**AIWIR – Lab Task1**

**Dataset:**  [https://www.kaggle.com/c/nlp-getting-started/data?select=train.csv](https://www.kaggle.com/c/nlp-getting-started/data?select=train.csv" \t "_blank)

**Tasks**

1. Tokenize each Tweet into sentences
2. Tokenize each tweet into words
3. Remove stopwords in each tweet - NLTK library

**Tokenization links:**

[https://towardsdatascience.com/tokenization-for-natural-language-processing-a179a891bad4](https://towardsdatascience.com/tokenization-for-natural-language-processing-a179a891bad4" \t "_blank)

[https://www.geeksforgeeks.org/nlp-how-tokenizing-text-sentence-words-works/](https://www.geeksforgeeks.org/nlp-how-tokenizing-text-sentence-words-works/" \t "_blank)

**Stopword links:**

[https://www.geeksforgeeks.org/removing-stop-words-nltk-python/](https://www.geeksforgeeks.org/removing-stop-words-nltk-python/" \t "_blank)

[https://www.analyticsvidhya.com/blog/2019/08/how-to-remove-stopwords-text-normalization-nltk-spacy-gensim-python/](https://www.analyticsvidhya.com/blog/2019/08/how-to-remove-stopwords-text-normalization-nltk-spacy-gensim-python/" \t "_blank)

**Create a folder containing the ipynb files for the above 3 tasks.  
------------------------------------------------------------------------------------------------------------**

**AIWIR – Lab Task 2**From the lab1, the list of terms from the twitter data set is ready.

1. **Construct inverted index** that will have terms in the dictionary and in the posting list, mention row no.s of the tweets where term occurs   
2. **Implement the merge algorithm.** ( so given terms as input eg. ‘election AND voter’, retrieve all tweets row no where election and voter both terms occurs ( merging posting list of ‘election’ and ‘voter’ ).

-----------------------------------------------------------------------------------------------------------