Assignment No :- 1

- Aim :- Write a programme for pre-processing of a text documents such as stop removal stemming

- Outcome :-

1) Understand and implement basic text preprocessing technique.
2) Remove stop words from a document
3) Apply stemming to words using NLP libraries.

- Hardware Requirement :-

1) Processor - Intel i3/i5 or above.
2) RAM :- Minimum 4GB.
3) Storage :- 2GB of free space.
4) OS :- Window / Linux / os.

Software Requirements :-

1) Python 3X.
2) Jupyter Notebook.
3) NLTK Library
4) Internet Connection.

Theory :-

Text Preprocessing in NLP - Detailed description.

1) Tokenization

It is a process of breaking do a large chunk of text into smaller units called tokens. The tokens can be words characte or sub words.

Tokenization :-

1) Word Tokenization - Split text int words.

2) Sentence Tokenization - split text sentences.

• Importance :-

- Tokenization is based step for N operation like parsing, pos taggi harmed entity recognition.

- Helps in feature extraction in ML models.

2. Stop word Removal :-

Stop words are commonly use

words in a language that are filtered out before passing because they usually dont carry significant meaning.

3. Stemming -

It is the process of reducing a word to its based root from by removing prefixes or suffixes!

eg - "running", "runs", "ran" → "run"
"studies", → "studi

Popular stemming Algorithms:

- •Porter stemmer - commonly used, rule based stemmer.
- •snowball stemmer - An improvement over the porter stemmer.
- •Lancaster stemmer :- more aggressive than Porter.

Advantages:-

- •Reduces the size of data for processing
- •Helps focus on meaningful words.
- •Improves model performance by reducing size.

Dis-Advantage :-

- Stemming may result in non-di
words.
- stop word in same contexts,
- Language-dependent - each langua
needs its own preprocessing.

- Conclusion :-

- Hence we, conclude, Text pre-pr
is a critical step in NLP, t
like stop words removal stemm
help in cleaning - standerdizing
data making it ready for ma
learning models.

Coding efficiency	Viva	Timely	Total	Sign
5	3	2	10	
4	2	2	8	

Assignment No:-2

- Aim :- To implement a program that retrives documents based on keyword queries using the inverted file indexing technique

- Outcome :-

1) Understand and implement inverted indexing
2) Store and retrieve document efficient based on keywords.
3) Apply the concept in search engines and information retrival system.

Hardware Requirement :-
1) Processor - Intel i3/i5/i7 or equivalent
2) RAM - minimum 4GB.
3) Disk space - Minimum 1GB free.
4) keyboard & Monitor.

Software Requirement :-
1) Operating System - Window/macos/linux.
2) Programming language - Python 3.6.
3) IDE :- VS code /pycharm.
4) Libraries :- No external libraries required.

Theory:-
Inverted file indexing is a method
storing mapping from content
to their location in a set of
documents. It is backbone of
modern search engines.

• Steps Involved :-

1) Tokenization - split documents
individual words.

2) Stop words removal - Remove comm
words.

3) Inverted Index construction - Map
word to the document ID's
which it occurs.

Conclusion:-

The inverted file inde
method was sucessfully implement
The basic mechanism behind se
engine worked using inverted ind

Coding	effi	Viva	Timely	Total
5		3	2	10
4		2	2	8