Weeks 3-4

Fundamentals of Natural Language Processing (NLP)

1. **Introduction to NLP: Key concepts and applications.**
2. **Text Processing: Tokenization, stemming, and lemmatization.**

* Tokenization:
  + Explain the concept of tokenization and its importance.
  + Demonstrate tokenization using NLTK and spaCy.
* Stopword Removal:
  + Discuss the significance of stopwords and their removal in text processing.
  + Implement stopwords removal using NLTK.
* Stemming and Lemmatization:
  + Define stemming and lemmatization and their differences.
  + Perform stemming and lemmatization using NLTK and spaCy.

1. **Basic NLP Tasks:**

* Part-of-Speech (POS) Tagging:
  + Explain POS tagging and its role in grammatical analysis.
  + Perform POS tagging using NLTK and spaCy.
* Named Entity Recognition (NER):
  + Introduce NER and its application in identifying named entities in text.
  + Implement NER using NLTK and spaCy.

Note: Use data set CoNLL 2003, IMDb Reviews

1. **Feature Extraction: Bag-of-words, TF-IDF, and word embeddings.**

* **Bag-of-Words (BoW) Representation**:
  + Define Bag-of-Words (BoW) model and its components.
  + Demonstrate the process of creating a BoW representation from text data.
  + Discuss the limitations of BoW representation, such as the loss of sequence information and the curse of dimensionality.
  + Implement BoW representation using Python and NLTK and spaCy.
* **Term Frequency-Inverse Document Frequency (TF-IDF):**
  + Explain the concept of TF-IDF and its significance in information retrieval.
  + Discuss the calculation of TF and IDF components.
  + Demonstrate how to compute TF-IDF scores for text data.
  + Compare TF-IDF representation with BoW representation.
  + Implement TF-IDF representation using Python and scikit-learn.
* **Word Embeddings:**
  + Introduce word embeddings and distributed representations of words.
  + Explain popular word embedding techniques such as Word2Vec, GloVe, and fastText.
  + Discuss the advantages of word embeddings over traditional methods like BoW and TF-IDF.
  + Demonstrate how to use pre-trained word embeddings.
  + Implement Word2Vec or GloVe embeddings using Python and gensim or spaCy.