Report for Week 7 – 13 May – 17 May

I. Summary

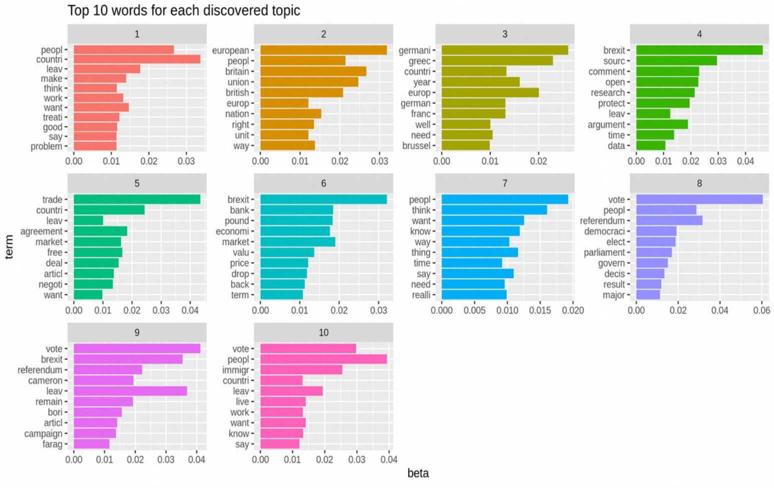
1. [**IMPORTANT!**] I did investigate more the issue of **LDA not being consistent across different runs**.
2. I analyzed the distances between **topics in each period.**
3. I analyzed the distances between **periods, in terms of topics.**
4. I continued the work with the hdbscan clustering and plotted the centroids of each cluster in the corresponding cluster.
5. I obtained the closest terms from the centroids and the closest posts from the centroids.

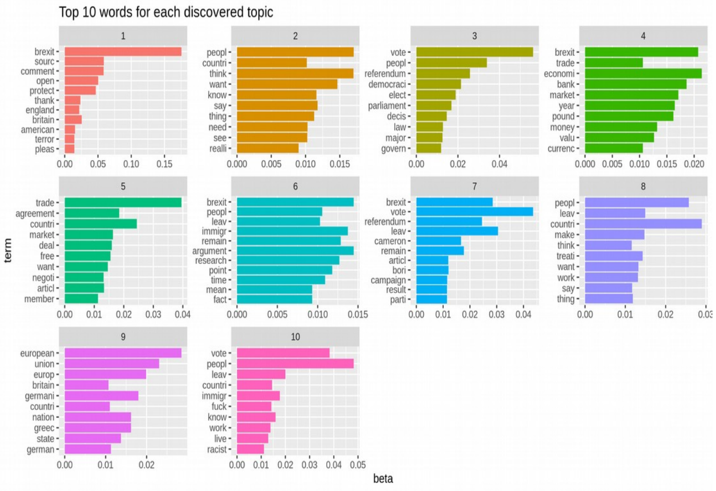
II. Actions performed

1. Inconsistency of LDA across consequent runs

After applying LDA and obtaining the topics, I perform the hierarchical clustering using hdbscan. Once I have the clusters, I compute the centroids. When trying to compute the closest posts to the centroids, I observed some inconsistency on the LDA models. This lead to further investigation which revealed the following fact: when running the LDA modeling algorithm several times, the topic models obtained tend to differ from time to time.

- The probability distribution of words over topics tend to remain similar over a number of different runs.

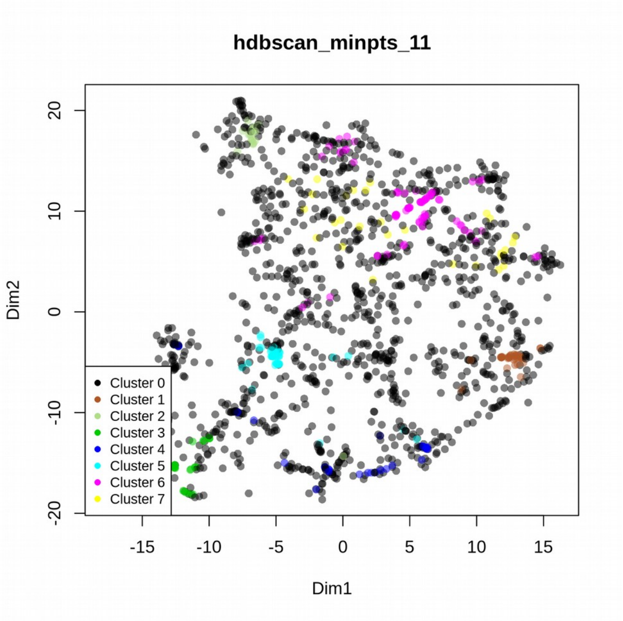


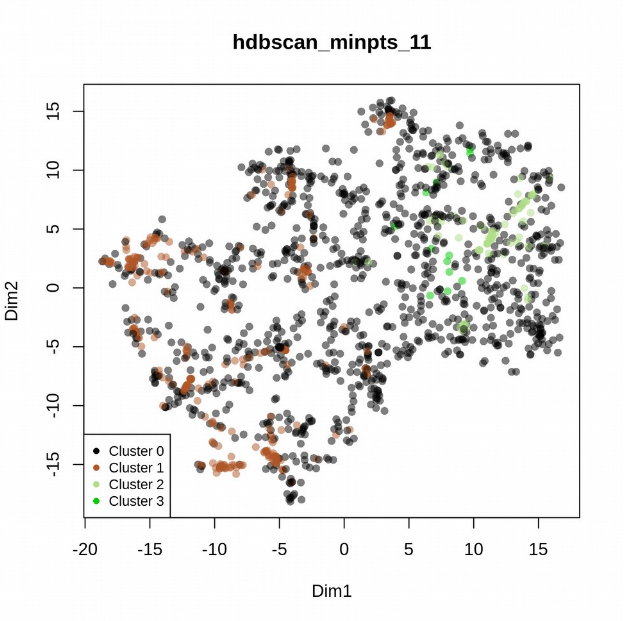


In the above figures we can observe that the distribution of terms over topics remains the same in two different runs. For more similar images from other runs check folder ***problem_with_lda.***

- The probability distribution of topics over documents (users posts) sometimes tends to change. I noticed this by running several times the clustering (which is performed among users, in the topic space).

The following figures are taken by repeatedly running topic modeling with 10 topics on timeframe 2, then performing clustering with dbscan.

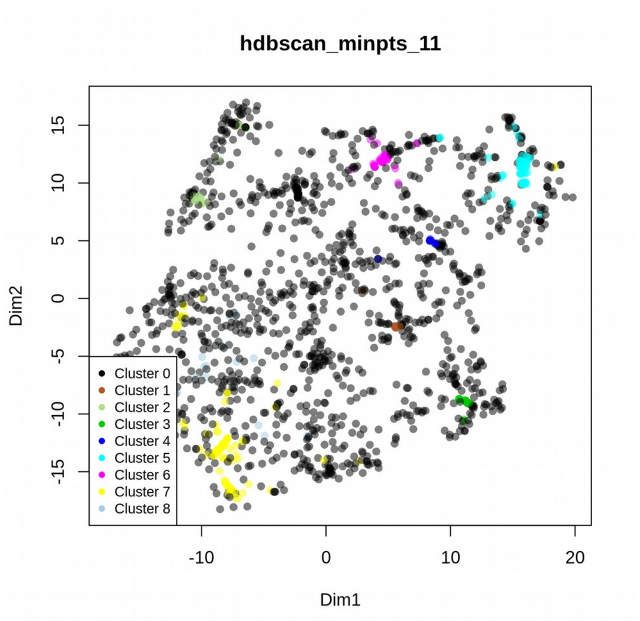




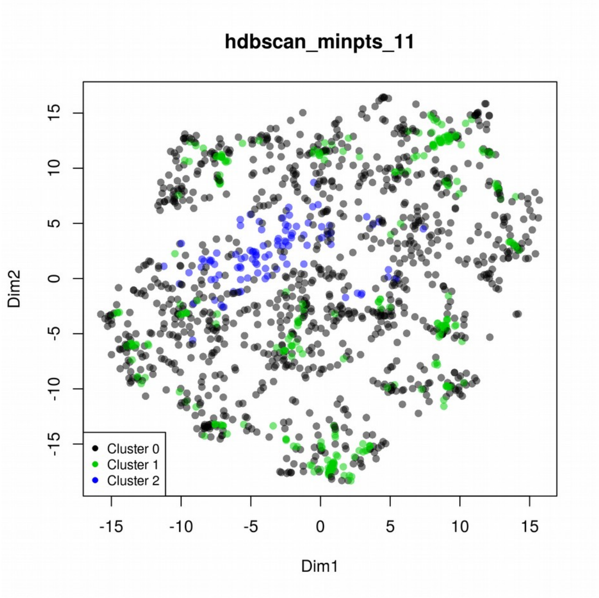
In the figures above we can see how the clustering algorithm gave pretty similar result for the first and last figures, but on the second one we only obtain 3 clusters.

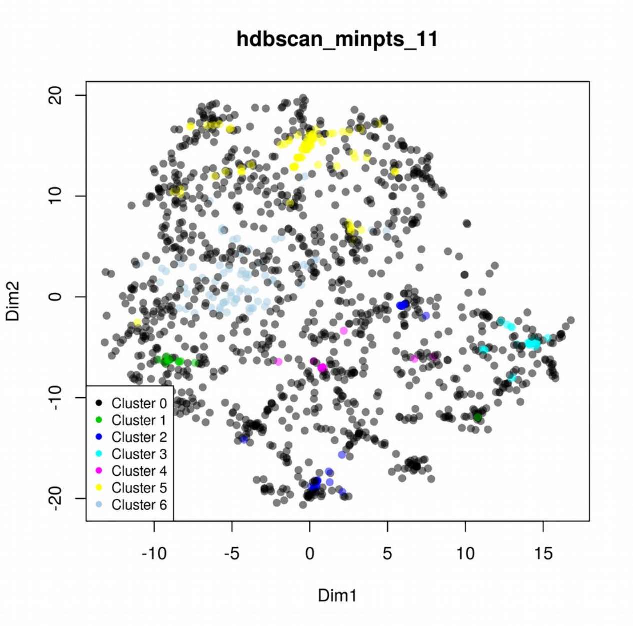
For more similar images from other runs check folder ***problem_with_lda.***

A decision needs to be made on whether we will use LDA in future, or not. To help in making this decision, I performed the above tests (LDA + clustering) on a couple of time periods, for each periods repeating the test for 20 times. The results can be observed in



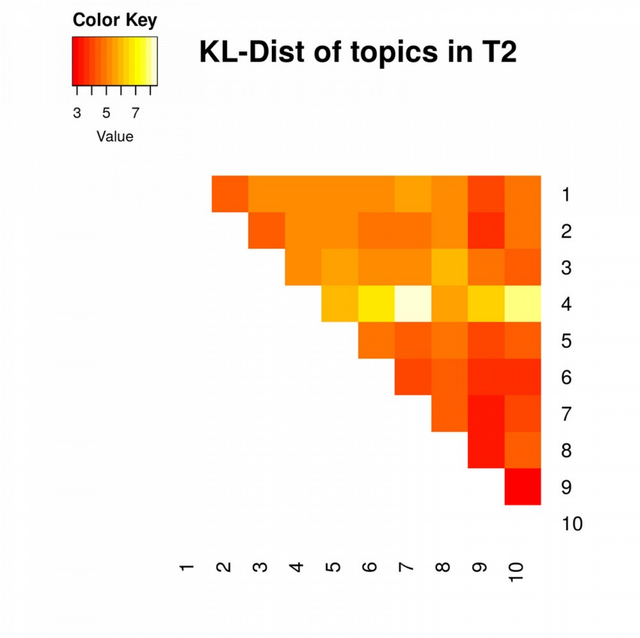
folder ***different_models/Figures_T*.*** My conclusion is that usually, even though the number of clusters obtained varies a bit (+- 1) there are however situations, when depending on the sampling performed in the LDA topic modeling algorithm, the results are totally different. For example, in the following figure, in T8, I obtained:

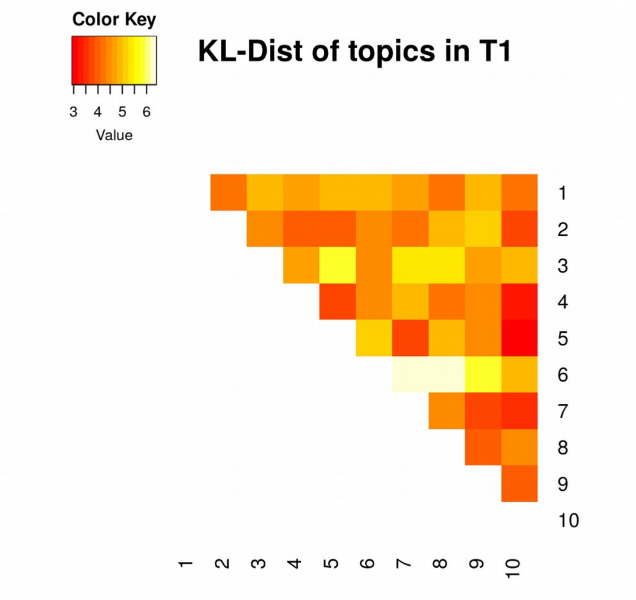




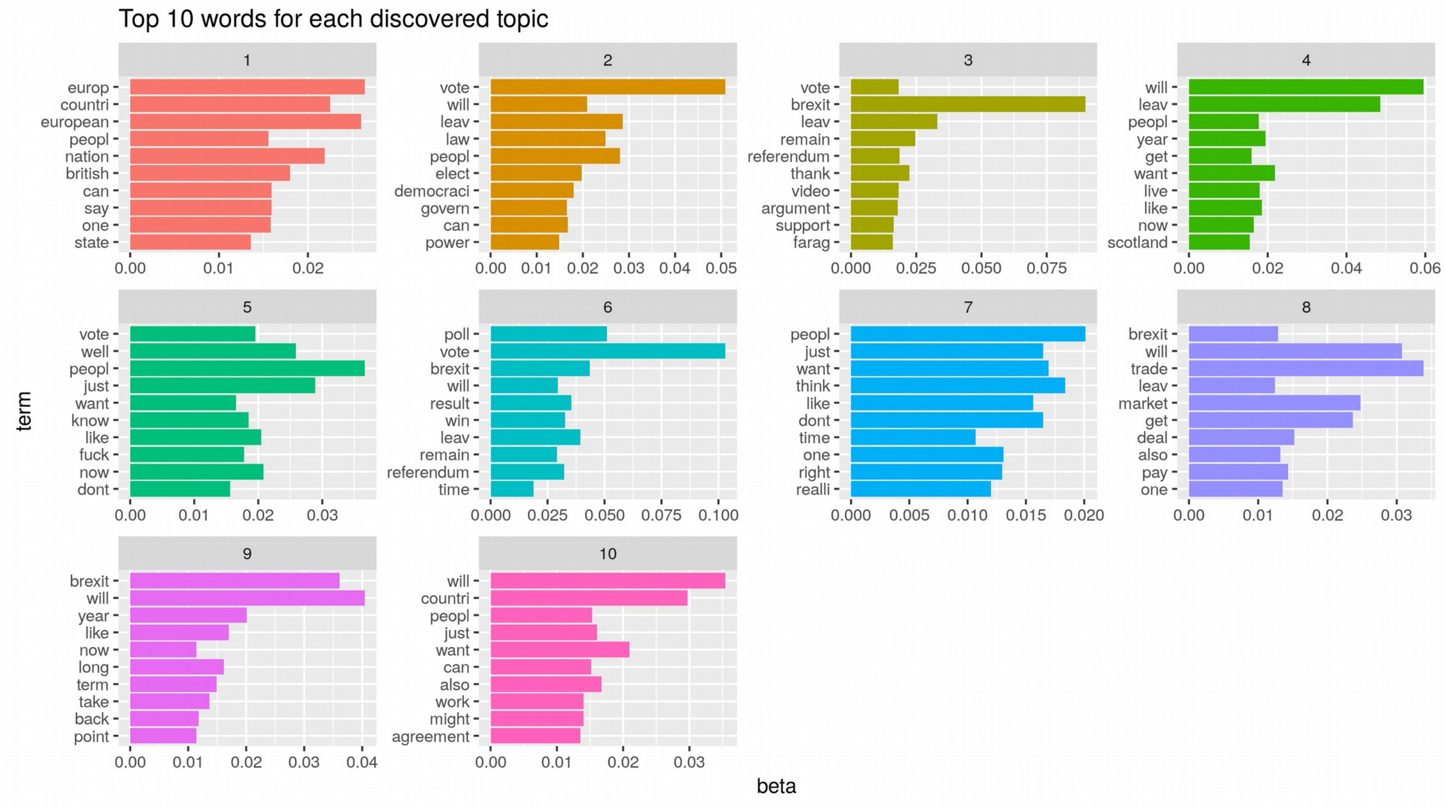
2. Distances between topics in each period

As we discussed, I computed the distance between the obtained topics for each time-frame. The notion of distance between topics refers to the KL divergence.





In the above figures we can see how topics differ in the first two timeframes. For the complet set of heatmaps, refer to folder ***Topic Distances/lda10_heatmap_T****

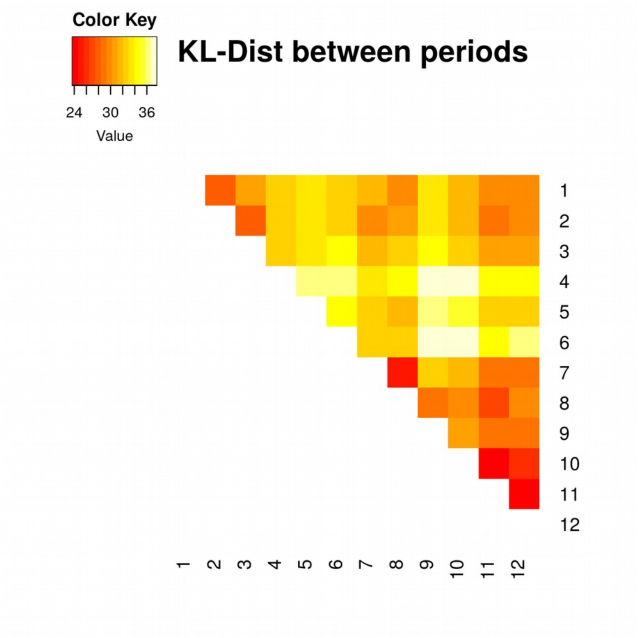


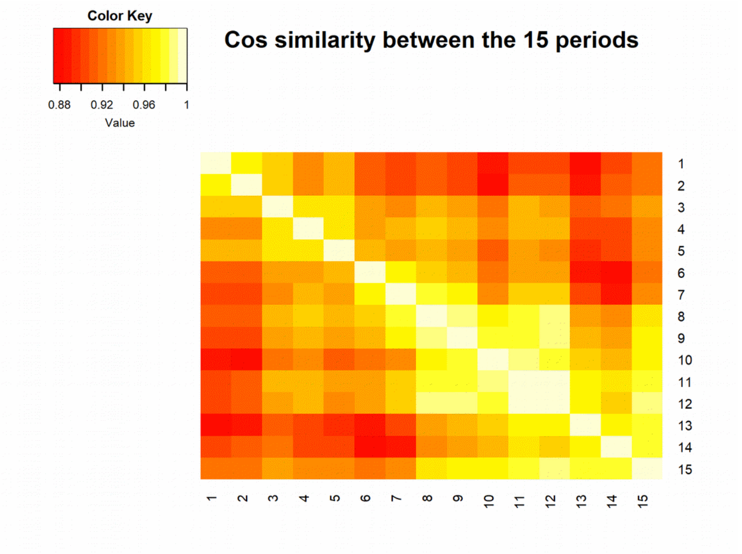
In figure above, we have the distribution of terms per topics for period 1. This should be regarded in correspondence with the first figure from the pair above. We ca observe that topic 3 for example is really different from topic 7 and 8, indeed from the top 10 most frequent words it seems to be so.

3. Distance between periods in terms of topics

The distance between periods in terms of topics was defined as: for each topic in P1, find the closest topic in P2 and add the distance between the two topic to the overall distance between periods. This way, the period distances are symmetrical and define how close is a period to another period, in terms of contained topics.

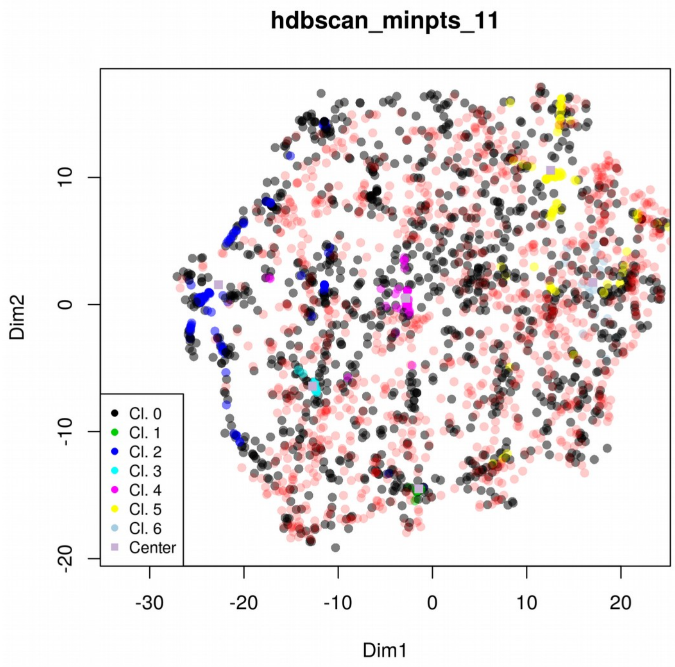
In the figure below, we can observe the heat-map obtained after computing the distances. Note that a red-sh color means small distance => periods that do not differ so much. In the right hand side figure, we can remember the heat-map obtained by using the cosine similarity. The two maps are inverted as one defined similarity and the other one distance, but we can still observe the similar results.

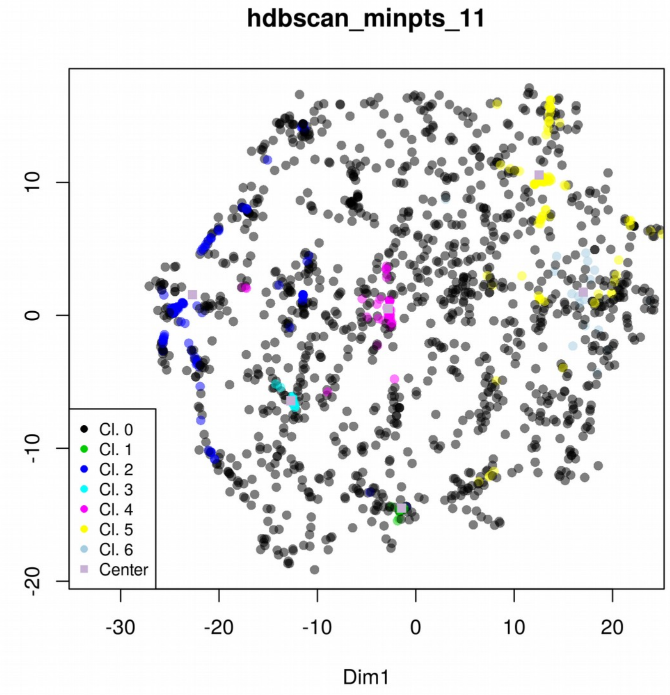




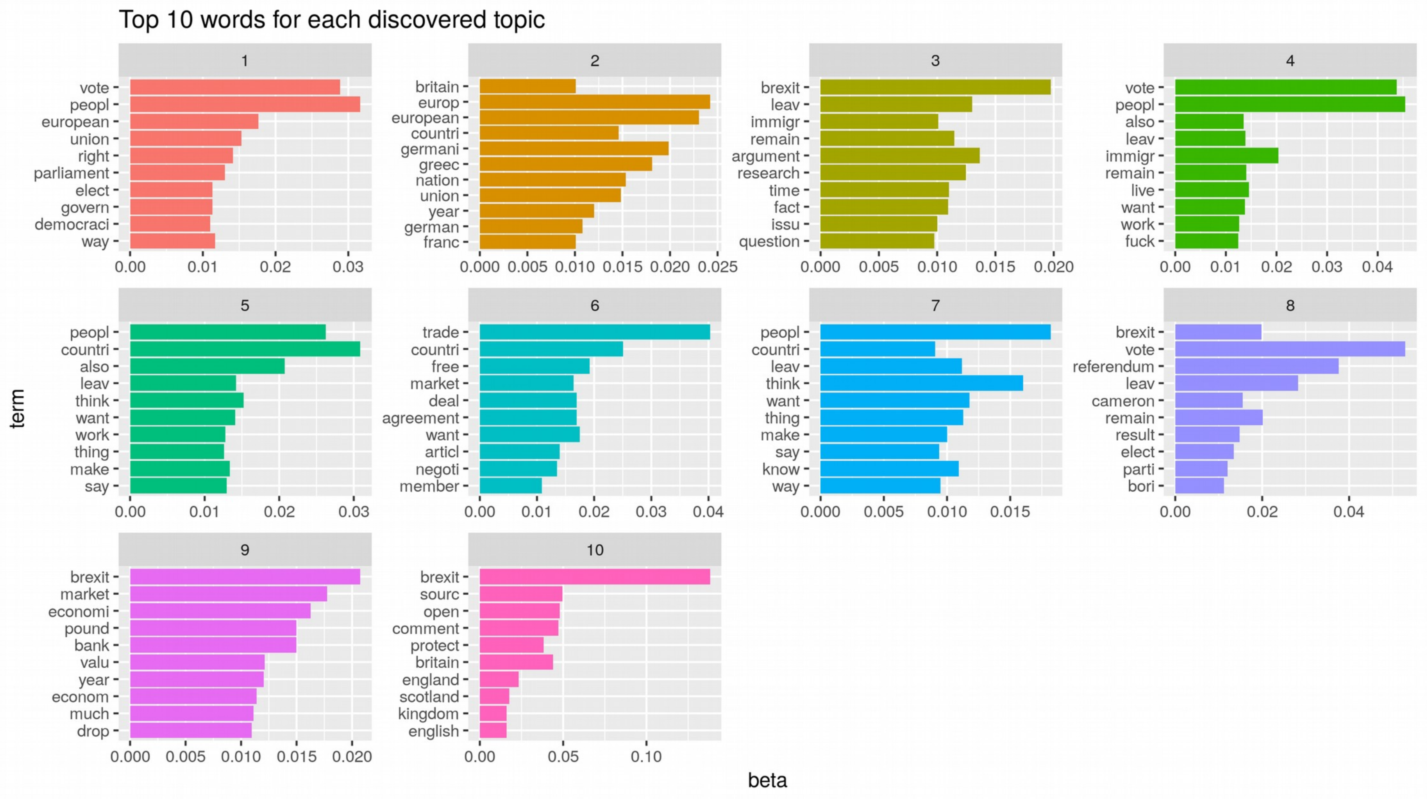
4. LDA + Dbscan + Centroids on the same figure

Continuing the work from last week, I performed LDA + clustering + computed the centroids of the clusters. This time I plotted the centroids on the same figure. The results can be observed in the following images:





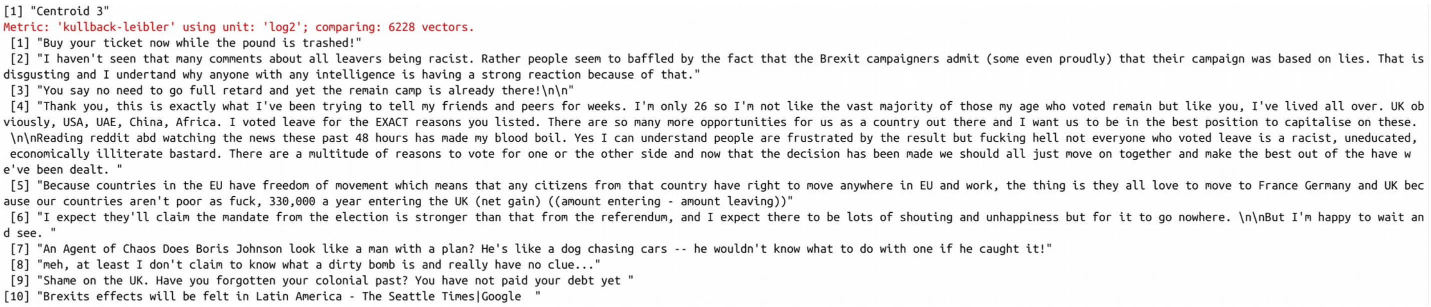
We can see that the centroids tend to be in the middle of the clusters obtained, thus the idea of averaging over the probabilities tends to prove a good idea.



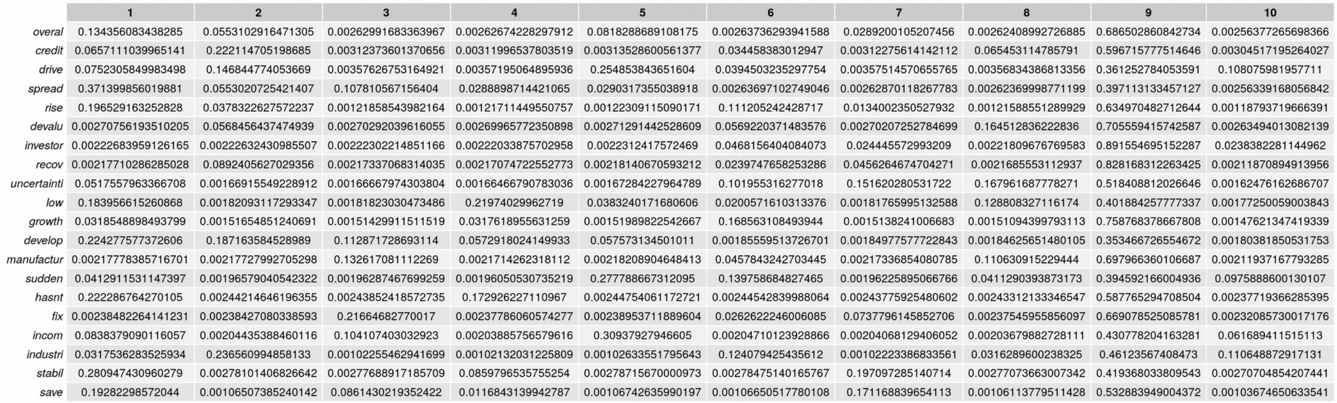
Perhaps more interesting than plotting the centroids are the closest terms and posts to these centroids, thus we can try to understand better the meaning of these clusters. In the next image, we have the 10 topics discovered by LDA.

5. LDA + Dbscan + Closest terms / posts to the centroids

Having the above figure, we can analyze each cluster centroids, and find which are the closest terms to the centroid / the closest posts.



```
[1] "Centroid 3"
Metric: 'kullback-leibler' using unit: 'log2'; comparing: 6228 vectors.
 [1] "Buy your ticket now while the pound is trashed!"
 [2] "I haven't seen that many comments about all leavers being racist. Rather people seem to baffled by the fact that the Brexit campaigners admit (some even proudly) that their campaign was based on lies. That is
disgusting and I undertand why anyone with any intelligence is having a strong reaction because of that."
 [3] "You say no need to go full retard and yet the remain camp is already there!\n\n"
 [4] "Thank you, this is exactly what I've been trying to tell my friends and peers for weeks. I'm only 26 so I'm not like the vast majority of those my age who voted remain but like you, I've lived all over. UK ob
viously, USA, UAE, China, Africa. I voted leave for the EXACT reasons you listed. There are so many more opportunities for us as a country out there and I want us to be in the best position to capitalise on these.
 \n\nReading reddit abd watching the news these past 48 hours has made my blood boil. Yes I can understand people are frustrated by the result but fucking hell not everyone who voted leave is a racist, uneducated,
 economically illiterate bastard. There are a multitude of reasons to vote for one or the other side and now that the decision has been made we should all just move on together and make the best out of the have w
e've been dealt. "
 [5] "Because countries in the EU have freedom of movement which means that any citizens from that country have right to move anywhere in EU and work, the thing is they all love to move to France Germany and UK bec
ause our countries aren't poor as fuck, 330,000 a year entering the UK (net gain) ((amount entering - amount leaving))"
 [6] "I expect they'll claim the mandate from the election is stronger than that from the referendum, and I expect there to be lots of shouting and unhappiness but for it to go nowhere. \n\nBut I'm happy to wait an
d see. "
 [7] "An Agent of Chaos Does Boris Johnson look like a man with a plan? He's like a dog chasing cars -- he wouldn't know what to do with one if he caught it!"
 [8] "meh, at least I don't claim to know what a dirty bomb is and really have no clue..."
 [9] "Shame on the UK. Have you forgotten your colonial past? You have not paid your debt yet "
[10] "Brexits effects will be felt in Latin America - The Seattle Times|Google  "
```



	1	2	3	4	5	6	7	8	9	10
overal	0.134356083438285	0.0553102916471305	0.00262991683363967	0.00262674228297912	0.0818288689108175	0.00263736293941588	0.0289200105207456	0.00262408992726885	0.686502860842734	0.00256377265698366
credit	0.0657111039965141	0.222114705198685	0.00312373601370656	0.00311996537803519	0.00313528600561377	0.034458383012947	0.00312275614142112	0.065453114785791	0.596715777514646	0.00304517195264027
drive	0.0752305849983498	0.146844774053669	0.00357626753164921	0.00357195064895936	0.254853843651604	0.0394503235297754	0.00357514570655765	0.00356834386813356	0.361252784053591	0.108075981957711
spread	0.371399856019881	0.0553020725421407	0.107810567156404	0.0288898714421065	0.0290317355038918	0.00263697102749046	0.00262870118267783	0.00262369998771199	0.397113133457127	0.00256339168056842
rise	0.196529163252828	0.0378322627572237	0.00121858543982164	0.00121711449550757	0.00122309115090171	0.111205242428717	0.0134002350527932	0.00121588551289929	0.634970482712644	0.00118793719666391
devalu	0.00270756193510205	0.0568456437474939	0.00270292039616055	0.00269965772350898	0.00271291442528609	0.0569220371483576	0.00270207252784699	0.164512836222836	0.705559415742587	0.00263494013082139
investor	0.00222683959126165	0.00222632430985507	0.00222302214851166	0.00222033875702958	0.0022312417572469	0.0468156404084073	0.024445572993209	0.00221809676769583	0.891554695152287	0.0238382281144962
recov	0.00217710286285028	0.0892405627029356	0.00217337068314035	0.00217074722552773	0.00218140670593212	0.0239747658253286	0.0456264674704271	0.0021685553112937	0.828168312263425	0.00211870894913956
uncertainti	0.0517557963366708	0.00166915549228912	0.00166667974303804	0.00166466790783036	0.00167284227964789	0.101955316277018	0.151620280531722	0.167961687778271	0.518408812026646	0.00162476162686707
low	0.183956615260868	0.00182093117293347	0.00181823030473486	0.21974029962719	0.0383240171680606	0.0200571610313376	0.00181765995132588	0.128808327116174	0.401884257777337	0.00177250059003843
growth	0.0318548898493799	0.00151654851240691	0.00151429911511519	0.0317618955631259	0.00151989822542667	0.168563108493944	0.0015138241006683	0.00151094399793113	0.758768378667808	0.00147621347419339
develop	0.224277577372606	0.187163584528989	0.112871728693114	0.0572918024149933	0.057573134501011	0.00185559513726701	0.00184977577722843	0.00184625651480105	0.353466726554672	0.00180381850531753
manufactur	0.00217778385716701	0.00217727992705298	0.132617081112269	0.0021714262318112	0.00218208904648413	0.0457843242703445	0.00217336854080785	0.110630915229444	0.697966360106687	0.00211937167793285
sudden	0.0412911531147397	0.00196579040542322	0.00196287467699259	0.00196050530735219	0.277788667312095	0.139758684827465	0.00196225895066766	0.0411290393873173	0.394592166004936	0.0975888600130107
hasnt	0.222286764270105	0.00244214646196355	0.00243852418572735	0.17292622711 0967	0.00244754061172721	0.00244542839988064	0.00243775925480602	0.00243312133346547	0.587765294708504	0.00237719366285395
fix	0.00238482264141231	0.00238427080338593	0.21664682770017	0.00237786060574277	0.00238953711889604	0.0262622246006085	0.0737796145852706	0.00237545955856097	0.669078525085781	0.00232085730017176
incom	0.0838379090116057	0.00204435388460116	0.104107403032923	0.00203885756579616	0.30937927946605	0.00204710123928866	0.00204068129406052	0.00203679882728111	0.430778204163281	0.061689411515113
industri	0.0317536283525934	0.236560994858133	0.00102255462941699	0.00102132031225809	0.00102633551795643	0.124079425435612	0.00102223386833561	0.0316289600238325	0.46123567408473	0.110648872917131
stabil	0.280947430960279	0.00278101406826642	0.00277688917185709	0.0859796535755254	0.00278715670000973	0.00278475140165767	0.197097285140714	0.00277073663007342	0.419368033809543	0.00270704854207441
save	0.19282298572044	0.00106507385240142	0.0861430219352422	0.0116843139942787	0.00106742635990197	0.00106650517780108	0.171168839654113	0.00106113779511428	0.532883949004372	0.00103674650633541

In the figures above, we can see that for example, for the centroid associated to cluster 3, the closest terms are: **overall, credit, drive, spread, rise, devalu, investor.** Moreover, if we look on the probabilities, we can see that the largest probabilities for these terms are in the 9th column, meaning the 9th topic. Indeed, looking at the previous figure, the one with words per topics, we can observe that topic 9 is about market, economi, poud, bank etc. In conclusion, the cyan cluster, cluster 3 might be related to economy. More similar tables like the ones presented previously, can be seen in folder ***dbscan_closest_words.***