CS 613- NLP

ASSIGNMENT-1
TEAM - 14
Language - French

Raw Dataset statistics:

https://docs.google.com/spreadsheets/d/1gQ-eICnAOyqB2F9NhmBiZYtZp_HCvC8SNulIDjmj3-U/edit?gid=0#gid=0

Dataset statistics after Cleaning:

https://docs.google.com/spreadsheets/d/1gQ-eICnAOyqB2F9NhmBiZYtZp_HCvC8SNulIDjmj3-U/edit?gid=1690993278#gid=1690993278

Dataset statistics after Deduplication:

https://docs.google.com/spreadsheets/d/1gQ-eICnAOyqB2F9NhmBiZYtZp_HCvC8SNulIDjmj3-U/edit?gid=1950151786#gid=1950151786

Individual Contributions of Group

Name	Roll NO.	Contribution
Husain Malwat	21110117	• Data scraping and crawling. • Data Cleaning and tabulation of cleaned data.
Amey Rangari	21110177	• Scraped data from various French websites to contribute to the raw dataset. • Data Cleaning and tabulation of cleaned data.

Netram Choudhary	21110138	• Crawled data • Deduplication of crawled data.
Vinay goud	21110125	• Scraped a significant amount of data from various websites and books. • Deduplication of crawled data.
Dhruv Patel	23210035	• Crawled Data for Haryanvi, Dzongkha and French as we changed our language a couple of times. • Deduplication of crawled data.