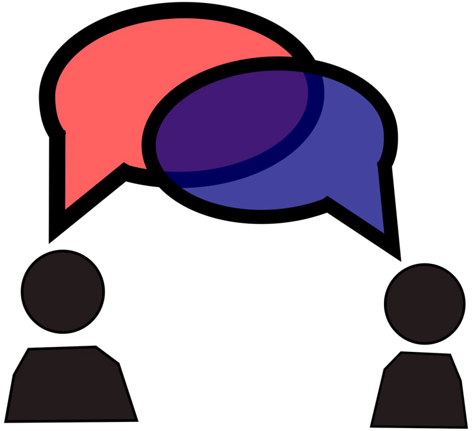
Interactive

- What is 1 thing you want to learn today?
- Where are you joining us from?



Introduction to NLP

Anya Mityushina - 03/23/2023

Agenda

1 **TL;DR**

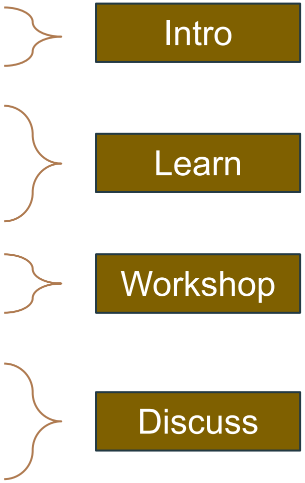
2 What is **NLP?**

3 **NLP** Techniques

4 NLP **Code** Examples

5 **Tips** and **outlook**

6 **Open** Mic



Intro

Learn

Workshop

Discuss

TL;DR

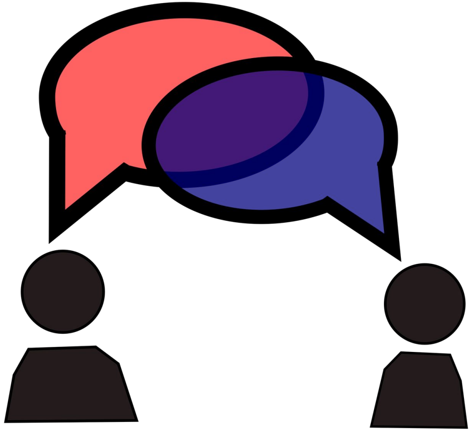
- NLP stands for **Natural Language Processing,** a branch of computer science and artificial intelligence that focuses on the interactions between humans and computers using natural language.
- **NLP** is a **suite of techniques** which help us **structure text data** understanding, and generating human language, including text and speech, using computer algorithms and statistical models.
- Some common **applications of NLP** include sentiment analysis, language translation, speech recognition, text summarization, and chatbot development.
- With any business question, start with your **use case**, **define** your problem space including your constraints, document them well and watch for changes, **choose tool LAST**
- **Data collection** matters!
- We use it to **bridge spoken language** into **machine language** to get insights
- The future is **bright** where everyone has access to information even faster, but WE all have to be active to shape it. It will not get more **fair or equitable** without our feedback and active work

Do you use it already?

- Virtual **assistants** (e.g., Siri, Cortana)
- Email **spam filter** (e.g., Spam, important)
- Customer service **chat**
- Auto suggestions in search and email (e.g., Hello and thank **you**)
- Spelling auto correct suggestions (e.g., tis is spelld corrctly)

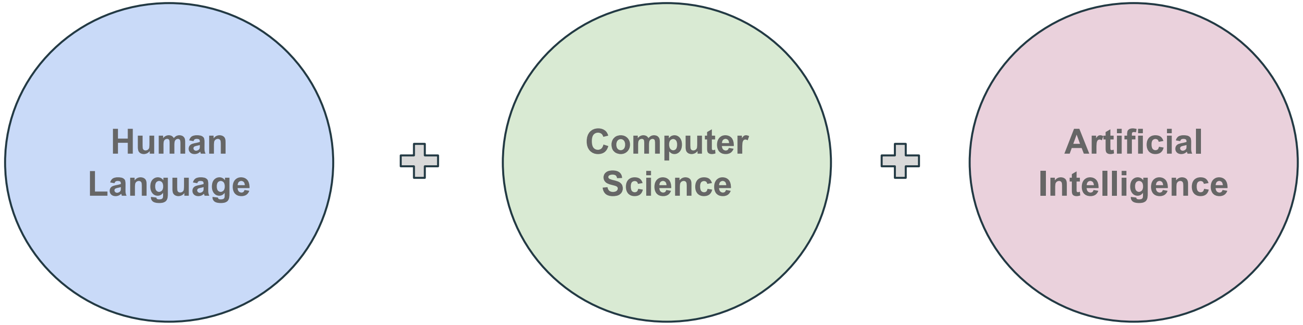
Interactive

- How do you figure out when someone is **happy, confused, upset**?
- How do you figure out what to get **someone for their birthday**?
- Ever been in a conversation where you didn’t know what someone **meant**?
- How do you fi**gure out if you should apply** feedback?



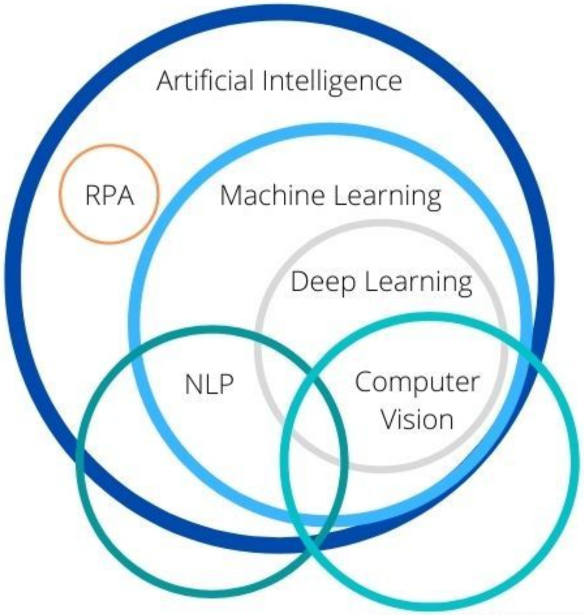
What is NLP: Definition

NLP stands for **Natural Language Processing,** a branch of computer science and artificial intelligence that focuses on the interactions between humans and computers using natural language.



What is NLP: AI and Use

- **NLP** is a **suite of techniques** which help us **structure text data** understanding, and generating human language, including text and speech, using computer algorithms and statistical models.
- Some common **applications of NLP** include sentiment analysis, language translation, speech recognition, text summarization, and chatbot development.



artificialintelligence.oodles.io

What is NLP: Before you start

- We generally think about **written** and **spoken languages** as a form of communication
- Albert Mehrabian, a researcher of body language, found that communication is **55% nonverbal, 38% vocal, and 7% words** only [link]

- Ways we get text data
    - [Video]
    - [Audio]
    - [Image]
    - [Text]
- It's **important** to know **where the data** is coming from, how it was **processed,** and how it's **stored**

Interactive

What is this?

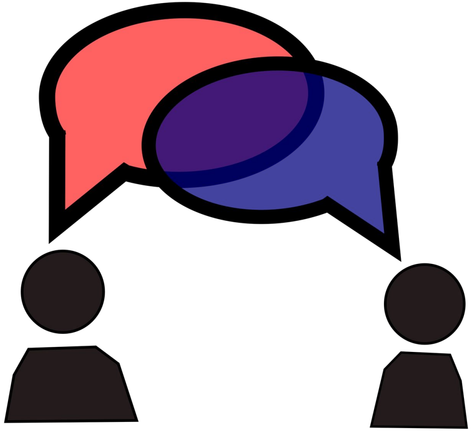
- xxx-xx-xxxx
- (xxx) xxx-xxxx

Does this sound positive?

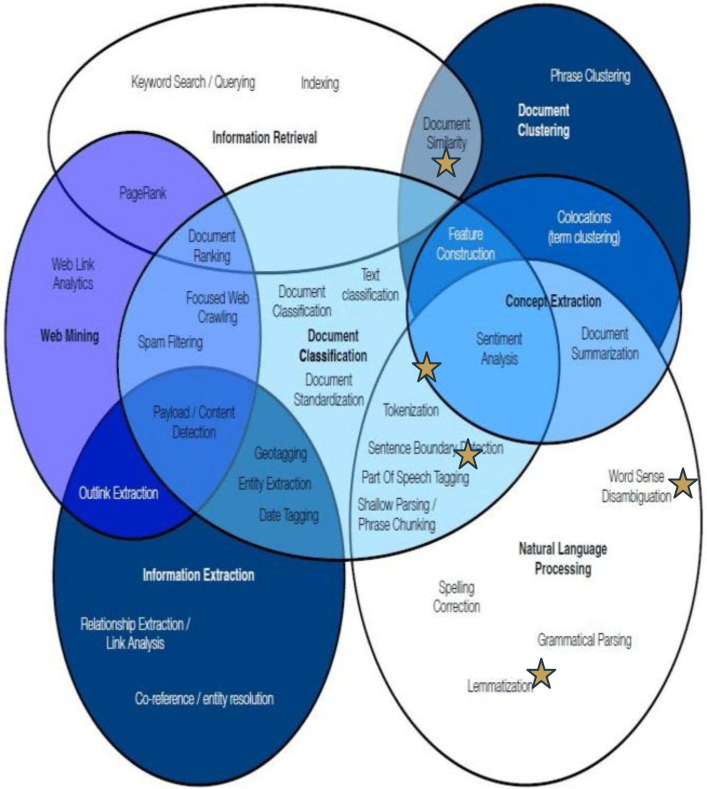
- Your book is **killer**
- They are on **fire**
- 🔥

Are these the same complaint?

- “Your customer service is **unresponsive**”
- “I get a to talk to someone once in a **blue moon**”
- “Necesito ayuda para **contactar a alguien**”
- “誰かに連絡を取るのに助けが必要です”



NLP Techniques: Examples



- Sentence Tokenization
- Word Tokenization
- Named Entity Recognition
- Lemmatization and Stemming
- Stop Words
- Regex
- Sentiment
- Topic Modeling

NLP Techniques: Terminology

- Corpus: A collection of written or spoken texts that are used to train and test NLP models.
- Document: A single text unit, such as an article, email, or tweet, that is analyzed by an NLP system.
- Tokenization: The process of breaking down a document or text into smaller units or tokens, such as words, phrases, or sentences, for further analysis.

NLP Techniques: Sentence Tokenization

[Definition] **Sentence Tokenization** is process of **splitting text** into **individual sentences**

[Example Input]

“Michelle Obama is an American lawyer, author, and former First Lady of the United States. She was born on January 17, 1964, in Chicago, Illinois, and graduated from Princeton University and Harvard Law School. During her time as First Lady, Michelle Obama focused on issues such as health and education, and she continues to be a prominent advocate for these causes.”

[Example Output]

- Michelle Obama is an American lawyer, author, and former First Lady of the United States.
- She was born on January 17, 1964, in Chicago, Illinois, and graduated from Princeton University and Harvard Law School.
- During her time as First Lady, Michelle Obama focused on issues such as health and education, and she continues to be a prominent advocate for these causes.

[Example Use]

- This may be useful to break down the information and track when we are switching context

NLP Techniques: Word Tokenization

[Definition] **Word Tokenization** is process of **splitting text** into **individual words**

[Example Input]

- She was born on January 17, 1964, in Chicago, Illinois, and graduated from Princeton University and Harvard Law School.

[Example Output]

- She - was - born - on - January - 17 - 1964 - in - Chicago - Illinois - and - graduated - from - Princeton - University - and - Harvard - Law - School

[Example Use]

- This may be useful to break down the information to apply changs at a **world level**

NLP Techniques: Named Entity Recognition

[Definition] **Named Entity Recognition (NER)** is a subtask of information extraction that seeks to locate and classify named entities mentioned in unstructured text into predefined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc.

[Example Input]

- She was born on January 17, 1964, in Chicago, Illinois, and graduated from Princeton University and Harvard Law School.

[Example Output]

- Chicago, Princeton - location
- January - date

[Example Use]

- May help disambiguate the text by knowing familiar entities like President, United States, etc.

NLP Techniques: Lemmatization and Stemming

[Definition] The goal of both stemming and lemmatization is to **reduce inflectional forms** and sometimes related forms of a word to a **common base** form [link]

[Example Input]

- She - was - born - on - January - 17 - 1964 - in - Chicago - Illinois - and - graduated - from - Princeton - University - and - Harvard - Law - School

[Example Lemma] graduated > **graduate |** was > **be** — *Other: having > have*

- She - was - born - on - January - 17 - 1964 - in - Chicago - Illinois - and - graduated - from - Princeton - University - and - Harvard - Law - School

[Example Stemming] graduated > **graduate |** was > **was** — *Other: having > hav*

[Example Use]

- Useful to reduce **ambiguity** in the variation of words and/or get to the base of the word. May be easier to draw comparisons between sentences at the cost of information loss

NLP Techniques: Stop Words

[Definition] Stop words are generally filler words. You can define more customization.

[Example Input]

- [**Input**] She - was - born - **on** - January - 17 - 1964 - **in** - Chicago - Illinois - **and** - graduated - **from** - Princeton - University - **and** - Harvard - Law - School
- [**Stop Words**] on, in, and, from

[Example Output]

- She - was - born - January - 17 - 1964 - Chicago - Illinois - graduated - Princeton - University - Harvard - Law - School

[Example Use]

- Ability to control what is processed by next steps. Can help in surfacing information which is different rather than common.

NLP Techniques: Regex

[Definition] A regular expression (shortened as regex or regexp) is a sequence of characters that specifies a match pattern in text. [regex generator link]

[Example Input]

- She - was - born - on - January - 17 - 1964 - in - Chicago - Illinois - and - graduated - from - Princeton - University - and - Harvard - Law - School.

[Example Output]

● ‘She’	= letter	s,h,e
● ‘17’	= numbers	1,7
● .	= punctuation	.

[Example Use]

- Helpful to perform operations on every single character and pattern.

NLP Techniques: Sentiment Analysis

[Definition] Sentiment analysis is used to systematically identify, extract, quantify, and study affective states and subjective information. (e.g., positive, neutral, negative)

[Example Input]

- She was born on January 17, 1964, in Chicago, Illinois, and graduated from Princeton University and Harvard Law School.

[Example Output]

- Positive - 98%

[Example Use]

- Helpful to bring in additional information from text about what the ‘emotion’ may be.

NLP Techniques: Topic Modeling

[Definition] Topic modeling is for discovery of hidden semantic structures in a text body.

[Example Input]

- Michelle Obama is an American lawyer, author, and former First Lady of the United States. She was born on January 17, 1964, in Chicago, Illinois, and graduated from Princeton University and Harvard Law School. During her time as First Lady, Michelle Obama focused on issues such as health and education, and she continues to be a prominent advocate for these causes.

[Example Output] 3 one word topics to describe the text above leveraging **word counts** :)

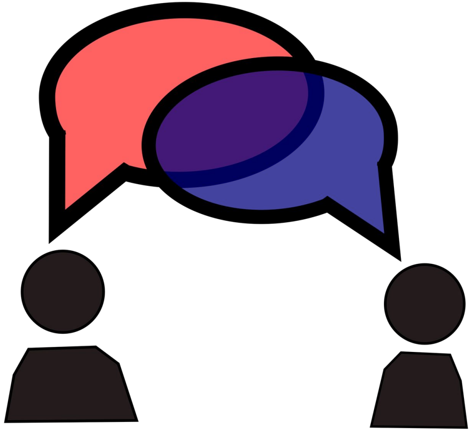
- Lady -2
- Michelle - 2
- Obama- 2

[Example Use]

- Helpful to summarize text based on observed words

Interactive

When do you think applying a custom stopword dictionary would be beneficial?



NLP Code Examples

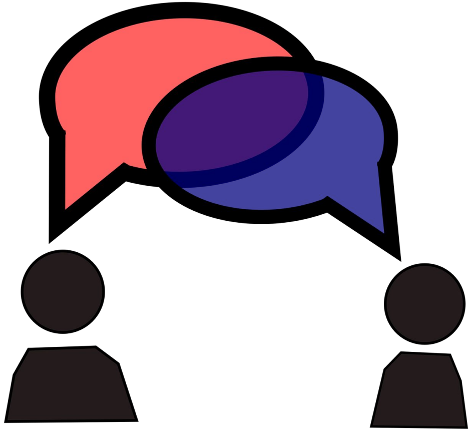
Tips and Outlook: Tips

- Understand your **use case**
- **Build hypothesis** on what you want to get from the text
- Evaluate **data origin** and **processing**
- **Pick** and **test** your NLP approach

Tips and Outlook: Outlook

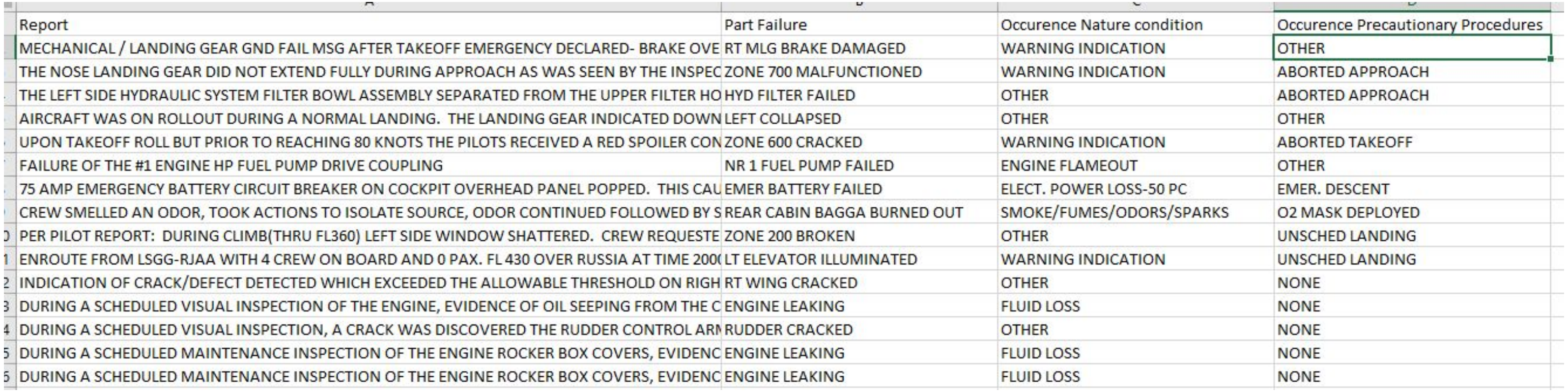
- The future is bright with lots of access to information even faster
- Interrogate the data and ask questions
- Think about how the data is processed, surfaced, and biased
- Have fun and try it :)

Open Mic



Appendix

[Data] Airline Safety



Report	Part Failure	Occurence Nature condition	Occurence Precautionary Procedures
MECHANICAL / LANDING GEAR GND FAIL MSG AFTER TAKEOFF EMERGENCY DECLARED- BRAKE OVE	RT MLG BRAKE DAMAGED	WARNING INDICATION	OTHER
THE NOSE LANDING GEAR DID NOT EXTEND FULLY DURING APPROACH AS WAS SEEN BY THE INSPEC	ZONE 700 MALFUNCTIONED	WARNING INDICATION	ABORTED APPROACH
THE LEFT SIDE HYDRAULIC SYSTEM FILTER BOWL ASSEMBLY SEPARATED FROM THE UPPER FILTER HO	HYD FILTER FAILED	OTHER	ABORTED APPROACH
AIRCRAFT WAS ON ROLLOUT DURING A NORMAL LANDING. THE LANDING GEAR INDICATED DOWN	LEFT COLLAPSED	OTHER	OTHER
UPON TAKEOFF ROLL BUT PRIOR TO REACHING 80 KNOTS THE PILOTS RECEIVED A RED SPOILER CON	ZONE 600 CRACKED	WARNING INDICATION	ABORTED TAKEOFF
FAILURE OF THE #1 ENGINE HP FUEL PUMP DRIVE COUPLING	NR 1 FUEL PUMP FAILED	ENGINE FLAMEOUT	OTHER
75 AMP EMERGENCY BATTERY CIRCUIT BREAKER ON COCKPIT OVERHEAD PANEL POPPED. THIS CAU	EMER BATTERY FAILED	ELECT. POWER LOSS-50 PC	EMER. DESCENT
CREW SMELLED AN ODOR, TOOK ACTIONS TO ISOLATE SOURCE, ODOR CONTINUED FOLLOWED BY S	REAR CABIN BAGGA BURNED OUT	SMOKE/FUMES/ODORS/SPARKS	O2 MASK DEPLOYED
PER PILOT REPORT: DURING CLIMB(THRU FL360) LEFT SIDE WINDOW SHATTERED. CREW REQUESTE	ZONE 200 BROKEN	OTHER	UNSCHED LANDING
ENROUTE FROM LSGG-RJAA WITH 4 CREW ON BOARD AND 0 PAX. FL 430 OVER RUSSIA AT TIME 200(	LT ELEVATOR ILLUMINATED	WARNING INDICATION	UNSCHED LANDING
INDICATION OF CRACK/DEFECT DETECTED WHICH EXCEEDED THE ALLOWABLE THRESHOLD ON RIGH	RT WING CRACKED	OTHER	NONE
DURING A SCHEDULED VISUAL INSPECTION OF THE ENGINE, EVIDENCE OF OIL SEEPING FROM THE C	ENGINE LEAKING	FLUID LOSS	NONE
DURING A SCHEDULED VISUAL INSPECTION, A CRACK WAS DISCOVERED THE RUDDER CONTROL ARM	RUDDER CRACKED	OTHER	NONE
DURING A SCHEDULED MAINTENANCE INSPECTION OF THE ENGINE ROCKER BOX COVERS, EVIDENC	ENGINE LEAKING	FLUID LOSS	NONE
DURING A SCHEDULED MAINTENANCE INSPECTION OF THE ENGINE ROCKER BOX COVERS, EVIDENC	ENGINE LEAKING	FLUID LOSS	NONE

[R Studio] Import: data, packages