**Title:** Authorship identification of a Book

**Problem statement:** It is important to have an automated author identification system to detect plagiarism and to identify anonymous authors.

**Data:**

We will use a corpus consisting of books written by the following authors: Enid Blyton, J.K.Rowling, Jane Austin, Shakespeare, Virginia Wolf , Mark twain for training our model.

We will then use about 10 most popular/signature books per author and draw around 10,000 sentences randomly per author.

We plan to use the data from the Gutenberg repository <https://www.gutenberg.org>

**Methodology /Experiments/ Metrics**:

After downloading the raw data, we will clean and process it, so that it is fit to be fed into NLP models.

Pre-trained word embeddings from Glove will be used. We first apply a Logistic regression for baseline. Then we will apply a basic neural network from scratch followed by, an RNN, GRU, LSTM and Bert model to see which gives the best results. We will perform hyperparameter tuning to get the best model based on the evaluation metric.

The dataset should be balanced with around same amount of text from each author, hence we will check for accuracy as a metric.

We would also like to classify the text based on the genre for which we will experiment with around 10,000 sentences per Genre.

**Miscellaneous**:

Pytorch framework will be used for building models.

Cloud: Google colab will be used.

**References**:

<https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2760185.pdf>

<https://www.computer.org/csdl/pds/api/csdl/proceedings/download-article/12OmNqFrGtS/pdf>

<https://www.gutenberg.org/policy/permission.html>

<https://arxiv.org/pdf/2011.06149.pdf>