**Problem Statement:**

#### Task 1: To break the twitter hashtags into proper words

#### Task 2:  To find out what people have said about these three personalities (Clint Eastwood, Bradley Cooper and Chris Kyle). Show 5 top trends per person. The tweet count should be in decreasing order.

**Data Understanding:**

1. Data consist of 100,000 rows which is scraped from twitter and possess the sentiments of people related to movie named American Sniper.
2. Data contains various Urls, Emoticons, Numbers, Extra spaces, Symbols, Hashtags and Mentions which needs to be processed.
3. Data also possess non-English language tweets.

**Data Exploration:**

1. Data is imported into the environment
2. Word count of each row is calculated
3. Character count of each row is calculated

**Data Preprocessing:**

1. Hashtags which are written together are separated initially (eg: word like americansniper is converted into american sniper)
2. Created a separate data frame for Clint Eastwood, Bradley Cooper and Chris Kyle from the twitter data
3. Created new columns for Hashtags and Mentions
4. Non-English words or Non ASCII words are removed from the data
5. Digits are removed
6. Punctuations are removed
7. Apostrophes/short words are processed *e.g. we’re converted to we are*
8. Whole dataset is converted to lower case
9. Stop words are removed e.g. *is*
10. Urls, mentions, hashtags etc. removed
11. Noise removed: Words like “rt” and “retweet” are replaced with blanks
12. Whitespaces are removed
13. Corrected the spellings *e.g. Goood converted to Good*
14. Slangs processed *e.g. luv converted to love*
15. Unigrams and Bigrams created to find out top trends per person
16. Word cloud created per person based on Unigrams

**Output:**

1. Clint Eastwood
   * Unigram:

Clint Eastwood

\*\*\*\*\*\*\*\*\*\*\*\*\*

interpretation (1364tweets)

great (333tweets)

director (280tweets)

oscar (224tweets)

box (163tweets)

office (152tweets)

* + Bigram:

Clint Eastwood

\*\*\*\*\*\*\*\*\*\*\*\*\*

box office (151tweets)

director guild (128tweets)

navy seal (113tweets)

guild selma (96tweets)

ignores history (62tweets)

sienna miller (51tweets)

1. Bradley Cooper
   * Unigram:

Bradley Cooper

\*\*\*\*\*\*\*\*\*\*\*\*\*

great (735tweets)

oscar (650tweets)

actor (284tweets)

amazing (250tweets)

performance (240tweets)

navy (225tweets)

* + Bigram:

Bradley Cooper

\*\*\*\*\*\*\*\*\*\*\*\*\*

navy seal (192tweets)

great actor (163tweets)

sensational portrayal (147tweets)

portrayal navy (141tweets)

sienna miller (122tweets)

oscar noms (118tweets)

1. Chris Kyle
   * Unigram:

Chris Kyle

\*\*\*\*\*\*\*\*\*\*\*\*\*

memorial (862tweets)

edition (845tweets)

author (803tweets)

honor (786tweets)

hero (743tweets)

great (567tweets))

* + Bigram:

Chris Kyle

\*\*\*\*\*\*\*\*\*\*\*\*\*

memorial edition (829tweets)

scott mcewen (427tweets)

navy seal (413tweets)

author scott (401tweets)

edition author (400tweets)

edition memorial (381tweets)

**Packages used:**

1. Pandas
2. NLTK
3. Matplotlib
4. Wordcloud
5. Ujson
6. Ekphrasis
7. tweet-preprocessor
8. textblob
9. re
10. sklearn (CountVectorizer)
11. string

Submitted by Suhas Pote.

Email: [suhaspote8@gmail.com](mailto:suhaspote8@gmail.com)

Mobile: +91-9702495051