» 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, . . .)

30/8/24

Paper 7 : AI snake oil

meaning fraud

· NLTK chapter 3

• Reading local files

```
                                    # example txt at the end
f = open('path of folder', 'w')
f.write("a quick brown fox jump over the lazy dog")   → to write
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 ⟶ ASCII ⟶ text ⟶ Vocab

tokanize the texts

- Regular expressions for detecting word patterns

→ ~~##~~ to find patterns from text using python

en → english

$ → word ending with this pattern

w is lower → w is of lower index

^ → word starting with this pattern

+aa → word having — aa in it

aa+ → of pattern aam, aali, etc.

- Stemmers: part of the word which gets modified into new word
  eg: li
- Prouss of stemming is lemmatisation.
- part of speech tagging (POS-tagging)
- tagged corpora
- dictionaries in Python