Sardar Vallabhbhai National Institute of Technology (SVNIT) Surat

Department of Artificial Intelligence

B.Tech. Artificial Intelligence

B. Tech. III (AI) Semester – V **NATURAL LANGUAGE PROCESSING** **AI357 Scheme**	**L**	**T**	**P**	**Credit**
	3	**0**	**2**	**04**

Assignment 1 Text Preprocessing

1. You need to complete 4 tasks.
    a. Visit https://huggingface.co/datasets/ai4bharat/IndicCorpV2 website and download the data from your language. Extract all the data.
    b. You need to write codes for a sentence tokenizer and word tokenizer. Tokenize each paragraph into sentences and words. Tokenize each word. Your tokenizer should tokenize punctuations, URLS, numbers (handle decimals), mail ids, dates.
    c. After your data is tokenized, save them into a file or multiple files.
    d. Then compute the following corpus statistics:
        i. Total number of sentences
        ii. Total number of words
        iii. Total number of characters
        iv. Average Sentence Length (Average number of words per sentence)
        v. Average word length (Average number of characters per word)
        vi. Type/Token Ratio (TTR) (Total number of unique tokens / Total number of tokens)

2. Repeat the same steps on a huge monolingual corpora available at https://huggingface.co/datasets/oscar-corpus/OSCAR-2301