1.what are corpora ?

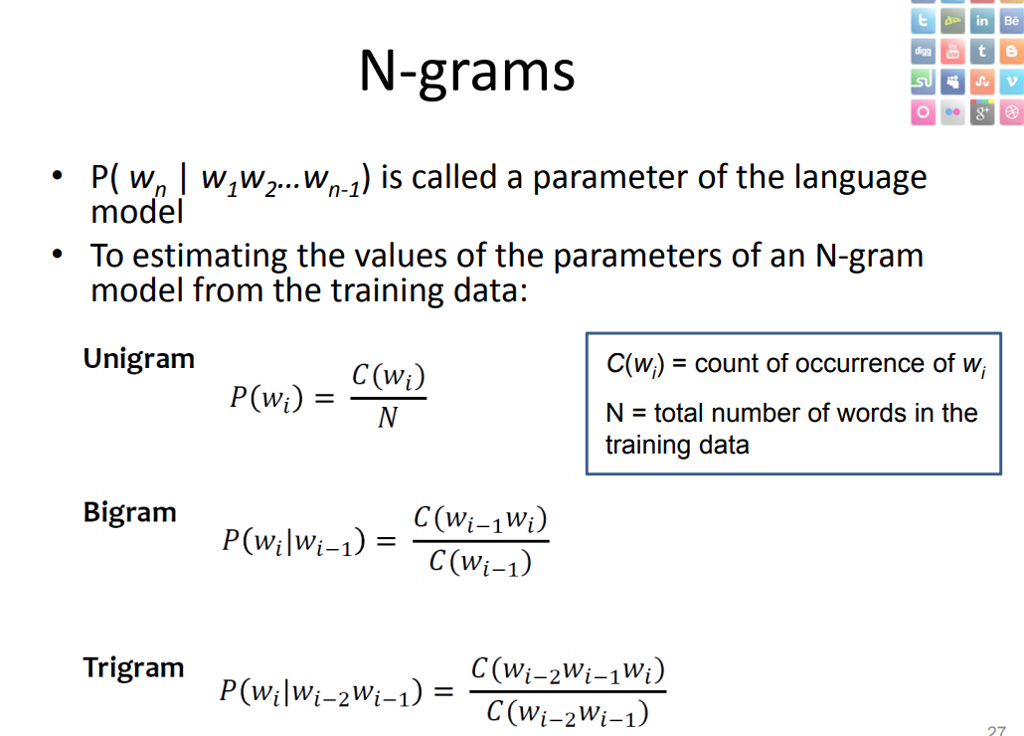
Ans : corpora can contain written or spoken (transcribed) language, modern or old texts, texts from one language or several languages. The texts can be whole books, newspapers, journals, speeches etc, or consist of extracts of varying length.

2.what are Tokens ?

Ans : A token is a instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing.

3.what are the unigrams,bigrams,trigrams ?

Ans : They include unigrams (one word), bigrams (two adjacent words), trigrams (three adjacent words) and 4-grams (four adjacent words).



4 .How to generate n grams from text ?

Ans :

N-gram can be defined as the contiguous sequence of n items from a given sample of text or speech. The items can be letters, words, or base pairs according to the application. The N-grams typically are collected from a text or speech corpus.

P(A|B) = \frac{P(A,B)}{P(B)}\\

P(A,B) = P(A|B)P(B)\\

5.Explain the Lemmatization.

Ans : Lemmatization refers to things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word.

Ex: good with words like better & best.

6.Explain the stemming.

Ans : Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma.

Ex : chocolates-choco.

7.Explain the part of speech(POS) tagging.

Ans : Part-of-speech (POS) tagging is a popular Natural Language Processing process which refers to categorizing words in a text (corpus) in correspondence with a particular part of speech, depending on the definition of the word and its context.

Ex : reading a sentence and being able to identify what words act as nouns, pronouns, verbs, adverbs, and so on..

8.Explain the chunking or shallow parsing.

Ans : Shallow parsing (also chunking or light parsing) is an analysis of a sentence which first identifies constituent parts of sentences (nouns, verbs, adjectives, etc.) and then links them to higher order units that have discrete grammatical meanings (noun groups or phrases, verb groups, etc.).

9.Explain the Noun phrase(NP)chunking.

Ans : They represent nouns and any words that depend on and accompany nouns.

10.Explain Named Entity Recognition.

Ans : Named Entity Recognition (NER) is the process of identifying and categorizing named entities in a given text. Examples of categories are organizations, locations, time, names, money, and rate. Other terms that are synonymous to NER are: Entity Identification.