30-8-24

Page No.
Date

Papers

④ → summary of ethics regulation
⑤ → More ethics less regulation
⑥ →
⑦ snake oil ⇒ fraud ;
identify AI snake oil ?

NLTK ch-3

→ Decode text from web using UTF-8.
→ First Parse html to text using beautiful soup.
→ pseudo code

```
f = open("    ", 'w')     }  replaces the
f.write( )                }  old Content
f.close( )                }

f = open("    ", 'a')     }  Appends the
f.write( )                }  new Content
  f.close( )              }
```

similarly 'r' is for read.

NLP Pipeline

→ Basic pipeline for kind of research.

① Obtaining Data: HTML → ASCII → Text → Vocab.
→ tokenised function / word punct function.

Page No.
Date

→ ~~cwd~~ Punct function takes care of punctuation.

Regular Expressions.

→ It is a tool to find patterns from the text.
e.g. words end with "ing"
e.g. words in english, → all words → english
w for w in nltk.corpus.words.words("en") if w.islower()] → stord with lower case

→ ends with "ing" ⇒ if re.search('ing$', w) ↑ ending with

→ starts with "win" ⇒ if re.search('^win', w)

re.search('+aa+', w). → aam, Paraam.

Stem

wrong sometimes e.g. listen → list.

lemmatiser

Segmentation
→ text into lines
⇒ Word segmentation → tokenisation.

NLTK Ch-4

Procedural vs Declarative

Procedural → sequence & operation

Declarative → variable, variable, variable

→ Can check type.

Program Development.

→ readability

→ easy to debug.

→ key checkpoints have output logs.

Algorithmic. Design.

→ Design / logic of the code defines performance.

→ using ~~bad~~ merge sort other than bubble sort save time. by n sec.

NLTK Ch-5

Parts-of-speech (POS) tagging

→ Given a sentence/text to find POS tag for every word!

→ used to find position and content of each word in whole text.

Code ① words = word_tokenise(sent)

② nltk.pos_tag(words).

Page No.
Date

Q How to make a POS - tagger

Default tagging → everything Noun

Regular Expression tagger ⇒ . ing →
e d → past
. es → present
. Etc.

Look up tagger .
→ check freq of tags and use it to tag ~~[illegible]~~ in text.

→ But without content we can't have high accuracy
→ for this we use Bigrams.
All the above are unigram tagger → they use one word

→ If we use two word ⇒ Bigram tagger.