Group Number : 15
Project Title : Hindi Authorship Attribution
Rohan Chhokra Shwetank Shrey

Dataset : We collected our own dataset from www.hindisamay.com . Which isn't friendly for scraping data - it went down everytime we tried to automate it. Needless to say it was tough to acquire it but in the end we managed to get data for around 48 authors. The data which we've used for our training is in .txt format. However some of the data in rest of the 43 authors has pdf format.

Code:
Main code in folder named scripts . Programs have beend divided into the type of features they work on - Lexical, N-Gram and Punctuation scripts
This folder contains the programs which can be run in python3. Some of them save plots in plots folder and some of them show a plot while running.