

Assignment-1

Natural Language Processing

Instructor: Muhammad Umair

Topics: Data Preprocessing

- Use the NLTK Python package for this assignment.

Q1: Text data is available to a great extent which is used to analyze and solve business problems. But before using the data for analysis or prediction, processing the data is important. To prepare the text data for the model building we perform text preprocessing. It is the very first step of NLP projects. Some of the preprocessing steps are:

1. Removing punctuations like., ! \$() * % @
2. Removing URLs
3. Removing Stop words
4. Lower casing
5. Tokenization
6. Stemming
7. Lemmatization
8. Sentence Segmentation

Apply these preprocessing steps on given datasets or you can take any public datasets of your choice.

Sample Datasets:

- [Large Movie Review Dataset](#)
- [Sentimental Analysis for Tweets](#)
- [E-Mail Spam classification](#)

Reference Material:

- [Natural Language Toolkit](#)
- [Google Colaboratory as Python IDE](#)
- [Regular Expression Documentation](#)
- [W3School for Sample RE Implementation](#)
- [Build and Test RE Online](#)