Anubhav Kundu
Midterm Project
CIS 4526

Features Designed:

- **Longest Subsequence:** Retrieves longest common subsequence of both sentences
- **# Words Overlapping:** Retrieves amount of overlapping words in both sentences
- **Synonym Amount:** Retrieves amount of possible synonyms comparing sentence 1 to sentence 2
- **Length Difference:** Retrieves the difference in amount of words present in both sentences
- **Similarity Ratio:** Used fuzzywuzzy library to calculate the similarity ratio of both sentence, using Levenshtein's distance
- **Cosine Similarity:** Computes cosine similarity of both sentences, inputting words present into vectors

Data Preprocessing:

- Data preprocessing was done using the Pandas libraries. The files were read through Panda's pd.read_csv() function
- NaN values were dropped using df.dropna()
- The gold label column was converted to integers from string object using .astype() function

Algorithms & Libraries:

- Libraries use include: **String, Scikit-Learn, Pandas, NLTK, Fuzzywuzzy**
- Scikit-Learn's **SVM** implementation was used to train the model

Experiences & Lessons:

- This project allowed for the experience of learning to solve a real world NLP problem using machine learning libraries
- Many hours were spent understanding how to use the Pandas and NLTK libraries as well as Scikit-Learn's SVM implementation
- Familiarized with defining features manually as well as comparing performances between different algorithms (SVM, Logistic Regression)
- Familiarized with using feature scaling to improve performance, as well as tuning hyperparameters for high accuracy, while avoiding overfitting of data