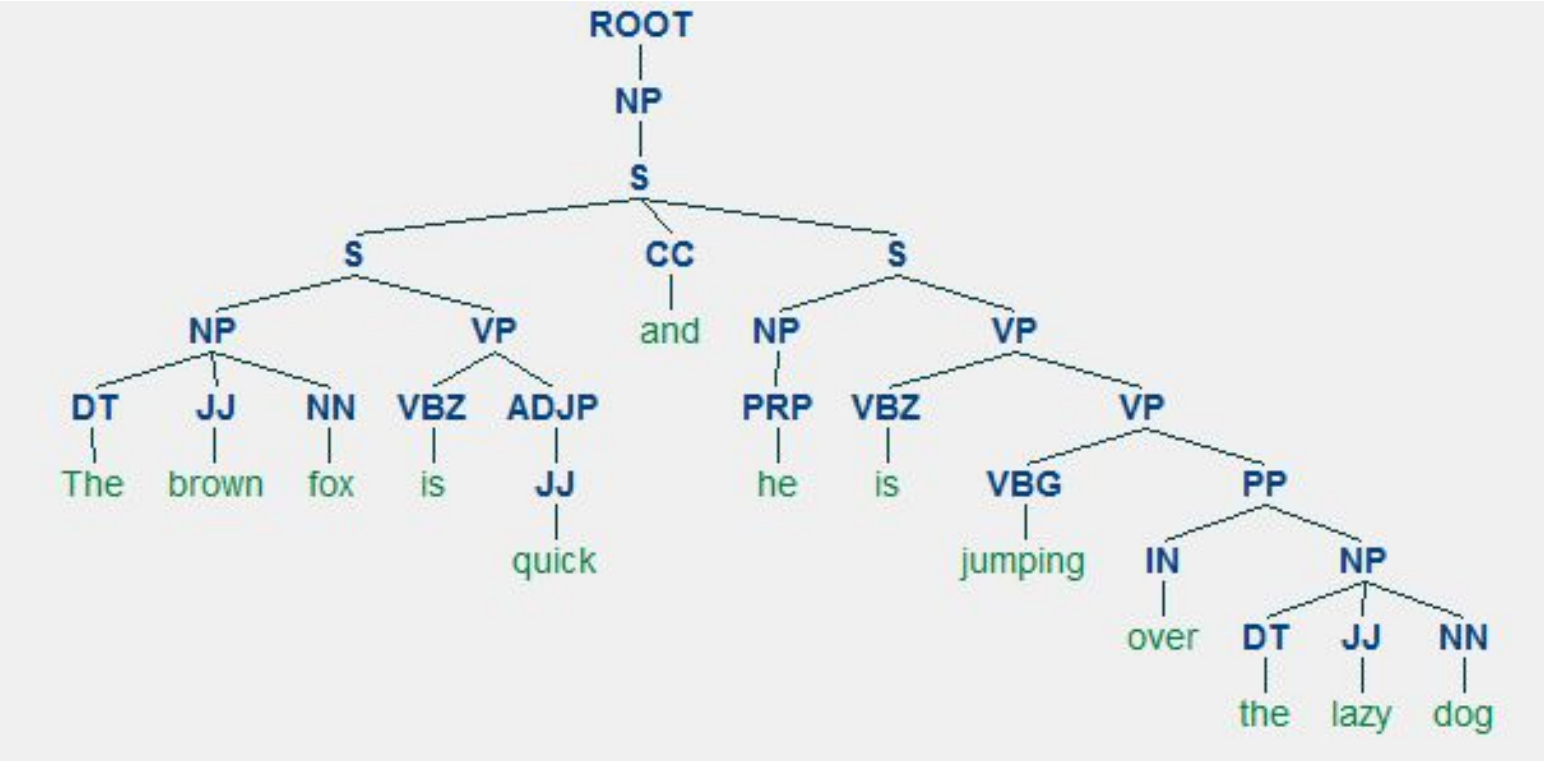
Syntactic Rules - Shallow Parsing/Chunking



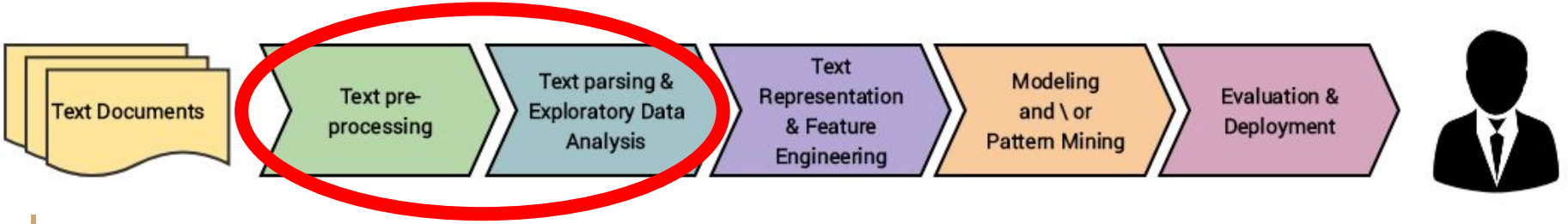
Syntactic Rules - Dependency Parsing



Syntactic Rules - Constituency Parsing



NLP Workflow



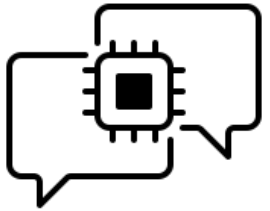
Pre-processing and EDA

- EDA steps to approach any NLP problems
  - Data describe, data info, basic visualization
- Pre-processing steps to approach any NLP problems with Colab Code
  - Step 1: Noise Cleaning - spacing, special characters
  - Step 2: Tokenization
  - Step 3: Spell Checking
  - Step 4: Contraction Mapping
  - Step 5: Stemming/Lemmatization
  - Step 6: ‘Stop Words’ Identification

Feature Extraction: Text to features

- The secret sauce to creating superior and better performing machine learning models
- Even more important for unstructured, textual data
- Main problem in working with NLP is that the algorithms cannot work on the raw text directly → we need some feature extraction techniques to convert text into a matrix (or vector) of features
- Some of the most popular methods of feature extraction are Bag-of-Words and TF-IDF (term frequency-inverse document frequency) Vectorizer.

Modeling



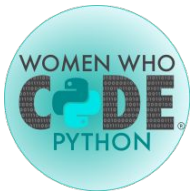
- System uses statistical methods to build its own ‘knowledge bank’, and is trained to make associations between a particular input and its corresponding output
- Also need to transform the text examples into something a machine can understand (vectors), a process known as feature extractor or text vectorization
- Once the texts have been transformed into vectors, they are fed to an algorithm together with their expected output (tags) to create a classification model
- This model can then discern which features best represent the texts, and make predictions on unseen data

Evaluation

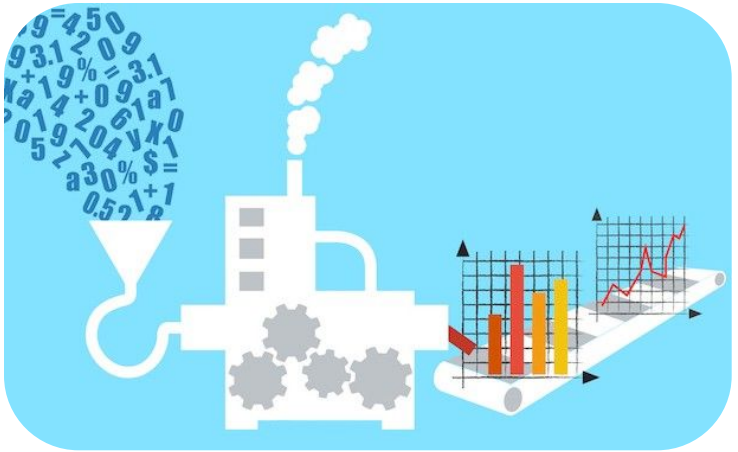
- Intrinsic vs. extrinsic:
  - Intrinsic: measure the performance of an NLP component on its defined subtask, usually against a defined standard in a reproducible laboratory setting
  - Extrinsic: focus on the component's contribution to the performance of a complete application, which often involves the participation of a human in the loop
- Automatic vs. manual
  - Automatic: evaluate an NLP system by comparing its output with the gold standard (or some desired) one, can be repeated as often as needed without much additional costs on the same input data. However, the definition of a gold standard is a complex task (objective evaluation)
  - Manual: performed by human judges, which are instructed to estimate the quality of a system, or most often of a sample of its output, based on a number of criteria; human judges can be considered as the reference for a number of language processing tasks (subjective evaluation)

5

Exploratory Data Analysis (EDA) & Pre-Processing Steps



Exploratory Data Analysis



EDA: Goal and Overview

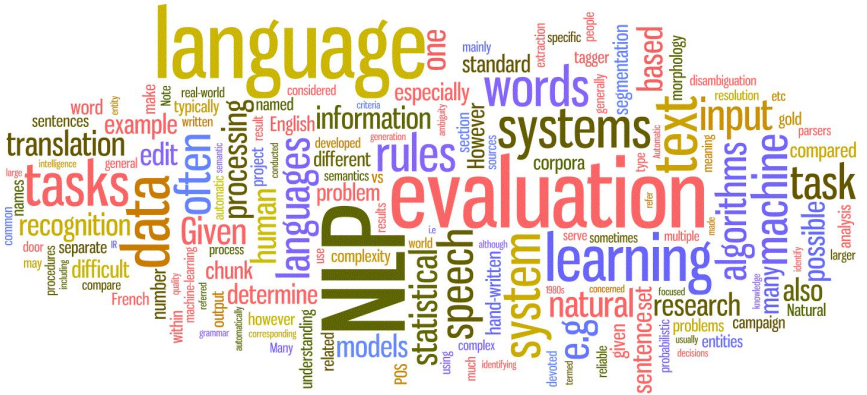
- Process of exploring data, generating insights, testing hypotheses, checking assumptions and revealing underlying hidden patterns in the data
- Through these goals, we can get a basic description of the data, visualize it, identify pattern in it, identify potential challenges of using the data, etc.

Dataset: SMS Spam/Ham

- SMS Spam Collection Data Set: public collection of SMS labeled messages that have been collected for mobile phone spam search.
- Refer to this website for more information on the dataset: https://archive.ics.uci.edu/ml/datasets/sms+spam+collection

Describe the Data

- A basic description of your data covers a broad spectrum
- You can interpret it as a quick and dirty way to get some information on your data, as a way of getting some simple, easy-to-understand information on your data, or to get a basic feel for your data
- Word clouds!



Source: https://blog.re-work.co/does-nlp-have-what-it-takes-to-fix-voice-text-recognition/

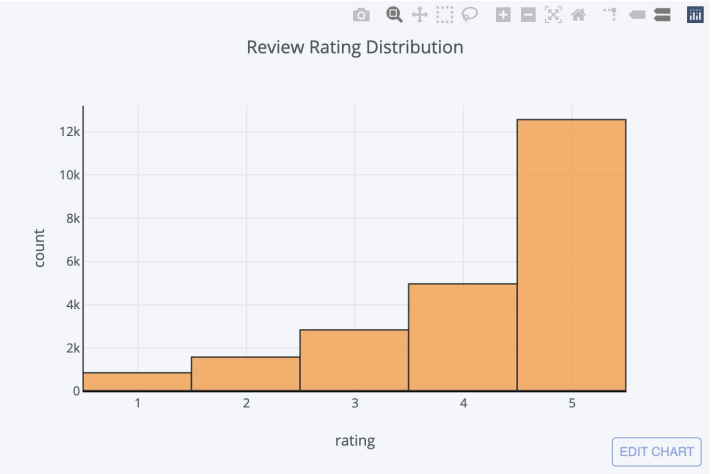
Data Information

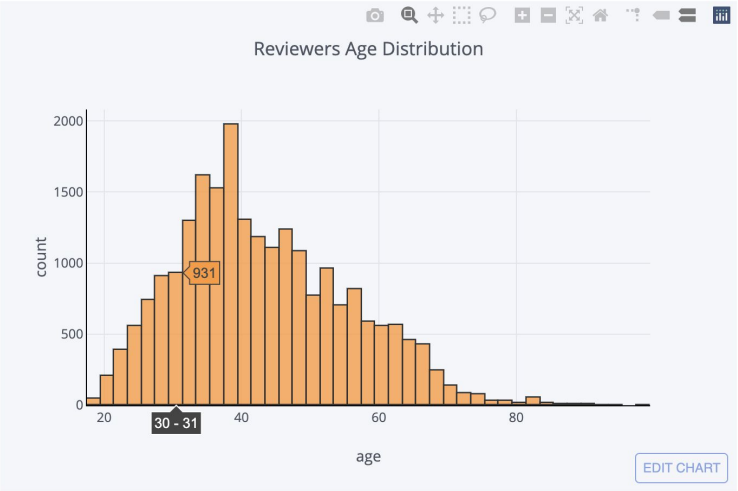
- Number of training vs. testing instances
- Missing data or missing labels of instances?
- Multi-dimensional data

	0	1
count	5572	5572
unique	2	5169
top	ham	Sorry, I'll call later
freq	4825	30

Basic Visualization

- Can help with identifying patterns in the data
- Python libraries Seaborn and Matplotlib are easy and quick ways to achieve this





Source: https://towardsdatascience.com/a-complete-exploratory-data-analysis-and-visualization-for-text-data-29fb1b96fb6a

Pre-Processing Steps

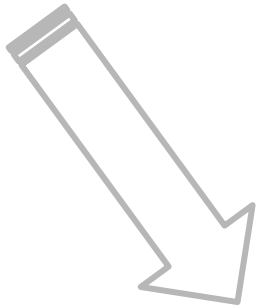
1. Noise Cleaning

	raw_word	cleaned_word
0	..trouble..	trouble
1	trouble<	trouble
2	trouble!	trouble
3	<a>trouble	trouble
4	1.trouble	trouble

2. Tokenization

from nltk.tokenize import sent_tokenize

text = "Hi, I would like to tokenize this sentence"



Output: ['Hi', 'I', 'would', 'like', 'to', 'tokenize', 'this', 'sentence']

3. Spell Checking

```
# find those words that may be misspelled
misspelled = spell.unknown(['something', 'is', 'hapenning', 'here'])
```

```
 happening
 {'penning', 'happening', 'henning'}
```

4. Contraction Mapping

17	5.0	It's not a startup anymore, but still ... learn so much from working w...	[it is, not, a, startup, anymore,, but, still, an, amazing, place, to, work!, You, learn, so, mu...
18	5.0	I learned a lot in this company about technology and navigation . This was a big opportunity for...	[I, learned, a, lot, in, this, company, about, technology, and, navigation, ., This, was, a, big...
19	4.0	Google is a great place to work. Respectful coworkers and management. The promotions can be ve...	[Google, is, a, great, place, to, work., Respectful, coworkers, and, management., The, promotion...

Source: https://towardsdatascience.com/preprocessing-text-data-using-python-576206753c28

5. Stemming/Lemmatization

	original_word	stemmed_words
0	connect	connect
1	connected	connect
2	connection	connect
3	connections	connect
4	connects	connect

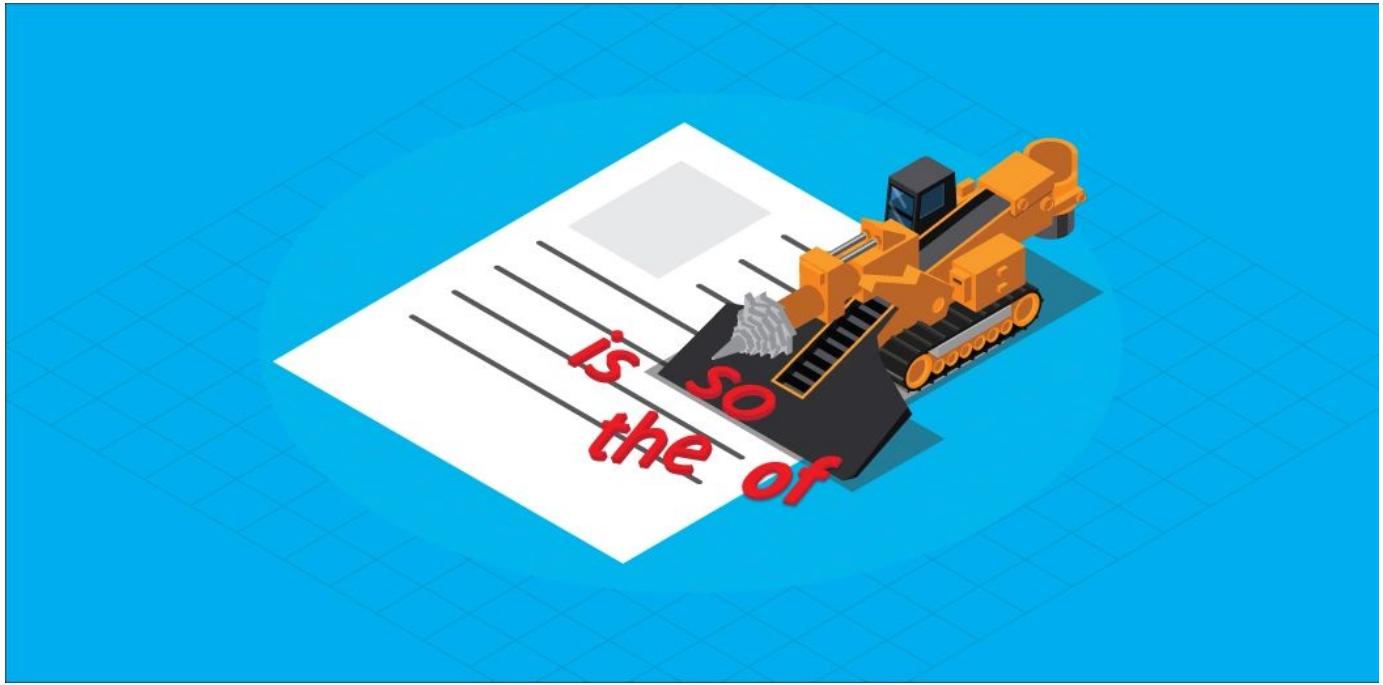
	original_word	stemmed_word
0	trouble	troubl
1	troubled	troubl
2	troubles	troubl
3	troublemsome	troublemsom

	original_word	lemmatized_word
0	trouble	trouble
1	troubling	trouble
2	troubled	trouble
3	troubles	trouble

	original_word	lemmatized_word
0	goose	goose
1	geese	goose

Source: https://www.quora.com/What-is-difference-between-stemming-and-lemmatization
Source: https://www.kdnuggets.com/2019/04/text-preprocessing-nlp-machine-learning.html

6. ‘Stop Words’ Identification



Source: https://www.analyticsvidhya.com/blog/2019/08/how-to-remove-stopwords-text-normalization-nltk-spacy-gensim-python/

Google Colab Coding!

https://bit.ly/3437M3I

7

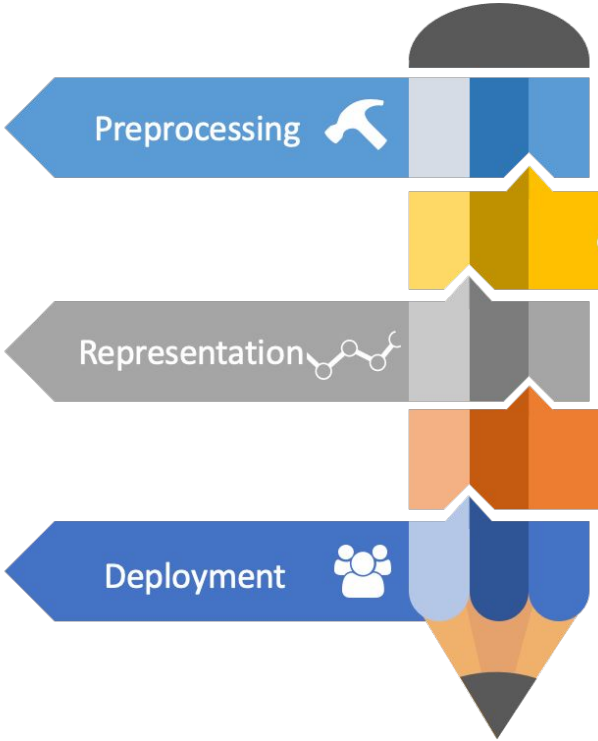
Recap & Next Steps

Recap

Text Preprocessing
Cleaning, Stemming, Contraction Removal, Special Char removal

Representation
Tokenization, text to sequence, padding sequences

Deployment
Prediction and model evaluation



EDA
Word2Vec Embedding enrichment, Misspell Removal, feature creation

Modelling
Models: Bi-LSTM/GRU, Attention, Capsule etc.

Source: https://mlwhiz.com/blog/2019/01/17/deeplearning_nlp_preprocess/

Next Session!

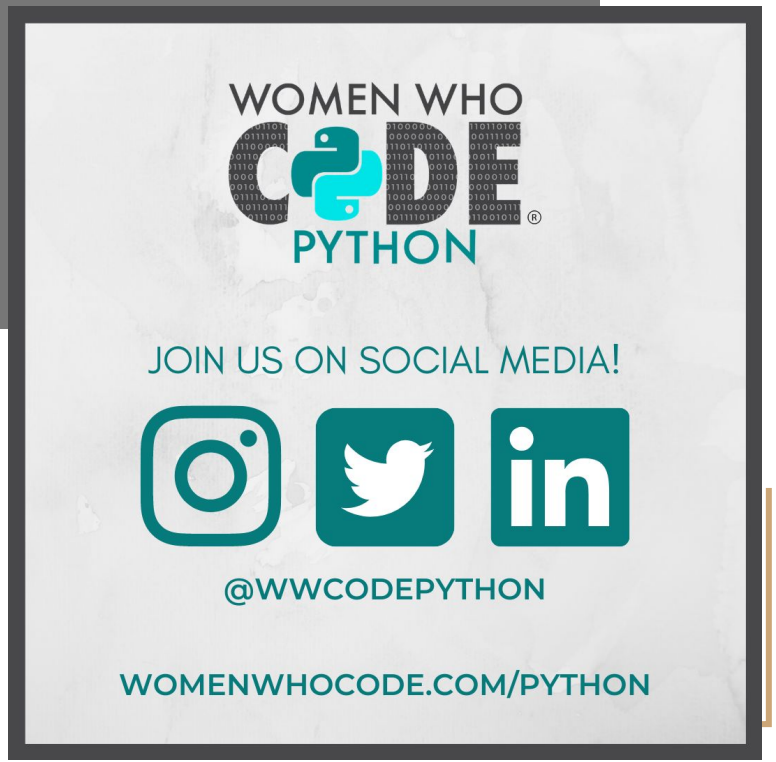
NLP and Deep Learning

October 25th, 8:00-9:30 pm (EDT)

Upcoming Events!

MON 28 SEP	✨ Open Source Contribution (feat. Hacktoberfest) ✨ *Featured* 12:00 PM – 1:00 PM (EDT) | Zoom	Register
WED 07 OCT	✨ Beginner Python Study Group ✨ Session 4: Data Types (Part 2) *Recurring* 8:00 PM – 9:30 PM (EDT) | Zoom	Register
WED 21 OCT	✨ Beginner Python Study Group ✨ Session 5: Programming Logic + Useful Functions *Recurring* 8:00 PM – 9:30 PM (EDT) | Zoom	Register
WED 04 NOV	✨ Beginner Python Study Group ✨ Session 6: Open Q&A/Review Session *Recurring* 8:00 PM – 9:30 PM (EST) | Zoom	Register
WED 18 NOV	✨ Beginner Python Study Group ✨ Session 7: Writing Your Own Python Module *Recurring* 8:00 PM – 9:30 PM (EST) | Zoom	Register

Stay Connected



Questions?

Join our Slack channel: #discover-nlp-with-python

Thanks everyone!