27/8/24

» nltk.corpus.gutenberg.fileids()

- frequency of using america increases over yrs due to slavery in 17s and 18s century.
- CoNLL data set widely used in Natural language processing
- no relation with other corpus is called isolated corpus, eg: gutenberg
- In code, wordnet.synset~~s~~(car.n.01)

  ↳ noun

  (similarly verb can also be used)
- raw = response.read().decode('utf8') —— to download text from int.

» wikipedia ka Text it cannot read, as structure is html (not .txt)

» Dealing with html

- use ~~beautiful~~ Beautifulsoup function
  ∴ from bs4 import Beautifulsoup
  raw = Beautifulsoup(html, ...)

<u>30/8/24</u>

Paper 7 : AI $\underbrace{\text{snake oil}}_{\text{meaning fraud}}$

· NLTK chapter 3

• **Reading local files**

//example.txt at the end
f = open('path of folder', 'w')
↳ to write
f.write("a quick brown fox jump over the lazy dog")
f. close

» [example named text file will be created with a quote written in it]

f. = open(same as earlier)
f. write ("a new line")
f. close

[earlier text will get deleted]

» write 'a' instead of ('w') to append it and not earase the earlier text.

» "\n a new line" with 'a' —— append with new line.

» 'r' used to read sentences from text file.

• **NLP Pipeline**

HTML $\longrightarrow$ ASCII $\xrightarrow[\text{tokanize the texts}]{}$ text $\longrightarrow$ Vocab