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

**The dataset used has 4 features:autor,content,poem name, age, type. The type here is the label, ie;the poetry genre.**

Preprocessing:

--Got rid of punctuation marks.

--removed the stop words ie; frequently used words

--Tokenization

--Stemming and lemmatization has not been done to keep the essence of writing style of different poetry.

-- Plotted the word clouds for different types of poetry to get intution of what kind

of words belong to different genre.

--plotted a pie chart to get distribution of different genre of poetry out of the entire

Dataset.

**MODEL 1:**

Text classification requires that text be represented as vectors of statistical features.

For the purpose of poetry classification we were interested in prosodic (the rhythm of the poem) and phonetic (the pronunciation of words) features. These two classes of features were used to identify poetic meter and rhyme.

We want accuracy of lexical stress detection.

I used GENSIM for this purpose. It’s a free python library..

It is used to detect scalable ststistical semantics.

It analyses the plain text for sematic structure and the model thus trained can be used to retrive semantically similar documents.

For genism , model is trained on entire dataset.(NOT USED HERE FOR PREDICTION)

Further , we have split the dataset into training and test set and used classifiers(SGD: This estimator implements regularized linear models with stochastic gradient descent (SGD) learning) to predict. The accuracy here couldn’t be improved further due to indufficient data.

We could also, alternatively, found the word map using tfidf vectorizer library.

**MODEL 2:**

For 2nd approach , I have used 2-layer LSTM, each layer has 512 hidden units.

For cost function, mean reduction is considered.

For optimization, Adam optimizer is used.

The coding has been done on jupyter notebook on lacal computer.