NLP Milestone 1

Report

Part 2

Nayera Mahran 52-26764
Mariam Wael 52-1647
Khadiga Yehia 52-1145

Overview:

In this milestone, the objective is to **pre-process and analyze** a YouTube dataset from **ElDa7ee7** . The dataset consists of transcriptions from podcast episodes, which require thorough **data analysis, cleaning, and preprocessing** before applying advanced Natural Language Processing (NLP) techniques.

To achieve this, we performed the following key steps:
1. **Data Analysis** – Understanding the structure and content of the dataset.
2. **Data Preprocessing** – Removing unnecessary elements, normalizing Arabic text, Tokenization, stopword removal, and preparing text for further NLP tasks.

Each of these steps will be discussed in detail in the following sections.

Steps:

1. **Data Analysis:**

   Before applying any Natural Language Processing (NLP) techniques, we conducted an exploratory data analysis on the dataset to understand its structure and content.

   1.1. Each text file was analyzed separately to extract key statistics:

   - **Total Sentences**: The number of sentences in each file.
   - **Total Words**: The total number of words found in each file.
   - **Unique Words**: The count of distinct words in each file.
   - **Average Sentence Length**: The average number of words per sentence.
   - **Average Word Length**: The average number of characters per word.
   - **Top 10 Frequent Words Per File**

   1.2. Overall dataset Statistics:

   - **Total Files Processed**: 47 files
   - **Total Sentences**: `Summed across all files`.
   - **Total Words**: `Total count of words in the dataset`.
   - **Unique Words**: `Number of distinct words`.
   - **Average Word Length**: `Average number of characters per word`.

   1.3. All files are stored in a Dataframe.

1.4. N-Gram Analysis (Bigrams, Trigrams, Quadgrams):

N-grams help identify common phrases and recurring word patterns.

Techniques Used:

- **Bigrams**:

('يا', 'عزيزي')

8178 times

- **Trigrams:**

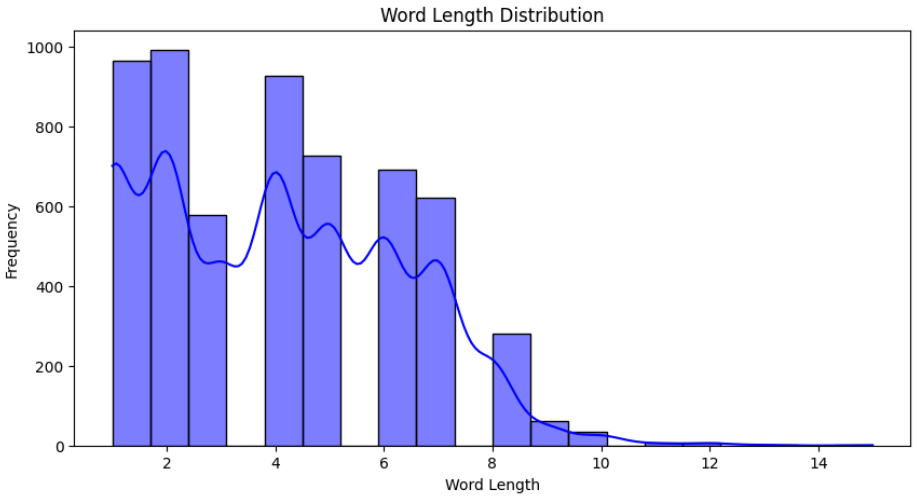
('يا', 'ابو', 'حميد')

1164 times

- **Quadgrams:**

('خلّيني', 'اقولّك', 'يا', 'عزيزي')

184 times



2. Data Preprocessing:

Data preprocessing is a crucial step in Natural Language Processing (NLP), ensuring that raw text is cleaned, structured, and ready.

2.1. Normalizing Data

We applied text normalization to standardize Arabic text. This process included removing punctuation and special characters, as well as unifying different forms of Arabic letters. Specifically, we converted all variations of "أ، آ، إ" to "ا", "ى" to "ي", "ة" to "ه", and similar adjustments to eliminate inconsistencies. This step was essential in improving the quality of the text for further NLP tasks such as topic modeling, and named entity recognition (NER).

2.2. Removing Stopwords

Initially we used the NLTK Arabic stopwords corpus, but we found that our dataset is in Arabic 3amy (dialectal Arabic) rather than Arabic Fusha (Modern Standard Arabic - MSA). As a result, many commonly used words in our dataset were not included in the NLTK stopword list. To address this, we expanded our stopword list by incorporating additional words from the Kaggle stopwords dataset. (https://www.kaggle.com/datasets/heeraldedhia/stop-words-in-28-languages?resource=download)

and further enriched it with frequently occurring words specific to our dataset, such as **"ده"**, **"بس"**, and **"ايه"**, ensuring more effective text preprocessing.

2.3. Removing unnecessary elements

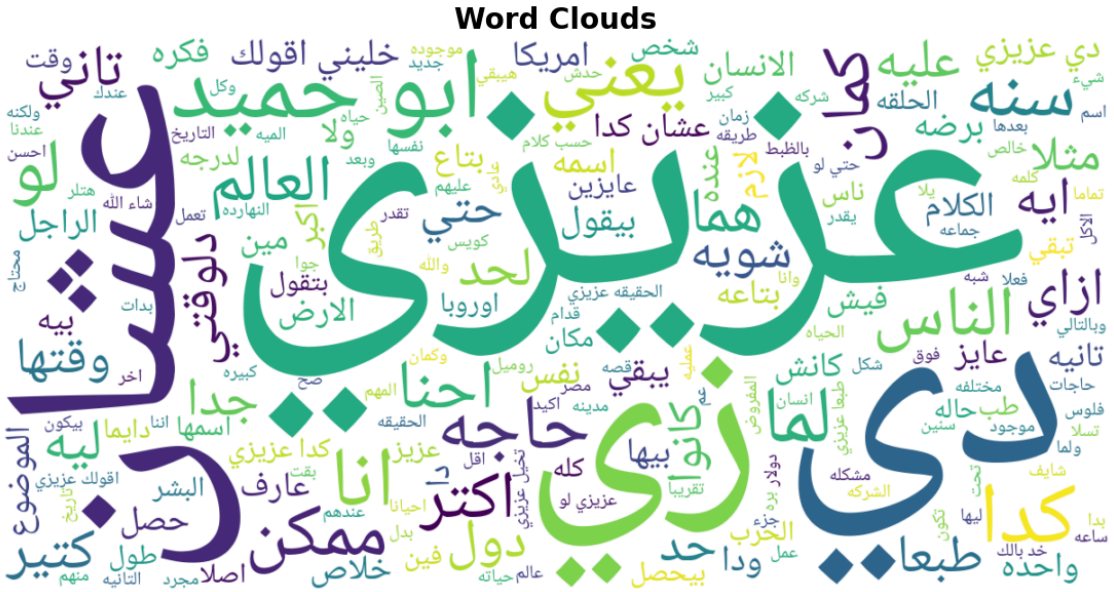
Since the podcast was sourced from YouTube videos, the transcriptions often included non-verbal annotations such as "[موسيقي]" to indicate background music, as well as timestamps marking different segments of the video. These elements were irrelevant to our analysis and introduced noise, so we removed them to focus on meaningful spoken content.

2.4. Word Cloud:

To better understand the most frequently occurring words in our dataset, we generated a word cloud that visually represents word frequency. In this visualization, larger words appear more frequently in the dataset, while smaller words are less common. Here we have 2 screenshots one before removing stopwords and unnecessary words and one after removing them.

Before removing stopwords:

After removing stopwords:



2.5. Named Entity Recognition(NER):

Named Entity Recognition (NER) was applied to extract important entities such as names , locations , and organizations from the podcast transcripts. we used a pretrained transformer-based Arabic NER model (`hatmimoha/arabic-ner`) from Hugging Face Transformers

```
(ORGANIZATION ,ماكدونالد) ,(PERSON ,امير توفيق) ,(LOCATION ,مصر)
(2005, DATE) , (عصام, PERSON), (ابو حميد, PERSON)
```

2.6. TF-IDF (Term Frequency-Inverse Document Frequency):

We applied TF-IDF to identify the most important words in the dataset by assigning higher weights to terms that appear frequently in a document but rarely across others. By limiting features to the top 20 words per document and extracting the top 5 highest-scoring words, we filtered out uninformative terms and highlighted key words for topic modeling and classification.

2.7. Translate English Words:

We used GoogleTranslate API to translate English words in the dataset into Arabic. Since the podcast transcripts contained many English words, and untranslated English words affected the clustering process negatively.

Task:

All these preprocessing steps were performed to clean and structure the dataset to ensure it was ready for clustering. After processing the text, we applied K-Means clustering, which grouped the podcast episodes into five distinct clusters based on content similarity. This helped in identifying patterns and categorizing discussions into meaningful groups.

```
==== TOP WORDS PER K-MEANS CLUSTER ====
Cluster 0: كلوب, رونالدو, كريستيانو, فيرجسون, الكوره, يورجن, النادي, الملعب, يورجن كلوب, يونايتد
Cluster 1: الحرب, خل, سنه, روميل, شركه, امريكا, البطاطس, الملح, العالم, هيس
Cluster 2: القنبله, الطياره, هيروشيما, الفضاء, كلاشنكوف, كوبا, اوبنهايمر, نوويه, قنبله, الذر يه
Cluster 3: الام, بيتهوفن, سنه, الاب, التشريح, الفيل, خل, التوحد, النباتات, اللبن
Cluster 4: سنه, الله, الانسان, العالم, خل, عارف, يبقي, دلوقتي, قوي, لحد
```

```
==== DOCUMENT CLUSTERING RESULTS ====
```

	Filename	Cluster_KMeans
1	Oppenheimer الدحيح.txt	2
2	آخر يوم في العالم الدحيح.txt	2
98	فورمولا ١ الدحيح.txt	4
99	فيلم رعب الدحيح.txt	2
100	كارت أحمر الدحيح.txt	0
101	كريستيانو رونالدو - الجزء الأول الدحيح.txt	0
102	كريستيانو رونالدو - الجزء الثاني الدحيح.txt	0