- Corpora: Multiple collections of body of text.

(i) Glutenberg Corpus

- It refers to project Glutenberg.
- World's largest collection of books whose copy-write ended.

(ii) Web & Chat text

- Text obtained from forum discussions. [Social Media, Media]

(iii) Brown corpus

- First million words corpus from 500 sources

(iv) Reuters Corpus.

- Largest news agency.
- 1.3 million words, 90 topics.
- Each document can have multiple categories.

(v) Inaugral address corpus.

- Types of Corpus: (i) Isolated (iii) Overlapping
  (ii) Categorized (iv) Temporal.

* WordNet: Type of dictionary but also provides with sense.

* Governance with Teeth: Models based on ethics should incorporate human rights.

* AI snake oil: Most AI research are snake oil.
  ↓
  Fraud. How to identify the snake oil.

- To open a document: open(<file_path>, <mode>)
  modes: 'w', 'a', 'r'.

- Regular Expression: To find patterns in a text.

  $ → words ending with. ; re.search(( ), w)
  ^ → words starting with. ↓
  +( ) → words having ( ). Expression

- Stemmer: Part of word which gets transformed into another word. Eg: Table → Tables.

- Sorting Algorithms: Bubble Sort, Merge Sort, etc.

- POS tagging : . Grammatical role of a word.
  . Supervised learning.

- Default Tagger: Give everything the same tag.
- R.E Tagger: Tagging according to the results of a regular expression.
- Lookup Tagger: Looking up the most frequent tag of a word from a labeled dataset.
  - In an experiment: ~45% accuracy.
    - We can increase the dataset for the words.
      - Increased acc. from ~45% to ~90% accuracy.
    - To go from 90 → 98/99%, we need context for the word. (which words precede/follow it).