

PROFESSIONAL CERTIFICATE IN MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Let's give everyone a couple of minutes to join...

Module 18

Natural Language Processing

Office Hours with Viviana Márquez
January 23, 2025



AGENDA

- Content review Module 18: Natural Language Processing
- Code examples
- Questions

AGENDA

- Content review Module 18: Natural Language Processing
- Code examples
- Questions

Structured data vs unstructured data

Mass (g)	Extension 1 (mm)	Extension 2 (mm)	Average Extension (mm)
0	0	1	0.5
100	5	6	5.5
200	9	9	9
300	15	15	15
400	20	21	20.5
500	24	25	24.5
600	30	31	30.5

- **Structured data** is highly organized and easily readable by machines. It is typically stored in tabular formats, such as spreadsheets (CSV, Excel) or relational databases (SQL).

Each observation is in a row, and its features are in predefined columns, making it easier to process and analyze.



- **Unstructured data** does not follow a specific format or structure, making it more challenging to organize and analyze. This type of data includes free text, images, videos, audio, and other multimedia formats.

Due to its nature, unstructured data often requires advanced techniques such as Natural Language Processing (NLP) or Convolutional Neural Networks (CNN).

Structured data vs unstructured data

Usually Machine Learning

Mass (g)	Extension 1 (mm)	Extension 2 (mm)	Average Extension (mm)
0	0	1	0.5
100	5	6	5.5
200	9	9	9
300	15	15	15
400	20	21	20.5
500	24	25	24.5
600	30	31	30.5

- **Structured data** is highly organized and easily readable by machines. It is typically stored in tabular formats, such as spreadsheets (CSV, Excel) or relational databases (SQL).

Each observation is in a row, and its features are in predefined columns, making it easier to process and analyze.

Usually Deep Learning

- **Unstructured data** does not follow a specific format or structure, making it more challenging to organize and analyze. This type of data includes free text, images, videos, audio, and other multimedia formats.

Due to its nature, unstructured data often requires advanced techniques such as Natural Language Processing (NLP) or Convolutional Neural Networks (CNN).

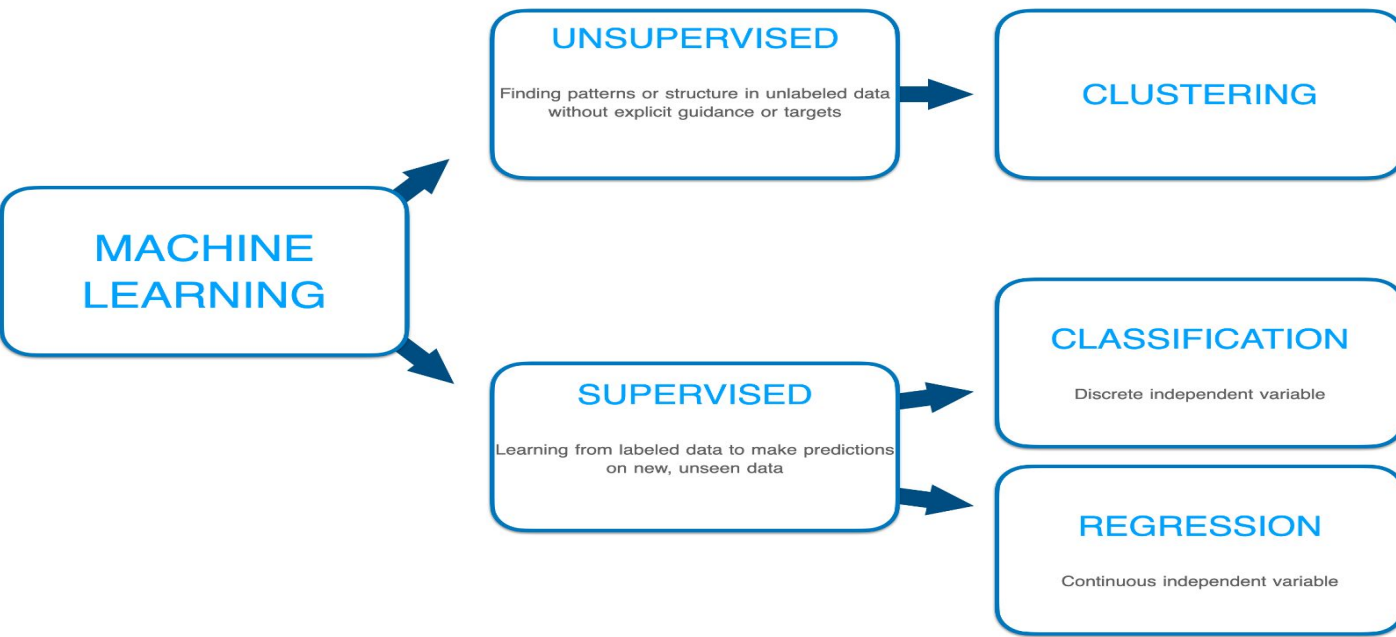




Section 3: Advanced Topics and Capstone

Do we have labels? Is my target variable discrete?

In NLP...



Unsupervised tasks:

- Topic modeling
- Document clustering
- Social media analysis

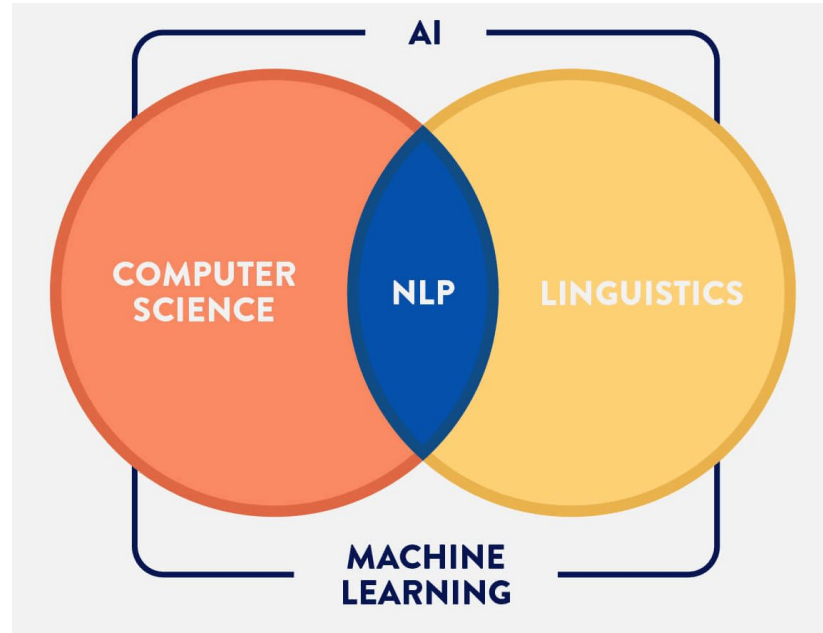
Classification tasks:

- Sentiment analysis
- Spam detection
- Topic classification

Regression tasks:

- Predict score of a review
- Predict age of author

What is NLP?



- It is the area of artificial intelligence that deals with human languages and derives valuable information from them
- There are dozens of methods and strategies to solve a given problem

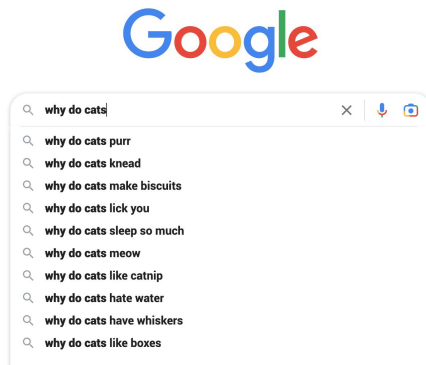
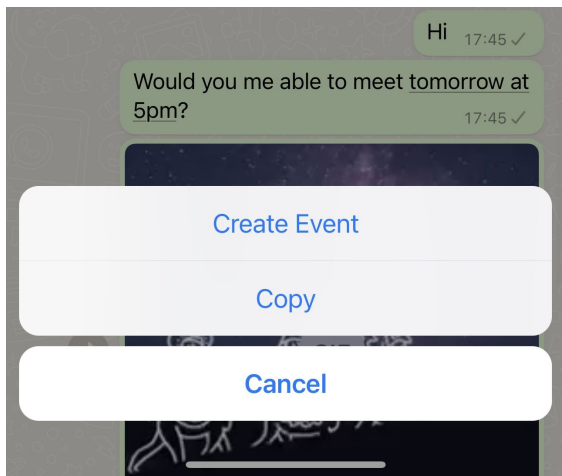
Text and Big Data



- From 80% to 90% of data generated and collected by organizations is unstructured → Most of it is in the form of text
- Millions of data are being generated right now: WhatsApp, Twitter, YouTube, Reddit, etc.

When was the last time you used NLP?

When was the last time you used NLP?



NLP and social media



- Companies want to understand their customers (targeted marketing campaigns)
- They want to be able to take advantage of explicit and implicit feedback from their customers
- Example: [Here](#)
- Example: Emoji analysis by Instagram ([More info here](#))

How do we talk to Siri? Does it speak English?



Representation of data in numerical form



What We See

```
08 02 22 97 38 15 00 40 00 75 04 05 07 78 52 12 50 77 91 08 08 02 22 97
49 49 99 40 17 81 57 40 87 17 43 98 43 49 48 04 56 42 00 49 49 99 40
81 49 31 73 55 79 14 29 93 71 40 47 53 88 30 03 49 13 56 65 81 49 31 73
52 70 95 23 04 60 11 42 49 24 68 56 01 32 56 71 37 02 36 91 52 70 95 23
22 31 16 71 51 67 63 89 41 92 36 54 22 40 40 28 66 33 13 80 22 31 16 71
24 47 32 60 99 03 45 02 44 75 33 53 78 36 84 20 35 17 12 50 24 47 32 60
32 98 81 28 64 23 67 10 26 38 40 47 59 54 70 66 18 38 44 70 32 98 81 28
87 24 20 68 02 62 12 20 95 63 94 39 63 08 40 91 66 49 94 21 67 26 20 68
24 55 38 05 64 73 89 24 97 17 78 78 86 83 14 88 34 89 63 72 24 55 38 05
21 36 23 09 75 00 76 44 20 45 35 14 00 61 33 97 34 31 33 95 21 36 23 09
78 17 53 28 22 75 31 47 15 94 03 80 04 62 16 14 09 53 56 92 78 17 53 28
16 39 05 42 96 35 31 47 55 88 88 24 00 17 54 24 36 29 85 57 16 39 05 42
86 56 00 48 35 71 89 07 05 44 44 37 44 60 21 58 51 54 17 58 86 56 00 48
19 80 81 68 05 94 47 69 28 73 92 13 86 52 17 77 04 89 55 40 19 80 81 68
04 52 08 83 97 35 99 16 07 97 57 32 16 26 26 79 33 27 98 66 04 52 08 83
88 56 48 87 57 62 20 72 03 46 33 47 46 55 12 32 63 93 53 69 88 36 48 87
04 42 16 73 38 25 39 11 24 94 72 18 08 46 29 32 40 42 76 36 04 42 16 73
20 49 36 41 72 30 23 88 34 42 99 49 82 47 59 85 74 04 36 16 20 49 36 41
20 73 35 29 78 31 90 01 74 31 49 71 48 86 81 16 23 57 05 54 20 73 35 29
01 70 54 71 83 51 54 69 16 92 33 48 61 43 52 01 89 19 47 48 01 70 54 71
```

What Computers See

- An image is represented on a computer in the form of a matrix where each cell represents a pixel of the image
- Similarly, a video is a collection of frames, where each frame is an image.
Therefore, any video can be represented as a collection of matrices
- (Un)fortunately, representing text numerically is not so simple

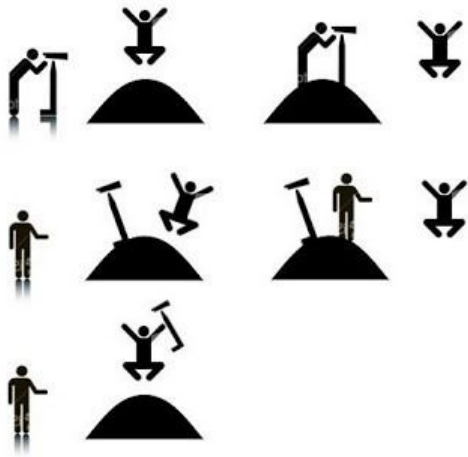


Why is NLP so hard?

I saw the man on the hill with a telescope

😞 🤯 Why is NLP so hard?

I saw the man on the hill with a telescope



- 1. I saw the man. The man was on the hill. I was using a telescope.
- 2. I saw the man. I was on the hill. I was using a telescope.
- 3. I saw the man. The man was on the hill. The hill had a telescope.
- 4. I saw the man. I was on the hill. The hill had a telescope.
- 5. I saw the man. The man was on the hill. I saw him using a telescope.

😞 🤯 Why is NLP so hard?

🤔 The issue at hand...



- Let's take as an example the word **bow**. Can we just give it a number to represent it?
- 🤖 Then... how do we represent the meaning of a word?

😞 🤯 Why is NLP so hard?

How do we, humans, know what a word means?



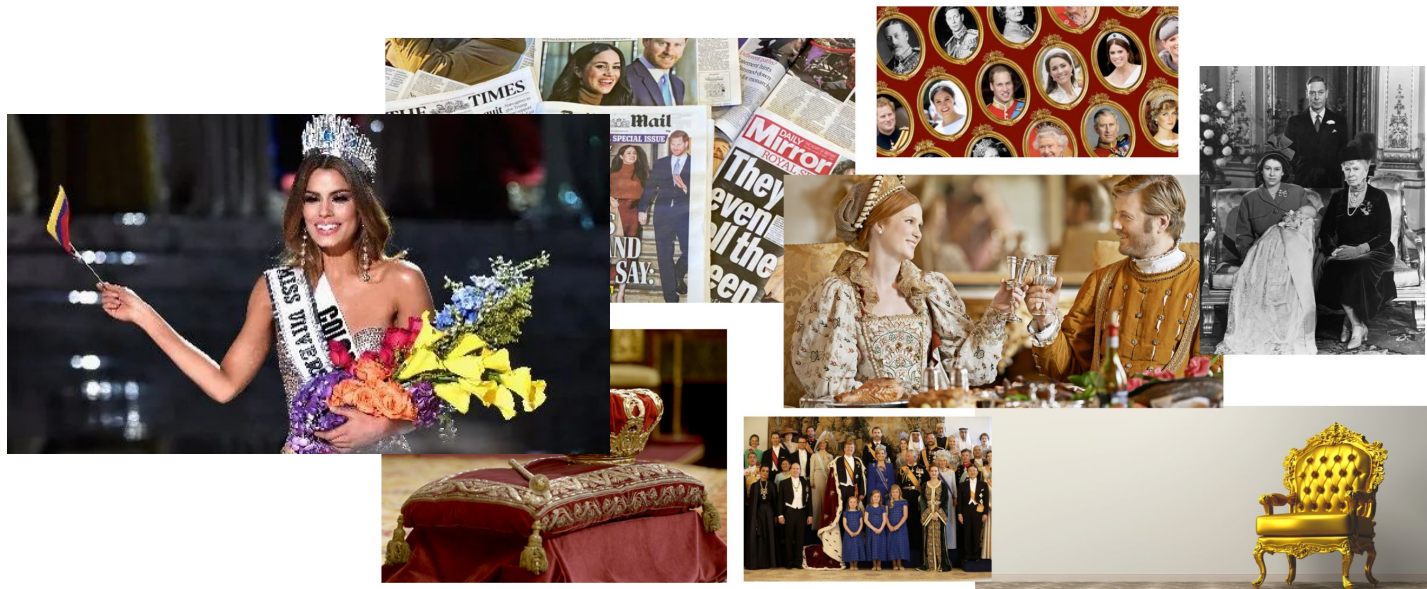
😞 🤯 Why is NLP so hard?

How do we, humans, know what a word means?
We learn through experience...



😞 🤯 Why is NLP so hard?

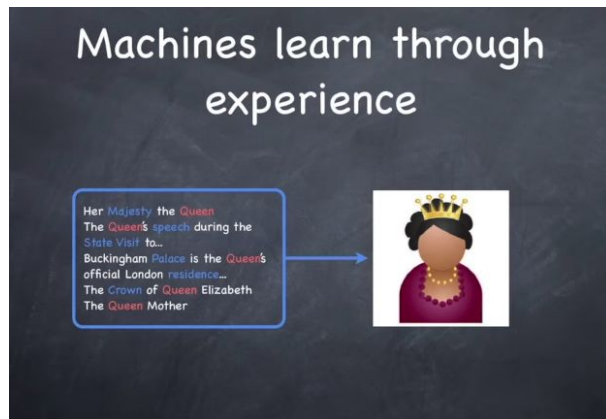
How do we, humans, know what a word means?
We learn through experience...



😞 🤯 Why is NLP so hard?

How do we, humans, know what a word means?
We learn through experience...

Good news: Machines also learn from experience





Why is NLP so hard?

What is tichiniky?

(Don't look it up on Google)



Why is NLP so hard?

What is tichiniky?

(Don't look it up on Google)

- Joseph offered a glass of **tichiniky** to his girlfriend.
- **Tichiniky** and steak make a great pairing for a nice meal.
- Charles stumbled, his face flushed from drinking too much **tichiniky**.
- Last night I had bread, cheese, and this excellent **tichiniky** for dinner.



Why is NLP so hard?

What is tichiniky?

(Don't look it up on Google)

- Joseph offered a glass of **tichiniky** to his girlfriend.
- **Tichiniky** and steak make a great pairing for a nice meal.
- Charles stumbled, his face flushed from drinking too much **tichiniky**.
- Last night I had bread, cheese, and this excellent **tichiniky** for dinner.



What is tichiniky?

An alcoholic beverage

Distributional semantics

- A bottle of **tichiniky** is on the table.
- Not everyone likes **tichiniky**.
- Don't drink **tichiniky** and drive.
- We make **tichiniky** with grapes.

Distributional semantics


- A bottle of _____ is on the table.
- Not everyone likes _____.
- Don't drink _____ and drive.
- We make _____ with grapes.

	1	2	3	4
tichiniky				
strong				
motor oil				
tacos				
wine				

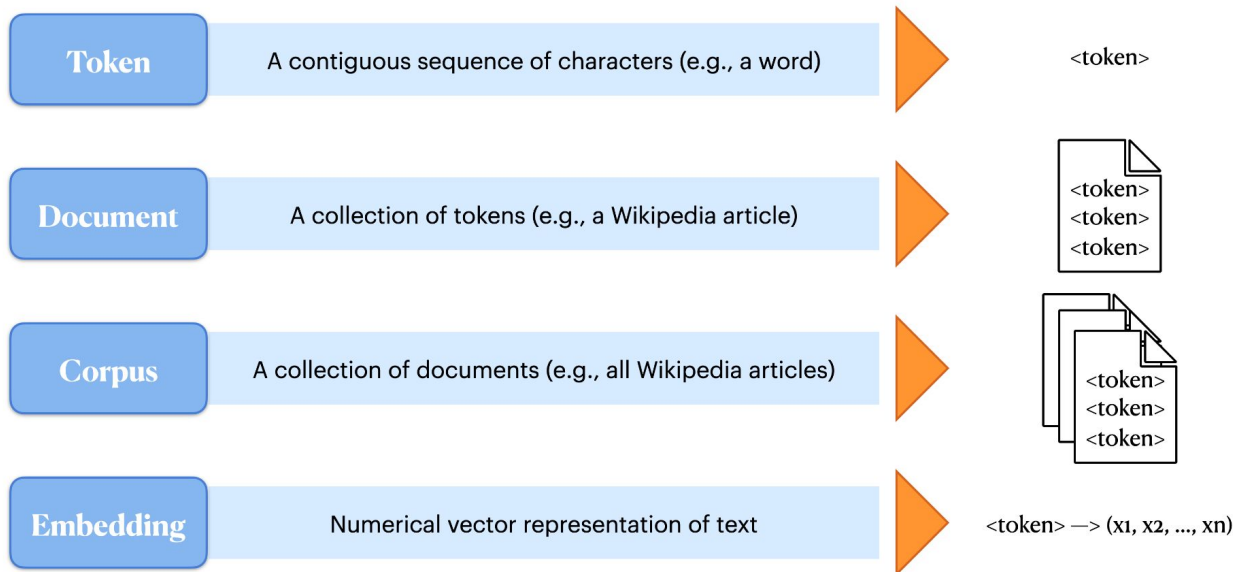
Distributional semantics

- A bottle of _____ is on the table.
- Not everyone likes _____.
- Don't drink _____ and drive.
- We make _____ with grapes.

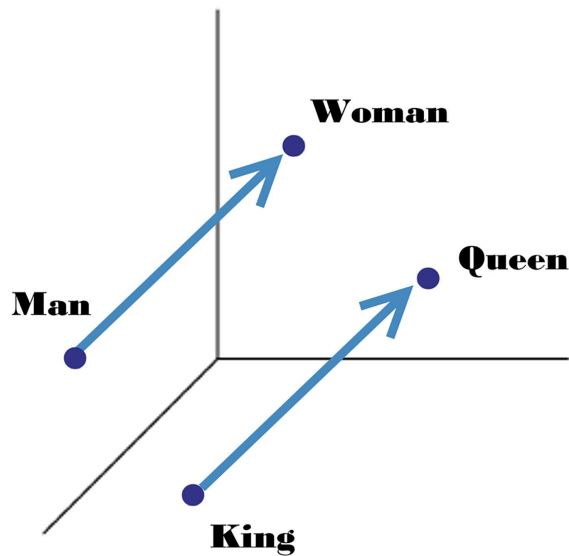
	1	2	3	4
tichiniky	1	1	1	1
strong	0	0	0	0
motor oil	1	0	0	0
tacos	0	1	0	0
wine	1	1	1	1

-  Tichiniky & wine have the same **vector representation**

Common NLP terminology



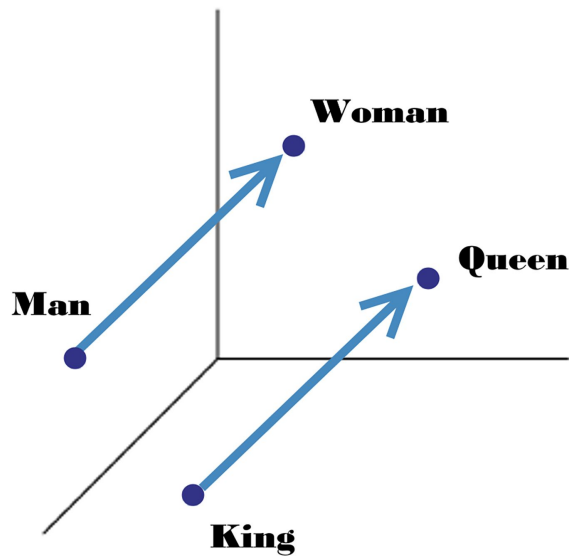
Once you have a numerical vector representation



You can do math!

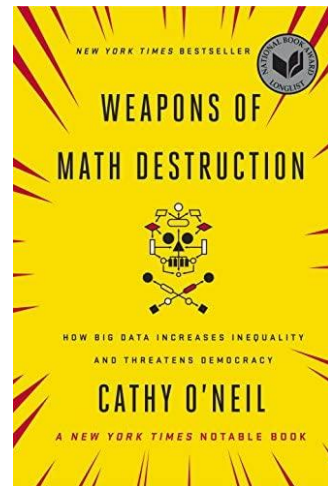
- $\text{King} - \text{Man} + \text{Woman} = \text{Queen}$
- Neural networks can make machines understand analogies like humans [Mikolov et al., 2013]

Once you have a numerical vector representation



You can do math!

- $\text{King} - \text{Man} + \text{Woman} = \text{Queen}$
- Neural networks can make machines understand analogies like humans [Mikolov et al., 2013]



The evolution of NLP



1950

Evolution of NLP models

2020

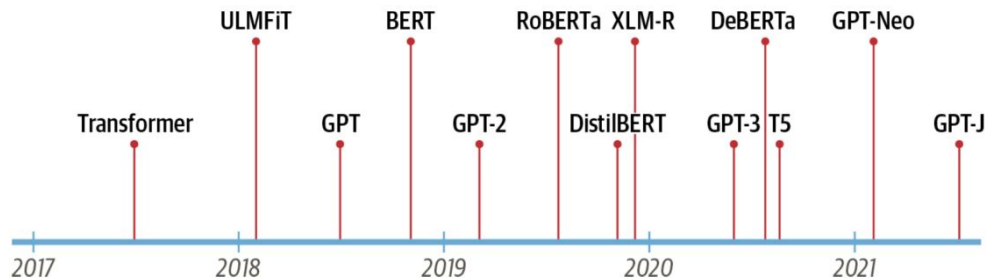
The evolution of NLP



1950

Evolution of NLP models

2020



The evolution of NLP



1950

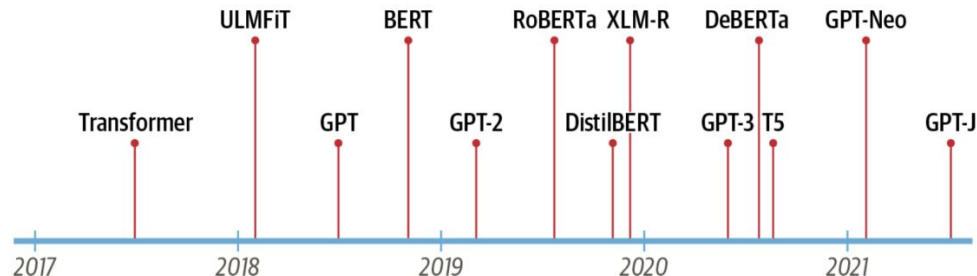
Evolution of NLP models

2020

LLMs

Large Language Models

deep learning models trained on vast amounts of text data to predict the next word in a sentence given the previous words



Training LLMs

1. Pre-training

- predicting the next token
- typically done by large organizations (OpenAI, Meta, Microsoft)
- across clusters of GPUs
 - GPT-4: 25,000 GPUs for 100 days
 - LLaMa-2 70B: 6,000 GPUs for 12 days
- costs millions of \$\$\$

=> to get a “base model”



How to take advantage of LLMs?



huggingface

- [Hugging Face](https://huggingface.co/docs/transformers/index) is a startup that offers 30 pre-trained modules in more than 100 languages and 8 architectures for NLU & NLG
- <https://huggingface.co/docs/transformers/index>
- https://huggingface.co/docs/transformers/main/en/main_classes/pipelines
- More slides: <https://docs.google.com/presentation/d/1yBWLNzlrIsfNprbEnmYdqckAMrwFZB-/edit#slide=id.p1>
- Example code: https://colab.research.google.com/drive/12_VkQMHiDw2VKcuWM3WL6wWDfljsgog_?usp=sharing



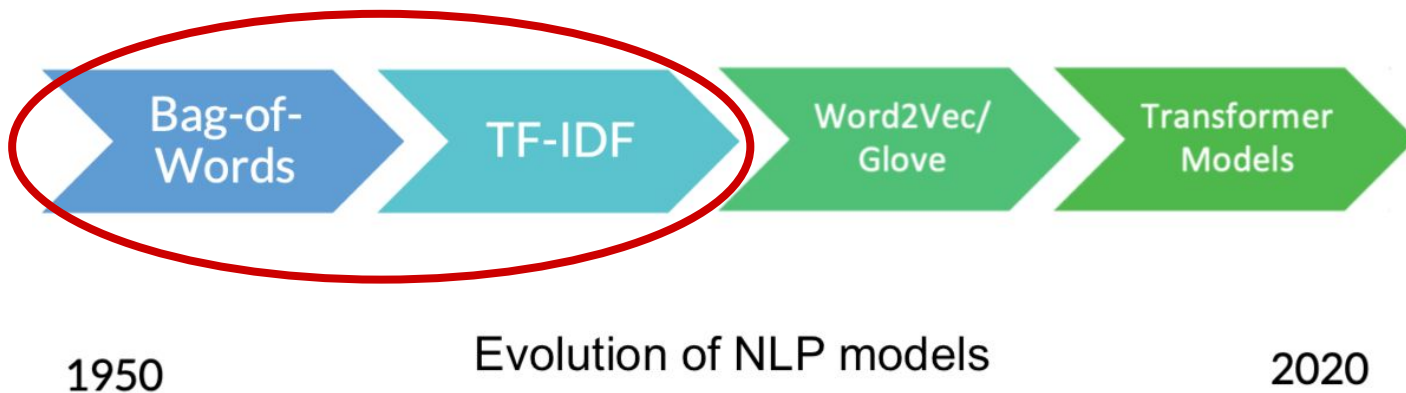
Retrieval
Query

Retrieved
Texts



Retrieval Augmented Generation.




The evolution of NLP



QUESTIONS?



AGENDA

-  Content review Module 18: Natural Language Processing
-  Code examples
-  Questions