- urlopen read text files & html files from web ( it use beautifulsoup to make web scrapping )
- feedparser is web scraping library
- ASCII text and HTML text are human-readable formats. Text often comes in binary formats—such as PDF and MSWord
- simple python basics about reading , write , list and string information
- codecs library use to encode the text and we have 3 types of encoding ( latina2 - utf-8 - GB2312
- NLTK tokenizers allow Unicode strings as input
- some python basics about regular expression
- Two or more words that are entered with the same sequence of keystrokes are known as textonyms. For example, both hole and golf
- tokenization is  is the process of transforming text into a single canonical form that it might not have had before
- Stemming (return words to its original ) is not a well-defined process, and we typically pick the stemmer that best suits the application we have in mind
- Stemming is a process that stems or removes last few characters from a word, often leading to incorrect meanings and spelling. Lemmatization considers the context and converts the word to its meaningful base form
- .The purpose of lemmatization is same as that of stemming but overcomes the drawbacks of stemming. In stemming, for some words, it may not give may not give meaningful representation such as “Histori”
- Stemming has its application in Sentiment Analysis while Lemmatization has its application in Chatbots, human-answering.
- We can use \W in a simple regular expression to split the input on anything other than a word character
- The function nltk.regexp_tokenize() is similar to re.findall() However, nltk.regexp_tokenize() is more efficient
- word segmentation is concept hard than tokenization because might be words not separated like thedog or doyou
- some simple basic info about list and strings