Rapport pour le projet de
**Deep Learning, NLP**
Version du 06 Juillet 2022

# Introduction to Natural Language Processing
# Catch-up 1

Sous la direction de

**Marc VON-WYL**

**Christopher DIAMANA LUTETE**
EPITA SCIA promotion 2022

# Table des matières

# 1 The dataset

## 1.1 Question 1.1

***How many splits does the dataset has ?***

The IMDB sentiment dataset has 3 splits of highly polar movie reviews. The first for training, the second for testing and the third for unsupervised learning.

## 1.2 Question 1.2

***How big are these splits ?***

- The train split contains 25 000 reviews.
- The test split contains 25 000 reviews.
- The unsupervised split contains 50 000 reviews.

## 1.3 Question 1.3

***What is the proportion of each class on the supervised splits ?***

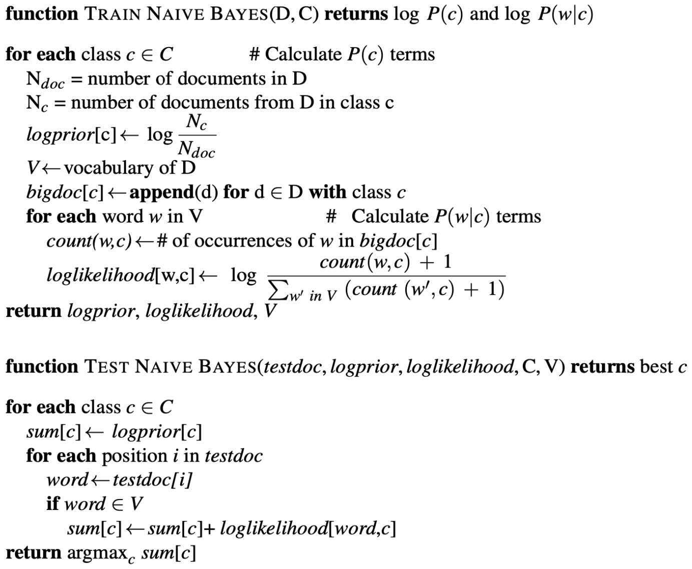
On the supervised splits there are :

- For the train split : 12 500 negative reviews and 12 500 reviews.
- For the test split : 12 500 negative reviews and 12 500 reviews.

There are a total of 25 000 negative reviews and 25 000 positive reviews.

# 2 Naive Bayes classifier

I chose to implement my own naive Bayes classifier following this algorithm :



```
function TRAIN NAIVE BAYES(D, C) returns log $P(c)$ and log $P(w|c)$

for each class $c \in C$            # Calculate $P(c)$ terms
   $N_{doc}$ = number of documents in D
   $N_c$ = number of documents from D in class c
   $logprior$[c] ← log $\frac{N_c}{N_{doc}}$
   $V$ ← vocabulary of D
   $bigdoc$[$c$] ← append(d) for d ∈ D with class $c$
   for each word $w$ in V                  #  Calculate $P(w|c)$ terms
      $count(w,c)$ ← # of occurrences of $w$ in $bigdoc$[$c$]
      $loglikelihood$[w,c] ← log $\frac{count(w,c) + 1}{\sum_{w' \text{ in } V} (count(w',c) + 1)}$
return $logprior$, $loglikelihood$, $V$

function TEST NAIVE BAYES($testdoc$, $logprior$, $loglikelihood$, C, V) returns best $c$

for each class $c \in C$
   $sum$[$c$] ← $logprior$[$c$]
   for each position $i$ in $testdoc$
      $word$ ← $testdoc[i]$
      if $word \in V$
         $sum$[$c$] ← $sum$[$c$]+ $loglikelihood$[$word$,$c$]
return argmax$_c$ $sum$[$c$]
```

## 2.1 Question 2.4

***Why is accuracy a sufficient measure of evaluation here ?***

# 3 FastText

## 3.1 Question 3.4

***Look at their attributes. How do the two models differ ?***

## 3.2 Question 3.6

***Why is it likely that the attributes minn and maxn are at 0 after an hyper-parameter search on our data ?***

## 4 Theoritical questions

### 4.1 Question 4.1

***Explain with your own words, using a short paragraph for each, what are :***

❖ ***Phonetics and phonology***

Phonetics and phonology are two linguistics subfields that study language sounds. Phonetics is the study and classification of speech sounds. It's concerned with the physical properties of speech sounds, their physiological production, acoustic properties and auditory perception. Phonology is the system of contrastive relationships between speech sounds that constitute the fundamental components of a language. In other words, phonology is the study of sounds, especially the different patterns of sounds in different languages.

❖ ***Morphology and syntax***

Morphology and syntax are two significant sub-disciplines in the field of linguistics. Morphology studies how words are formed whereas syntax studies how sentences are formed.

❖ ***Semantics and pragmatics***

Both semantics and pragmatics are two main branches of study in linguistics. Semantics is the study of the meaning of words and their meaning within sentences. Pragmatics is the study of the same words and meanings but with emphasis on their context as well.

### 4.2 Question 4.2

***What is the difference between stemming and lemmatization ?***

❖ ***How do they both work ?***
❖ ***What are the pros and cons of both methods ?***

### 4.3 Question 4.3

***On logistic regression :***

❖ ***How does stochastic gradient descent work ?***
❖ ***What is the role of the learning rate ?***
❖ ***Will it always find the global minimum ?***

### 4.4 Question 4.4

***What problems does TF-iDF try to solve ?***

❖ ***What the is the TF part for ?***

- ❖ ***What is the iDF part for ?***

## 4.5 Question 4.5

***Summarize how the skip-gram method of Word2Vec works using a couple of paragraphs.***

- ❖ ***How does it uses the fact that two words appearing in similar contexts are likely to have similar meanings ?***