**NLP 2**

****

### **Introduction to the Series:**

Welcome to the exciting series on Natural Language Processing (NLP)! In this journey, we will unravel the secrets behind the techniques and models that power language-based applications. Here’s what you can look forward to:

1. **Text Pre-processing**
2. **Text Representation (This Article)**
3. **Word Embedding**
4. **Language Modeling**
5. **Topic Modeling**
6. **Recurrent Neural Network (RNNs) for NLP**
7. **LSTM for NLP**
8. **Transformer Models**
9. **BERT**

#### **Prerequisites:**

1. **Knowledge of Python**
2. **Text Pre-processing (Part 1)**

### **In This Article:**

Today, we’ll dive into the text representation techniques used in NLP restricted to the following topics:

* **Why do we need to represent text as numbers?**
* **How do we represent text as numbers?**
  + **Sparse representation of text**
  + **Dense representation of text**
* **One Hot Encoding**
* **Bag of Words**
* **TF IDF**
* **Cosine Similarity**
* **Case Study: Medium Article Recommendation System**

**What is Text Representation?**

**The process of converting a text into numerical vectors that the machine can understand is called as text representation in NLP.**

**Why do we need to represent text as numbers?**

**Computers know only number (binary), whatever the activity that need to be performed using it should be done through numbers. The computer understands the text using ascii codes.**

**Not like human, the machine learning model learns to do an activity based on the numbers.**

**Hence it is required to represent the text as numbers.**

**How do we represent text as numbers?**

**Broadly the texts are represented as a number in two categories.**

**Sparse Representation Techniques:**

**It refers to a way of representing data where most of the elements in the data structure are zero.**

**Dense Representation Techniques:**

**There are variety of ways that we can represent at text as a number or vector based on the requirements. The representation should be capable of represent the semantic and syntactic meaning of the word or text.**

**In the previous article, we have seen a simple word frequency dictionary based text representation technique. Here we will some advanced methods.**

**One hot encoding:**

One-hot encoding is a straightforward text representation technique where each word or token is represented by a binary vector with a single "1" indicating the presence of the word and all other positions set to "0"

**Example:**

Consider the following sentences

Sentence 1: “**it was the best of times**”

Sentence 1: “**it was the worst of times**”

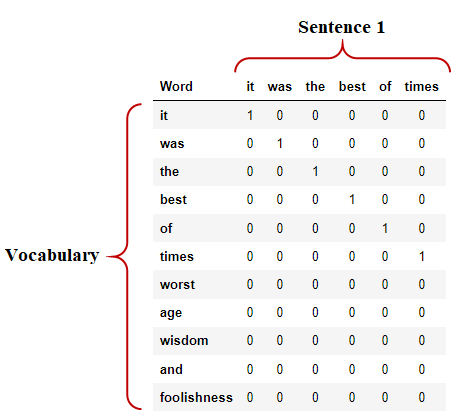
Sentence 3: “**it was the age of wisdom and the age of foolishness**”

1. **Build the vocabulary:**

* Create a list of all unique words across the entire corpus (Here 3 sentences)
* Vocabulary: [it, was, the, best, of, times, worst, age, wisdom, and, foolishness]

1. **Create one hot encoded representation:**

Each word in the vocabulary is represented by a binary vector with a size equal to the number of unique words.



**For the sentence 1,**

"it" → [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

"was" → [0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]

"the" → [0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0]

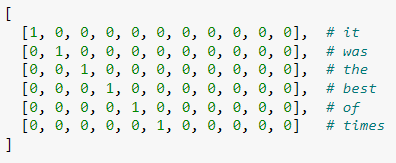
"best" → [0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0]

"of" → [0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0]

"times" → [0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0]

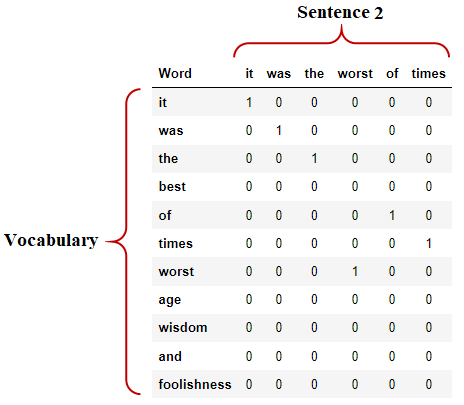
**Hence each word is represented with the size of 1 x N vector, where N is the size of the vocabulary.**

**On combining these vectors, the sentence is represented as:**



**Hence the sentence is transformed into numerical vector of size M x N, where M is the number of words in the sentence.**

**Similarly for the sentence 2, one hot encoded representation will be as,**

****

**Drawbacks:**

* **All the words in the sentence were given equal importance**
* **High dimensional sparse matrix representation**
* **The vector representation of each word is orthogonal and hence the relationship between different words cannot be measured**

**Bag of Words:**

**Some of the drawbacks are addressed in the bag of words technique.**

**Follow the same steps as that of one hot encoding technique, instead of representing the word with 1 or 0 represent it with count of the word in the sentence.**