By

Aditya Gupta

[adityagupta8052@gmail.com](mailto:adityagupta8052@gmail.com)

Problem:

Make a dataset of poems and try categorising them in 4 genres using two different classifiers

Approach:

1. Searched for websites which provide great poetry resources differentiated by genre.
2. Selected Four genres – Nature, Love, Humor, Spiritual from allpoetry.com
3. Wrote a python script(main.py) to extract poem URLs for different genres and saved them to a csv file “dataset.csv”.
4. Then wrote a python script (ExtractContent.py) to extract the poem text from those URLs (“dataset.csv”) and saved them as “MyDataset.csv”.
5. After getting the dataset, I have to pre-process it because most of the poems are duplicated (allpoetry.com was written badly).
6. Wrote a script on Jupyter Notebook (“DatasetPreprocessing.py”) for pre-processing the dataset(“MyDataset.csv”) and finally saved the dataset as   
   “Final\_Dataset.csv”).
7. Now again used jupyter notebook to write a script for pre-processing text data to feed it into classifier and used two classifiers – SVM and Multinomial Naïve Bayes and saved it as “Classifiers.py”.

Software Used:

1. Python – 3
2. Anaconda – 4.5.12
3. ChromeDriver - 73.0.3683.68

Packages Used:

1. Scikit-Learn
2. Selenium
3. Beautiful Soup
4. Pandas
5. NumPy