Tweets clustering

Sources/references we used in this project:

- Lecture slides
- Reference Jaccard distance: https://en.wikipedia.org/wiki/Jaccard_index
- Dataset: https://archive.ics.uci.edu/ml/datasets/Health+News+in+Twitter

Tweets Preprocessing:

Firstly, the tweets are preprocessed using the following steps:

- tweet ids and timestamps are removed
  - words that start with the symbol '@', e.g., @AnnaMedaris, are removed
- hashtag symbols are removed, e.g., #depression is converted to depression
- any URL are removed
- every word is converted to lowercase

- Here, the tweets are clustered using Jaccard distance metric and K-means clustering algorithms.

the best efficient K value here will be: k = 7

- We have reached our minimum SSE (sum of squared error) by using the number of clusters to 7 and it will be: SSE -> 3292.2105529424507

Experiment Analysis on"bbchealth"dataset:

```
------ Running K means for experiment no. 1 for k = 3
running iteration 0
running iteration 1
converged
1:  1613 tweets
2:  1348 tweets
3:  968 tweets
--> SSE : 3397.920622955551


------ Running K means for experiment no. 2 for k = 4
running iteration 0
running iteration 1
converged
1:  836 tweets
2:  1489 tweets
3:  872 tweets
4:  732 tweets
--> SSE : 3346.372127852338


------ Running K means for experiment no. 3 for k = 5
running iteration 0
running iteration 1
converged
1:  1140 tweets
2:  856 tweets
3:  589 tweets
4:  521 tweets
5:  823 tweets
--> SSE : 3349.610365334829


------ Running K means for experiment no. 4 for k = 6
running iteration 0
running iteration 1
converged
1:  513 tweets
2:  838 tweets
3:  502 tweets
4:  664 tweets
5:  889 tweets
6:  523 tweets
--> SSE : 3308.2658638448484


------ Running K means for experiment no. 5 for k = 7
running iteration 0
running iteration 1
converged
1:  709 tweets
2:  371 tweets
3:  509 tweets
4:  449 tweets
5:  1132 tweets
6:  407 tweets
7:  352 tweets
--> SSE : 3265.6399514003638
```

Credits:
- Ahmed Osama Gouda (Klevy)
- Nour Mohamed Abdelaziz
- Ibrahim Mohamed Ibrahim
- Mohamed Rafat
- Abdelrhman Amr Mohamed