Lecture 2

- **What is common Components of a Generic pipeline?**

  Data Acquisition => Text Cleaning => Pre-Processing => Feature Engineering => Modeling => Evaluation => Deployment => Monitoring and Model Updating .

- **What mean by data acquisition in NlP?**

  Is the process of collecting and preparing data for use.

- **Unlabeled data is heart of ML systems (False)**
- **Mentation some of Data Acquisition approaches?**

  ⇨ **Use public dataset.**

  ⇨ **Scrap data (collecting data from websites (customer reviews))**

  ⇨ **Product intervention (use user feedback to improve the user interface of their product)**

  ⇨ **Data augmentation (adding new data points to an existing dataset)**

  ⇨ **Synonym replacement (This approach involves replacing words in a text with synonyms to create variations of the original text.)**

  ⇨ **Back Translation**

  ⇨ **TF-IDF based word replacement.**

  ⇨ **Bigram Flipping**

  ⇨ **Replacing entities**

  ⇨ **Adding noise to data**

- **Mentation examples for Text Extraction and cleanup?**

  Html parsing and cleanup(extract information from html code cleanup for example remove <br>)

  Unicode Normalization

  Spelling Correction<hllo => hello >

  Fat finger Problem(asham => adham)

- **What is meant by Preliminaries?**

  ⇨ **Sentence Segmentation**

  Breaking the whole text up into sentences and tokens

  ⇨ **Word Tokenization**

  Converts the sentence into set of tokens.

- **Issues in tokenization?**

  ⇨ **Finland's capital?**

  ⇨ **Finland AND s? Finland's? Finland's? Hewlett-Packard ?**

  ⇨ **Hewlett and Packard as two tokens? San Francisco: one token or two?**

  ⇨ **state-of-the-art: break up hyphenated sequence.**

  ⇨ **co-education**

  ⇨ **lowercase, lower-case, lower case?**

- **What is frequent Steps (preprocessing)?**

  ⇨ **Stop words removal (at, In, or, the, for).**

  ⇨ **Converts to lower case.**

  ⇨ **Removing punctuation(# , . , / ,$,@)**

  ⇨ **Digits removal(remove numbers)**

- **Stemming and Lemmatization?**

  Stemming: removing suffixes and reducing a word to base form ex: adjustable=> adjust || airliner => airlin.

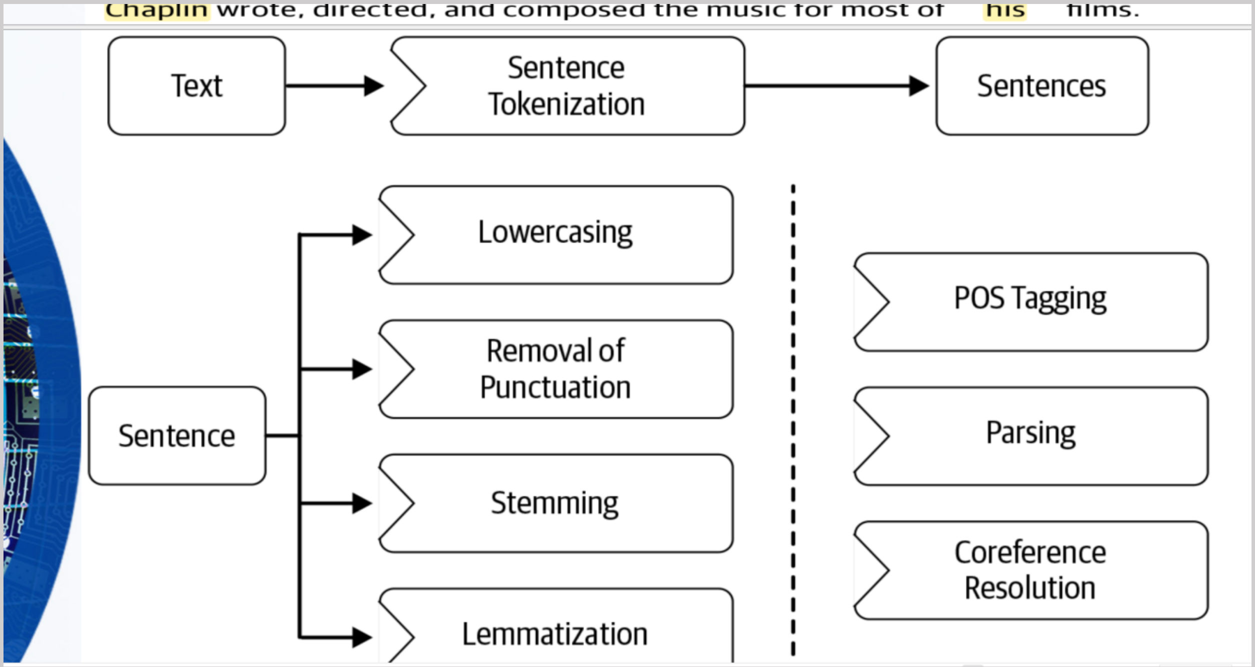
  Lemmatization: mapping all the different forms of a word to its base word or lemma was => (to) be || better => good

- **What is advanced preprocessing examples?**

  Text Normalization
  Language Detection
  Code mixing (استخدام اكثر من لغة فى نفس المحادثة)
  Affected language.
  Transliteration

- **What is advanced Processing phases**

  Part of speech (Pos) tagging => verb noun Determiner
  Parsing
  Coreference Resolution



Chaplin wrote, directed, and composed the music for most of his films.