
IRT
SAINT
EXUPÉRY

Text vectorization Exercices

Master ValDom – 10/01/2025

In blue: statements

In black: correction

Exercice

Create the vocabulary (size 5) for the corpus "banana bandana banner" **using Byte Pair Encoding (BPE) technical:**

Step1: Initialize Tokenization

banana -> b a n a n a
bandana -> b a n d a n a e r
banner -> b a n n e r

Vocabulary = [‘b’,’a’,’n’,’d’,’r’,’e’]
Size = 6

Step2: Count Pair Frequencies

(b, a) -> 3
(a, n) -> 5
(n, a) -> 3
(n, d) -> 1
(d, a) -> 1
(a,e) -> 1
(e, r) -> 2
(n, n) -> 1
(n, e) -> 1

Exercice

Create the vocabulary (size 5) for the corpus "banana bandana banner" **using Byte Pair Encoding (BPE) technical:**

Step 3: Merge the Most Frequent Pair

b a n a n a -> b an an a
b a n d a n a -> b an d an a e r
b a n n e r -> b an n e r

Vocabulary = [‘b’,’an’,’a’,’d’,’n’,’e’, ’r’]
Size = 7

Step4: Count New Pair Frequencies

(b, an) -> 3
(an, an) -> 1
(an, a) -> 2
(an, d) -> 1
(d, an) -> 1
(a, e) -> 1
(e, r) -> 2
(an, n) -> 1
(n, e) -> 1

Exercice

Create the vocabulary (size 5) for the corpus "banana bandana banner" **using Byte Pair Encoding (BPE) technical:**

Step 5: Merge the Most Frequent Pair

b an an a -> ban an a
b an d an a -> ban d an a e r
b an n e r -> ban n e r

Vocabulary = [‘ban’,’an’,’a’,’d’,’n’,’e’, ’r’]
Size = 7

Step6: Count New Pair Frequencies

(ban, an) -> 1
(an, a) -> 2
(ban, d) -> 1
(d, an) -> 1
(a, e) -> 1
(e,r) -> 2
(ban, n) -> 1
(n,e) -> 1

Exercice

Create the vocabulary (size 5) for the corpus "banana bandana banner" **using Byte Pair Encoding (BPE) technical:**

Step 7: Merge the Most Frequent Pair

ban an a -> ban ana
ban d an a –> ban d ana e r
ban n e r -> ban n e r

Vocabulary = [‘ban’,’ana’,’d’,’n’, ’e’,’r’]
Size = 6

Step8: Count New Pair Frequencies

(ban, ana) -> 1
(ban, d) -> 1
(d, ana) -> 1
(ana, e) -> 1
(e,r) -> 2
(ban, n) -> 1
(n,e) -> 1

Exercice

Create the vocabulary (size 5) for the corpus "banana bandana banner" **using Byte Pair Encoding (BPE) technical:**

Step 8: Merge the Most Frequent Pair

ban an a -> ban ana
ban d an a –> ban d ana
ban n e r -> ban n er

Vocabulary = [‘ban’,’ana’,’d’,’n’,’er’]
Size = 5

Step8: Final Vocabulary size 5

ban ana d n er

Exercice – Text Cleaning

- **Remove punctuation and emoji.**
- **Lowercase**
- **Use stopwords: [“a”, “in”, “the”, “to”]**
- **Use Stemming**

Corpus

Document 1:
"A dog runs in the park every morning."

Document 2:
"Every dog loves to run."

Document 3:
"I have 3 very nice dogs, I love them. 😀"

Exercice – Text Cleaning

- **Remove punctuation and emoji.**
- **Lowercase**
- **Use stopwords: [“a”, “in”, “the”, “to”]**
- **Use Stemming**

Cleaned corpus

Document 1:
"dog run park every morning"

Document 2:
“every dog love run"

Document 3:
“i have 3 very nice dog i love them”

Exercice – Tokenization

Tokenize documents with Whitespace Tokenization

Cleaned corpus

Document 1:
"dog run park every morning"

Document 2:
“every dog love run"

Document 3:
“i have 3 very nice dog i love them”

Exercice – Tokenization

Tokenize documents with Whitespace Tokenization

Tokenized corpus

Document 1:
["dog”, “run”, “park”, “every”, “morning”]

Document 2:
[“every”, “dog”, “love”, “run”]

Document 3:
[“i”, “have”, “3”, “very”, “nice”, “dog”, “i”, “love”, “them”]

Exercice – Bag of Words (BoW)

Create Bag of Words representation for each document

Tokenized corpus

Document 1:
["dog", “run”, “park”, “every”, “morning”]

Document 2:
[“every”, “dog”, “love”, “run”]

Document 3:
[“i”, “have”, “3”, “very”, “nice”, “dog”, “i”, “love”, “them”]

Exercice – Bag of Words (BoW)

Create Bag of Words representation for each document

Vocabulary: ["dog”, “run”, “park”, “every”, “morning”, “love”, “i”, “have”, “3”, “very”, “nice”, “them”]

Vectorized corpus

Document 1:
[1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0]

Document 2:
[1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0]

Document 3:
[1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1]

Exercice – TF-IDF

Calculate de TF-IDF value for “dog” and “run” for each document.

Tokenized corpus

Document 1:
["dog”, “run”, “park”, “every”, “morning”]

Document 2:
[“every”, “dog”, “love”, “run”]

Document 3:
[“i”, “have”, “3”, “very”, “nice”, “dog”, “i”, “love”, “them”]

$$\mathrm{TF}(t) = \frac{\text{Count of term } t \text{ in the document}}{\text{Total terms in the document}}$$

$$\mathrm{IDF}(t) = \log\left(\frac{\text{Number of documents}}{\text{Number of documents containing } t}\right)$$

$$\text{TF-IDF}(t) = \mathrm{TF}(t) \times \mathrm{IDF}(t)$$

Exercice – TF-IDF

Calculate de TF-IDF value for “dog” and “run” for each document.

Tokenized corpus

$$\mathrm{TF}(t) = \frac{\text{Count of term } t \text{ in the document}}{\text{Total terms in the document}}$$

Document 1:
["dog”, “run”, “park”, “every”, “morning”] TF(dog) = 1/5

Document 2:
[“every”, “dog”, “love”, “run”] TF(dog) = 1/4

Document 3:
[“i”, “have”, “3”, “very”, “nice”, “dog”, “i”, “love”, “them”] TF(dog) = 1/9

Exercice – TF-IDF

Calculate de TF-IDF value for “dog” and “run” for each document.

Tokenized corpus

Document 1:
["dog”, “run”, “park”, “every”, “morning”]

Document 2:
[“every”, “dog”, “love”, “run”]

Document 3:
[“i”, “have”, “3”, “very”, “nice”, “dog”, “i”, “love”, “them”]

$$\mathrm{IDF}(t) = \log\left(\frac{\text{Number of documents}}{\text{Number of documents containing } t}\right)$$

IDF(dog) = log(3/3) = 0

Exercice – TF-IDF

Calculate de TF-IDF value for “dog” and “run” for each document.

$$\text{TF-IDF}(t) = \text{TF}(t) \times \text{IDF}(t)$$

Tokenized corpus

Document 1:
["dog”, “run”, “park”, “every”, “morning”] TF-IDF(dog) = 1/5 x 0 = 0

Document 2:
[“every”, “dog”, “love”, “run”] TF-IDF(dog) = 1/4 x 0 = 0

Document 3:
[“i”, “have”, “3”, “very”, “nice”, “dog”, “i”, “love”, “them”] TF-IDF(dog) = 1/9 x 0 = 0

Exercice – TF-IDF

Calculate de TF-IDF value for “dog” and “run” for each document.

Tokenized corpus

$$\mathrm{TF}(t) = \frac{\text{Count of term } t \text{ in the document}}{\text{Total terms in the document}}$$

Document 1:
["dog”, “run”, “park”, “every”, “morning”] TF(run) = 1/5

Document 2:
[“every”, “dog”, “love”, “run”] TF(run) = 1/4

Document 3:
[“i”, “have”, “3”, “very”, “nice”, “dog”, “i”, “love”, “them”] TF(run) = 0/9

Exercice – TF-IDF

Calculate de TF-IDF value for “dog” and “run” for each document.

Tokenized corpus

Document 1:
["dog”, “run”, “park”, “every”, “morning”]

Document 2:
[“every”, “dog”, “love”, “run”]

Document 3:
[“i”, “have”, “3”, “very”, “nice”, “dog”, “i”, “love”, “them”]

$$\mathrm{IDF}(t) = \log\left(\frac{\text{Number of documents}}{\text{Number of documents containing } t}\right)$$

IDF(run) = log(3/2) = 0.176

Exercice – TF-IDF

Calculate de TF-IDF value for “dog” and “run” for each document.

Tokenized corpus

$$\text{TF-IDF}(t) = \text{TF}(t) \times \text{IDF}(t)$$

Document 1:
["dog”, “run”, “park”, “every”, “morning”]

TF-IDF(run) = 1/5 x 0.176 = 0.0352

Document 2:
[“every”, “dog”, “love”, “run”]

TF-IDF(run) = 1/4 x 0.176 = 0.44

Document 3:
[“i”, “have”, “3”, “very”, “nice”, “dog”, “i”, “love”, “them”]

TF-IDF(run) = 0/9 x 0.176 = 0

Exercice – N-Grams

Create N-grams representation for each document with n = 2

Cleaned corpus

Document 1:
"dog run park every morning"

Document 2:
“every dog love run"

Document 3:
“i have 3 very nice dog i love them”

Exercice – N-Grams

Create N-grams representation for each document with n = 2

Step 1: Tokenization

Tokenized corpus

Document 1:
["dog run”, “run park”, “park every”, “every morning”]

Document 2:
[“every dog”, “dog love”, “love run”]

Document 3:
[“i have”, “have 3”, “3 very”, “very nice”, “nice dog”, “dog i”, “i love”, “love them”]

Exercice – N-Grams

Create N-grams representation for each document with n = 2

Step 2: Vectorization

Vocabulary: ["dog run”, “run park”, “park every”, “every morning”, “every dog”, “dog love”, “love run”, “i have”, “have 3”, “3 very”, “very nice”, “nice dog”, “dog i”, “i love”, “love them”]

Vectorized corpus

Document 1:
[1,1,1,1,0,0,0,0,0,0,0,0,0,0,0]

Document 2:
[0,0,0,0,1,1,1,0,0,0,0,0,0,0,0]

Document 3:
[0,0,0,0,0,0,0,1,1,1,1,1,1,1,1]