**Lecture 2**

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

**Data Acquisition الحصول على البيانات=> Text Cleaning تنظيف النص => Pre-Processing ما قبل المعالجة => Feature Engineering هندسة الخصائص=> Modeling النمذجة=> Evaluationتقييم => Deployment التوظيف=> Monitoring and Model Updating .المرقبة وتحديث النموذج**

**(Breaking the major problem down into subproblems => divide and conquer)**

* **What does data acquisition mean in NlP?**

**Is the process of collecting and preparing data for use.**

**هى عملية تجميع وتجهيز البيانات للاستخدام**

* **Unlabeled data is heart of ML systems (False)**
* **Mentation some of Data Acquisition approaches?**
* **Use public dataset. استخدام مجموعة البيانات العامة**

**Kaggle -**

* **Scrap data (collecting data from websites (customer reviews))خردة البيانات**
* **Product intervention (use user feedback to improve the user interface of their product)**
* **Data augmentation (adding new data points to an existing dataset)زياده البيانات**

**Some data Augmentation forms.**

* **Synonym replacement (This approach involves replacing words in a text with synonyms to create variations of the original text.)استبدال المرادفات**
* **Back Translation (translate to another language )**
* **TF-IDF based word replacement.** "I feel happy today" or "I feel good today ".

**(identify words have high tf-idf scores and replace them with less frequent synonyms that have similar meanings)**

* **Bigram Flipping(happy today => today happy )**
* **Replacing entities (replacing entites ) ex: I go to home , I go to my house | I love human , I love person**
* **Adding noise to data for example => I feel good today , I feel god tody**
* **Mentation examples for Text Extraction and cleanup?**

**Any correction or remove is text cleaning.**

**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)**

**System specific error correction (Original text: "I bought an iPhne and it is amzing!")**

* **What is preprocessing phases ?** 
  + **Preliminaries**
  + **Frequent steps**
  + **Other steps**
  + **Advanced processing**
* **What is meant by Preliminaries?**
* **Sentence Segmentation تجزئة الجملة**

**Breaking the whole text up into sentences and tokens**

* **Word Tokenization تجزئة الكلمات**

**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)?**
  + **I feel AMAZING in 2023 !!!**
* **Stop words removal (at, In, or, the, for).=> in,i**
* **Converts to lower case.=> amazing**
* **Removing punctuation(# , . , / ,$,@ ),=> !!!**
* **Digits removal( remove numbers )=> 2023**
* **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 others steps preprocessing examples?**

**Text Normalization (map all different forms for word to one form)**

**Language Detection(التعرف على اللغة المستخدمة)**

**Code mixing (استخدام اكثر من لغة فى نفس المحادثة)**

**Transliteration** **نسخ كتابة لغة بحروف لغة اخرى**

**Affected language.(كوفيرتا و بوسطا)**

**Transliteration(copy the right of letters to other language letters but save their pronunciation)**

* **What is advanced Processing phases**

**Part of speech (Pos) tagging => verb noun Determiner**

**Parsing(syntactic mean)**

**Coreference Resolution (ahmed work with ali ,he is good friend)**

**Diagram

Description automatically generated**