

Assignment 1

B.Prabhakar
201505618

Points of Interest:

- ✓ Handling the **punctiation** symbols correctly, as the punctuation sybols present in email-ids, 's, 'nt, urls and mentions form a hurdle.
- ✓ Handling the **ellipses** at the end of the tweet.

Solution proposed:

1. Ignore the ellipses that are present at the end of the tweet. Looses the continuity if the tweets are segmented.
 2. Tokenize the ellipses into 3 individual tokens so that let the consuming application can handle the segmented tweets.
- ✓ Different types of emoticons.
 1. Handle the unicode emojis and textual emojis seperately.