- Corpora is huge large body of text (book corpus or firefox research , conversations ..etc), we can import them by using nltk.corpus and we can choose which text we would want to deal with it (sents - words - raws)
- raw take the text without splitting it (without token) but words make tokens
- some corpus example (`gutenberg - brown - Reuters - webtext -Inaugural Address Corpus - Annotated Text Corpora - and other language )`
- There are different structure of corpus like (isolated - overlapping - categorized - temporal)
- we can upload our own data to use it as nltk corpus using PlaintextCorpusReader or BracketParseCorpusReader
- instead we process sequence of words we can process sequence or paris Each pair has the form (condition, event)
- FreqDist() takes a simple list as input, ConditionalFreqDist() takes a list of pairs
- The bigrams() function takes a list of words and builds a list of consecutive word pairs
- we can generate text using distribution frequency and bigrams by defining the max and reset every word in the loop , example in notebook
- some python recap on functions , IDLE and module
- A collection of variable and function definitions in a file is called a Python module. A collection of related modules is called a package
- wordlist corpora used by some spellcheckers
- stopwords is fucken corpora !!!!!!!!! Thus, with the help of stopwords, we filter out a third of the words of the text
- another corpora called names which contain male & female names
- NLTK includes the CMU Pronouncing Dictionary for U.S. English like ('fireball', ['F', 'AY1', 'ER0', 'B', 'AO2', 'L'])
- Swadesh wordlists, lists of about 200 common words in several languages and there is Toolbox Corpora
- WordNet is a semantically oriented dictionary of English
- Hypernyms and hyponyms are called lexical relations because they relate one synset to another
-