


Applied Text Analytics & Natural Language Processing



with Dr. Mahdi Roozbahani
& Wafa Louhichi

One Hot Encoding



Learning Objectives

This week you will be introduced to different discrete text representations. At the end of this lesson, you will be able to:

- Explain **text data** into numerical format using three methods: One Hot Encoding, Bag of words, and TF-IDF
- Explain **documents** into numerical format
- Understand the advantages and disadvantages of each method

Why Numerical Text Representation?

- The goal of NLP is to be able to design algorithms to allow computers to "understand" natural language in order to perform some task
- Computers are good with numbers, so how do we convert text data to numerical data that can be used in a model?

Representing Words

- Every word can be represented by a vector of 0 except for one position that has a value of 1
- Example: This is a simple sentence
- "This" -> [1,0,0,0,0]
- "is" -> [0,1,0,0,0]
- "a" -> [0,0,1,0,0]
- "simple" -> [0,0,0,1,0]
- "sentence" -> [0,0,0,0,1]

From Word to Document Representation

- Our sentence is thus represented as:

"This is a simple sentence" -> $\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$

```
1 from sklearn.preprocessing import LabelEncoder
2 from sklearn.preprocessing import OneHotEncoder
3
4 text="This is a simple sentence"
5 # integer encode
6 label_encoder = LabelEncoder()
7 integer_encoded = label_encoder.fit_transform(text.split())
8 # binary encode
9 onehot_encoder = OneHotEncoder()
10 onehot_encoded = onehot_encoder.fit_transform(integer_encoded.reshape(-1, 1))
11 onehot_encoded.toarray()

array([[1., 0., 0., 0., 0.],
       [0., 0., 1., 0., 0.],
       [0., 1., 0., 0., 0.],
       [0., 0., 0., 0., 1.],
       [0., 0., 0., 1., 0.]])
```


One-Hot Encoder

- In a corpus with a vocabulary V of size d (number of unique words or dimensions), a word w is represented as a vector X of size d such as:

$$X_i^w = \begin{cases} 1, & \text{if } idx(w) = i \\ 0, & \text{otherwise} \end{cases}$$

- A document can be represented as matrix of size $n \times d$, where n is the number of words in the document, or a single vector of dimension d , with multiple values of 1 where the words from the vocabulary are present.

Document: $D = \text{this is a sentence}$

Vocabulary: $V = [\text{aardvak}, \dots, \text{this}, \dots, \text{is}, \dots, \text{a}, \dots, \text{sentence}, \dots, \text{zyther}]$

One Hot Encoding: $X^D = [0, \dots, 1, \dots, 1, \dots, 1, \dots, 1, \dots, 0]$

Advantages and Disadvantages of One-Hot Encoding

- Advantages of One-Hot Encoding:
 - Simple and easy to implement
- Disadvantages of One-Hot Encoding:
 - Every word is represented as a vector of the size of the vocabulary: not scalable for a large vocabulary (100,000 words)
 - High dimensional sparse matrix which can be memory & computationally expensive
 - Every word is represented independently: there is no notion of similarity/meaning in one-hot encoding. All the vectors are orthogonal
 - Example: Despite the words "*good*" and "*great*" carry similar meaning, and the word "*bad*" carries the opposite meaning, the dot product is 0

$$(W^{good})^T \cdot W^{great} = (W^{good})^T \cdot W^{bad} = 0$$

Advantages and Disadvantages of One-Hot Encoding

- Advantages of One-Hot Encoding:
 - Simple and easy to implement
- Disadvantages of One-Hot Encoding:
 - Every word is represented as a vector of the size of the vocabulary: not scalable for a large vocabulary (100,000 words)
 - High dimensional sparse matrix which can be memory & computationally expensive
 - Every word is represented independently: there is no notion of similarity/meaning in one-hot encoding. All the vectors are orthogonal
 - Example: Despite the words "*good*" and "*great*" carry similar meaning, and the word "*bad*" carries the opposite meaning, the dot product is 0

$$(W^{good})^T \cdot W^{great} = (W^{good})^T \cdot W^{bad} = 0$$

Summary

- Use One-Hot Encoding as a discrete text representation method to represent words and documents
- Understand the advantages and disadvantages of this method