Module 3

1. **Text Preprocessing Basics:**
   * What are some **common steps** in text preprocessing?

The way we can summarize text preprocessing is via the following:

* Tokenization: This breaks down text into smaller units called tokens. There are two main types: Sentence tokenization, which splits text into individual sentences. Word tokenization, which splits text into words or smaller pieces (tokens).
* Tagging: Parts of speech tagging assigns grammatical categories to each word in a text (noun, verb, adjective, etc.)
* Chunking: identifies and segments phrases (noun phrases, verb phrases, etc.) in sentences without specifying their internal role.
* Stemming: This step reduces words to their root or base form by removing affixes. This is a simple, rule-based approach that can sometimes produce non-words.
* Lemmatization: Reduces words to their dictionary base form (lemma) using a vocabulary and morphological analysis. Unlike the stemming step, it always produces valid words. Lemmatization is more accurate but also computationally more expensive than stemming
  + Discuss different approaches for handling **emojis, emoticons, colloquialisms**, and other modern text elements. Provide examples where these may require special handling.

There are a few ways we can approach handling emojis, emoticons, and colloquialisms. Here are some ideas:

* + - Embedding-based approach: There are some embeddings like emoji2vec that can be used alongside word embeddings. Emji2vec is useful to make vector representations of emojis, which allows models to understand semantic relationships between emojis and words.
    - Conversion to textual descriptions: We can also convert emojis and emoticons to their textual descriptions. It could be something like “I love you! [red heart]”. This way, we can improve sentiment analysis and regular NLP techniques.
    - When it comes to Colloquialisms, we can normalize informal text to their standard forms. For instance, “gonna” and “u” will be converted to “going to” and “you” respectively.