Text Classification using Naïve Bayes:

1. Data:

The data are attachment 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 :

1st class : Metaheuristics
2nd class : Machine & Deep Learning
3rd class : Combination of Metaheuristics & Machine/Deep Learning
4th 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** $\boldsymbol{d}$, **identifies** among all possible **classes** $\boldsymbol{c} \in \boldsymbol{C}$ **the one** that is the **most probable**, i.e., the **class** $\boldsymbol{c}$ having the **highest probability**:

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

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

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

  - The document $\boldsymbol{d}$ can be represented as a **set of words**:

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

  - The probability $\boldsymbol{P}(\boldsymbol{c})$ represents the **proportion of documents** in the **training set** that **belong** to **class** $\boldsymbol{c}$:

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

Where:

$\boldsymbol{N_c}$ is the **number of documents** in the **training set belonging** to class $\boldsymbol{c}$.
$\boldsymbol{N}$ 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:

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

Where:

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

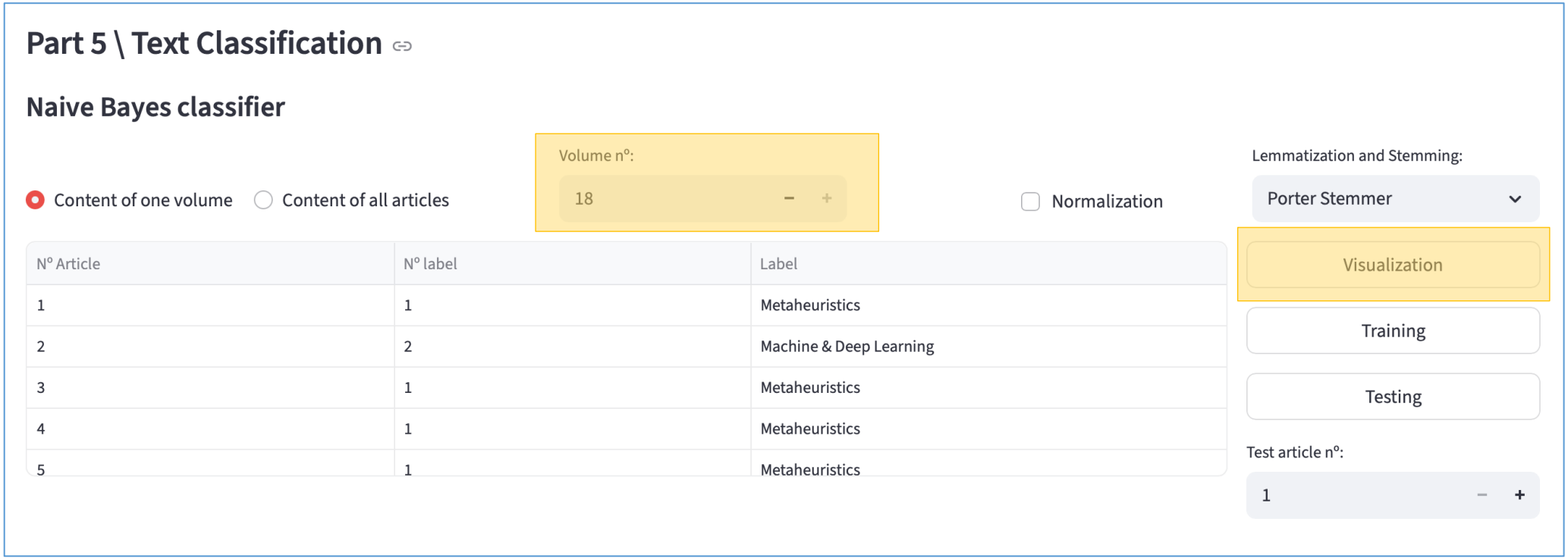
- The **final equation** for **determining** the **most probable class** c according to the **Naive Bayes classification** is as follows:

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

- **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**.



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

Naive Bayes classifier

Volume n°: 18

Content of one volume / Content of all articles

Normalization

Lemmatization and Stemming: Porter Stemmer

Visualization

Training

Testing

Test article n°: 1

Estimating the class probabilities P(c)

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

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