{EPITECH}

T-AIA-901

<Kickoff>

01

NLP

Natural Language processing

The Graal of UX

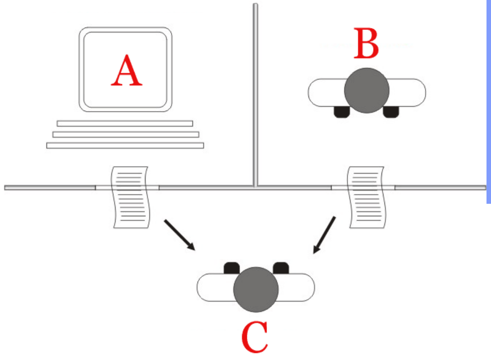
<Most natural way to interact with computer ?/>

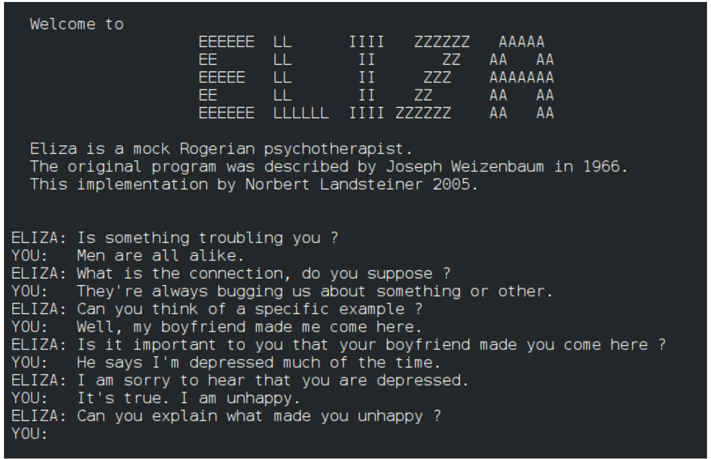
Talking to computer using natural language

1950 Turing test
1964 Eliza 1st chatbot

(very) common in sci-fi (2001 a space odyssey)

2023 ?? ChatGPT : in our perception, IA and NLP are nearly synonymous.





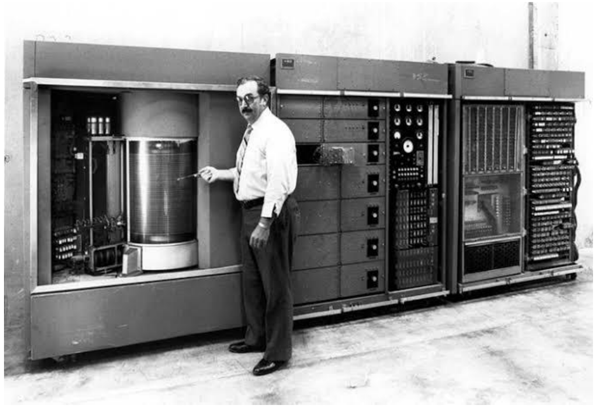
```
  Welcome to
                    EEEEEE  LL      IIII   ZZZZZZ   AAAAA
                    EE      LL       II        ZZ  AA   AA
                    EEEEE   LL       II      ZZZ   AAAAAAA
                    EE      LL       II     ZZ     AA   AA
                    EEEEEE  LLLLLL  IIII  ZZZZZZ   AA   AA

  Eliza is a mock Rogerian psychotherapist.
  The original program was described by Joseph Weizenbaum in 1966.
  This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

A winding road

<the « we are nearly there » syndrome />

1954 IBM + Georgetown translation experiment
Huge federal investment
« within three or five years, machine translation could well be a solved problem »

1966 ALPAC senate report stops the funding



Very diverse categories of problems

< Not surprising as natural language structures our though />

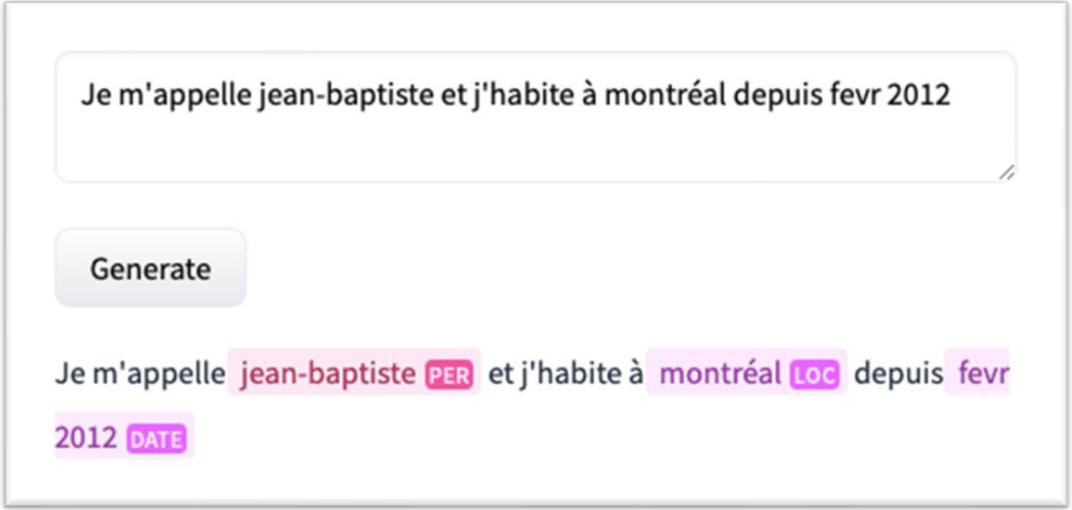
- Syntactic and grammatical analysis
- Sentiment analysis & Classification (e.g: spam/not spam, book categories...)
- Information extraction (entity linking, NER, etc...)
- Text generation (genAI, translation, answer to questions, résumé, reformulation...)
- Searching (document database query)
- Etc..

NER

< Named entity recognition />

Some items in sentences may be a date, a person, a location, an organization... and it is valuable to identify this.

How about "*Florence est née à Paris et habite Albert, alors qu'Albert est né à Florence et habite Paris*" ?



Je m'appelle jean-baptiste et j'habite à montréal depuis fevr 2012

Je m'appelle jean-baptiste PER et j'habite à montréal LOC depuis fevr 2012 DATE

SUCCESS STORY_

While studying at Epitech and drawing inspiration from Noam CHOMSKY’s work, Thomas SOLIGNAC developed a sovereign and energy-efficient NLP technology.

In 2016, a year after graduating alongside Killian VERMERSCH, they co-founded Golem.ai, a company offering NLP solutions to automate repetitive and time-consuming tasks.

< Thomas & Killian Forbes interviews>

MORE SUCCESS STORIES_

Company	Known for	Foundation	2025 valuation
Mistral AI	Large Language Models	2023	$14 billions
Hugging Face	Transformers library	2016	$4.5 billions
Deepl	Translation API	2017	$2 billions
Deepset	Haystack	2018	$60 millions

02 Processing words

When you only know how to process « 0 » and « 1 »

Numerous techniques

< Plenty of methods, algorithms and libraries />

- Decades of research
- In mathematics, linguistic and of course computer science

- Many different problems addressed by NLP

- ... Leads to a great variety of tools

- In all cases, computer software can handle numerical value, but not really text values

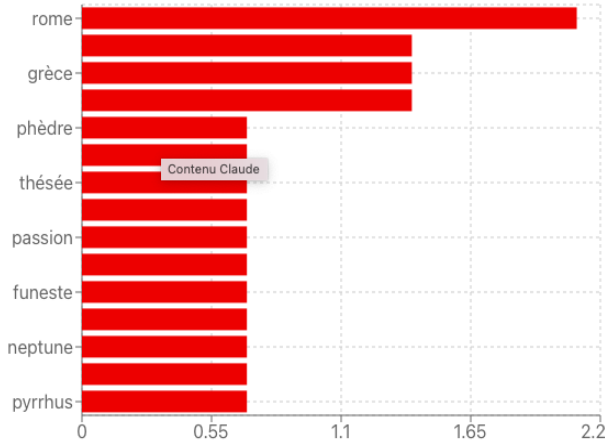
Statistical analysis

< for instance Bag of Words, TF-IDF />

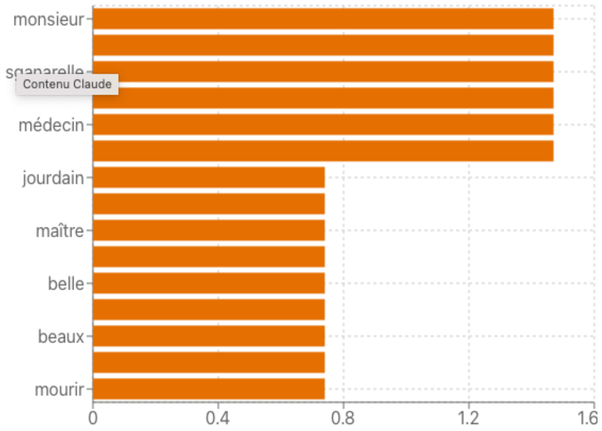
Sometime presence of some words are sufficient to classify text or extract information.
Obviously very frequent words (in fr : “et”, “ou”, “car”) have less impact (TF-IDF
Structural information is lost

E.g. to classify between Racine and Molière

👑 Mots caractéristiques de Racine



🎭 Mots caractéristiques de Molière



Semantic relationship

“être” et “sera” : two different words but obviously related : same lemme -> lemmatization

But obviously we are missing something :

- "manger" et "mangera" have same lemme and are closely related
- "diner", "se goinfrer", "boire" : not identical but yet related

Semantic relationship

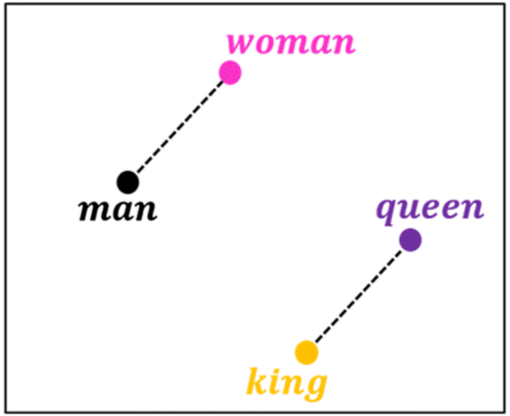
< word embedings />

Words represented by vectors
Relatively low number of dimensions

Vectorial operation
"King – man + woman = queen"

Distance (angle) -> semantic similarity

Complex task !! (Neural networks, trained on large text corpus).
"Words are known by their context"



"You shall know a word by the company it keeps",
linguist John Rupert Firth 1957

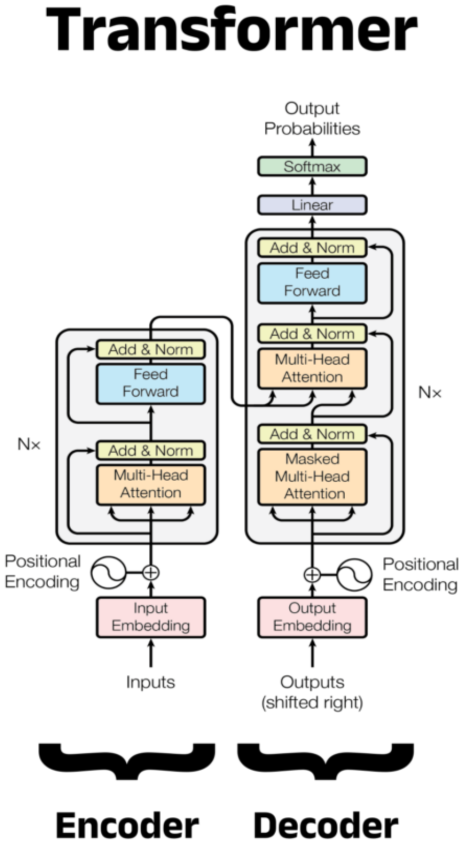
03 Transformers

State of the art NLP

Transformers & BERT

< the ‘T’ of ‘GPT’ />

- Bidirectional Encoder Representation from Transformer
- 2018 (Google)
- Ubiquitous in NLP just a couple of years after
- Many variants (french : CamemBERT)
- Highly structured neural network (around 100M parameters)



04

T-AIA-901 project

NER on origin / destination town

Prioirity and delivery

< Secure your delivery />

Proposed steps :

- Dataset (create your sentences)
- Define metrics
- Simple implementation first (Spacy ?)
- Experiment transformer (Encoder model, like BERTor one of BERT variants)

- Don't forget evaluation of your model(s) : metrics
- Don't forget a synthetic report (PDF)

{EPITECH}

Thank you