Report for Week 3 – 15 April – 19 April

I. Actions performed

1. **[Folder t-SNE**] - Built t-SNE representation for all timeframes of 16-cut version (see folder **t-SNE** (in most of the cases we can observe that we have about 2 clusters excepting T3) *maybe T3 is too small, this could explain the heatmap problem*. Maybe it's worth trying to merge it with T2 or T4. For T5 I did more tries (see folder T5) but there I couldn't get a nice clustered representation.
2. **[Folder T2**] – After obtaining a nice representation with grouped items, I selected a certain timeframe, ie T2, and tried to see if I can obtain a clustering / partition. For this step, I used k-Medoids and results can be seen in kmedoids_tsne_t2.png. The partition of the users seems to be quite good, I obtained 2 main clusters. Moreover, in *cluster_i* images, we can see the topics in the 2 clusters. One of them is about Brexit (cluster 2) , the other one is about voting, referendum, leaving / remaining, people decision.
3. **[Folder T1_T14_Same_Figure**] – I plotted the two most different periods of time in the same figure (T1 and T14), to see how groups evolve. The target was to detect the pro / against groups (or another polarization) and see if we can detect these groups in multiple time frames, despite the drifting topics. To obtain figure *tsne_T1_T14* :

1) I tried to get term matrix (rows are users, cols are frequent words) for T1 and for T2. Then I merged the two matrices, and where a matrix contained a word that the other one didn't contain it, I put 0 in the one that didn't contain that word. (eg: matrix for T1 is :

```
   w1 w2 w3
u1...
u2 ...
```

while matrix for T2 is :

```
   w2 w3 w4
u3
u4
```

the resulting matrix would be:

```
    w1 w2 w3 w4
u1  ....  ....  ....  0
u2  ....  ....   ....  0
u3  0   ....  ....   ....
u4  0    ....  ....  .....
```

4. In figures tsne_T1 and tsne_T14 we can see the clusters in the 2 time frames (the 2 groups of people).Figures *wordcloud_cluster_i_T_j* show the word clouds and the distribution of words in all 4 situation (T1 – cluster 1, cluster 2 and T2 – cluster 1, cluster 2)
5. **[Folder Clustering**] – we can see a comparison, done on T1 respectively T14, between k-medoids and dbscan. The latter is based on the density of points.
6. **[Folder Transitions**] – Back to T1 vs T14, plotted in the same figure. There are 21 common authors between the 2 periods of time. For each of them, I created a figure, with it's starting location and the ending location.