SMS Spam or Ham project

❖ Project description

We are a communications company that cares about fulfilling our customers' desires and listening to their requirements.

Thus, we received many complaints about the large number of unclear messages they had, which caused a lack of distinction between ham and spam messages.

To solve this problem due to the desire of our customers, we used a set of data and artificial intelligence algorithms to create a model and training to classify messages in terms of their type, whether they are ham or spam.

❖ Data description

- "Target": indicating the class of message as ham or spam
- "Text": column is the string of text.

❖ Tools

- Python libraries
- Jupyter notebook

❖ Steps for execute project

- ✓ Import All libraries that will be used.
- ✓ Load and read data :
  - Spam.csv
- ✓ Verifying Basic Data Integrity:
  - data.shape
  - data.info()
  - data.head()
  - data.describe()
- ✓ Dropped the column that will not be used.
- ✓ Data Exploration :
  - evaluate the target and find out if our data is imbalanced or not (spam or ham).
  - Adding a column of numbers of characters, words, and sentences.
  - Dropping the outliers.
- ✓ Pre-preprocessing text data “NLP” :
  - Clean data
    - Duplicate
    - Null value
    - Messing
    - Convert to Lower case.

- Tokenization :
  - is a way of separating a piece of text into smaller units called tokens. Here, tokens can be either word, characters, or subwords. Hence, tokenization can be broadly classified into 3 types – word, character, and subword (n-gram characters) tokenization.

- Stopwords Step :
  - Are the most common words in any natural language. For the purpose of analyzing text data and building NLP models, these stopwords might not add much value to the meaning of the document. Generally, the most common words used in a text are “the”, “is”, “in”, “for”, “where”, “when”, “to”, “at” etc... add no meaning to the statement while parsing it.

- Lemmatization:
  - In simpler forms, a method that switches any kind of a word to its base root mode is called Lemmatization. In other words, Lemmatization is a method responsible for grouping different inflected forms of words into the root form, having the same meaning. It is similar to stemming, in turn, it gives the stripped word that has some dictionary meaning. The Morphological analysis would require the extraction of the correct lemma of each word.

- Vectorization:
  - The process of converting words into numbers is called Vectorization.

✓ Build Model (Clustering):

○ What’s a cluster ?

Group of data points that are close to each other to make this computer-friendly, need a mathematical definition of “close ”.
Close (most common definitions): based on distance or density.
the clustering algorithm is many but we will use three algorithm for our project:

- **First:** K-mean

  K-mean is: A partitioning algorithm that divides the data into k clusters Points are assigned to a cluster based on the metric
  (such as Euclidean distance) to the nearest cluster centroid
  The value of k is chosen by the user

- **Second:** Hierarchical Clustering

  There are two types of hierarchical clustering: Agglomerative and Divisive:

  1. Agglomerative Hierarchical Clustering: In Agglomerative Hierarchical Clustering, Each data point is considered as a single cluster making the total number of clusters equal to the number of data points. And then we keep grouping the data based on the similarity metrics, making clusters as we move up in the hierarchy. This approach is also called a bottom-up approach.

  2. Divisive hierarchical clustering: Divisive hierarchical clustering is opposite to what agglomerative HC is. Here we start with a single cluster consisting of all the data points. With each iteration, we separate points that are distant from others based on distance metrics until every cluster has exactly 1 data point.

- **Third:** DBScan

  DBSCAN, or density-based spatial clustering used for clustering data points based on density, i.e., by grouping together areas with many samples. This makes it especially useful for performing clustering under noisy conditions: as we shall see, besides clustering, DBSCAN is also capable of detecting noisy points, which can – if desired – be discarded from the dataset.

Group names:

Nada alqabbani

Shahad almubki

Nada alhamad

Sarah alameer

Hala almulhim

Instructor:

Mohammed Baddar