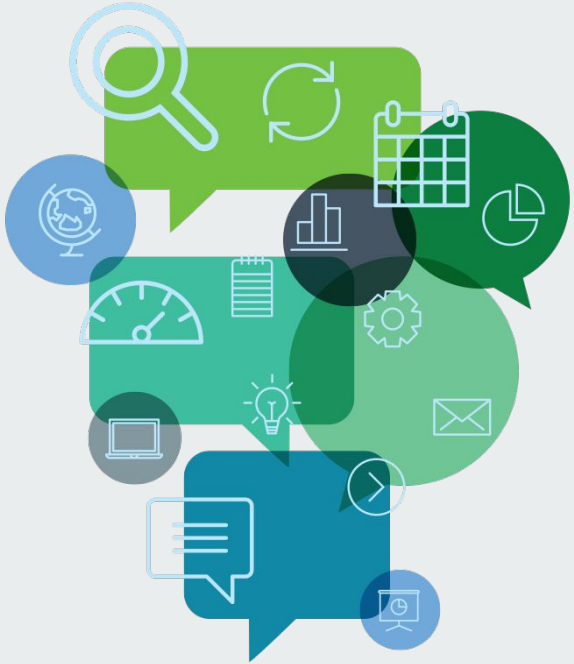
WOMEN WHO CODE DATA SCIENCE

NLP 101

Challenges of text data, workflow, EDA, Pre-processing, Concepts



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

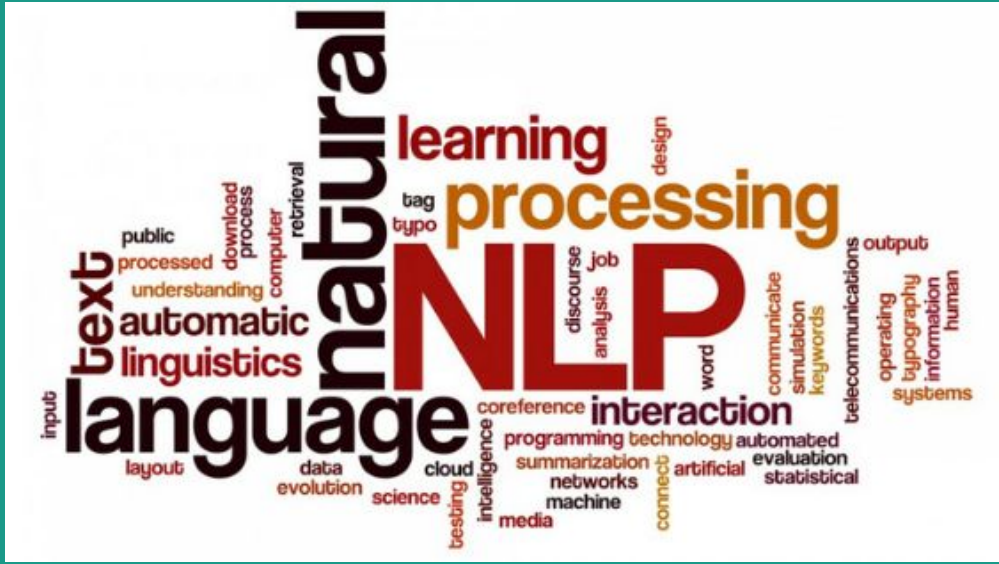
WOMEN WHO CODE DATA SCIENCE

Agenda

1. What is NLP?
2. Where is NLP used?
3. Challenges in understanding natural language text
4. NLP Workflow Description
5. Exploratory Data Analysis (EDA)
6. Pre-processing steps
7. Wrap-up and Next Steps
8. Implement with Google Colab

1

What is **NLP**?



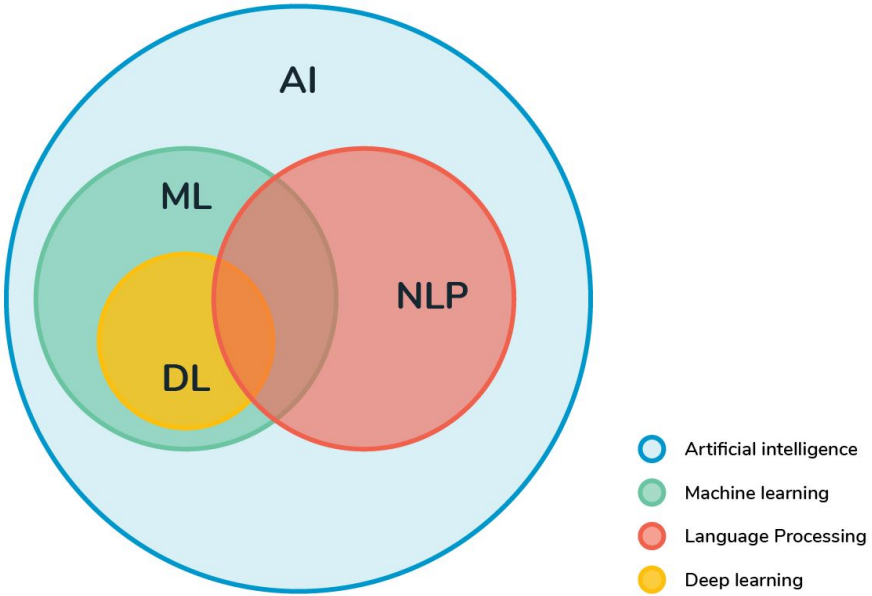
Source: https://dzone.com/articles/natural-language-processing-1

NLP Overview

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

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

Source: raindrop74 / Shutterstock

WOMEN WHO CODE DATA SCIENCE

NLP Timeline

1600 - 1957

Part 1 of 3

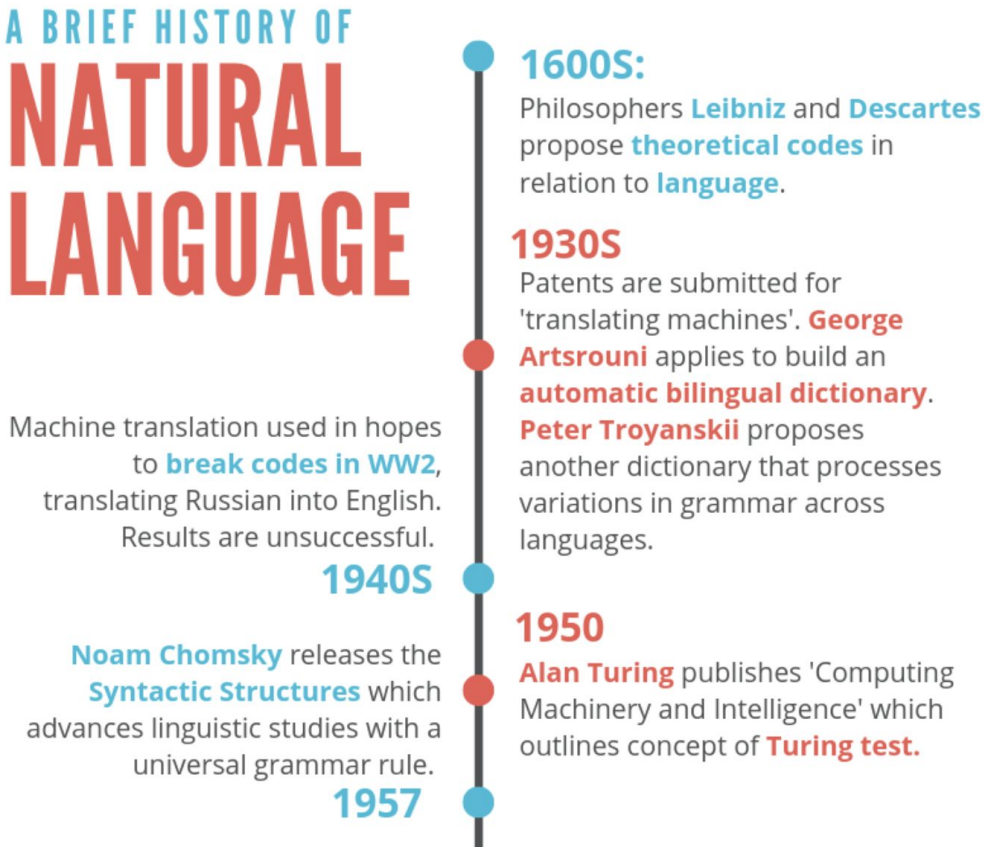


Image Source
https://www.hostcomm.co.uk/blog/2019/evolution-natural-language-infographic

WOMEN WHO CODE DATA SCIENCE

NLP Timeline

1966 - 2006

Part 2 of 3

1966

ELIZA, a computer psychotherapist and **first bot**, is created by **Joseph Weizenbaum**.

1968-1970

SHRDLU, an **early NLP** program, developed by **Terry Winograd** at **MIT** which allows computers and people to converse but with restrictions.

1970-1980

Roger Schank introduces conceptual dependency theory for NLP. **William A. Woods** releases the **augmented transition network** to show natural language inputs. A wealth of bots are written including **PARRY**.

1980S

The first **statistical machine translation systems** are developed. Strict and complex hand-written rules are swapped for **newly-developed algorithms** which increase a computer's understanding.

1990-2000S

Programmers develop **models** to increase the capabilities of computers using NLP.

2006

IBM create **AI software**, Watson which goes on to win competition against best human contestants in 2011.

Image Source
https://www.hostcomm.co.uk/blog/2019/evolution-natural-language-infographic

WOMEN WHO CODE DATA SCIENCE

NLP Timeline

2010 - now

Part 3 of 3

2010-2020

People introduce technologies that utilise **NLP into their homes**, such as mobile assistant **Siri** (2011) and Amazon assistant, **Alexa** (2014). 2017 marks the **rise in chatbot integration** into business operations.

2020 +

Rising **adoption rates of AI-powered bots** for customer-facing roles. NLP will continue to develop so communication with computers will be as effortless as human interactions.

Image Source
https://www.hostcomm.co.uk/blog/2019/evolution-natural-language-infographic

2

Where is NLP used?

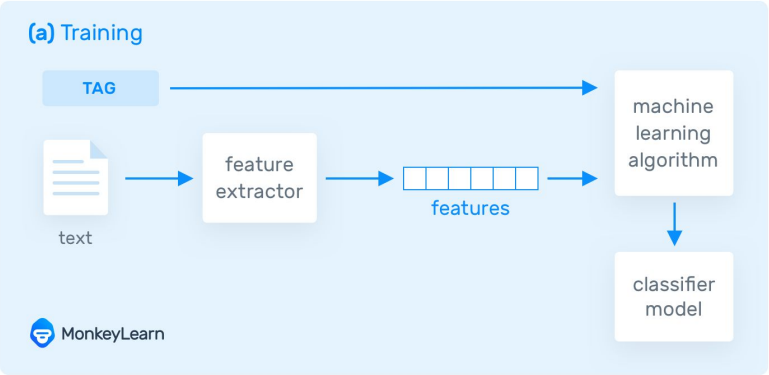
WOMEN WHO CODE DATA SCIENCE

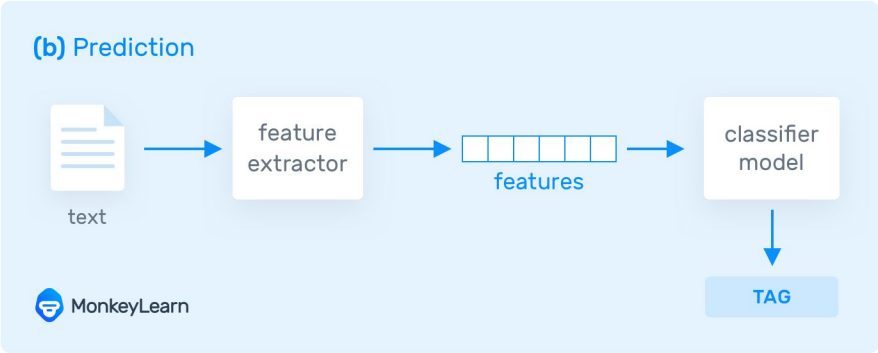
Machine Translation

- 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 to a sequence is harder for a computer than working with numbers only
  - 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).

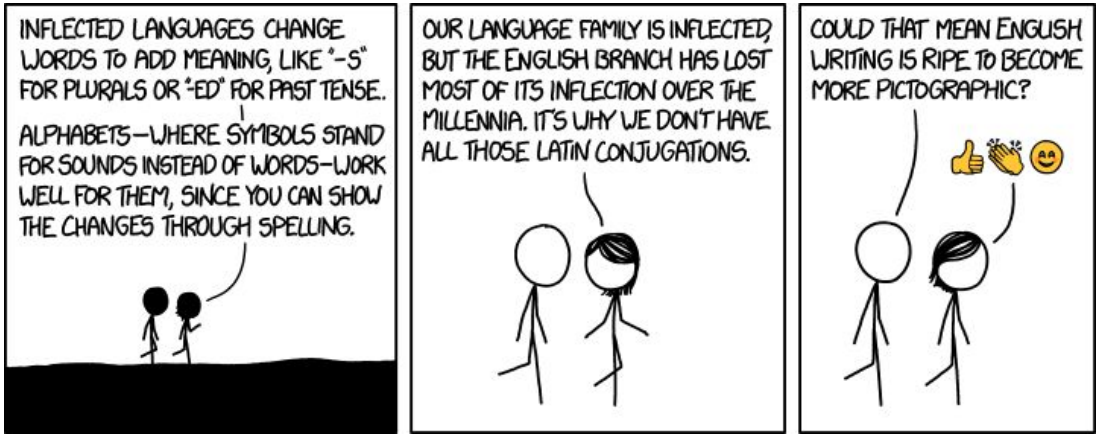
Source: https://monkeylearn.com/sentiment-analysis/

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

WOMEN WHO CODE DATA SCIENCE

3

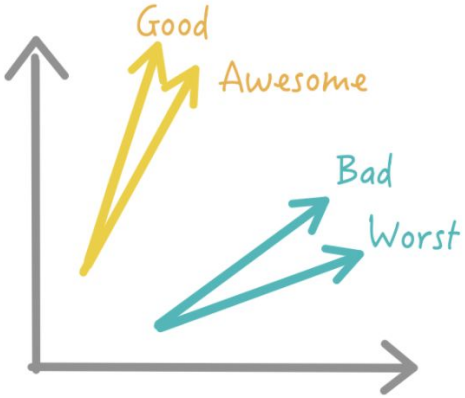
Challenges in understanding Natural Language Text

Ambiguity

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

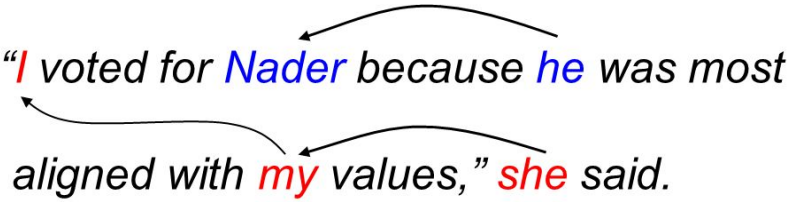
Synonymity

- 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

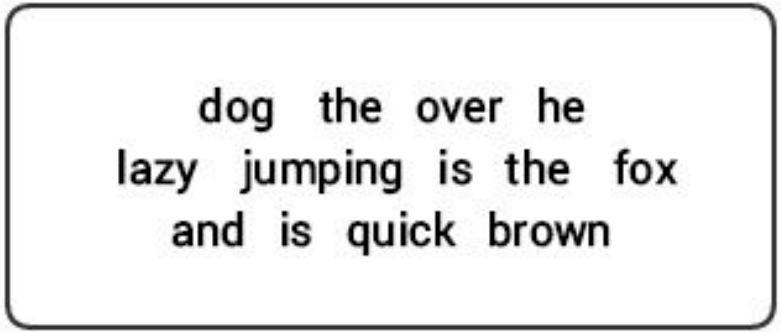
- Process of finding all expressions that refer to the same entity in a text
- Important step for a lot of higher-level NLP tasks that involve natural language understanding
- Notoriously difficult for NLP researchers, revived recently with the advent of cutting-edge techniques of deep learning and reinforcement 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

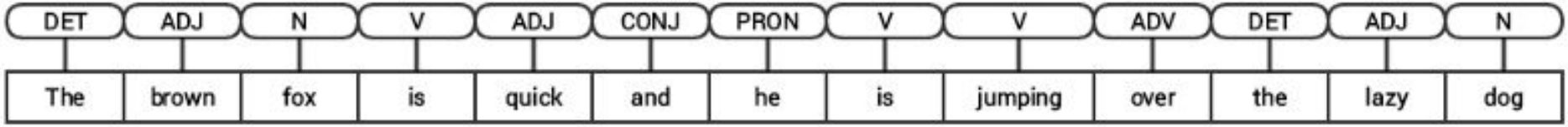
Syntactic Rules

- Knowledge about the structure and syntax of 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



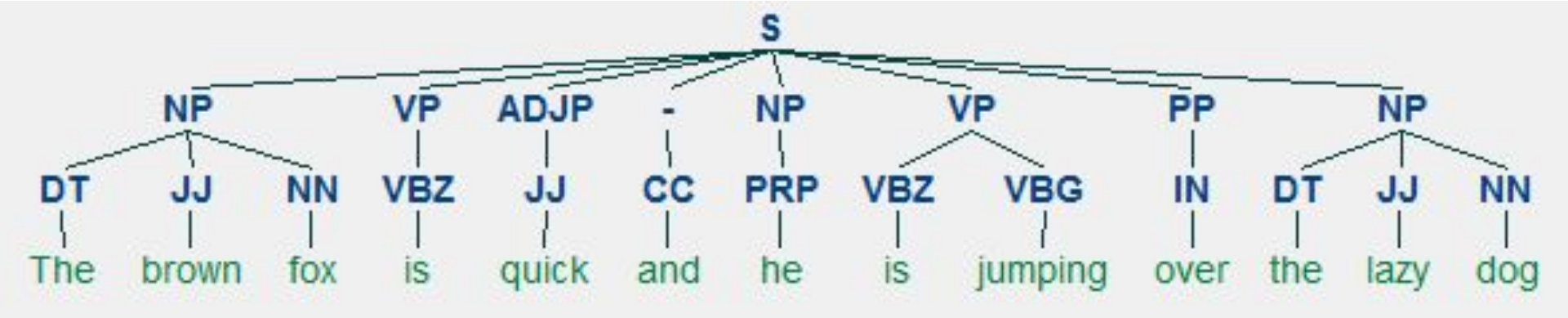
Source: https://www.kdnuggets.com/2018/08/understanding-language-syntax-and-structure-practitioners-guide-nlp-3.html

1. Parts of Speech Tagging

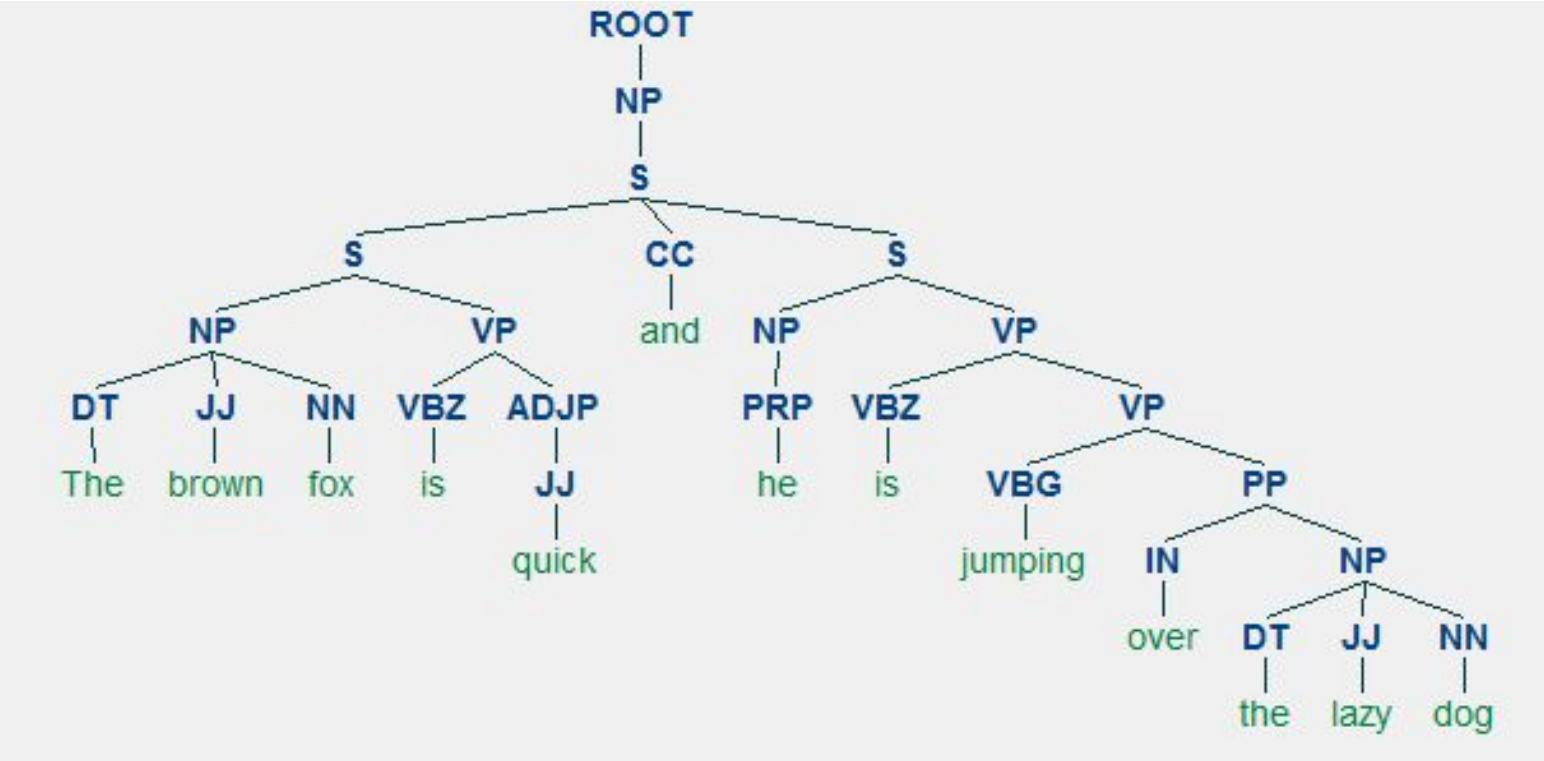


DET: Dependency tag
ADJ: Adjective
N: Noun
V: Verb
CONJ: Conjunction (coordinating)
PRON: Pronoun
ADV: Adverb

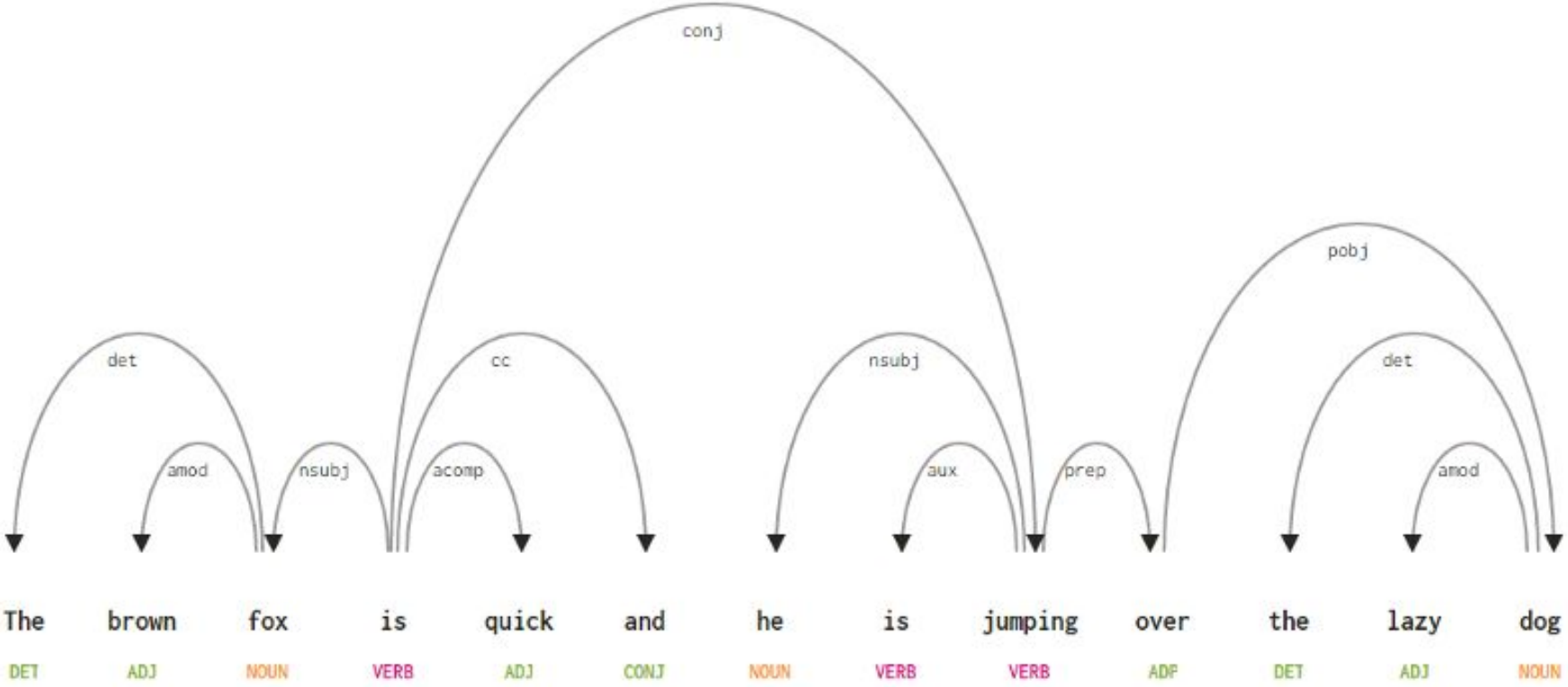
2. Shallow Parsing / Chunking



3. Constituency Parsing

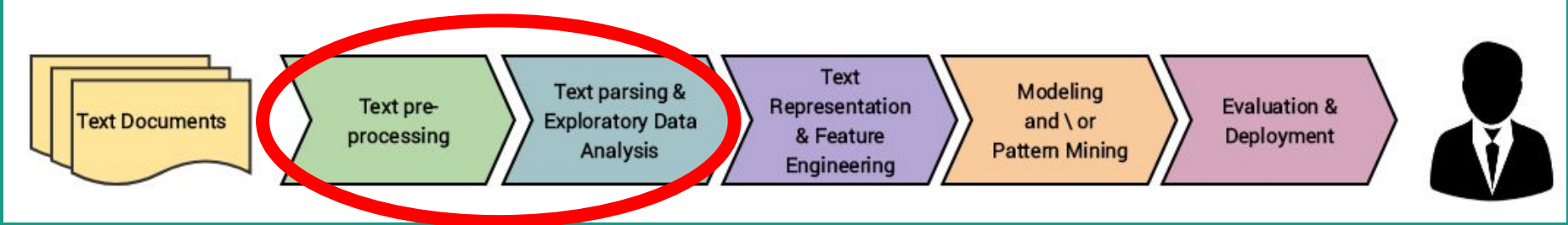


4. Dependency Parsing



4

NLP Workflow



Source: https://towardsdatascience.com/a-practitioners-guide-to-natural-language-processing-part-i-processing-understanding-text-9f4abfd13e72

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

WOMEN WHO CODE DATA SCIENCE

5

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/

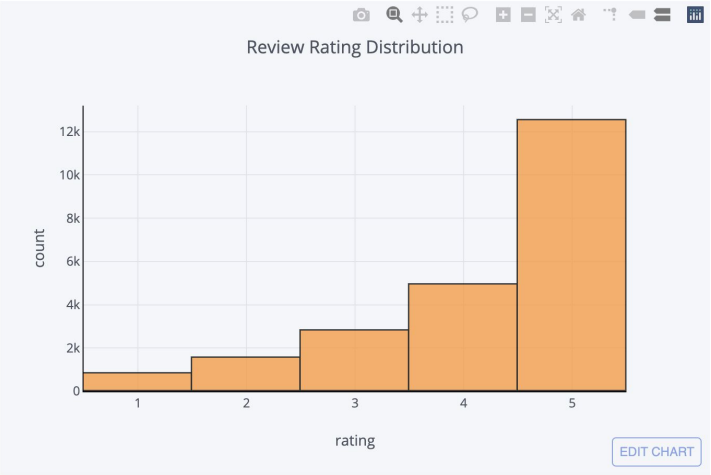
Data Information

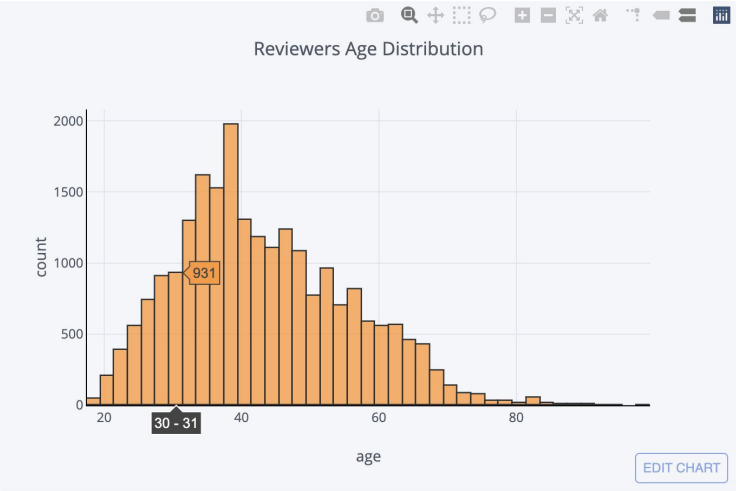
- Number of training vs. testing instances
- Missing data or missing labels of instances?
- Multi-dimensional data

	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

6

Pre-processing

WOMEN WHO CODE DATA SCIENCE

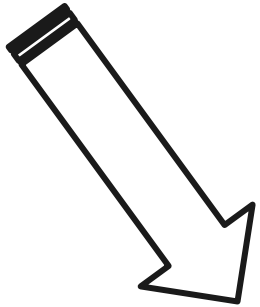
1. Noise Cleaning

	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 an amazing place, 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

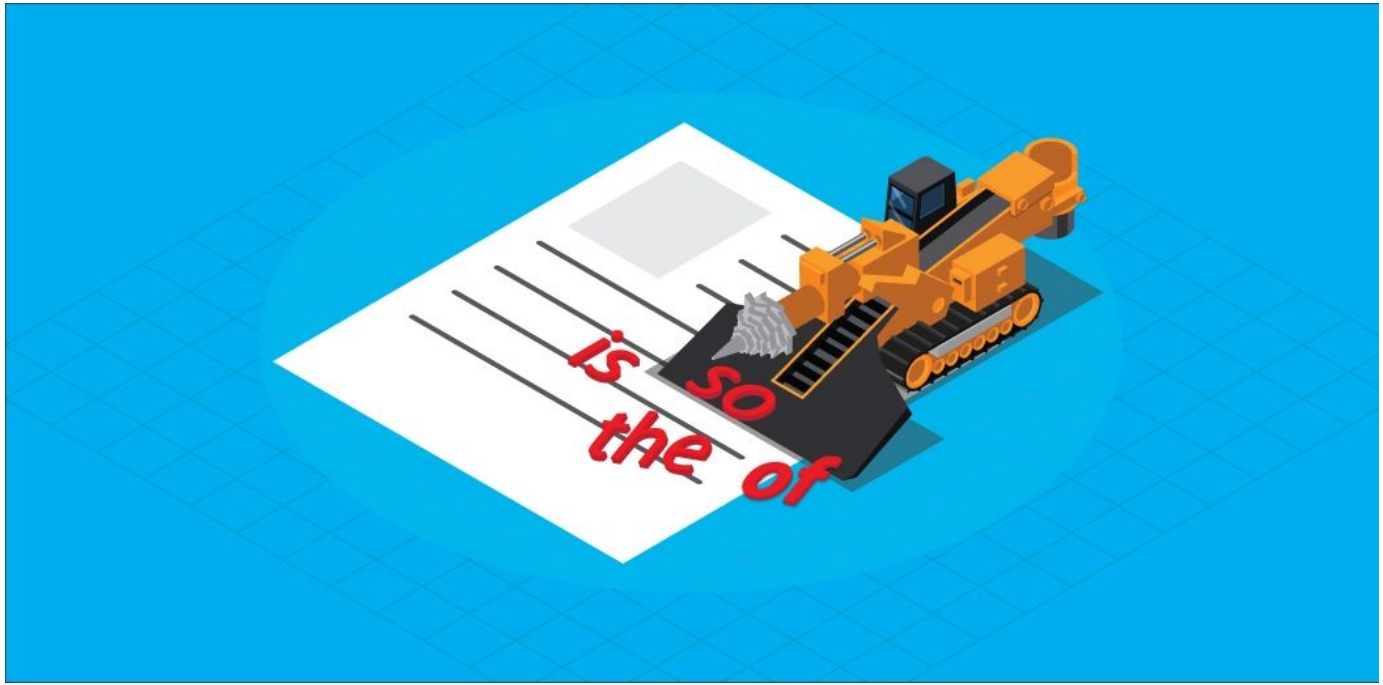
	original_word	stemmed_word
0	trouble	troubl
1	troubled	troubl
2	troubles	troubl
3	troublemsome	troublemsom

	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/

WOMEN WHO CODE DATA SCIENCE

7

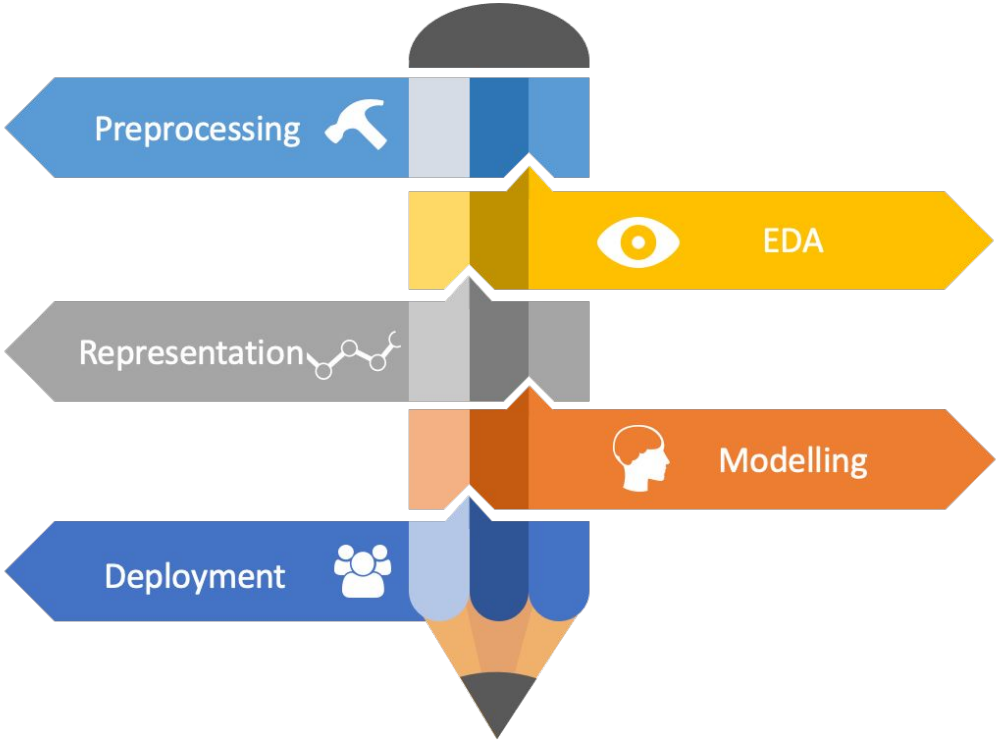
Theory Wrap-up & Next Steps

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/

8

Google Colab Project

https://bit.ly/introtonlp-week1-notebook

WOMEN WHO CODE DATA SCIENCE

Homework #1

- **(DataCamp)** Python Exploratory Data Analysis Tutorial
- **(TowardsDataScience)** A Complete Exploratory Data Analysis and Visualization for Text Data
- (**TowardsDataScience)** NLP Part 2| Pre-Processing Text Data Using Python
- **(AnalyticsVidhya)** A Beginner's Guide to Exploratory Data Analysis (EDA) on Text Data (Amazon Case Study)

WOMEN WHO CODE DATA SCIENCE

Homework #2

Additional Datasets

- **General Datasets**
  - NLTK Corpora
  - Google Blogger Corpus
  - Recommender Systems
- **Sentiment Analysis Dataset**
  - Yelp Reviews
- **Text Classification Dataset**
  - Jeopardy!

WOMEN WHO CODE DATA SCIENCE

See you next week!

Questions?

Join us on Slack and post your questions to the #help-me channel