CS 6375
ASSIGNMENT 2

Names of students in your group:
Aditya Kulkarni (axk230069@utdallas.edu)

Number of free late days used: 0

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

Report:

1. Data File Used

From the given data, we are using "gdnhealthcare.txt" file. The file has *2997* samples.
Format of file: <id>|<date time information>|<tweet>
The tweets contain URLs, hashtags, and user ids.
Note: any other data file can be used in the same code via command line argument.

2. Pre-Processing

For pre-processing the data we have used simple string functions like replace(), split() and strip(), as well as regular expressions for removing URLs and other symbols.

i. Remove the tweet id and timestamp.

```
20          delim = lines[i].split("|")[2:]
21          lines[i] = " | ".join(delim)
```

ii. Remove any word that starts with the symbol @ e.g. @AnnaMedaris.

```
24          lines[i] = " ".join(filter(lambda x: x[0] != '@', lines[i].split()))
```

iii. Remove any hashtag symbols e.g. convert #depression to depression.

```
27          lines[i] = lines[i].replace('#', '')
```

iv. Remove any URL.

```
30          lines[i] = re.sub(r"http\S+", "", lines[i])
31          lines[i] = re.sub(r"www\S+", "", lines[i])
32          lines[i] = lines[i].strip()
```

v. Convert every word to lowercase.

```
35          lines[i] = lines[i].lower()
```

vi. Removed all punctuations and other symbols.

```
38          lines[i] = re.sub('[^A-Za-z0-9 ]+', '', lines[i])
39          lines[i] = " ".join(lines[i].split())
```

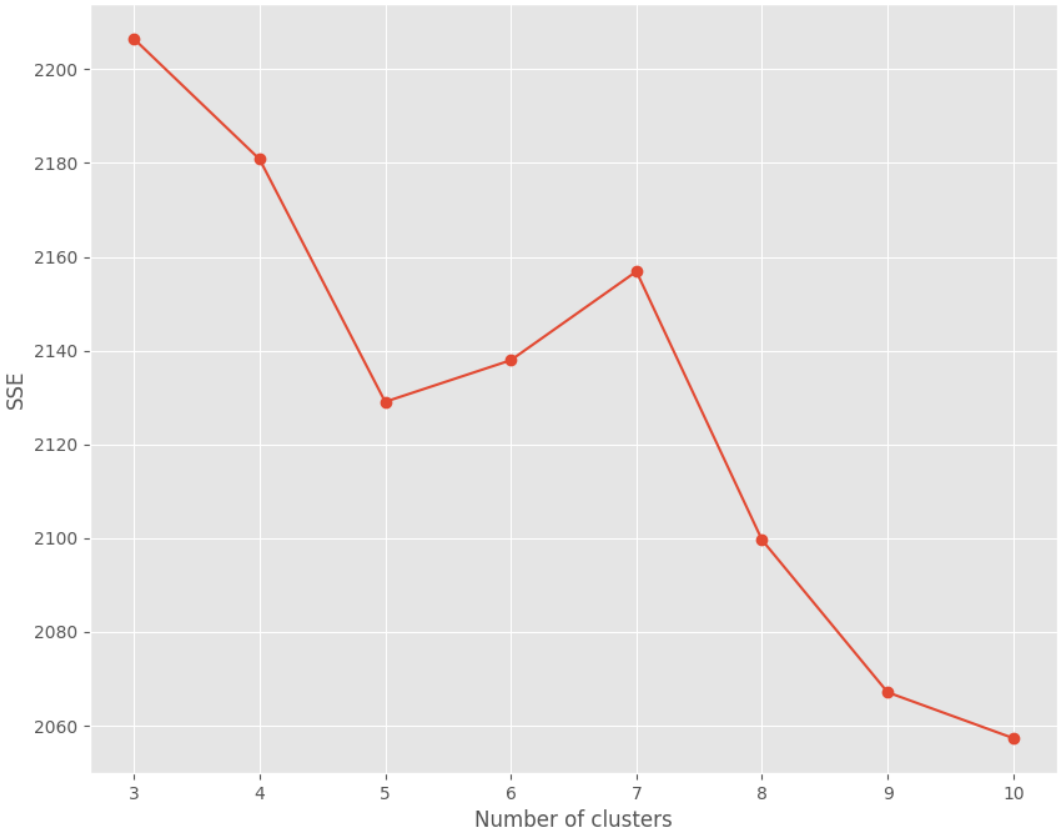
3. Results

The results are generated on a single file.

Jaccard distance is used to calculate distance between two sentences:

```
72    def jaccard_distance(t1, t2):
73            return 1 - (len(set(t1).intersection(t2)) /
74                        len(set().union(t1, t2)))
```

For experiments, we have set value of k from 3 to 10. The SSE and cluster distribution is noted below:



k	SSE	Clusters
3	2206.47	Cluster 0 Length: 1069 Cluster 1 Length: 1444 Cluster 2 Length: 327
4	2180.80	Cluster 0 Length: 583 Cluster 1 Length: 796 Cluster 2 Length: 401 Cluster 3 Length: 1060
5	2129.08	Cluster 0 Length: 483 Cluster 1 Length: 392 Cluster 2 Length: 649 Cluster 3 Length: 885 Cluster 4 Length: 431
6	2137.95	Cluster 0 Length: 993 Cluster 1 Length: 528 Cluster 2 Length: 216 Cluster 3 Length: 296 Cluster 4 Length: 553 Cluster 5 Length: 254
7	2156.88	Cluster 0 Length: 296 Cluster 1 Length: 79 Cluster 2 Length: 167 Cluster 3 Length: 828 Cluster 4 Length: 607 Cluster 5 Length: 221 Cluster 6 Length: 642
8	2099.62	Cluster 0 Length: 218 Cluster 1 Length: 308 Cluster 2 Length: 352 Cluster 3 Length: 265 Cluster 4 Length: 173 Cluster 5 Length: 429 Cluster 6 Length: 68 Cluster 7 Length: 1027
9	2067.13	Cluster 0 Length: 178 Cluster 1 Length: 51 Cluster 2 Length: 208 Cluster 3 Length: 283 Cluster 4 Length: 445 Cluster 5 Length: 543 Cluster 6 Length: 245 Cluster 7 Length: 373 Cluster 8 Length: 514
10	2057.36	Cluster 0 Length: 332 Cluster 1 Length: 718 Cluster 2 Length: 102 Cluster 3 Length: 353 Cluster 4 Length: 329 Cluster 5 Length: 344 Cluster 6 Length: 218 Cluster 7 Length: 52 Cluster 8 Length: 210 Cluster 9 Length: 182