NLP ASSIGNMENT 2

Q1) corpora is a large collection of documents and structured texts used in NLP for analysis and hypothesis testing.

Q2) reducing the piece of text into smaller parts like words, characters or subtexts.

Q3) unigram is one word, bigrams is two words and trigrams is three words.

Q4) n grams are generated by splitting text with space.

Q5) lemmatisation gives base word of a word.

Q6) stemming means chopping off some end part of word.

Q7) it tags the part of speech of word like verbs, nouns etc in much deeper detail.

Q8) chunking identifies parts of speech, links them to their grammatical groups and tries to make sense of grammar.

Q9) noun phrase chunking identifies groups of words that go together to form symbolic meaning.

Q10) NER classifies named entities based on predefined groups like persons, organisations, countries etc.