

|  |   |   |
|--|---|---|
| Année Universitaire : 2025/2026<br>Master 2 : SII<br>Module : TALN | Université des Sciences et de la Technologie Houari Boumediene<br>Faculté d'Informatique<br>Département d'Intelligence Artificielle et Sciences des Données | TP N°4<br>Text Classification<br>Part 1 |
|--|---|---|

## Text Classification using Naïve Bayes:

### 1. Data:

The data are attached with these folders:

#### Training Data:

**Folder:** All-in-many/

**Description:** 117 articles (titles and abstract) of Volume 18 of Evolutionary Intelligence Journal.

#### Training Data Labeling:

**Folder:** All-in-many\_classification/

**Description:** Four classes available :

**1<sup>st</sup> class :** Metaheuristics

**2<sup>nd</sup> class :** Machine & Deep Learning

**3<sup>rd</sup> class :** Combination of Metaheuristics & Machine/Deep Learning

**4<sup>th</sup> class :** Others

### 2. Naive Bayes classifier for text classification :

- Naive Bayes classification is a probabilistic classifier based on supervised learning which, for a given document  $d$ , identifies among all possible classes  $c \in \mathcal{C}$  the one that is the most probable, i.e., the class  $c$  having the highest probability:

$$\hat{c} = \arg \max_{c \in \mathcal{C}} P(c|d)$$

- The most probable class  $\hat{c}$  for a given document  $d$  is computed by selecting the class with the highest product of two probabilities:
  - o The prior probability of the class  $P(c)$
  - o The probability of the document given the class  $P(d|c)$

$$\hat{c} = \arg \max_{c \in \mathcal{C}} P(c) P(d|c)$$

- o The document  $d$  can be represented as a set of words:

$$\hat{c} = \arg \max_{c \in \mathcal{C}} P(c) P(w_1, w_2, w_3 \dots, w_n | c)$$

- o The probability  $P(c)$  represents the proportion of documents in the training set that belong to class  $c$ :

$$P(c) = \frac{N_c}{N}$$

Where:

$N_c$  is the number of documents in the training set belonging to class  $c$ .

$N$  is the total number of documents in the training set.

- The probability  $P(w_1, w_2, w_3 \dots, w_n | c)$  is computed as the product of the individual probabilities:

$$P(w_1, w_2, w_3 \dots, w_n | c) = \prod_{i=1}^n P(w_i | c)$$

Where:

The probability  $P(w_i | c)$  corresponds to the **relative frequency** of the word  $w_i$  among all words in documents belonging to class  $c$ :

$$P(w_i | c) = \frac{\text{count}(w_i, c) + 1}{\sum_{w \in V} \text{count}(w, c) + V}$$

Where:

$V$  represents the **vocabulary** of the **training set**.

- The final equation for determining the most probable class  $c$  according to the **Naive Bayes classification** is as follows:

$$\hat{c} = \arg \max_{c \in C} P(c) \prod_{i=1}^n P(w_i | c)$$

- Note:**

- From the training set, we need to **remove** the symbols `<s>`, `</s>`, as well as stop words using `nltk.corpus.stopwords.words('english')`

### 3. Implementation :

#### A. Visualization of the article numbers from volume 18 along with their labels.

## Part 5 \ Text Classification ↵

### Naive Bayes classifier

Volume n°:  - +

Content of one volume  Content of all articles

Normalization

Lemmatization and Stemming:

| Nº Article | Nº label | Label                   |
|------------|----------|-------------------------|
| 1          | 1        | Metaheuristics          |
| 2          | 2        | Machine & Deep Learning |
| 3          | 1        | Metaheuristics          |
| 4          | 1        | Metaheuristics          |
| 5          | 1        | Metaheuristics          |

Visualization

Training

Testing

Test article n°:  - +

The screenshot shows a user interface for a Naive Bayes classifier. At the top, there are three input fields: 'Volume n°:' with a value of '18', a 'Content of one volume' radio button (selected), and a 'Normalization' checkbox. Below these are two dropdown menus for 'Lemmatization and Stemming' and 'Porter Stemmer'. A large table lists five articles with their labels: article 1 is 'Metaheuristics', article 2 is 'Machine & Deep Learning', and articles 3, 4, and 5 are 'Metaheuristics'. To the right of the table is a vertical stack of three buttons: 'Visualization' (highlighted in yellow), 'Training', and 'Testing'. At the bottom is a 'Test article n°:' field with a value of '1' and a +/- button.

## B. Learning (or estimating) the probabilities $P(c)$ and $P(w_i|c)$

### Naive Bayes classifier ↗

Volume n°:

Content of one volume    Content of all articles

18   -   +

Normalization

Lemmatization and Stemming:

Porter Stemmer

Estimating the class probabilities  $P(c)$

| 1st Class | 2nd Class | 3rd Class | 4th Class |
|-----------|-----------|-----------|-----------|
| 0.4359    | 0.265     | 0.1453    | 0.1538    |

Visualization

Training

Testing

Test article n°:

1   -   +

Estimating the conditional probabilities  $P(w|c)$

| $P(w c)$     | 1st Class | 2nd Class | 3rd Class | 4th Class |
|--------------|-----------|-----------|-----------|-----------|
| optimization | 0.014     | 0.0012    | 0.0054    | 0.0019    |
| algorithm    | 0.0131    | 0.0023    | 0.0066    | 0.0015    |
| algorithms   | 0.0068    | 0.0016    | 0.0047    | 0.0007    |
| performance  | 0.0056    | 0.0022    | 0.003     | 0.0016    |
| proposed     | 0.0056    | 0.0035    | 0.0038    | 0.0031    |