Echahid Hamma Lakhdar University - El Oued

Faculty of Exact Sciences

Department of Computer Science ***2nd master's degree***

Duration: 1 hour 30 minutes ***Date 07/01/2025***

Natural Language Processing (NLP) correction Exam

Questions (5 pts)

1. What are traditional approaches to text representation ?
   1.1. One-Hot Encoding
   1.2. Bag of Word (Bow)
   1.3. Term frequency-inverse document frequency (TF-IDF)
2. How are Word Embeddings used in Natural Language Processing ?
   2.1 They are used as input to machine learning models.
   Take the words →Give their numeric representation →Use in training or inference.
   2.2 To represent or visualize any underlying patterns of usage in the corpus that was used to train them.
3. Define Rule-Based Natural Language Processing Methods.
   A rule-based NLP model is a system that relies on a set of ules to perform a specific task. The rules can be based on syntax, semantics, morphology, phonology, or pragmatics of the language.
4. What are the advantages of Recurrent Neural Network (RNN) Models?

RNN can capture the context provided by particular text sequence. They learn the sequential structure of the data, where every word is dependent on the previous word or a word in the previous sentence.

5. How did transformer architecture revolutionized NLP ?

6. Transformers rely entirely on a self-attention mechanism that processes all words in parallel (instead of sequentially), which can significantly increase training speed and reduce inference cost

Exercice 1 (5 pts)

The following formula shows the Naive Bayes's theorem.

P(H | D) = (P(D | H) * P(H)) / P(D)

1. Give a short description of the formula.

- P(H | D) = Probability of hypothesis (H) given the data (D)
- P(D | H) = Probability of data (D) given the hypothesis (H)
- P(H) = Prior probability of hypothesis (H)
- P(D) = Prior probability of data (D)

2. Give steps for applying Multinomial Naive Bayes to NLP problems.

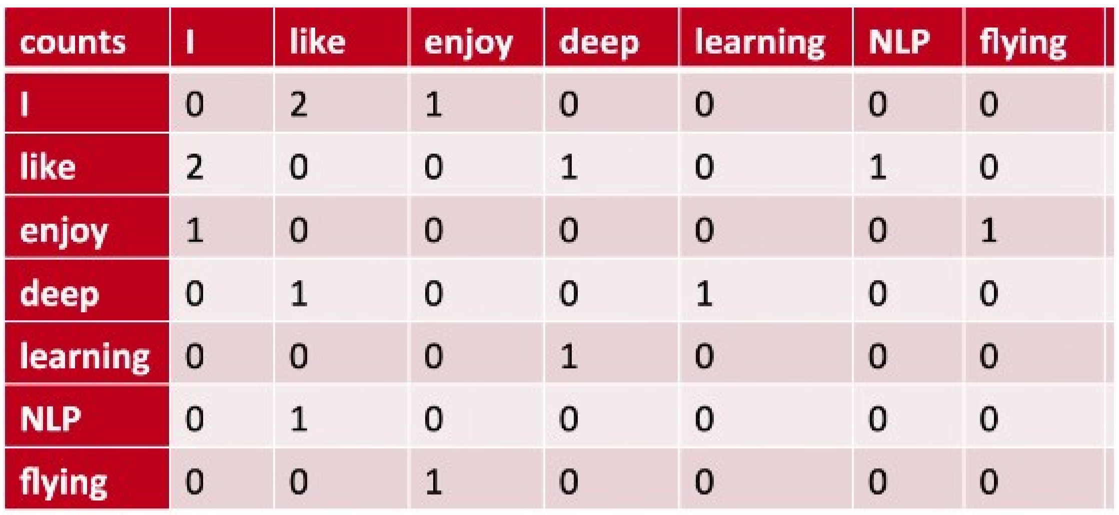
– Preprocessing the text data: The text data needs to be preprocessed efore applying the algorithm. This involves steps such as tokenization, stop-word removal, stemming, and lemmatization.
– Feature extraction: The text data needs to be converted into a feature vector format that can be used as input to the MNB algorithm. The most common method of feature extraction is to use a bag-of-words model, where each document is represented by a vector of word frequency counts.
– Splitting the data: The data needs to be split into training and testing sets. The training set is used to train the MNB model, while the testing set is used to evaluate its performance.
– Training the MNB model: The MNB model is trained on the training set by estimating the probabilities of each feature given each class. This involves calculating the prior probabilities of each class and the likelihood of each feature given each class.
– Evaluating the performance of the model: The performance of the model is evaluated using metrics such as accuracy, precision, recall, and F1-score on the testing set.
– Using the model to make predictions: Once the model is trained, it can be used to make predictions on new text data.

Exercice 2 (5 pts)

Consider the following sentences:
Example Corpus: I like deep learning. I like NLP. I enjoy flying.

1. Find the co-occurrence matrix for the above corpus.

From the above corpus, the list of unique words present are as follows:
Dictionary: ['I', 'like', 'enjoy', 'deep', 'learning', 'NLP', 'flying']



counts	I	like	enjoy	deep	learning	NLP	flying
I	0	2	1	0	0	0	0
like	2	0	0	1	0	1	0
enjoy	1	0	0	0	0	0	1
deep	0	1	0	0	1	0	0
learning	0	0	0	1	0	0	0
NLP	0	1	0	0	0	0	0
flying	0	0	1	0	0	0	0

2. Breifly explain how to find similarity between words using.

   Find distance between two term vectors.

   A) Lp Norm approach

   Lp Norm approach: This is basically the pth root of the summation (difference of each element of two term vectors i,j for term k) ^ p

   – Manhattan distance with p = 1

   – Euclidean distance with p =2

   B) Cosine similarity

   Cosine similarity: to find angular distance between two term vectors

Exercice 3 (5 pts)

Let « text » represent the variable containing the corpus

Write python code that :

1. Remove :
   A) Punctuation
   B) Special Characters
   C) HTML Tags
   D) Unwanted Words: stop Words
2. Changing to lowercase.
3. Word tokenize the text.

```
import re


import nltk

nltk.downoad('stopwords')

from nltk.corpus import stopwords


def clean_text(text):
```

```
# 1. Remove HTML tags
text = re.sub(r'<[^>]*>', '', text)

# 2. Remove punctuation and special characters
text = re.sub(r"[^\w\s]", '', text)

# 3. Convert to lowercase
text = text.lower()

# 4. Split the text into words (basic tokenization)
words = text.split()

# 5. Remove stop words
stop_words = set(stopwords.words("english"))
filtered_words = [word for word in words if word not in stop_words]

return filtered_words
```