

# NLP & Analyse de sentiments

Loïc, Stan & JB



# BRIEF

## • Contexte du projet

- La base de données Internet Movie Database (IMDb) est considérée le "Hello World" du traitement automatique du langage naturel. Les données d'entrée sont des critiques de film rédigées en anglais par les spectateurs. La critique est labellisée 0 si elle est négative, 1 si elle est positive. On parle alors d'analyse de sentiments. L'objectif de ce travail est d'utiliser cette base de données pour acquérir certains fondamentaux du traitement automatique du langage naturel. Pour cela, il va falloir tout d'abord construire une présentation de type Powerpoint se déroulant selon le plan suivant :
- Qu'est-ce que le traitement automatique du langage naturel ?
- Quelles sont ses applications ? (une diapositive)
- Qu'est-ce que l'analyse de sentiments ? (une diapositive)
- Décrire la base de données IMDb (une diapositive). La base de données IMDb peut se charger selon deux manières :
  - Avec les données d'entrée stockées sous forme textuelle ;
  - Avec les données d'entrée stockées sous forme de suite de nombres entiers, chaque nombre entier encodant un mot. C'est cette version que nous allons utiliser dans un premier temps avec la ligne TensorFlow/Keras suivante : `(train_samples, train_labels), (test_samples, test_labels) = imdb.load_data(num_words=10000)`. Décrire dans la présentation ce principe d'encodage de mots par des nombres entiers.
- Cette séquence de nombres entiers peut ensuite être encodée à son tour en utilisant l'encodage One-Hot. Décrire dans la présentation ce principe.
- Entraîner ensuite dans un Jupyter Notebook un perceptron multicouche en utilisant la librairie TensorFlow/Keras sur les données IMDb que vous aurez encodées au format One-Hot.

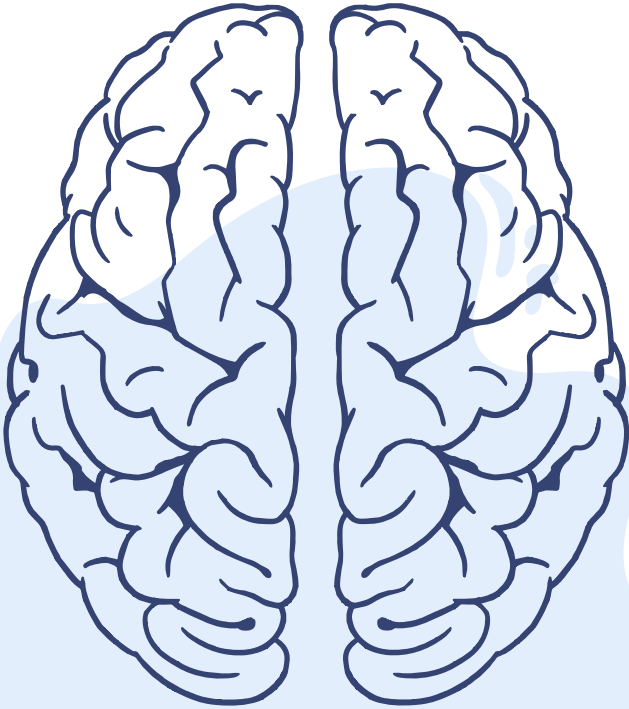
# Une présentation en plusieurs étapes

**01** NLP ? Vous avez  
dit NLP

**03** IMDB

**02** Quelques applications

# Le Traitement automatique en quelques mots



Le traitement naturel du langage, ou Natural Language Processing (NLP) en anglais, est une technologie d'intelligence artificielle visant à **permettre aux ordinateurs de comprendre le langage humain.**



02

# **Quelques applications**



### Information Retrieval

Doc A



Doc 1

Doc 2

Doc 3

### Sentiment Analysis



### Information Extraction



### Machine Translation



# Natural Language Processing

### Question Answering



Human: When was Apollo sent to space?



Machine: First flight -  
AS-201,  
February 26,  
1966

$$\text{NLP} = \text{NLU} + \text{NLG}$$

## NLG

- Discourse Generation
- Lexical Choice
- Sentence Planning/Generation
- Realization
- Document Structuring

## NLP

- Named entity Recognition (NER)
- Part-of-Speech Tagging (POS)
- Syntactic Parsing
- Coreference Resolution
- Machine Translation

## NLU

- Lexical Ambiguity/Analysis
- Sentiment Analysis
- Topic Classification
- Information Extraction
- Entity Detection
- Summarization
- Semantic Parsing
- Syntactical Analysis



02

# **L'analyse de sentiments**







1.0



0.5



0.0

03

## La BDD IMDB

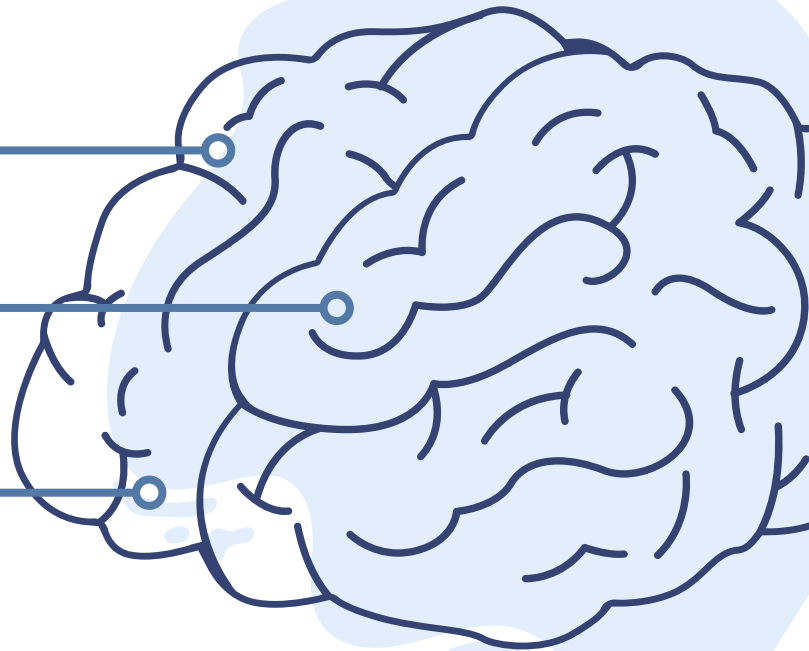


# Pour mieux comprendre le NLP : un exemple

IMDB en  
quelques  
mots

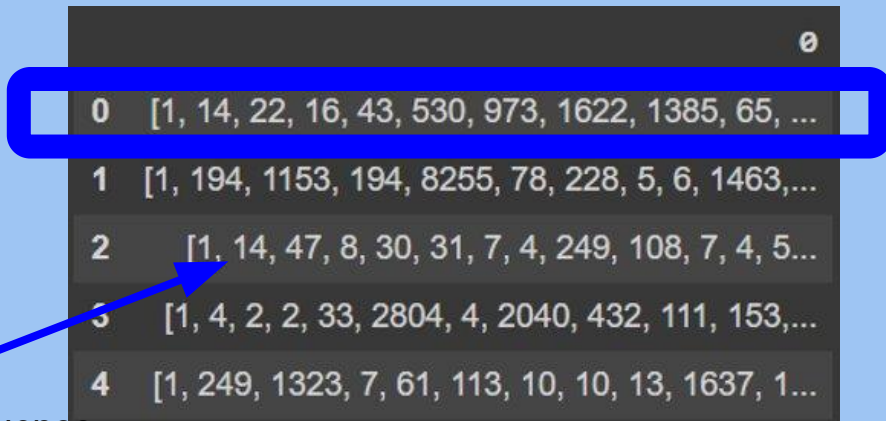
IMDB par  
Keras

One Hot  
Encoding



# IMDB Dataset imported with Keras

this film was just brilliant casting location scenery story  
direction everyone's really suited the part they played and  
you could just imagine being there robert # is an amazing  
actor and now ...



0	[1, 14, 22, 16, 43, 530, 973, 1622, 1385, 65, ...]
1	[1, 194, 1153, 194, 8255, 78, 228, 5, 6, 1463, ...]
2	[1, 14, 47, 8, 30, 31, 7, 4, 249, 108, 7, 4, 5...]
3	[1, 4, 2, 2, 33, 2804, 4, 2040, 432, 111, 153, ...]
4	[1, 249, 1323, 7, 61, 113, 10, 10, 13, 1637, 1...]

Ce nombre indique la fréquence  
d'apparition d'un mot



# Entrainement du modèle



01

Vectorize

Entrainement du modèle

02



# Powered by Simplon



M. Stan



fantastic3



M. Jean Baptiste



M. Loïc



3fantastic@simplon.world



<http://youtube/frtf3>