Command to run :- python3 Assn1.py <File_Path>

Ex - python3 Assn1.py 20_newsgroups/rec.motorcycles/102616

Assumptions

Sentences:-

- '\n' is replaced by ' ' in the whole of the input file since the message body constitutes a large portion of the text.
- After the above step, **nltk.sent_tokenize(text)** is used which breaks the given text into sentences.

Word:-

- **word_tokenize** is used from **nltk.tokenize** to break the sentence into probable words and those words are removed from the list which does not start with **alphanum.**

Q1:- Print the number of words and sentences contained in text.

- Count the number of words and sentences using **above assumption** for Sentences and Words.

Q2:- Print the number of words starting with consonants and the number of words starting with vowels.

- Broke the text into words as done in **Q1**.
- Checked for condition with **case-insensitive**.

Q3:- List all the email ids.

- Broke the text into sentences as done in **Q1**.
- ID starts with alphanum.
- Only contain alphanum, !#$%&'*+-/=?^_{|}~ in prefix (part before @)
- Prefix is of length atleast 1.
- @ can be followed by only alphanum(no special characters).
- only alphanum is expected before @.

- Exaclty one @.
- Atleast a dot after @.
- no dot before @.
- No consecutive dots in ID
- Max length of ID is 64
- Ends with alphanum.
- Duplicate IDs are not printed and counted.
- If a ID is invalid but if its substring is a valid ID then it is chosen.

Q4-5:- Print the sentences and number of sentences starting with a given word.

- Case-Insensitive.
- Numerical search allowed.
- If entered word is a sentence then we split it into words and then take searched word to be first word.

Q6:- Print the count of that word and sentences containing that word.

- Numerical search allowed.
- Case-insensitive.

Q7:- Given an input file, print the questions present, if any, in that file.

- Broke the text into sentences as done in **Q1**.
- Print those which contains question-mark at end.

Q8:- List the minutes and seconds mentioned in the date present.

- **\bHH:MM:SS\b** format is assumed which can be anywhere in the text.
- List all times and not just one.

Q9:- List the abbreviations present in a file given as input.

- All uppercase.
- Atleast length 2

Word to Number conversion:-

- “03” is not considered as 3.
- #3 is considered as 3.
- #03 is not considered as 3.