

Vidyavardhini's College of Engineering & Technology

Department of Computer Science and Engineering (Data Science)

Experiment No.4
Apply Stemming on the given Text input
Date of Performance:
Date of Submission:



Vidyavardhini's College of Engineering & Technology

Department of Computer Science and Engineering (Data Science)

Aim: Apply Stemming on the given Text input.

Objective: Understand the working of stemming algorithms and apply stemming on the

given input text.

Theory:

Stemming is a process of linguistic normalization, which reduces words to their word root

word or chops off the derivational affixes. For example, connection, connected, connecting

word reduce to a common word "conect".

Stemming is the process of producing morphological variants of a root/base word. Stemming

programs are commonly referred to as stemming algorithms or stemmers. A stemming

algorithm reduces the words "chocolates", "chocolatey", "choco" to the root word,

"chocolate" and "retrieval", "retrieved", "retrieves" and reduces to the stem "retrieve".

Stemming is an important part of the pipelining process in Natural language processing. The

input to the stemmer is tokenized words.

Applications of stemming:

1. Stemming is used in information retrieval systems like search engines.

2. It is used to determine domain vocabularies in domain analysis.

Porter's Stemmer Algorithm:

It is one of the most popular stemming methods proposed in 1980. It is based on the idea that

the suffixes in the English language are made up of a combination of smaller and simpler

suffixes. This stemmer is known for its speed and simplicity. The main applications of Porter

Stemmer include data mining and Information retrieval. However, its applications are only

limited to English words. Also, the group of stems is mapped on to the same stem and the

output stem is not necessarily a meaningful word. The algorithms are fairly lengthy in nature

and are known to be the oldest stemmer.

CSDL7013: Natural Language Processing Lab



Vidyavardhini's College of Engineering & Technology

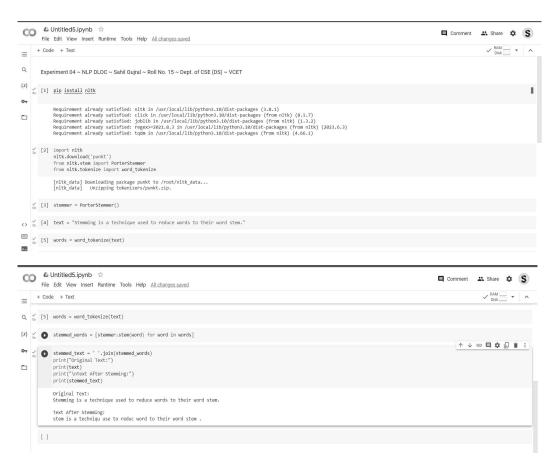
Department of Computer Science and Engineering (Data Science)

Example: EED -> EE means "if the word has at least one vowel and consonant plus EED ending, change the ending to EE" as 'agreed' becomes 'agree'.

Advantage: It produces the best output as compared to other stemmers and it has less error rate.

Limitation: Morphological variants produced are not always real words.

Implementation:



Conclusion: