# NLP Milestone1

Sohaila Ibrahim
**ID:** 52-21225

Nada Elbehery
**ID:** 52-8973

March 10, 2025

## Contents

# 1 Overview

In Milestone 1, we focused on conducting a thorough data analysis and review of the dataset. This process involved exploring the structure and characteristics of the data, identifying patterns, trends, and any potential inconsistencies or irregularities. Additionally, we applied preprocessing and cleaning techniques to transform the raw dataset into a structured and well-organized format suitable for machine learning (ML) tasks.

# 2 Dataset Description

Our dataset consists of multiple text files containing transcripts in the Egyptian dialect. These texts are primarily sourced from various episodes of *El Mokhbir El Iqtisadi* (The Economic Informant), a popular Arabic-language YouTube channel that discusses global economic and political issues. The dataset covers a diverse range of topics, including:

- International trade and financial crises.
- Energy policies and resource control.
- Geopolitical conflicts and their economic implications.
- Technological advancements and digital economy dynamics.
- The economic rise and fall of corporations and industries.
- Social and economic trends in different regions.

Some files discuss historical events and economic strategies of major powers such as the United States, China, and Russia, while others focus on inflation, supply chain disruptions, and corporate financial maneuvers.

# 3 Methodology

Natural Language Processing (NLP) for Arabic text involves several preprocessing steps to enhance text representation for further analysis. This report details the sequential flow of preprocessing techniques applied to Arabic text, followed by TF-IDF weighting, Vector Space Model representation, and LDA topic modeling.

# 4 Preprocessing Flow

Farasa is a library designed for Arabic Natural Language Processing (NLP). It provides various tools for tasks such as tokenization, part-of-speech tagging, named entity recognition, stemming, diacritization, and segmentation. The preprocessing pipeline consists of multiple steps that leverage Farasa's modules to prepare Arabic text for analysis.

### 4.1 Segmentation

Segmentation is the process of breaking down words into smaller meaningful units, such as prefixes, stems, and suffixes. The `FarasaSegmenter` module is responsible for segmenting Arabic words into their morphemes.

```
from farasa.segmenter import FarasaSegmenter

def segment(text):
    segmenter = FarasaSegmenter()
    segmented = segmenter.segment(text)
    return segmented
```

### 4.2 Stemming

The `FarasaStemmer` module reduces words to their root forms.

```
from farasa.stemmer import FarasaStemmer

def stem(text):
    stemmer = FarasaStemmer()
    stemmed = stemmer.stem(text)
    return stemmed
```

### 4.3 Part-of-Speech Tagging

The `FarasaPOSTagger` module assigns part-of-speech (POS) tags to words in Arabic text.

```
from farasa.pos import FarasaPOSTagger

def tag(text):
    tagger = FarasaPOSTagger()
    tagged = tagger.tag(text)
    return tagged
```

### 4.4 Named Entity Recognition (NER)

The `FarasaNamedEntityRecognizer` module detects named entities such as persons, organizations, and locations in Arabic text.

```
from farasa.ner import FarasaNamedEntityRecognizer

def recognize(text):
    named_entity_recognizer = FarasaNamedEntityRecognizer()
    named_entity_recognized = named_entity_recognizer.recognize(text)
    return named_entity_recognized
```

## 4.5 Stopword Removal

Eliminating frequently occurring words with little semantic meaning. Predefined Arabic stop words were collected from the library arabicstopwords and extended with additional stop words to handle the Egyptian dialect.

## 4.6 Removing Non-Arabic Characters

Filtering out non-Arabic symbols to clean the text.

## 4.7 Sample text

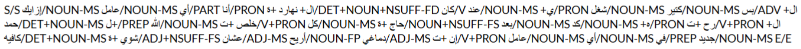
For example, given the input text:

إزايك عامل إيه أنا النهاردة كان عندي شغل كتير بس الحمد لله خلصت كل حاجة بعد كده رحت الكافيه شوية عشان أريح دماغي إنت عامل إيه في جديد

- After preprocessing, the text is transformed into:

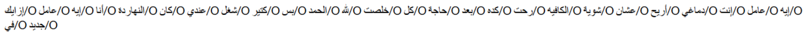
  إزايك عامل أي أنا نهارد ي شغل كتير حمد ل الله خلص كل حاج كد ه رح كافي شوي أريح دماغ إن عامل أي جديد

- After tagging, the text is transformed into:



```
S/S إزايك/NOUN-MS عامل/NOUN-MS أي/PART أنا/PRON +ه نهارد +ال/DET+NOUN+NSUFF-FD كان/V عند/NOUN-MS +ي/PRON شغل/NOUN-MS كتير/NOUN-MS بس/ADV +ال
حمد/DET+NOUN-MS +ل/PREP الله/NOUN-MS خلص +ت/V+PRON كل/NOUN-MS حاج +ة/NOUN+NSUFF-FS بعد/NOUN-MS كد/NOUN-MS +ه/PRON رح +ت/V+PRON +ال
كافيه/DET+NOUN-MS +ة شوي/ADJ+NSUFF-FS عشان/ADJ-MS أريح/NOUN-FP دماغي/ADJ-MS إن +ت/V+PRON عامل/NOUN-MS أي/NOUN-MS في/PREP جديد/NOUN-MS E/E
```

- After entity recognition, the text is transformed into:



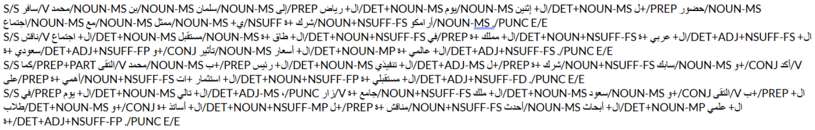
```
إزايك/O عامل/O إيه/O أنا/O النهارده/O كان/O عندي/O شغل/O كتير/O بس/O الحمد/O لله/O خلصت/O كل/O حاجة/O بعد/O كده/O رحت/O الكافيه/O شوية/O عشان/O أريح/O دماغي/O إنت/O عامل/O إيه/O
في/O جديد/O
```

Another example, input text:

سافر محمد بن سلمان إلى الرياض يوم الإثنين لحضور اجتماع مع ممثلي شركة أرامكو. ناقش الاجتماع مستقبل الطاقة في المملكة العربية السعودية وتأثير الأسعار العالمية. كما التقى محمد بالرئيس التنفيذي لشركة سابك وأكد على أهمية الاستثمارات المستقبلية. في اليوم التالي، زار جامعة الملك سعود والتقى بالطلاب والأساتذة لمناقشة أحدث الأبحاث العلمية.

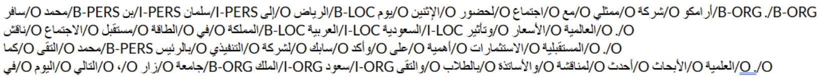
- After preprocessing, the text is transformed into:
  سافر محمد سلمان روضة يوم إثن حضور اجتماع ممثل شرك أرامكو ناقش اجتماع مستقبل مملك عربي سعودي تأثير سعر عالمي التقى محمد رئيس تنفيذي شرك سابك أكد أهمي استثمار مستقبلي يوم تالي زار جامع ملك سعود التقى طالب أساتذ مناقش أحدث بحث علمي

- After tagging, the text is transformed into:



```
S/S سافر/V محمد/NOUN-MS بن/NOUN-MS سلمان/NOUN-MS إلى/PREP ال+ رياض/DET+NOUN-MS يوم/NOUN-MS ال+ إثنين/DET+NOUN-MS ل+/PREP حضور/NOUN-MS
اجتماع/NOUN-MS مع/NOUN-MS ممثل/NOUN-MS ي+/NSUFF ة+ شرك/NOUN+NSUFF-FS أرامكو/NOUN-MS ./PUNC E/E
S/S ناقش/V ال+ اجتماع/DET+NOUN-MS مستقبل/NOUN-MS ة+ طاق ال+/DET+NOUN+NSUFF-FS في/PREP ة+ مملك ال+/DET+NOUN+NSUFF-FS ة+ عربي ال+/DET+ADJ+NSUFF-FS ال+
ة+ سعودي/DET+ADJ+NSUFF-FP و+/CONJ تأثير/NOUN-MS ال+ أسعار/DET+NOUN-MP ة+ عالمي ال+/DET+ADJ+NSUFF-FS ./PUNC E/E
S/S كما/PREP+PART التقى/V محمد/NOUN-MS ب+/PREP ال+ رئيس/DET+NOUN-MS ال+ تنفيذي/DET+ADJ-MS ل+/PREP ة+ شرك/NOUN+NSUFF-FS سابك/NOUN-MS و+/CONJ أكد/V
على/PREP ة+ أهمي/NOUN+NSUFF-FS ات+ استثمار ال+/DET+NOUN+NSUFF-FP ة+ مستقبلي ال+/DET+ADJ+NSUFF-FD ./PUNC E/E
S/S في/PREP ال+ يوم/DET+NOUN-MS ال+ تالي/DET+ADJ-MS ،/PUNC زار/V ة+ جامع/NOUN+NSUFF-FS ال+ ملك/DET+NOUN-MS سعود/NOUN-MS و+/CONJ التقى/V ب+/PREP ال+
طلاب/DET+NOUN-MS و+/CONJ ة+ أساتذ ال+/DET+NOUN+NSUFF-MP ل+/PREP ة+ مناقش/NOUN+NSUFF-FS أحدث/NOUN-MS ال+ أبحاث/DET+NOUN-MP ال+ علمي
ة+/DET+ADJ+NSUFF-FP ./PUNC E/E
```

- After entity recognition, the text is transformed into:



```
سافر/O محمد/B-PERS بن/I-PERS سلمان/I-PERS إلى/O الرياض/B-LOC يوم/O الإثنين/O لحضور/O اجتماع/O مع/O ممثلي/O شركة/O أرامكو/B-ORG ./B-ORG
ناقش/O الاجتماع/O مستقبل/O الطاقة/O في/O المملكة/B-LOC العربية/I-LOC السعودية/I-LOC وتأثير/O الأسعار/O العالمية/O ./O
كما/O التقى/O محمد/B-PERS بالرئيس/O التنفيذي/O لشركة/O سابك/O وأكد/O على/O أهمية/O الاستثمارات/O المستقبلية/O ./O
في/O اليوم/O التالي/O ،/O زار/O جامعة/B-ORG الملك/I-ORG سعود/I-ORG والتقى/O بالطلاب/O والأساتذة/O لمناقشة/O أحدث/O الأبحاث/O العلمية/O ./O
```

## 5 TF-IDF Weighting

TF-IDF (Term Frequency-Inverse Document Frequency) measures the importance of words in a corpus.

## 6 Vector Space Model

The Vector Space Model represents documents numerically, facilitating similarity measurements. For example, given a query vector:

```
Query vector: [0.19348638238589283]
Document ____________.txt vector: [0]
```

This allows us to identify the most relevant documents.

## 7 LDA Topic Modeling

Latent Dirichlet Allocation (LDA) is an unsupervised machine learning algorithm used for topic modeling, which aims to discover hidden topics within a collection of documents. It assumes that each document is a mixture of multiple topics, and each topic is characterized by a distribution of words. The function

**perform_lda(dic)** handles the creation of the model. It takes as an input a dictionary of word counts per document, structured as:

$$\texttt{dic} = \{\text{file\_name}_i : \{\text{word}_j : \text{count}_{ij}\}\} \tag{1}$$

where each document (file) contains words mapped to their frequency.

### 7.1 Processing Steps

1. Extracts unique words from each document:

$$\texttt{processed\_docs} = [\text{list of words from each document}] \tag{2}$$

2. Creates a Gensim dictionary mapping words to unique IDs.

3. Converts documents into a Bag-of-Words (BoW) representation:

$$\texttt{corpus} = [(\text{word\_id}, \text{frequency}) \text{ tuples}] \tag{3}$$

4. Trains an LDA model with:
   - `num_topics` = 6 (number of topics)
   - `passes` = 10 (iterations for better topic assignment)
   - `random_state` = 46 (for reproducibility)

### 7.2 Enhancing Topic Modeling with Named Entity Recognition (NER)

Named Entity Recognition (NER) can improve topic modeling by refining the textual representation of documents before applying Latent Dirichlet Allocation (LDA). We explore three different approaches to integrating NER:

1. **Use Only Named Entities:** In this approach, topic modeling is performed solely on named entities extracted from the documents.

2. **Boost Named Entities:** Here, named entities are given higher weights in the document representation, making them more influential in topic generation.

3. **Entities with Top 20 Words:** This approach combines named entities with the 20 most frequent words per document to balance contextual and specific information.

## 8 Conclusion

This NLP pipeline successfully processes Arabic text, extracts key information using TF-IDF and NER, and represents documents numerically for further analysis using the Vector Space Model and LDA topic modeling.