# Reflective Diary Group 113

Group Number: \_\_113\_\_\_\_\_\_\_

|  |  |
| --- | --- |
| Full Name | Akshatha Shivashankar Chindalur |
| Role in group: | Developer |
| Role responsibilities: | Task 1: Sparse Feature Generation   1. Extraction of Bigrams 2. Remove Context Independent and Dependent stop words 3. Remove tokens less than 3 4. Remove rare tokens   Identified the correct order of pre-processing steps  Generated the output files:   1. Features with their respective index 2. Sparse Matrix   Cleaning and documentation of the code. |
| Contribution to Group: | 50 % |
| Learnings from Project | Learnt the importance of pre-processing before feature generation and how to parse PDF files to extract raw data. |
| Learning Techniques | Going through the lecture and tutorial notes helped to get a foundational understanding of text pre-processing. Looked at few more examples from online resources. |
| What went well | Team Work |
| What went wrong? | Initial struggle to identify the correct sequence of pre-processing. Confused during bigram extraction and stemming. |
| Resolution to solve issues: | Understood the requirement and output format better after which the code was altered accordingly. |
| **Overall Conclusion:** The assignment help us gain knowledge required for text preprocessing and how data can be collected from raw text in PDF files. | |

|  |  |
| --- | --- |
| Full Name | Pradnya Rajendra Alchetti |
| Role in group: | Developer |
| Role responsibilities: | * Downloading the 200 URLs programmatically * Reading the PDF file into text   **Task 1 - Sparse Feature Generation:**   * Retrieving paper body and initial cleaning of the PDF text * Sentence segmentation and case normalization * Word tokenization * Multiprocessing using pool * Initial development for removing bigrams and context-independent stopwords (later cleaned up by my partner)   **Task 2- Statistics Generation:**   * Retrieving paper abstract, title, and authors * Word Tokenization * Removing context-independent words from abstract and title * Generating statistics CSV for most frequent terms |
| Contribution to Group: | 50% |
| Learnings from Project | Learned how to parse PDF files, extract data and pre-process using various techniques and convert it into a suitable format so that it can be useful for further topic modeling tasks. |
| Learning Techniques | Used the website ​ https://regex101.com​ to design regular expressions to extract suitable content from the file.  Reading through the python documentation of the nltk library helped to  understand how tokenization and bigram extraction methods can be applied to the given content. |
| What went well | The group member's contribution to the assignment went well. |
| What went wrong? | We were struggling to understand the sequence of techniques to be applied to the extracted data to retrieve meaningful information from the content. But this helped us to explore and understand different ways of natural language processing. |
| Resolution to solve issues: | More research and exploration of how to apply different NLP techniques coupled with trial and error. |
| **Overall Conclusion: The given assignment had several interesting tasks that helped us learn to**  **parse raw into structured data and apply various natural language processing techniques to extract meaningful information which can be used for analysis.**  **How would you do it, if asked to do it again? I would first decide the sequence in which different text extraction techniques should be applied and then attempt to start coding.** | |