Lecture 9

-NLP is all about words, their arrangement and their meaning

-NLP is a branch of AI that tries to emulate/understand the way a human speaks to another human

-Lecture 9 is about Feature extraction from text

Dr. Seba Susan, Delhi Technological University

BoW (Bag of Words) feature representation

-One-hot encoding, TF, TF-IDF (3 TYPES OF BoW feature vectors)

- Consider a text document. (INPUT)
- Your dataset contains lot of such documents
- Task: Construct a feature vector for each document
- After tokenization and stopword removal we obtain keywords from a document.
- Repeat for all documents
- Collect all the keywords from all documents (with no repetitions)
- Arrange the keywords along columns and documents along rows
- Each row is a BoW representation for a document

One-hot encoding feature vector

- Binary features 0 or 1
- Document x Keyword matrix
- Each cell in this matrix contains either:
- 0 (keyword is not there in the document)
- 1 (keyword is there in the document)

TF (Term Frequency) feature vector

- Features (d)= count of a keyword w in a document d
- Document x Keyword matrix
- Each cell in this matrix contains either:
- Count

$$tf(w,d) = count(w)\big|d$$

TF – IDF feature vector

- TF: Term Frequency
- IDF: Inverse document Frequency
- Let Nt be total number of documents
- Let Nw be the total number of documents containing the keyword w
- TF-IDF formula:

$$tfidf(w,d) = tf(w,d) \times idf(w)$$

$$idf(w) = \log \frac{N_t}{N_w}$$

Other popular feature representations

- Bag of Characters
- Bag of Phrases (n-gram models) [Note: BoW is unigram model]
- Word2vec
- BERT
- GLOVE

Answer the Programming Assignment of this week related to Lecture 9 in the link provided in MOODLE (by Monday 14th Sept’20 11 pm).