**Lecture 5**

* **What mean by text Representation?**

**How do we transform a given text into numerical form so that is can be fed into ML algorithms**

* **What is Text Representation Approaches?**

**Basic Vectorization**

**Distributed Representation**

**Universal Language**

**Handcrafted Features**

* **How To Extract Meaning Of The Sentence (NLU)?** 
  + **Break the sentence into** **lexical units such as words, and phrases.**
  + **Derive the meaning for each of the lexical units.**
  + **Understand the syntactic(grammatical) structure of the sentence.**
  + **Understand the context in which the sentence appears.**
* **What mean by Vector Space Model ?**

**It’s a mathematical model that represent text units as vectors such as index numbers in a corpus (Cosine Similarity)**

* **What is the basic vectorization approaches ?**

**One-Hot Encoding**

**Bag of words (Bow)**

**Bag of N-grams(BoN)**

**Term Frequency – inverse Document Frequency (TF-IDF)**

**One-Hot Encoding**

**A picture containing text, screenshot, font, number

Description automatically generated**

**Disadvantages:**

**A sparse representation (many zero) inefficient to store**

**Not fixed length representation for most learning**

**Have no idea about (dis)similarity between words run,ran have the same meaning and apple can be fruit or apple technology**

**Out of vocabulary this mean the word appears in the test does not** **represent in training.**

**Bag of words (bow)**

**A picture containing text, screenshot, font, number

Description automatically generated**

**Advantages**

**Box is simple to understand and implement**

**Document having the same word will have representation closer to each other**

**The vector space captures the semantic similarity of document**

**We have a fixed-length encoding for any sentence of arbitrary length**

**Disadvantages**

**Sparsity(control by limiting number of the most frequent words)**

**Does not capture the similarity between different words.**

**Have no way to handle oov**

**Words order information is lost in this representation.**

**Bag of n-Grams(BoN)**

**It breaking text into chunks of n-contiguous words tokens capture some context which is the earlier approaches could not do**

**A screenshot of a computer

Description automatically generated with low confidence**

**Advantages :**

* **Captures some context and word-order information.**
* **Vector space captures some semantic similarity.**
* **Documents have the same n-grams will have vectors closer to each other.**

**Disadvantages**

* **N increase sparsity rapidly**
* **Its sill provides no way to handle oov**