Lab-Assignment-1: Social Media Data Tokenization & Normalization

Writeup

**Dataset**

* The dataset for the tokenizer and normalizer is tweets that were scrapped from twitter
* These tweets were scrapped using GetOldTweets3 which is a Python 3 library for accessing tweets.
* GetOldTweets3 allows you to scrape tweets based on username, query search, username along with bound dates as well as the top tweets from a particular user
* I scrapped the top 50 tweets from the ‘timesofindia’ twitter handle and stored it in a text file to create the dataset

**Creation of Tokenizer and Normalizer**

* The first step was to consider each punctuation mark as a separate token. After observing the dataset this included: {‘.’, ‘?’, ‘!’, ‘,’, ‘:’}. These were separated by checking if a word ends with any of these characters and separating the two if present
* The next step was to handle single quotation marks in the sentence. These could indicate either a clitic or a possessive apostrophe or a quote.
  + Clitics:

Clitics had to be handled in both normal format like you’re as well as slang format like youre. Both must be converted to you are.

**Challenge:** There is a problem in the case of we’re as the slang can be were which may refer to past tense form of “to be” instead of “we are” hence “were” cannot be considered as a slang clitic.

* + Quotes:

Quotes indicating a certain person’s statement could be represented via single or double quotes this needs to be separated

**Challenge**: A word may end with a punctuation mark but be followed by a quote. This would not detect it when we check if it ends with a punctuation mark

**Challenge**: A word may contain a quotation mark followed by a punctuation mark like a colon or semi colon.

Problems : “These are apples”