Report for Week 5 – 06 May – 09 May

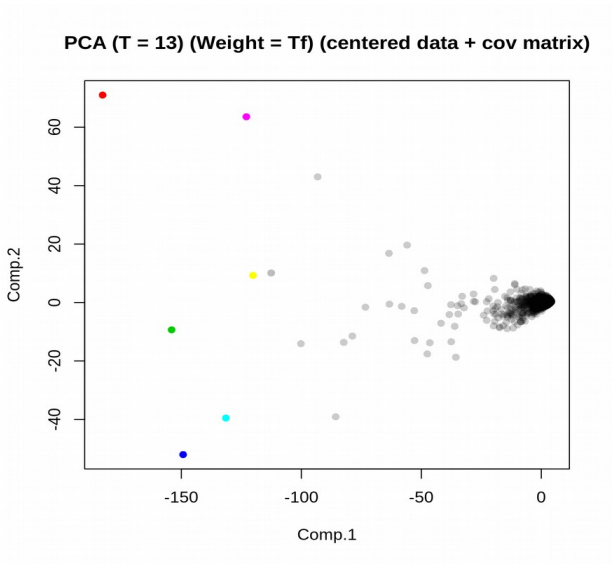
I. Summary

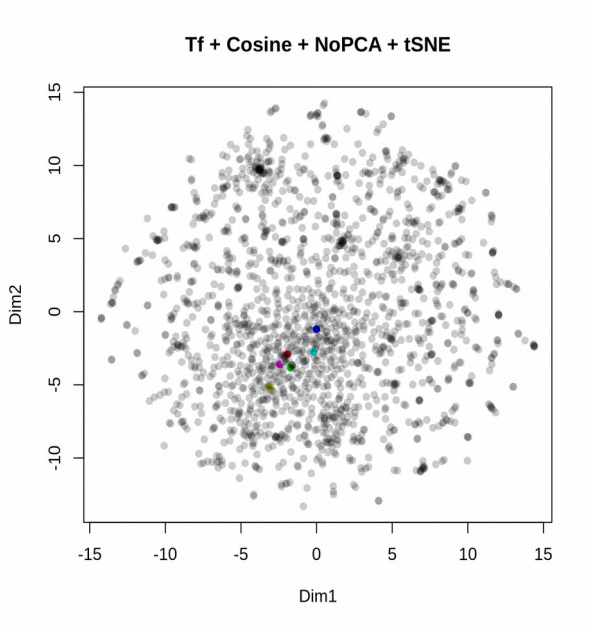
1. I tried to find a link between the PCA and t-SNE TF representation (w.r.t those points that were far away in PCA representation)
2. I remade the Periods graph, enhanced with extra information.
3. I performed 5-Cross Validation for LDA on all time-frames to find out which values for k = no. of topics would suit our needs).
4. I ran LDA on more time-frames, with different values for k and performed clustering on them. **Important discovery: The number of minPts (needed for hdbscan, does not affect the clustering, provided that it is greater than the number of dimensions of the data (no. of topics) + 1.**
5. I ran LDA, Clustered, aggregated the distributions inside a cluster, and analyzed the distances between the aggregated distributions.

II. Actions performed

1. PCA vs t-SNE analysis

As we discussed last time, I tried to plot the outliers from the PCA 2D representation in the t-SNE representation. The results can be observed in the following 2 images.





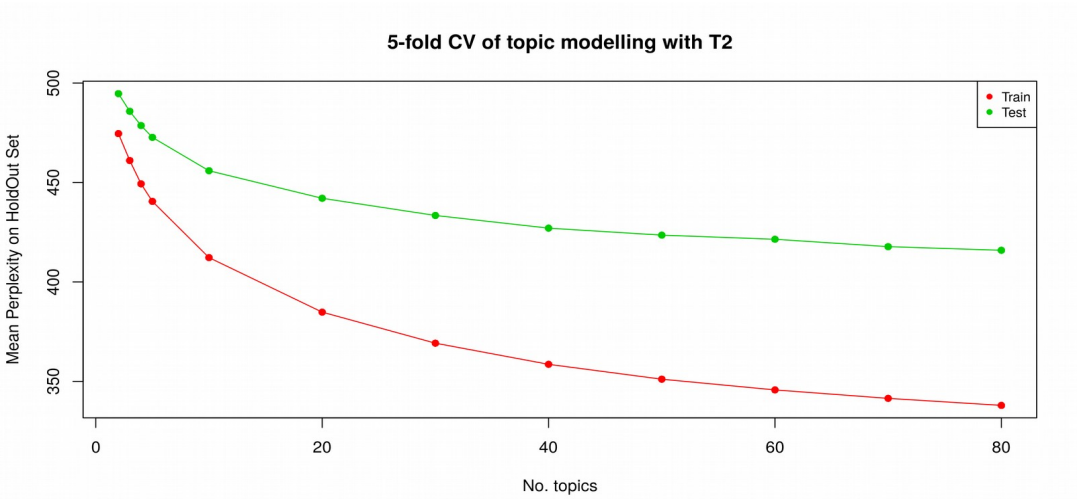
In the picture on the left, we have the PCA representation and the points that are far away might be interesting to analyse. However, when plotted using a t-SNE representation, in the TF space, using cosine similarity, they tend to be in the central part of the group.

2. Timeframe table

I remade the periods time table. This time, with equal sized cells for the equal-cut strategy. Moreover, I added extra details about the number of users in each period and about the number of posts and the events in that period. The new splitting tables are in ***Periods.pdf***

3. K-Cross Validation for the LDA

As we talked last time, to asses the quality of the LDA representation, I performed a k-Cross validation process, with k = 5. (training on 4 parts, and testing on the 5th). I used as a evaluation metric the perplexity. This is a well-known metric used in LDA for assessing quality of the models on new unseen data. Basically, it is a decreasing function of the log likelihood (the smaller the better). I performed the tests on most of the time-frames and the results can be noticed in the following figure.



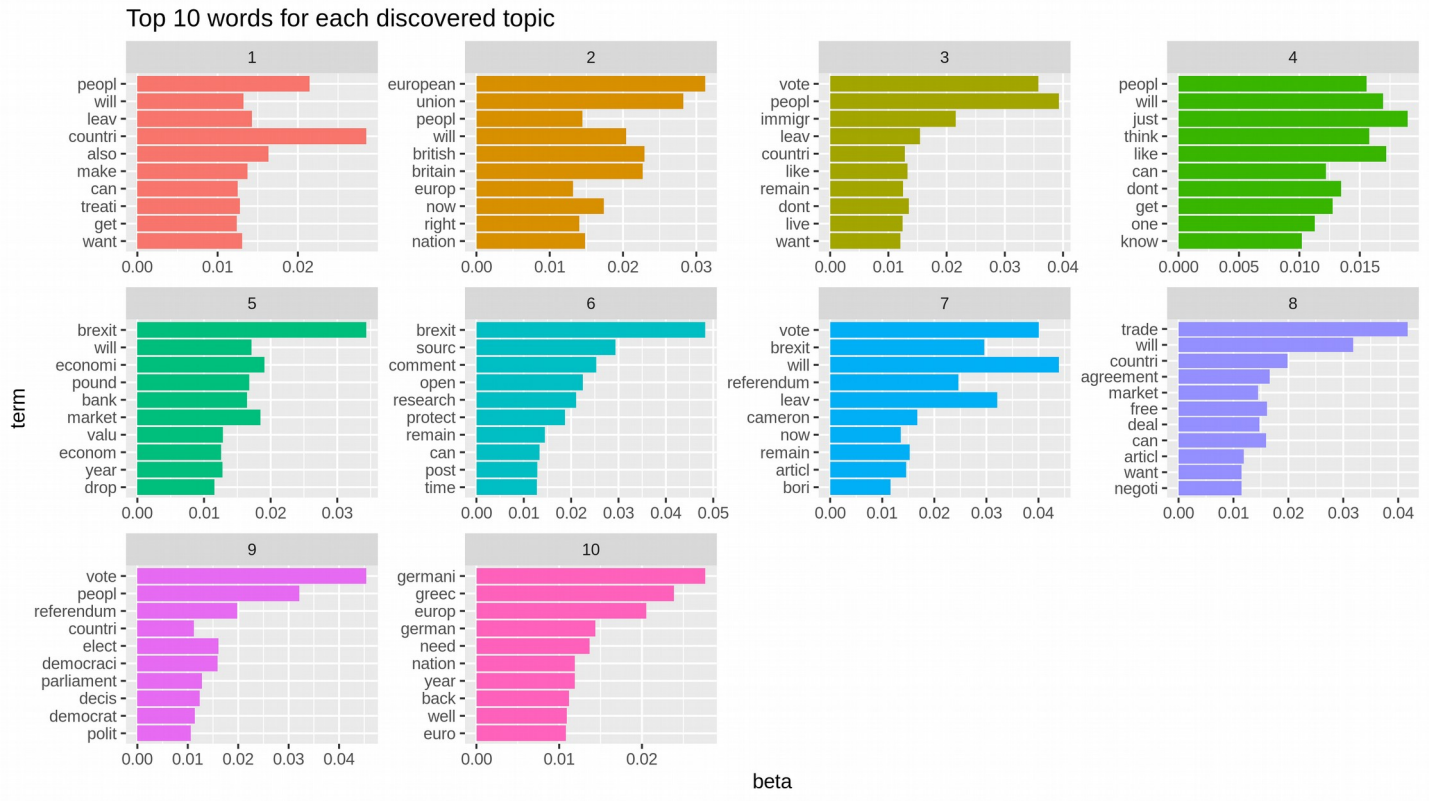
In the above figure, with red we have the training perplexity and with green the test one (the one measured on unknown data). We can observe that the more we increase k = number of topics, the more we overfit the data, which is reflected by the high difference between the green and the red spots for high values of the no of topics. **Based on an analysis on all the timeframes which were tested, I would conclude that a value of k between 5 and 10 would be ok for our needs.** Values over 10, tend to have good results at train time, but start to overfit, however, values smaller than 5 do not offer a good quality of the results (test time perplexity is still large). The results of the runs on all time-frames are available in the folder ***LDA_CrossValidation_Periods.***

4. LDA on multiple time-frames

The main purpose at this moment is to find a good representation space in which to perform clustering, so that we can advance the study. To do this, I spent time on analysis of the different representation spaces.

a) Period 2 – 10 topics

I ran LDA on period 2, using a number of **10 topics** as I explained in the previous subsection.

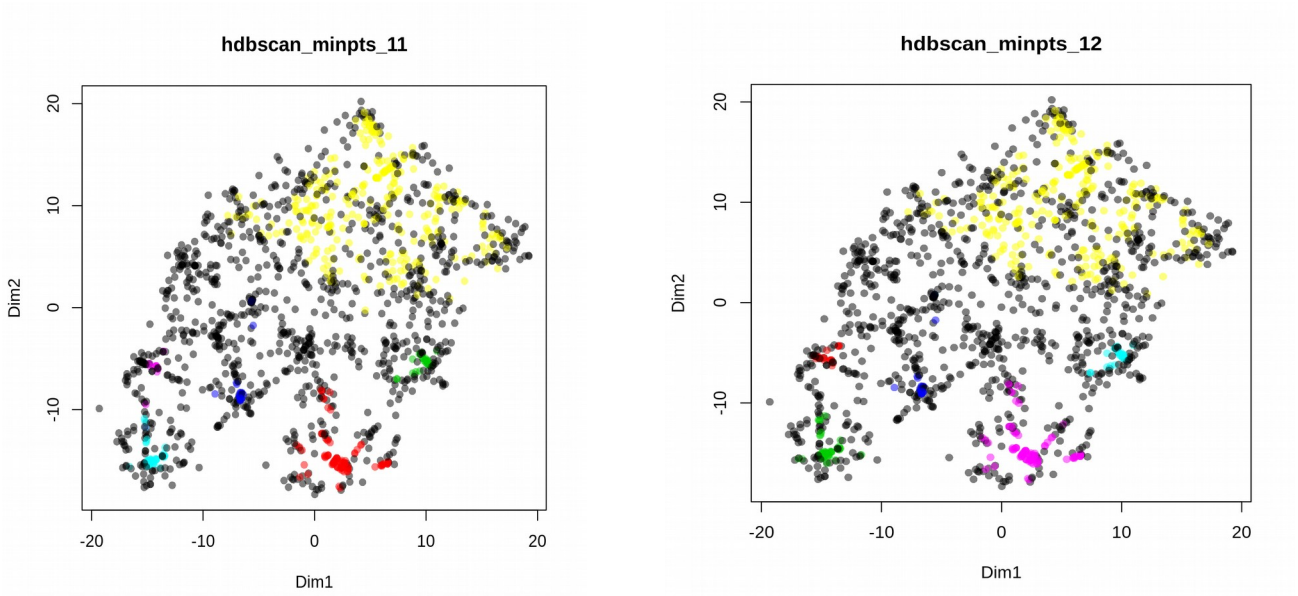


The results can be observed in the above picture. We can notice that the groups of words (topics) tend to look better. As we will see in the clustering part, there is always some noise in the replies. However, for example in topic 2, we can see that people talk about european union, britain etc. In the 3rd one about immigration etc.

The second part, after obtaining the topic representation, I tried to perform clustering, using dbscan. **I would suggest using dbscan over K-means as this one deals with noise as well, and it doesn't force me to pick a number of clusters.**

The only parameter to be tuned is the minPts (to look around), but according to the paper and to the R package implementation, this minPts can be set to the number of dimensions + 1, thus it's the number of topics + 1.

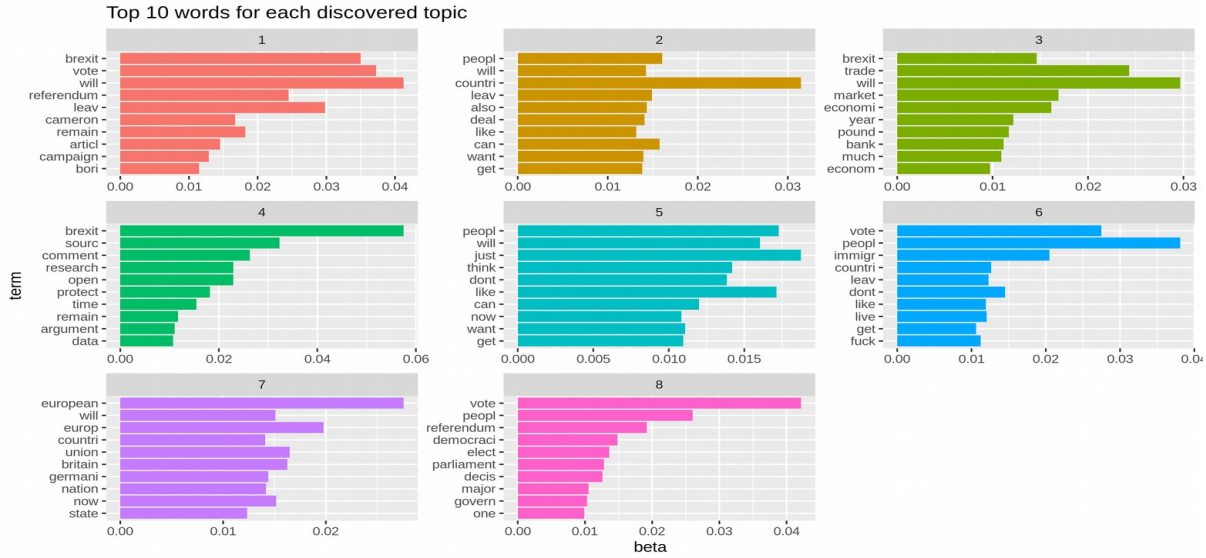
This is good news, as the number of topics is known to be around 5-10 for all periods, this has been exploited in the K-Cross validation task.



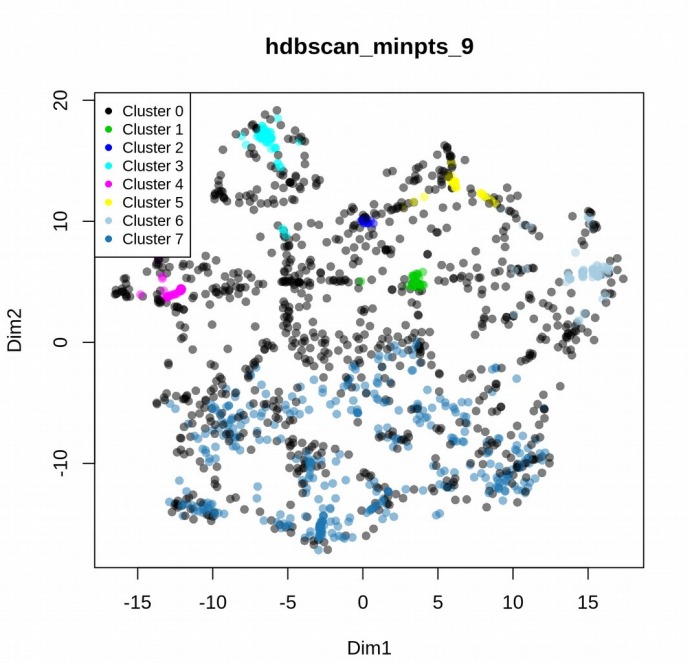
In the above figures, we can see that the results do not differ so much, when modifing the minPts from 11 to 12. In the folder ***LDA_Period 2/k10*** there are also examples for minPts 13 and 14.

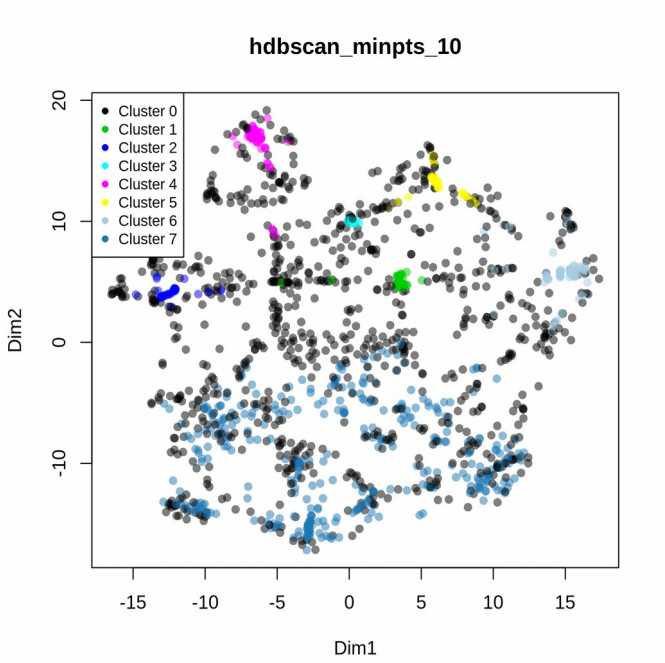
b) Period 2 – 8 topics

I also performed some tests on the same period, with a different number for the embedded topic space. ***LDA_Period 2/k8*** contains all the figures for this setup.

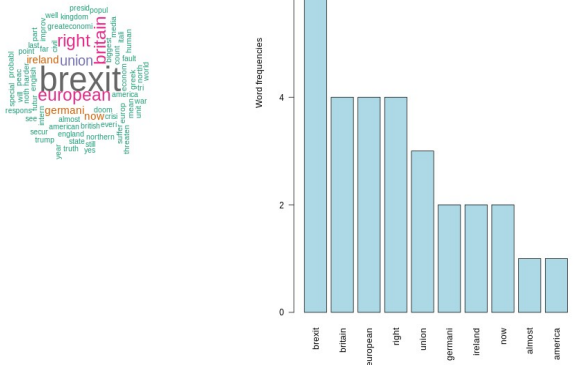


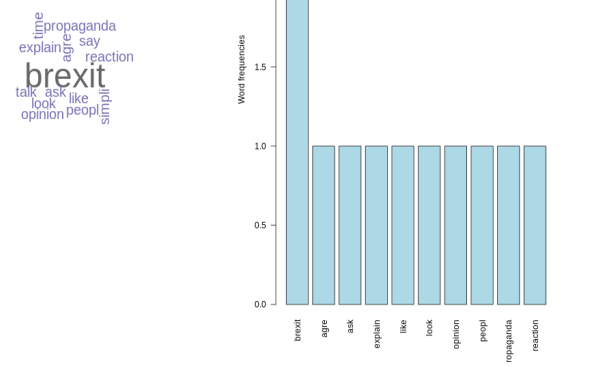
We can observe that some topics are discovered by both 8-topics and 10 topics LDA. This is a good sign, because it proves some stability to the choice of the number of topics. Having a look at the clusters obtained, using dbscan, we can see that the number of minPts does not affect that much the clustering. For instance:





In the two above figures, we can see that the number of clusters remains the same, and the way they are distributed is very similar. For the setup: 8 topics, minPts 9, I also analyzed each cluster composition in terms of words used. Some of the clusters are similar to the topics found by LDA, while other ones find other subjects of discussions. For example the cluster presented in the following figure on the left, has the same subject as topic 7 presented in the previous figure. However, the one on the right, presented a new topic, the one about opinions and propaganda.





In a similar manner, I analyzed the setup for 6 topic, and the results can be observed in folder ***LDA_Period 2/k6***

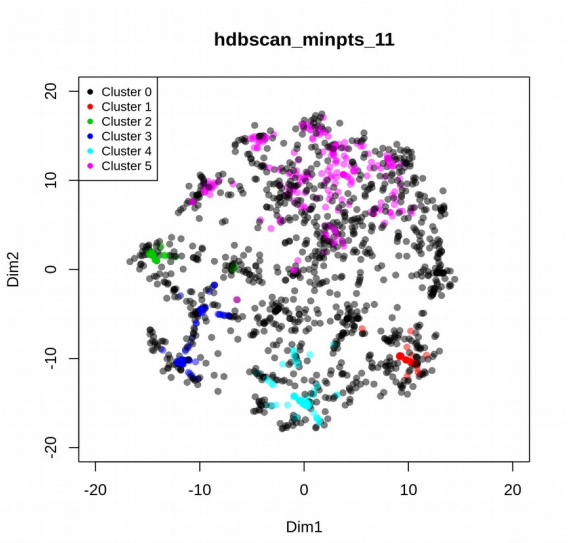
The same analysis was done for Periods 3 and 14 and we can discuss the results in the meeting.

5. LDA – Analysis of the distribution of topics inside clusters and between clusters

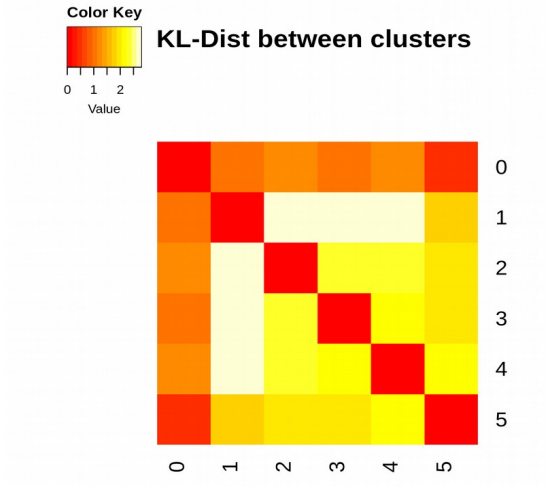
As we discussed, I ran LDA, then in the topic space performed clustering with hdbscan and then I aggregated the distributions of probability inside each found cluster and performed distance analysis on them. The results bring confirmation that we can trust the LDA representation as presented in the following figures, from folder ***LDA_Period_2_Distribution_of_topics.***

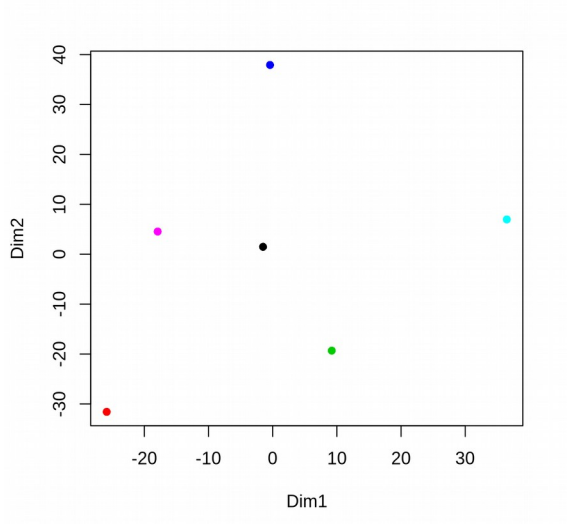
I did some tests on both 5 topic scenario and on 10 topics scenario, each time having the minPts set up at no_of_topics + 1.

a) **topics10_minPts11**



In the above figure, we can see that we have 5 clusters + one noise group (black ones). So, I aggregated all the probabilities in every cluster by averaging them and obtained a 10 dimension probability vector. Then I computed the mean standard deviation of the new average vector.





After obtaining the mean vectors, I computed the dissimilarities between them using KL Divergence. In the above figure on the left, we have a heatmap – the main diagonal is 0 and we can also observe that the 5th cluster, is on average close to the 0 one. This claim is also supported by the figures depicting clusters – black one is close to the violet one. Moreover, in a t-SNE representation, the 2 are also close to each other.

The same experiment was done also on 5-topics results are available in folder ***LDA_Period_2_Distribution_of_topics/topics5_minPts6.***