Lab Assignment

Submit to: submit2aam@gmail.com

⚠ COPY IS NOT ACCEPTED

If anyone found copied from someone both will be severely penalized. I will neither show any mercy nor accept any request. Please BE CAREFUL in this regard.

👍 TRY IT YOURSELF

If anyone does not copy from someone but tries herself/himself will be appreciated and rewarded.

1. Accomplish your assigned task on dataset. This is your bilingual corpus or parallel corpus.
2. Develop a python program to split sentences in a Bangla dataset. Save the program as *nlp_sent_<roll>.py*. Write some key observations in *nlp_sent_<roll>.pdf*
3. Develop a python program to tokenize a Bangla dataset. Save the program as *nlp_tokenizer_<roll>.py*. Write some key observations in *nlp_tokenizer_<roll>.pdf*
4. Develop a python program to accomplish the following tasks: (Save the program as *nlp_histogram_<roll>.py*.). We will provide access to your parallel dataset when you finish so that you can perform following tasks.

    - Extract following statistics from your parallel corpus:

	English side	Bangla side
Corpus size (in words) excluding punctuation		
Corpus size (in chars) excluding spaces		
Average sentence length (in words)		
Vocabulary size (no. of unique words)		
Lexical diversity*		
Corpus size (in lines)		

* How frequently on average each vocabulary item appears in the corpus.

    - Top ten frequent words in your parallel corpus:

English side			Bangla side		
Words	Frequency	(%)	Words	Frequency	(%)

- Draw some contrastive pictures between two languages from the statistics you observed and write in *nlp_histogram_<roll>.pdf*