#### **I m mahnoor , i m in atomcamp.**

I M MAHNOOR , I M IN ATOMCAMP

ATOMCAMP/atomcamp/atomCAMP→ LOWERCASE

#### **NLP Pipeline**

#### An NLP pipeline is a sequence of steps used to process and analyze text data. Here’s a detailed breakdown of each step:

#### **Text Preprocessing**:

#### **Tokenization**: Splitting text into tokens (words or sentences).

#### **Lowercasing**: Converting all characters to lowercase to ensure uniformity.

#### **Stop Words Removal**: Removing common words that do not contribute much meaning (e.g., "and", "the").

#### **Stemming and Lemmatization**: Reducing words to their root form (e.g., "running" to "run").

#### **Feature Extraction**:

#### **Bag of Words (BoW)**: Representing text as a set of words without considering grammar or order.

#### **Term Frequency-Inverse Document Frequency (TF-IDF)**: Measuring the importance of a word in a document relative to a collection of documents.

#### **Word Embeddings**: Using pre-trained models like Word2Vec, GloVe, or BERT to convert words into numerical vectors that capture semantic meaning.

#### **Model Building**:

#### **Supervised Learning**: Using labeled data to train models for specific tasks like sentiment analysis or NER.

#### **Unsupervised Learning**: Finding patterns in unlabeled data, such as clustering similar documents.

#### **Deep Learning**: Employing neural networks for tasks like machine translation and text generation.

#### **Model Evaluation**:

#### **Accuracy**: The percentage of correctly predicted instances.

#### **Precision and Recall**: Measuring the relevance and completeness of the model's predictions.

#### **F1 Score**: The harmonic mean of precision and recall.

#### **Deployment**:

#### **API Integration**: Making the NLP model available through an API for other applications to use.

#### **User Interface**: Creating interfaces like chatbots or dashboards for end-users to interact with the NLP system.

#### **Monitoring and Maintenance**:

#### **Performance Monitoring**: Continuously tracking the model's performance in production.

#### **Retraining**: Updating the model with new data to maintain its accuracy and relevance.

#### 

#### **1. Text Preprocessing**

Text preprocessing is the first step in the NLP pipeline, crucial for cleaning and preparing raw text data for further analysis. This involves several sub-steps:

1. **Tokenization**:
   * **Word Tokenization**: Splitting a sentence into individual words. For example, "I love NLP" becomes ["I", "love", "NLP"].
   * **Sentence Tokenization**: Splitting a paragraph into individual sentences. For example, "I love NLP. It is fascinating." becomes ["I love NLP.", "It is fascinating."].
2. **Lowercasing**:
   * Converting all characters in the text to lowercase to ensure uniformity. For example, "I Love NLP" becomes "i love nlp".
3. **Stop Words Removal**:
   * Removing common words that do not contribute significant meaning to the text. These include words like "and", "the", "is". For example, "I love NLP and it is fascinating" becomes "love NLP fascinating".
   * The cat/cat -> same thing
   * I went to pesh and pesh was hot
4. **Stemming and Lemmatization**:
   * **Stemming**: Reducing words to their base or root form. For example, "running" becomes "run".
     1. Do- i am doing my work
     2. Do/ Doing--- state
   * **Lemmatization**: Reducing words to their dictionary form, considering the context. For example, "better" becomes "good".
     1. **Good** worse worst
5. **Punctuation Removal**:
   * Removing punctuation marks from the text. For example, "Hello, world!" becomes "Hello world".
6. **Handling Special Characters**:
   * Removing or converting special characters (like emojis, currency symbols) as per the requirement of the analysis.
7. **Text Normalization**:
   * Converting different forms of text into a standard format. This can include converting all numbers to a specific format or expanding contractions (e.g., "don't" to "do not").
   * hain/hyn/ hayn ---> hyn

* Good
  + Beautiful, happy, nice good(btter best)
  + Good
  + Better
  + Best
  + 🥳
  + 😍 🙂
* Bad
  + Unhappy , ugly, bad ,worst ,worse

Good bag

Better bag

2- bag, Good/better-good

Did matric

Doing matric

2- matric, did/doing

Handbag/bag