SciPy Project – Sarcasm Detection

General

1. Data Preparation
    a. Kaggle: "News Headlines Dataset for Sarcasm Detection" by Rishabh Misr
    b. contains two files → training / testing → split it
2. Feature Extraction
    a. convert text data into numerical features for the machine learning model
    b. TF-IDF?
3. Model Training
    a. Suitable Classifier from sklearn
        i. Different Naïve Bayes
        ii. Random Forest
        iii. Support Vector Machines
        iv. KNN
4. Evaluation
    a. accuracy, precision, recall, and F1 score to measure
5. User Input Classification
    a. preprocess and pass it through the trained classifier
6. User Interaction
    a. plot the results: for each value and one plot for all results
    b. + written output of results

Project Structure

1. Root Directory
    a. Main.py
        i. Main script
        ii. User interactions, classification, plotting
    b. Preprocessing.py
        i. Preprocessing functions
        ii. Cleaning data, converting data to numerical
    c. InputPreprocessing.py
        i. Preprocessing of user input
    d. Model.py

        i. Model training, evaluation
    e. Requirements.txt
        i. Listing required libraries
    f. README.md
        i. Explaining project

2. Data Directory
    a. Train.csv
        i. Contains training data
    b. Test.csv
        i. Contains testing data

3. Results Directory (if user decides to save and plot results)
    → but user decides what he wants to see
    a. Prediction_results.csv
        i. save user input and results
    b. Sth to save the plots at
        i. Save plots

Libraries

- Scikit-learn
    - classifiers
- Pandas
    - Data manipulation, reading & processing csv files
- NumPy
    - computations
- Matplotlib
    - plotting
- nltk
    - text processing & tokenization