MES Wadia College of Engineering Pune-01
Department of Computer Engineering

Name of Student:	Class:
Semester/Year:	Roll No:
Date of Performance:	Date of Submission:
Examined By:	Subject: LP_VI (EL-V NLP)

Assignment No. 2

Aim: Perform bag-of-words approach (count occurrence, normalized count occurrence), TF-IDF on data. Create embeddings using Word2Vec.
Dataset to be used: https://www.kaggle.com/datasets/CooperUnion/cardataset

Theory:

Problem Statement: Perform bag-of-words approach (count occurrence, normalized count occurrence), TF-IDF on data. Create embeddings using Word2Vec.

Bag of Words (BoW)

Bag of words is a Natural Language Processing technique of text modelling.

A problem with modeling text is that it is messy, and techniques like machine learning algorithms prefer well defined fixed-length inputs and outputs.Machine learning algorithms cannot work with raw text directly; the text must be converted into numbers. Specifically, vectors of numbers. This is called feature extraction or feature encoding.

The bag-of-words model is a way of representing text data when modeling text with machine learning algorithms. It is a popular and simple method of feature extraction from text data.

A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves two things:

1. A vocabulary of known words.
2. A measure of the presence of known words.

It is called a “bag” of words, because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document.

The most common kind of characteristic, or feature calculated from the Bag-of-words model is term frequency, which is essentially the number of times a term appears in the text. Term frequency is not necessarily the best representation for the text, but it still does find successful applications in areas like email filtering. Term frequency isn’t the best representation of the text because common words such as "the", "a", "to" are almost always the terms with highest frequency in the text. This shows that having a high raw count does not necessarily indicate that the corresponding word is more important.

Advantages of BoW Approach

The most significant advantage of the bag-of-words model is its simplicity and ease of use. It can be used to create an initial draft model before proceeding to more sophisticated word embeddings.

Disadvantages of BoW Approach

- **Vocabulary**: The vocabulary requires careful design, most specifically in order to manage the size, which impacts the sparsity of the document representations.
- **Sparsity**: Sparse representations are harder to model both for computational reasons (space and time complexity) and also for information reasons, where the challenge is for the models to harness so little information in such a large representational space.
- **Meaning**: Discarding word order ignores the context, and in turn meaning of words in the document (semantics). Context and meaning can offer a lot to the model, that if modeled could tell the difference between the same words differently arranged ("this is interesting" vs "is this interesting"), synonyms ("old bike" vs "used bike"), and much more.

TF-IDF

Term Frequency - Inverse Document Frequency (TF-IDF) is a widely used statistical method in natural language processing and information retrieval. It measures how important a term is within a document relative to a collection of documents (i.e., relative to a corpus). Words within a text document are transformed into importance numbers by a text vectorization process. There are many different text vectorization scoring schemes, with TF-IDF being one of the most common.

Term Frequency: TF of a term or word is the number of times the term appears in a document compared to the total number of words in the document.

$$Term\ Frequency = \frac{number\ of\ instances\ of\ word\ w\ in\ document\ d}{total\ number\ of\ words\ in\ document\ d}$$

Inverse Document Frequency: IDF of a term reflects the proportion of documents in the corpus that contain the term. Words unique to a small percentage of documents (e.g., technical jargon terms) receive higher importance values than words common across all documents (e.g., a, the, and).

$$IDF = log\left(\frac{total\ number\ of\ documents\ (N)\ in\ text\ corpus\ D}{number\ of\ documents\ containing\ w}\right)$$

The TF-IDF of a term is calculated by multiplying TF and IDF scores.

TF-IDF = TF * IDF

Importance of a term is high when it occurs a lot in a given document and rarely in others. In short, commonality within a document measured by TF is balanced by rarity between documents measured by IDF. The resulting TF-IDF score reflects the importance of a term for a document in the corpus.

TF-IDF is useful in many natural language processing applications. For example, Search Engines use TF-IDF to rank the relevance of a document for a query. TF-IDF is also employed in text classification, text summarization, and topic modeling.

Objective:-

The objective of the experiment is to understand the fundamental concepts and techniques of natural language processing (NLP)

Procedure:-

Bag-of-words example

Let's assume we have three sentences in our vocabulary.

Sentence 1: Data science is fun and interesting

Sentence 2: Data science is fun

Sentence 3: science is interesting

The unique words in the sentences are : [data, science, is, fun, and, interesting]. Hence, the bag of words vectors for the above sentences will be

Sentence 1: [1, 1, 1, 1, 1, 1]

Sentence 2: [1, 1, 1, 1, 0, 0]

Sentence 3: [1, 1, 1, 1, 0, 0]

Questions:

1. Compare syntactic analysis with semantic analysis. 4
2. Elaborate syntactic representation of natural language. 3
3. Describe parsing algorithms in detail. 3
4. Write short note on: 6

4.1 Probabilistic Context-Free Grammar
4.2 Statistical Parsing
4.3 Lexical semantic
4.4 Dictionary based approach

5. Discuss relations among lexemes and their senses: 6

a. Homonymy
b. Polysemy
c. Synonymy
d. Hyponymy
e. Wordnet
f. Word Sense Disambiguation(WSD)