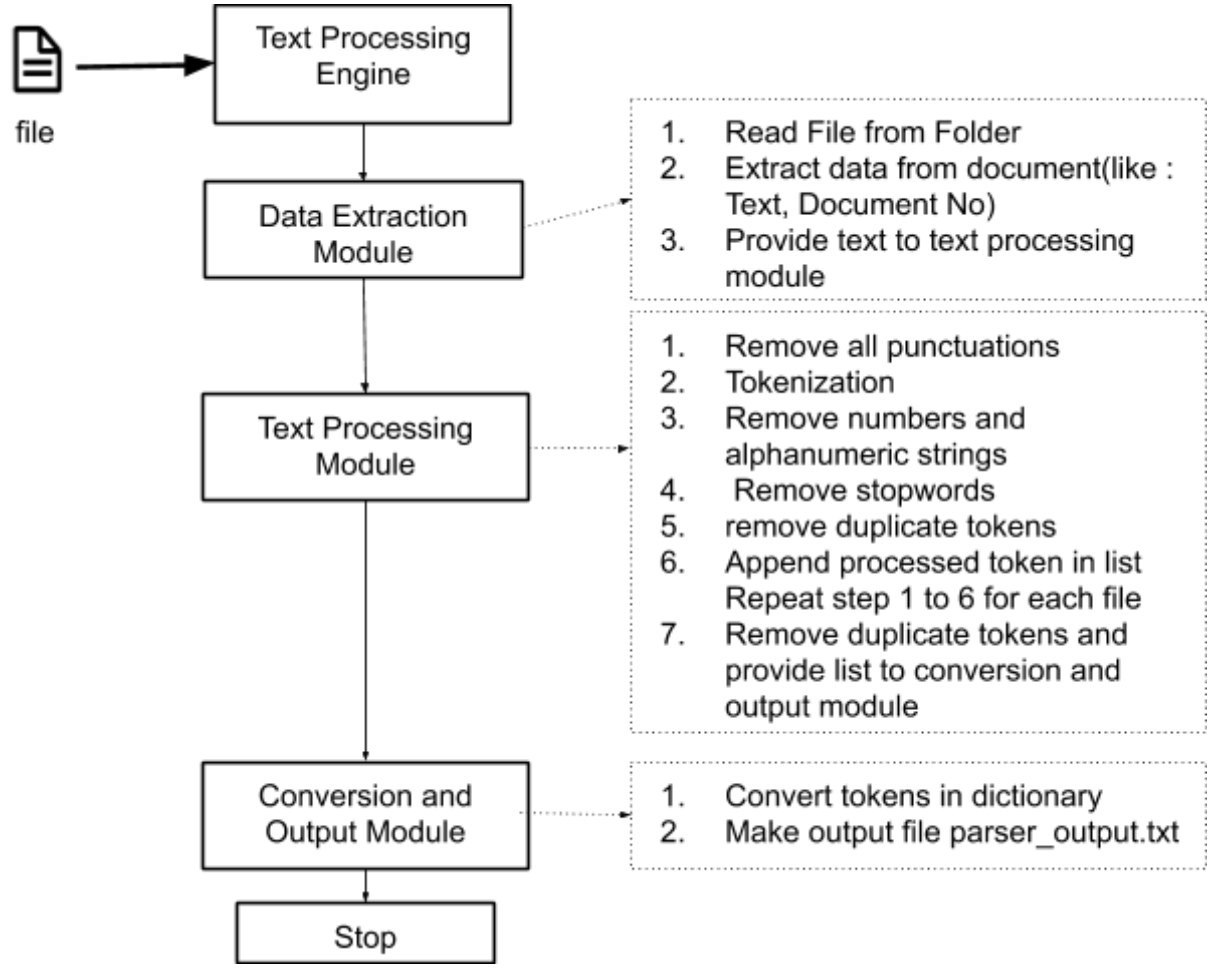
Text Processing Demo:

Architecture & High level Flow:



Library and its use :

```python
import glob
import os
from bs4 import BeautifulSoup
import nltk
from itertools import chain
import json
#nltk.download()
import re
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
```

Glob :
Its python library to find all files in folder
In project :
In this project we have dataset folder ft911 we are using go get each file path from this folder

BeautifulSoup:
Its 3rd party python library basically used to get data from html.
In project :
So We have data between Opening and closing tag same as html. To extract document data we are using this library.

Example :

<DOC>
<DOCNO>FT911-374</DOCNO>
<PROFILE>_AN-BEMAUAADFT</PROFILE>
<DATE>910513
</DATE>
<HEADLINE>
FT 13 MAY 91 / Survey of Cardiff (6): Principality House may set pace - The property sector
</HEADLINE>
<BYLINE>
  By GARETH GRIFFITHS
</BYLINE>
<TEXT>
MANY PEOPLE describe Cardiff as a major international city waiting to happen. After the confusion in parliament last month, the wait may be longer than they expect.

</TEXT>
<PUB>The Financial Times
</PUB>
<PAGE>
London Page 18

</PAGE>
</DOC>
<DOC>
<DOCNO>FT911-375</DOCNO>
<PROFILE>_AN-BEMAUAACFT</PROFILE>
<DATE>910513
</DATE>
<HEADLINE>
FT 13 MAY 91 / Survey of Cardiff (3): The barrage meets its first storm - Cardiff Bay / The future of the plan, after a parliamentary jolt
</HEADLINE>
<BYLINE>
  By STEWART DALBY
</BYLINE>
<TEXT>
Progress of a locally-sponsored private bill, to allow a barrage to be built across Cardiff Bay, creating a non-tidal 500-acre freshwater lake, was blocked at an all-night sitting. A group of Labour MPs who oppose the scheme, largely on environmental grounds, made long speeches and raised persistent objections.
<PUB>The Financial Times
</PUB>
<PAGE>
London Page 16 Map (Omitted). Photograph (Omitted). Photograph Above, the low-tide mudflats of Cardiff Bay, which environmentalists would preserve, but which planners say would discourage development. Below, two citizens study a model of the alternative registration project, at the visitor centre (Omitted).
</PAGE>
</DOC>

Methods :

Find_all (‘doc’) : provide data between each <DOC> …. </DOC> tag and store in the list
Example : [<DOC> {data} </DOC> , <DOC> {data} </DOC>]

find(‘text’).get_text() : will provide data between single <Text>...</Text> tag store in string
Example : “{textData}”

NLTK :
(Analyzing Text with the Natural Language Toolkit.)
This library is a python library to do natural language stuff . so it contains existing methods for stem , stopword removal and best library to do research in natural language processing So it provides functions for stem and remove stop words.

PorterStemer:
Example : https://www.geeksforgeeks.org/python-stemming-words-with-nltk/

Re:
Library for regular expression (regex)
In project:

```
# remove all punctuations
document_text_without_punctuation =  re.sub(r'[^\w\s]', '', document_text)
# Convert in tokens (split by white space)
document_token_list = document_text_without_punctuation.split()
# remove numbers and alphanumeric strings
removed_numbers_tokens = [item for item in document_token_list if not any(data.isdigit() for data in item)]
# get stop words from nltk stop word list
stop_words_nltk = set(stopwords.words("english"))
# remove stop words
removed_stopword_tokens = [word for word in removed_numbers_tokens if not word in stop_words_nltk]
# stem words using PorterStemmer
ps = PorterStemmer()
stemmed_tokens = [ps.stem(word) for word in removed_stopword_tokens]
# remove duplicate tokens
clean_tokens = list(dict.fromkeys(stemmed_tokens))
return clean_tokens
```

1. re.sub(r'[^\w\s]','',document_text) : replace all punctuations with '
    a. r'[^\w\s] : ^\w : not a word , \s: for punctuations
2. data.split() : split all data using space (" ") and store words/tokens in array ,(it is used to tokenize)
3. Other things can be easily understandable by comments