**HPE**

**Topic: DATA EXTRACTION AND THEME BASED NLP SUMMARIZER**

|  |  |
| --- | --- |
| **Task given** | **Task Completed/Status** |
| Explore the tools for gathering data from multiple sources | Pypdf2, Pdfminer, Tabula, Tesseract, Scrapy, Beautiful Soup, Parse hub, Selenium |
| Identify the best tools for data extraction | Pdfminer, Scrapy, Beautiful soup |
| Explore algorithms to summarize data | Gensim, LexRank, TextRank, Luhn, Bert, Skip Thought |
| Identify the best method to summarize data | To be decided |
| Optimization of algorithm to be more theme based | Not completed |
| Integration | On going |