NLP

Word Representation and Vectorization

Week 1: Natural Language Processing

Learning Objectives

Upon completion, you should be able to:

- Define Word Representation and Vectorization
- Demonstrate Vectorization hands-on
- Explain Text classification and Sentiment Analysis
- Demonstrate Sentiment Analysis hands-on
- Illustrate Dense Encoding
- Explain Word2Vec Vectorization
- Define GloVe
- Demonstrate Dense Embeddings hands-on

Agenda

- Context and background
- Introduction to Vectorization

Introduction to Vectorization

- Machines, as we all know, **cannot really understand text as input**.

- In order to perform Machine Learning on text, we need to **convert text into a numerical format that machines can understand in order to find patterns and make predictions.**

  So, can't we use one-hot encoding to convert text data into integers?

Introduction to Vectorization

- **Can't we use one-hot encoding to convert text data into integers?**
  Let's consider the below example of a one-hot encoded representation of the text:

"The Queen has entered the room" —After Preprocessing→ **"The Queen has entered room"**

The	Queen	has	entered	room
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

We can see that in one-hot encoding, **every single word in our text corresponds to a vector element.**
For ex: Here "Queen" is [0,1,0,0,0] and "room" is [0,0,0,0,1]

Introduction to Vectorization

- **Can't we use one-hot encoding to convert text data into integers?**

  Therefore, the result of one-hot encoding is a **sparse matrix**, or in other words, a matrix where the vast majority of the elements are zeroes, such as the matrix we have just created for the sentence below.

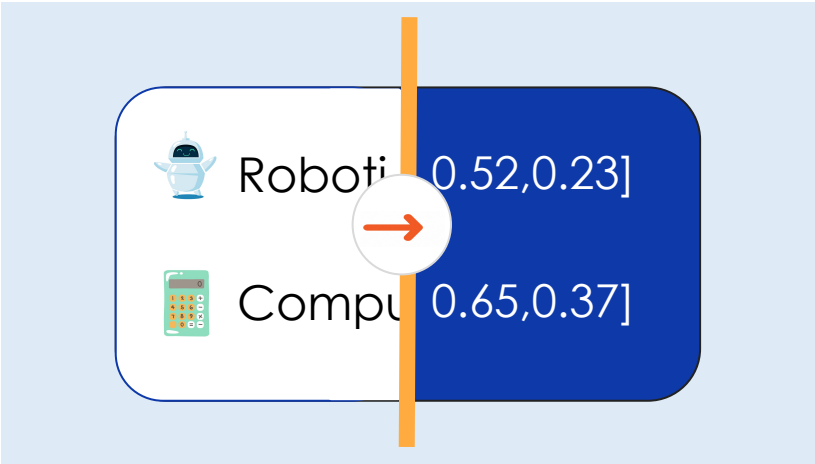
"The Queen has entered room"

The	Queen	has	entered	room
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

Introduction to Vectorization

- **Can't we use one-hot encoding to convert text data into integers?**

  One disadvantage of working with sparse matrices is that they end up **consuming a lot of memory, especially as the size of the corpus increases**, making it **computationally expensive** to work with them.
  This is where **vectorization comes into the picture.**



What is Vectorization?

- In programming, a **vector is a data structure similar to an array** used for the purposes of input representation that **computers can easily process**.
- The **process of converting text data into vector format** can be referred to as **Vectorization**.
- There are various vectorization methods available for turning text into numerical features. In this lecture, we'll take a quick look at the following **traditional count-based approaches.**

1. **Bag of Words (BoW)**
2. **Term Frequency - Inverse Document Frequency (TF-IDF)**

Introduction to Vectorization

Before diving deeper into vectorization techniques, let's understand the below terms:

- **What is Vocabulary?**

- The set of unique words used in the corpus after preprocessing the given text data is called the vocabulary.

- **What is the size of the Vocabulary?**

- The number of unique words in the vocabulary after preprocessing is called the size of the vocabulary.

Introduction to Vectorization

Before diving deeper into the vectorization techniques, let's understand the below terms:

- **What is the size of the Vocabulary?**

- The number of unique words in the vocabulary after preprocessing is called the size of the vocabulary.

    - Each word in a vocabulary will be represented by a vector which will have as many elements as the vocabulary size.

    - Each word gets associated with a unique index in the vector. To represent a word as a vector, the element at the index associated with a word gets a 1 while other elements get a 0.

Now, let's have a look at different vectorization techniques.

Summary

A brief recap:

- We have learned the importance of **converting text data into a numerical representation and why we need vectorization**.
- We have learned **what vectorization is,** and we have discussed some important terms in vectorization, such as **vocabulary and the size of a vocabulary.**

Great Learning

Happy Learning !