Name: Jainisha Choksi
Assignment: Text Processing

Text Processing:
Unstructured text data can be automatically analyzed and sorted by text processing to obtain useful information. Text processing systems can automatically comprehend human language and derive value from text data using Natural Language Processing and Machine Learning, a branch of artificial intelligence.

Why is Text Processing essential?
Text processing is essential for extracting insights from unstructured text data and enabling machines to understand and generate human-like language, underpinning a wide range of applications from information retrieval to natural language understanding.

Text Processing Methods:

Text Preprocessing:

1) Tokenization: Splitting text into individual words, phrases, or sentences.
2) Lowercasing: Converting all text to lowercase to ensure consistency.
3) Stopword Removal: removing common words that don’t carry significant meaning.
4) Punctuation Removal: Eliminating punctuation marks from the text.
5) Stemming and Lemmatization: Reducing words to their base or root form to normalize variations.

Feature Extraction:

6) Bag of Words: Representing text as a frequency count of words occurring in a document, ignoring grammar and word order.
7) Term Frequency- Inverse Document Frequency: Assigning weights to words based on their frequency in a document relative to their frequency across all documents, emphasizing rare terms.
8) Word Embeddings: Representing words as dense, low-dimensional vectors capturing semantic relationships. Techniques like Word2Vec, GloVe, and FastText are commonly used for this purpose.

Text Representation:

9) Vectorization: Converting textual data into numerical vectors suitable for machine learning algorithms.
10) Document-Term Matrix: Constructing a matrix where rows represent documents, columns represent documents, columns represent terms, and each cell contains the frequency of a term in a document.
11) Word Embedding Matrix: Building a matrix where rows correspond to words, and each row contains the word embedding vector.

Text Analysis:

12) Text Classification: Assigning predefined categories or labels to documents based on their content.
13) Sentiment Analysis: Determining the sentiment expressed in a piece of text.
14) Named Entity recognition: Identifying and categorizing entities mentioned intext, such as people, organizations, and locations.
15) Topic Modeling: Uncovering latent topics present in a collection of documents.