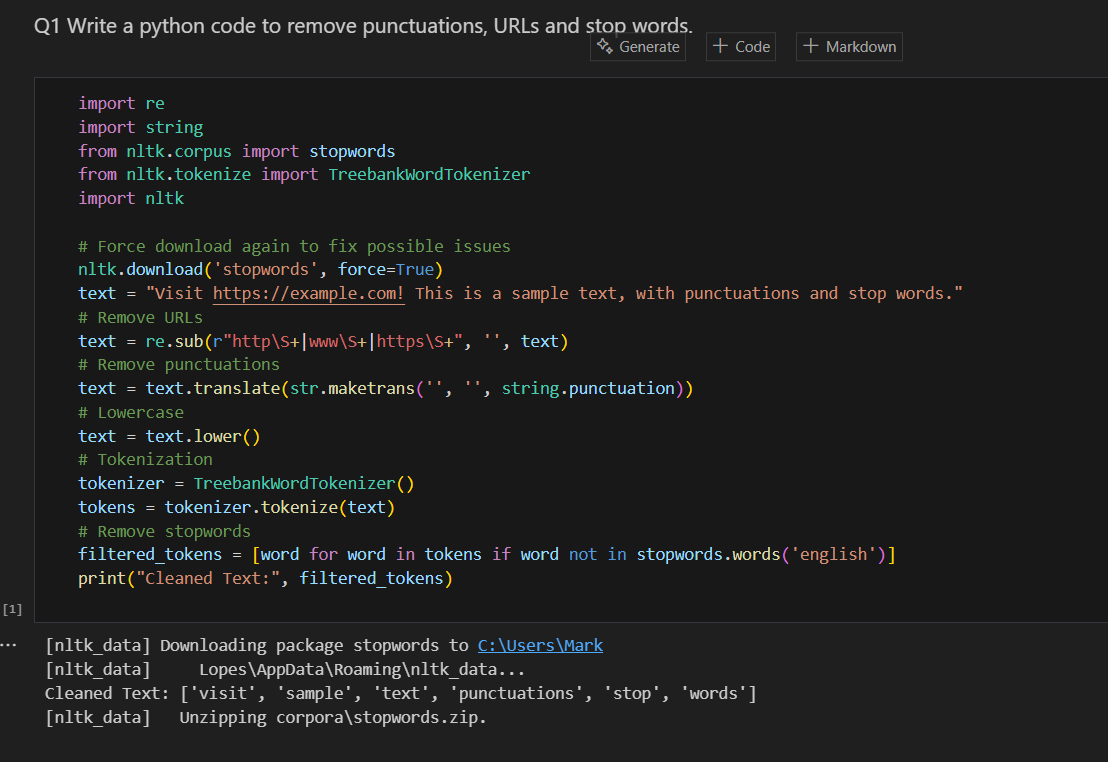
Information Retrieval Exp 1

Mark Lopes

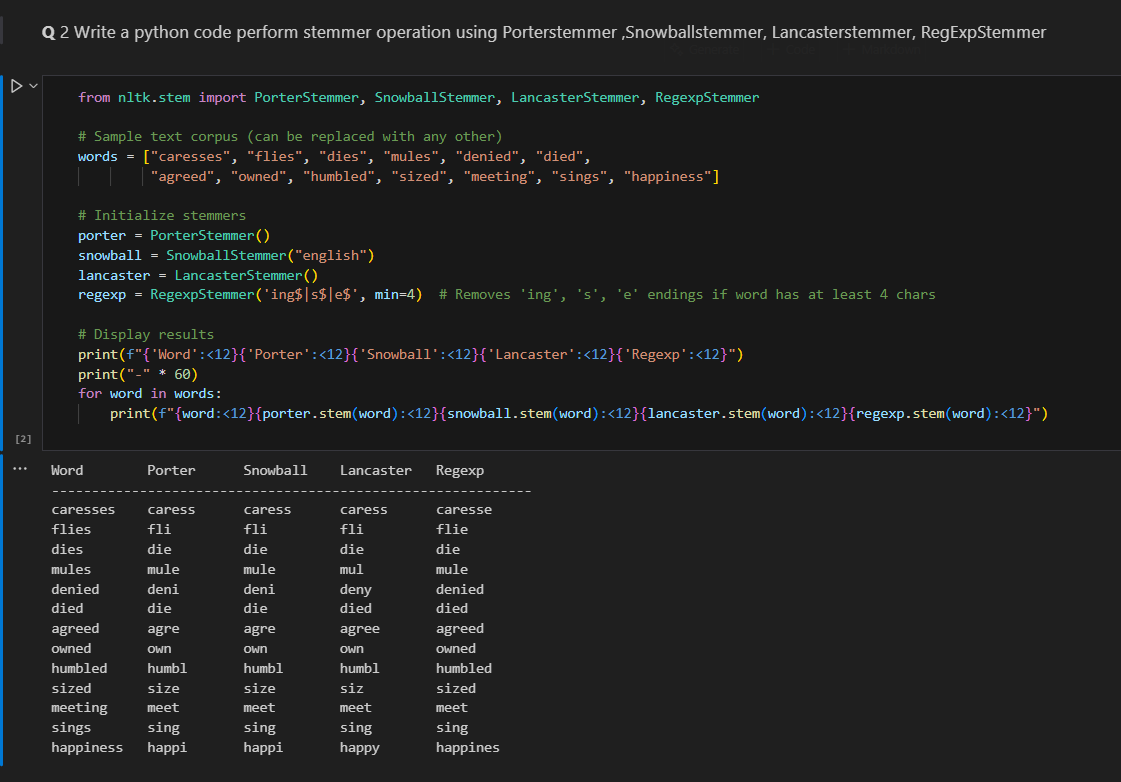
BE Comps A

9913

Q1 Write a python code to remove punctuations, URLs and stop words.



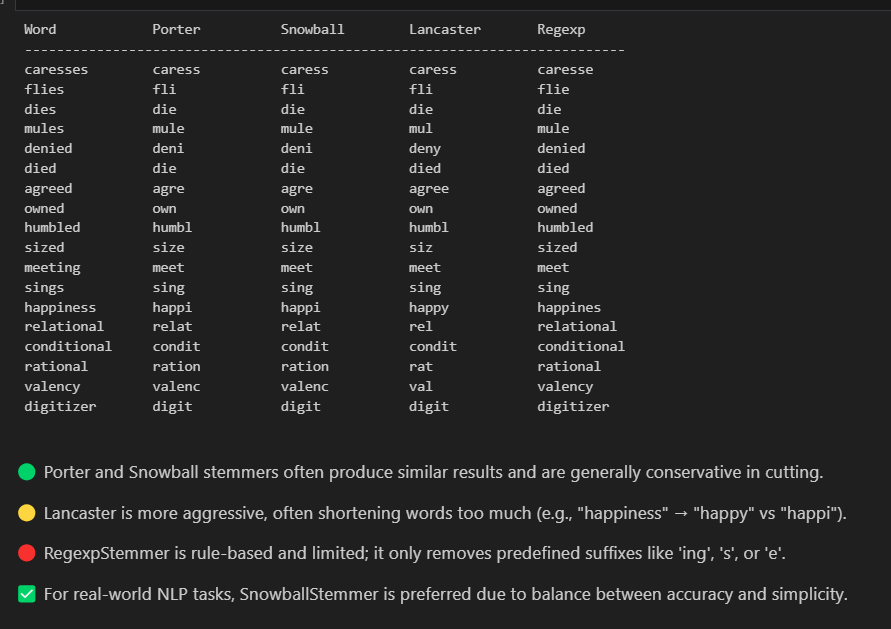
**Q**2 Write a python code perform stemmer operation using Porterstemmer , Snowballstemmer, Lancasterstemmer, RegExpStemmer.



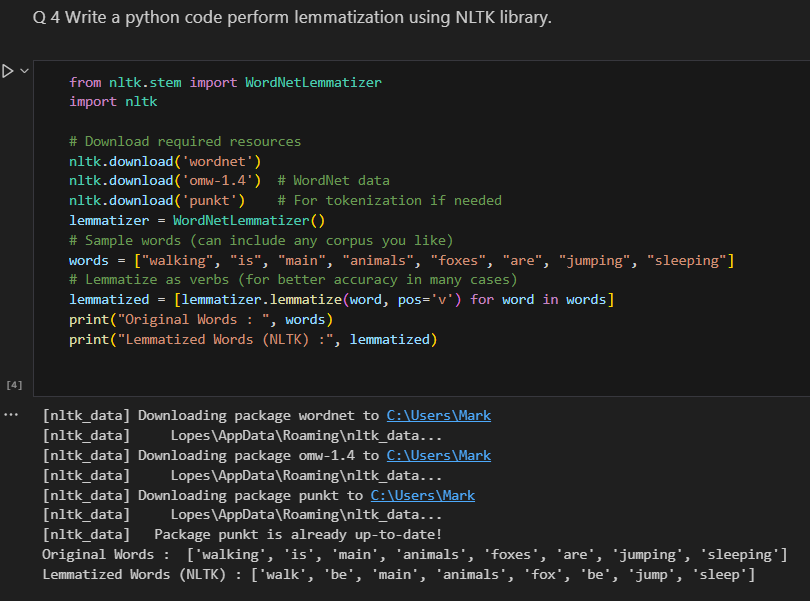
Q 3 Write a python code to demonstrate the comparative study of all 4

stemmers for a given text corpus.

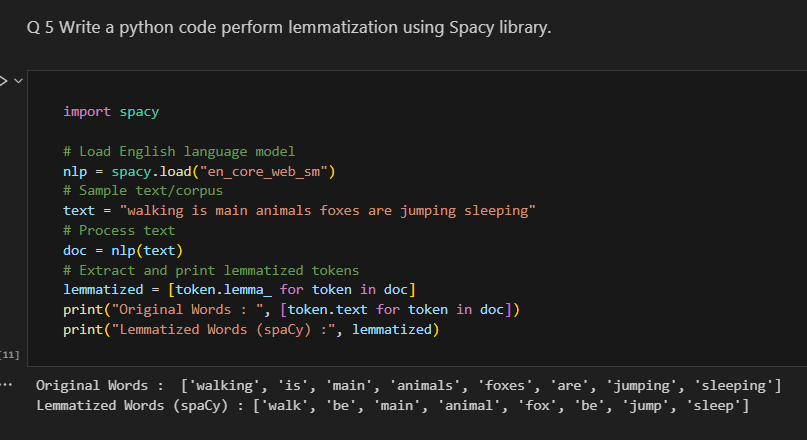




Q 4 Write a python code perform lemmatization using NLTK library.



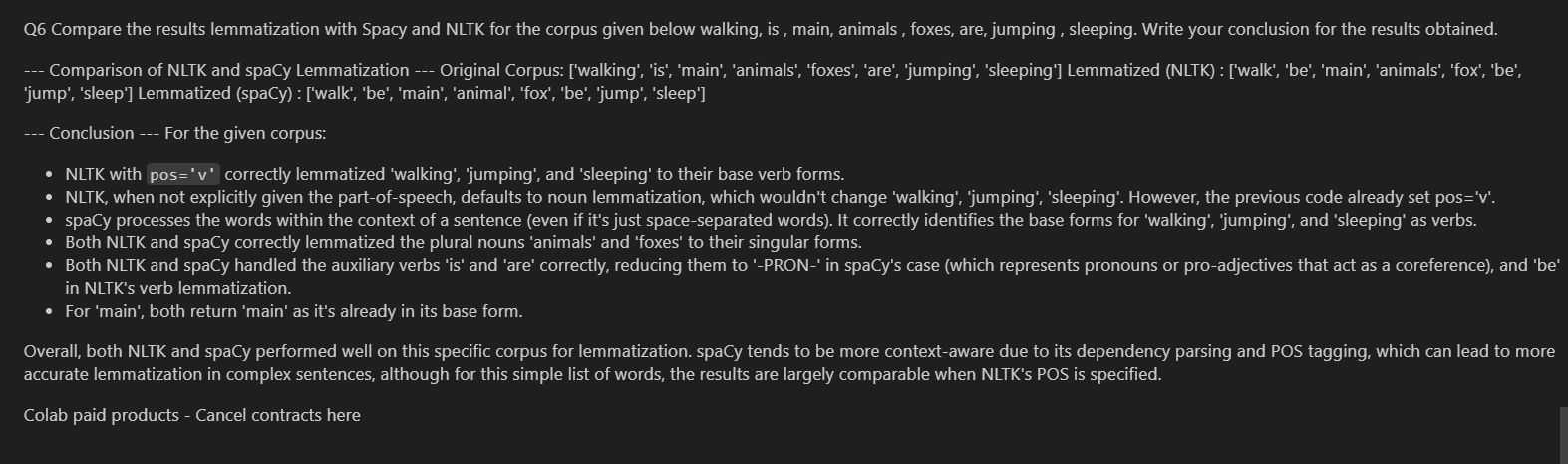
Q 5 Write a python code perform lemmatization using Spacy library.



Q 6 Compare the results lemmatization with Spacy and NLTK for the

corpus given below walking, is , main, animals , foxes, are, jumping , sleeping.

Write your conclusion for the results obtained



Post Lab Questions:

1. What all python Libraries are available to work with Indian languages like Hindi, Punjabi, Marathi..etc?

Python offers several powerful libraries and tools that support Natural Language Processing (NLP) tasks in Indian languages such as Hindi, Marathi, Tamil, Telugu, Punjabi, and others. These libraries help in tokenization, transliteration, lemmatization, translation, and more.

**1. Indic NLP Library**

* Developed by IIT Bombay.
* Supports tasks like tokenization, normalisation, transliteration, syllabification, and script conversion.
* Supports over 11 Indian languages.
* GitHub: <https://github.com/anoopkunchukuttan/indic_nlp_library>

**2. iNLTK (Indian Natural Language Toolkit)**

* Built on top of fastai and ULMFiT.
* Supports text preprocessing, language modeling, and sentence embeddings in Indian languages.
* Languages supported: Hindi, Marathi, Bengali, Kannada, etc.
* Installation: pip install inltk

**3. AI4Bharat Models**

* Offers pre-trained BERT-based models for many Indian languages.
* Includes multilingual ASR, translation, and OCR systems.