Classifier for Poems

By: Riya Jain

Problem Statement:
Construct a dataset of poems and try to classify them into 4 genres using 2 classifiers.

My approach to the problem:

1. Data Collection:
   Data has been majorly collected from the following sources:
   - https://www.familyfriendpoems.com/
   - https://www.poetryfoundation.org/collections
   - https://bookriot.com/2018/01/19/love-poems/

   The collected data has been filtered (according to word limit),and stored into a csv file.

   The dataset contains poems belonging to four genres: Love, Death, Nature and Spiritual. Distribution is as follows:

Type	No. of instances
Death	84
Love	86
Nature	83
Spiritual	95

2. Splitting dataset into train and test set:
   Dataset has been split as follows:
   Training set : 80%
   Test set : 20%

3. Countvectorizer and tfidf:
   Countvectorizer has been used to convert the poems to a matrix of token counts.
   Tfidf has been used to calculate inverse document frequencies.
   The goal of using tf-idf instead of the raw frequencies of occurrence of a token in a given document is to scale down the impact of tokens that occur very frequently in a given corpus and that are hence empirically less informative than features that occur in a small fraction of the training corpus.
   The output of tfidf is given as features to the classifiers.

4. Classifier:
   The submitted python script uses two classifier : Logistic Regression and LinearSVC to classify the given dataset.

   - Logistic Regression : Using sigmoid hypothesis function, the data points are classified using one vs all technique.

   - LinearSVC : A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. LinearSVC is a Support Vector Classifier(SVC) having linear kernel. It is flexible in the choice of penalties and loss functions and scales better to large numbers of samples.

5. Classification Results:

- Using Logistic Regression:
    - Weighted F1 score: 0.84
    - Accuracy: 84.3%

- Using LinearSVC:
    - Weighted F1 score: 0.85
    - Accuracy: 85.7%

Softwares/packages used:

- python 3
- Numpy
- Pandas
- Nltk
- Sklearn

All the needed packages can be installed by running the following command: ‘pip install -r requirements.txt’