**IT412: Natural Language Processing**

# Assignment 1: Preprocessing in NLP

*Instructor:* Prasenjit Majumder

## 1 Problem description

Data preprocessing is an essential step for building robust and reliable machine learning models. In NLP, preprocessing consists of a pipeline which converts the corpus into a format that allows a model to efficiently solve a given task.

## 2 Implemetation

### 2.1 Dataset

We will be using the datasets provided in the following tutorial for our implementation:
https://www.kaggle.com/l3nnys/useful-text-preprocessing-on-the-datasets

### 2.2 Exercise

The primary objective of this assignment is to help you gain familiarity with loading and cleaning text data using python modules like Pandas, NLTK, RegEx etc.. After loading the dataset you are required to,

1. Remove punctuation, unwanted tags, i.e. noise removal
2. Expand contractions, for example, replace ”could’ve” with ”could have”,
3. Tokenization
4. Convert the text to lowercase
5. Remove stop-words
6. Perform Stemming and Lemmatization (separately) and see how they differ.

## 3 References

- NLTK documentation: https://www.nltk.org/
- NLTK tutorial: https://www.youtube.com/watch?v=X2vAabgKiuM
- RegEx tutorial: https://www.w3schools.com/python/python_regex.asp
- Pandas tutorials:
    1. https://data36.com/pandas-tutorial-1-basics-reading-data-files-dataframes-data-selection/
    2. https://pandas.pydata.org/pandas-docs/stable/getting_started/10min.html#min
    3. https://pandas.pydata.org/pandas-docs/stable/user_guide/cookbook.html#cookbook
- https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html