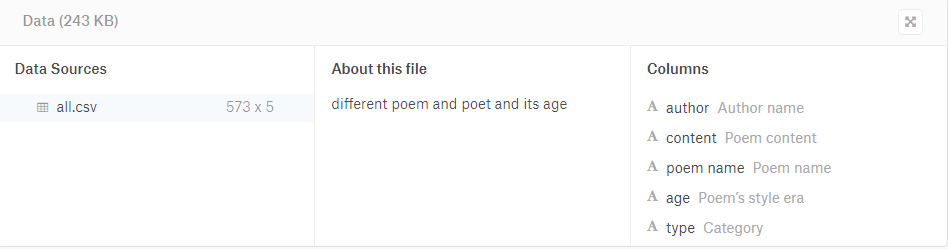
**REPORT**

**Software/Packages**

Jupyter notebook

Libraries – sklearn, nltk, textblob, numpy, pandas

**Data**



<https://www.poetryfoundation.org/> has the copyright for the data.

Data have 5 column and 573 rows as shown in figure. Column ‘content’ contains poems or corpora of poetry with various genres. Column ‘type’ contains genre of respective poem. Genres are of 3 types Nature, Love, Mythology & Folklore.

**Data processing**

As an initial step to get high quality input data, the corpora of poetry data was passed through a data cleansing process by removing all other secondary data in the texts. I converted all the character in “content” to lower case. The lower Case Tokens ensured that all words were in a uniform case, since poems have a lot of variance in their word-capitalizations which may prove detrimental to the study. Removed punctuation from the data, removed stop words from the data.

Stemming of the words was necessary so that all forms of word will have same meaning.

Next I converted “content” feature in to the vector form using TfidVectorizer so we can input this feature in to our models.

Next step was Splitting of Data into x\_train and x\_test which are 80% and 20% of the original data.

**Models used for classification are SVM and Naïve Bayes.**

Support Vector machine is a supervised learning model and is mainly used for classiﬁcation, pattern recognition and regression analysis. SVM with the ‘linear’ kernel provided an accuracy of 73.913% on our test data.

Naive Bayes is one of the simplest probabilistic algorithm. The basic idea behind this algorithm is to predict the occurrence of some event when certain other event is occurring. Naïve Bayes model provided an accuracy of 61.739% on our test data.