NLP ASSIGNMENT 1

Q1) onehot encoding converts categorical data into binary columns which allows us to treat categorical data as binary classification. Many machine learning models can now be applied.

Q2) a sentence or document is represented just as bag of words irrespective of grammar and order. But multiplicity of words is counted and kept.

Q3) a sentence or document is represented as bag of ngrams irrespective of order.

Q4) tf-idf gives importance of words in documents based on frequency of words.

Q5) words not found in vocabulary create out of vocabulary problem.

Q6) converting words into vectors so that words which are closer in vector space have similar meaning is called word embedding. Since words cannot be used as they are, they have to be converted into some form for machine learning or natural language processing.

Q7) continuous bag of words also tries to understand the context of words in addition to collecting sentences or documents as bag of words.

Q8) skipgram finds most relevant words for a given word.

Q9) glove is an algorithm for generating word embeddings by aggregating global word concurrence matrices.