Intro to NLP and Text Preprocessing

Here is an article to get you started on the core NLP principles - Intro to NLP and Text Processing. These have already been covered in Week0. You can revise it once again using these resources. Focus on how it's done now and the stepwise approach.

Few videos for your better understanding:

Tokenization- NLP Demystified 2: Text Tokenization

Basic Preprocessing (Stop Word Removal, Stemming, Lemmatization)-

NLP Demystified 3: Basic Preprocessing (case-folding, stop words, stemming...

Advanced Preprocessing (POS tagging, NER, parsing)-

NLP Demystified 4: Advanced Preprocessing (part-of-speech tagging, entity...

Text Representation

Computers don't speak English; they speak Math. Here is how we translate:

1. **Bag of Words (BoW)-** Creates a vocabulary list and a vector for each document based on word counts.
   Make sure to read this article: Bag of Words
   And watch this video-
   Text Representation Using Bag Of Words (BOW): NLP Tutorial For Begi...
2. **One-Hot Encoding-** every word (even symbols) which are part of the given text data are written in the form of vectors, constituting only of 1 and 0
   Read this: One Hot Encoding
3. **TF-IDF-** TF (Term Frequency) measures how often a word appears in a document, while IDF (Inverse Document Frequency) gauges the word's rarity across documents.
   Formula Overview: (Term Frequency) x (Inverse Document Frequency).
   Read More: TF-IDF Vectorizer
   Text Representation Using TF-IDF: NLP Tutorial For Beginners - S2 E6

Word Embeddings-

NLP Demystified 12: Capturing Word Meaning with Embeddings

1. **The Word2Vec Family**
   - **CBOW (Continuous Bag of Words):**
     Focus: Predicting the word from context.
   - **Skip-gram:** The inverse of CBOW.
     Focus: Predicting context from a word.

Must read : Word2Vec stanford lecture, Word2Vec Tutorial
Word Embedding and Word2Vec, Clearly Explained!!!

2. **Global Vectors (GloVe)**
    - **Insight:** Learns relationships from co-occurrence statistics. Word2Vec looks at local neighbours; GloVe looks at the entire data.

    Essential Reading: Intro to Glove

3. **FastText-** FastText breaks words into sub-parts (like prefixes and suffixes), meaning it can understand words it has never even seen before.

4. **BERT (Bidirectional encoder representations from transformers)**
    Function: Uses the Transformer architecture for deep, contextualized understanding. Reads and understands the context from both the directions

Additional Resources:

- What is FastText?
- BERT Embeddings