**Part1:**

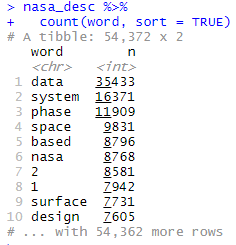
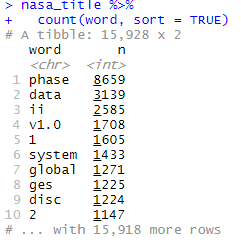
The connection between the datsets is first needed to understand the content and information provided in the metadata. The various titles, descriptions and keywords are observed to find the most common keywords,wordpairs and titles that are discussed and published in the files. But for the analysis to be done the dataset within the metadata did not contain unique IDs. Therfore, a sequencial colunm was added to metadata$dataset table.

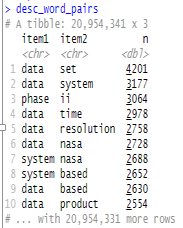
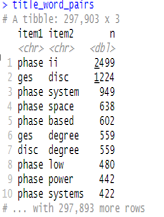
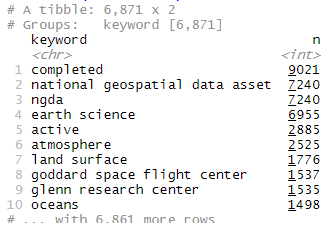
Also during the connecting topic modeling with keywords was done the document type in lda\_gamma nad to be changed to numeric. Wrangling and cleaning the data is done for a better view.



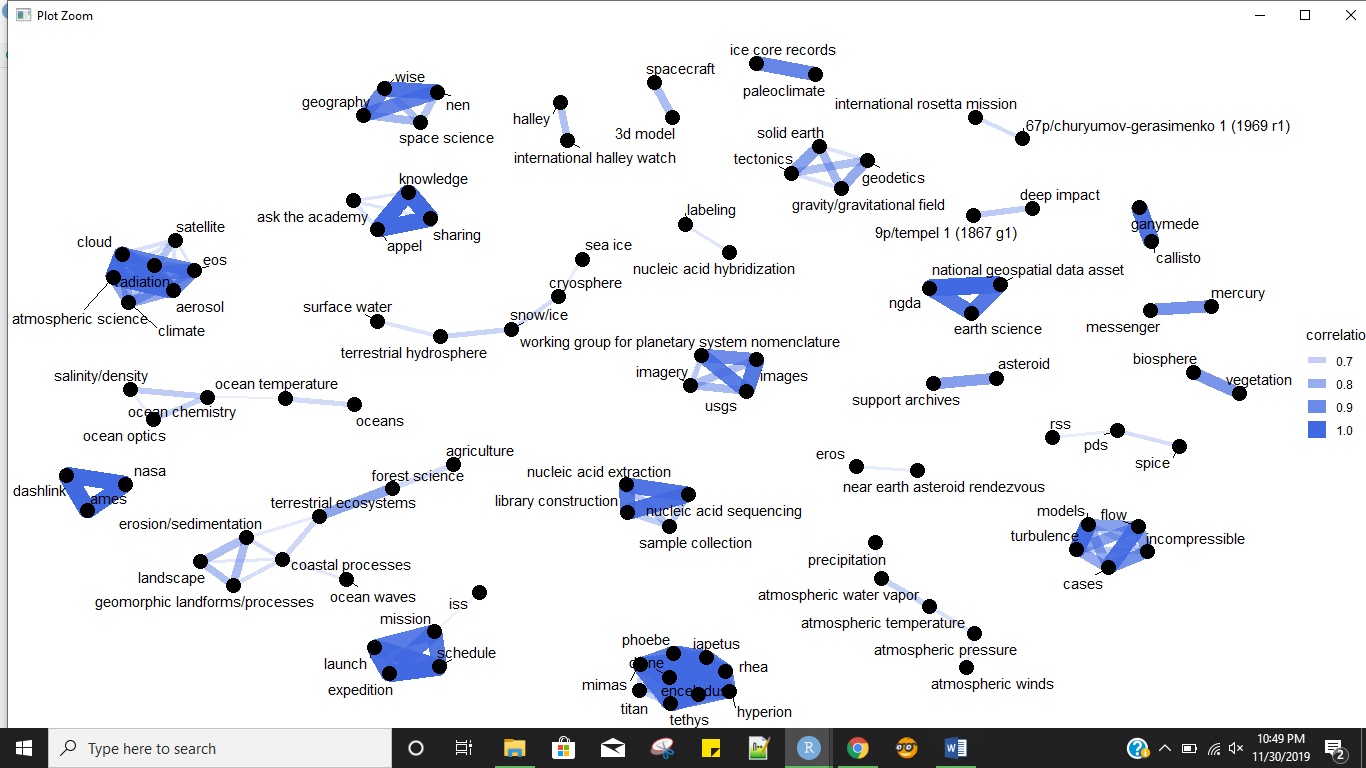
1.Initial Analysis:

Tokenisation is performed. Word count is done to check the frequency of each word.

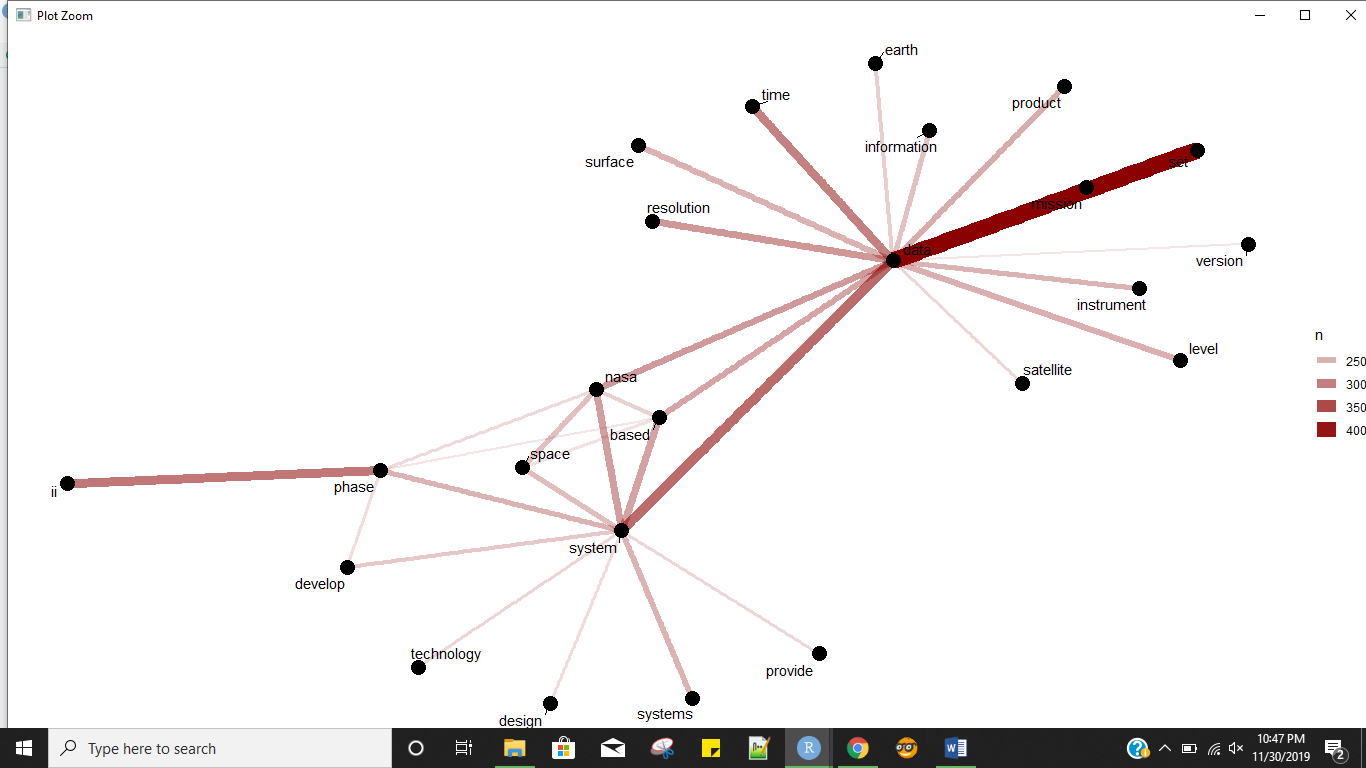
Various unwanted words are removed using stop\_words feature. The highest used keywords is checked. Completed seems to be a very highly used word which might have been used to decribe various stage completeions within a single project. “earth science”,”atmpsphere”,”land surface” and “oceans” must be the heirarcy in research topics apart from the geospatial data. For topics, various phases seem to be one of the most commonly paired word. For descriptions, data has the highest count of pairing.



2. Network of Decription and titles



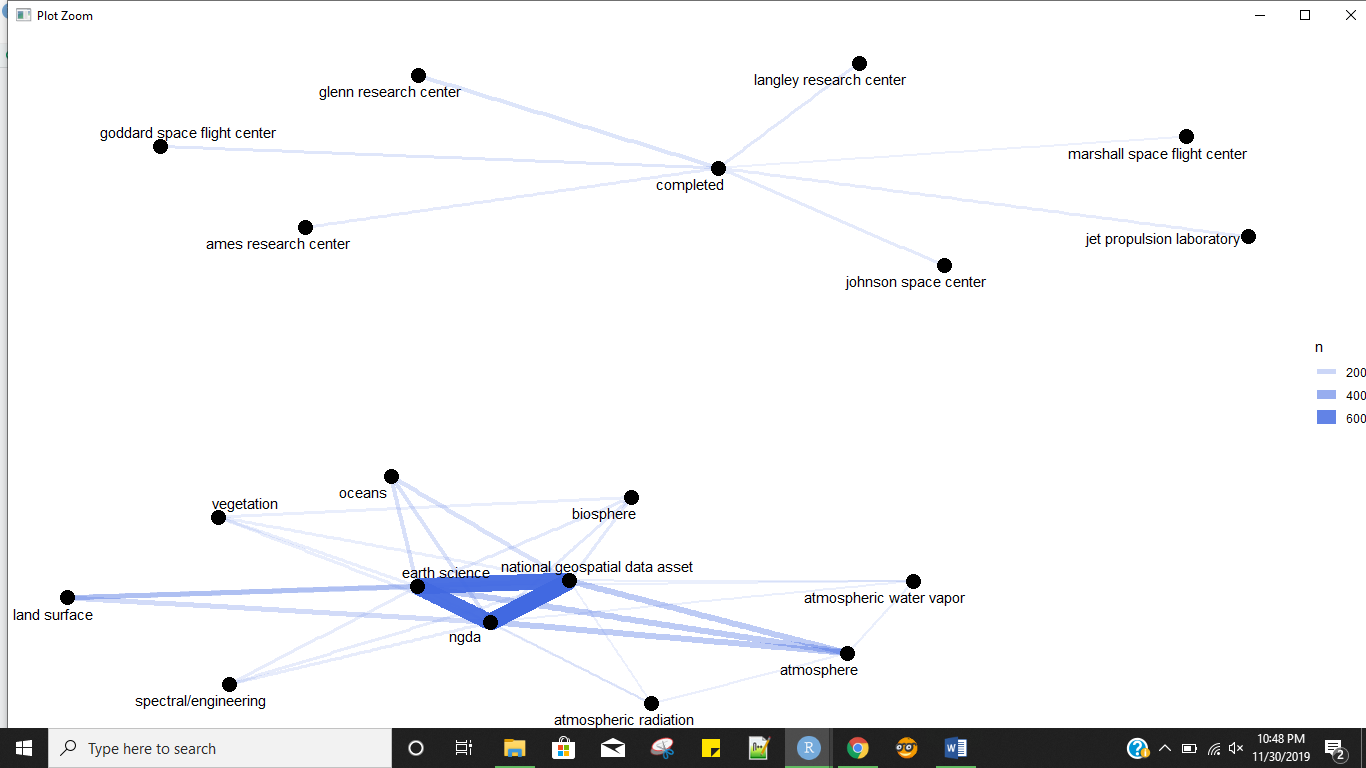
We can see a very high correlation between titles like, “ice core records with paleoclimate”, “biosphere with vegetation”, “knowledge sharing and appel connected to ask the academy”, “mission with launch, schedule and expedition”, etc. Also a connection is drawn for sea ice, cryosphere,snow/icoe, terrestrial hydrosphere and surface water. Such a network graph makes it easier to see the links between the titles.



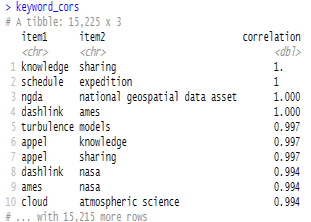
Data which was a very commonly used word in word pairs, is highly co-occurred with mission and set. Also has a strong connect with satellite, system, resolution, surface, information, time and various others. System is also a word which is quite commonly paired word with space, phase and nasa among other words.

3. Network of keywords

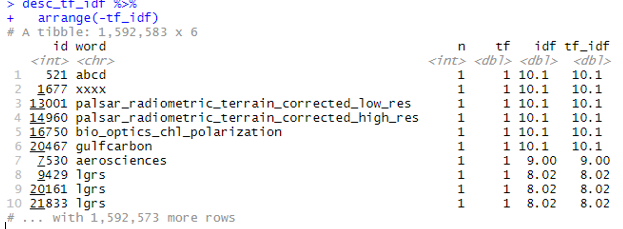
The network of keywords is pretty straightforward as shown in the figure.



The highest correlated words is shown below



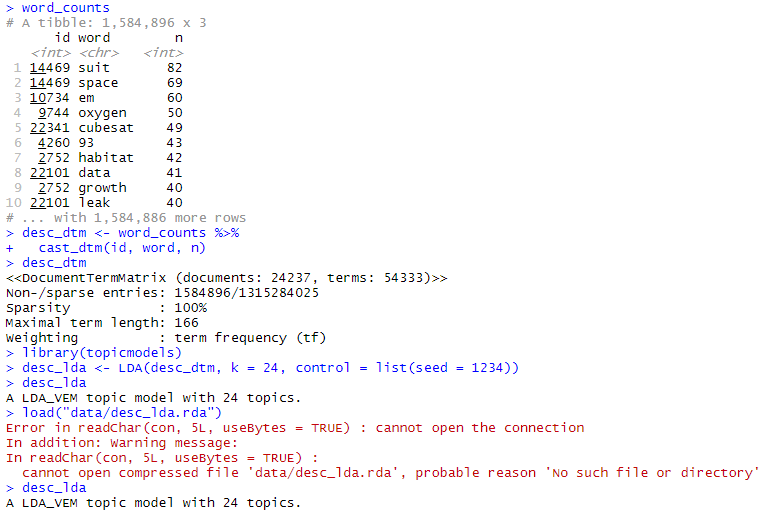
4. Calculating td-idfs



Apart from unknown words like abcd, xxxx, palsar\_radiometric\_terrian\_corrected\_low\_res/high\_res, bio\_optics, gulfcarbon and aerosciences and lgrs are the common words but not too common as they have a very high tf-idf score.

5. Topic modeling

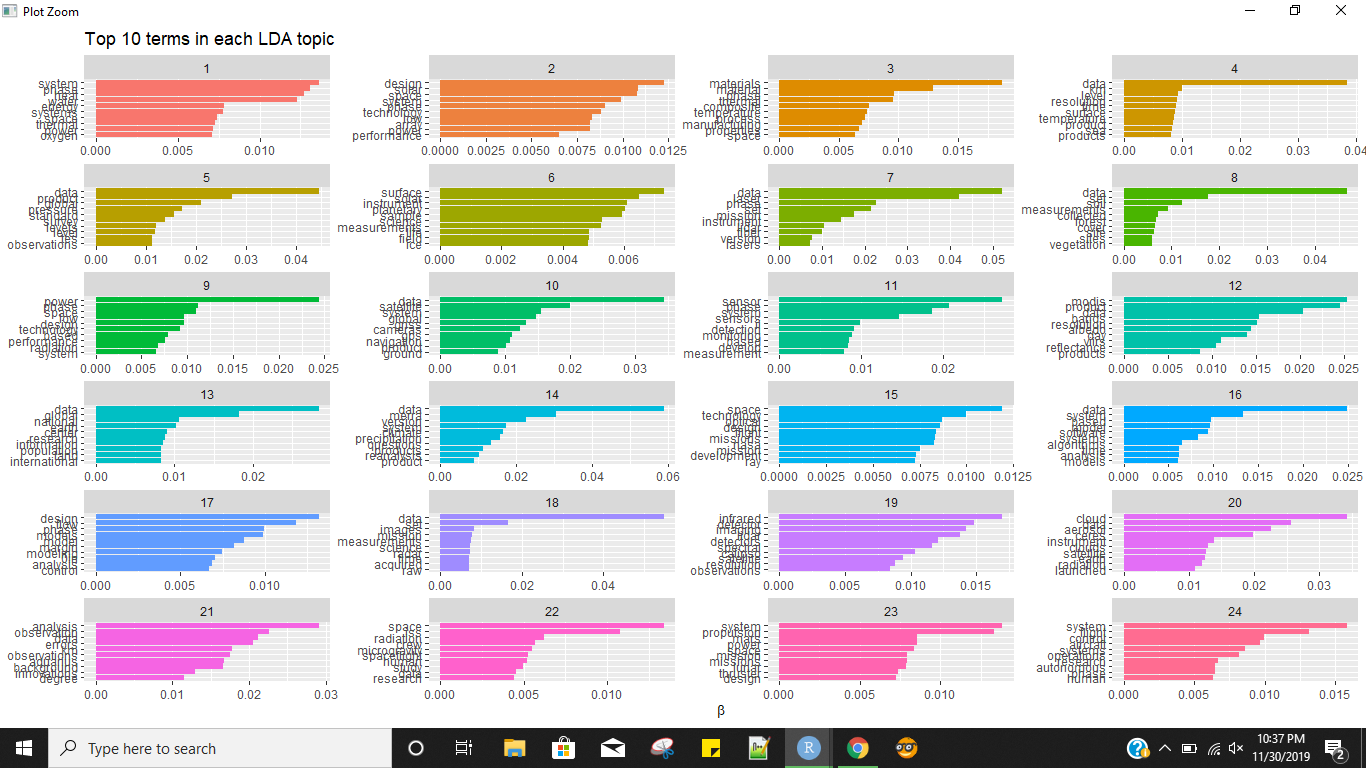
For the topic modeling various steps like creating document-term matrix was created, and cleaned, and a word count was done. Where suit, space and oxygen where the top 3 words if “em” was removed as we do not know what it means. The document term matrix is cast and the sparsity is 100%. It has 24237 documents and 54333 terms in it. An lda is created with 24 topics which was time consuming as it took more than 3 and half hours to complete.

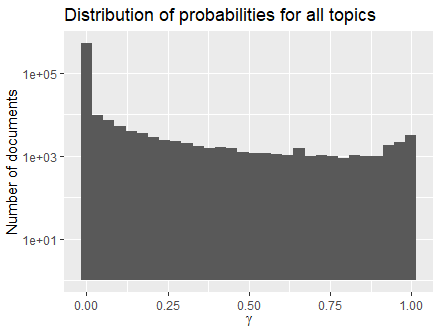


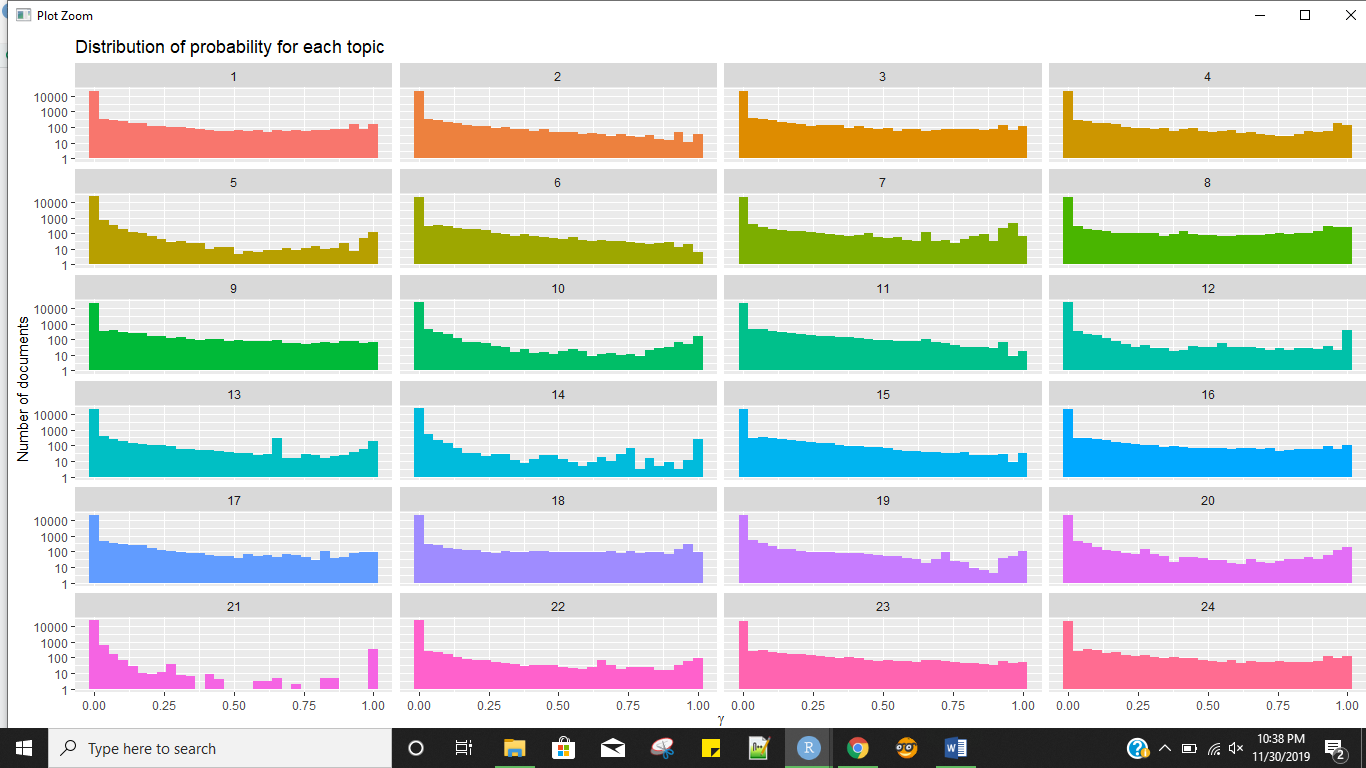




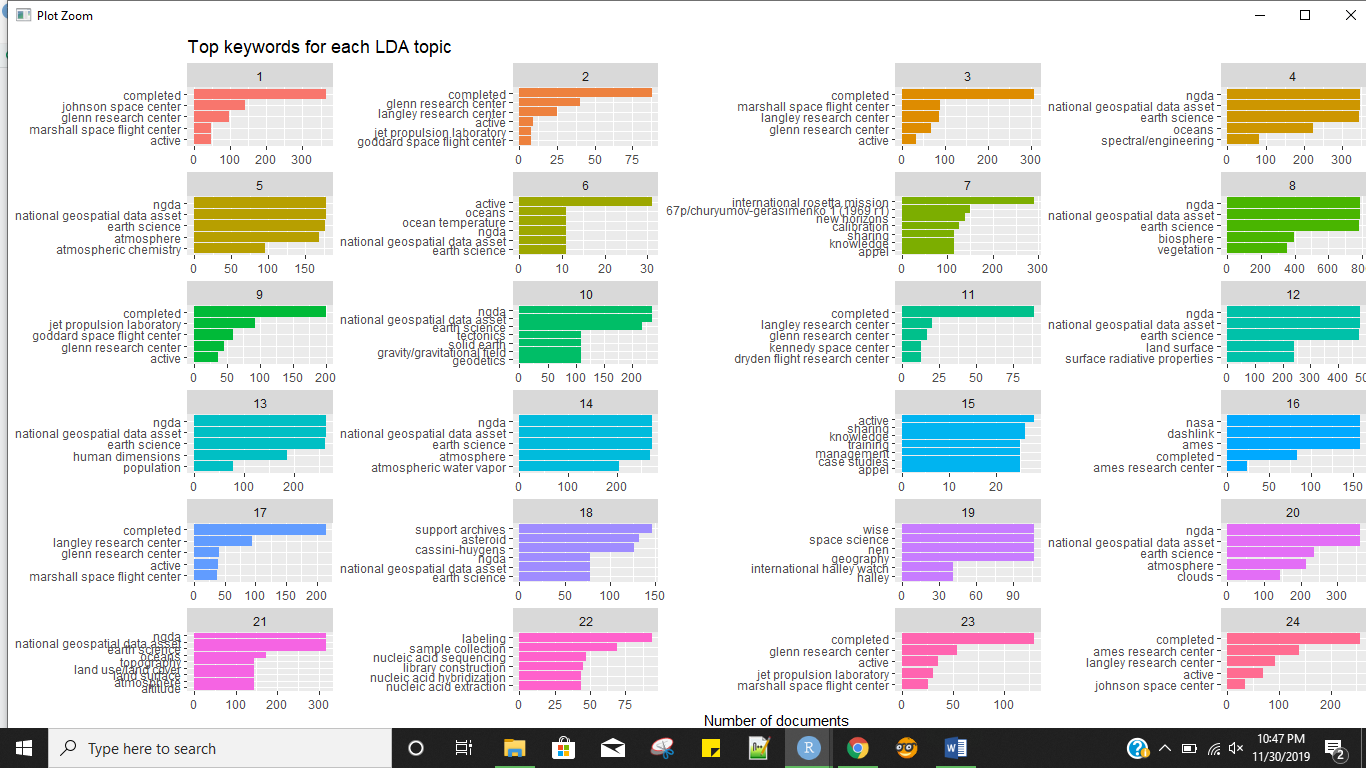
The top terms of each 24 topics is plotted for a better understanding. Data, research, space and system are most common top terms in all the topics.



There are definitely more wrongly placed words than correctly placed words in all the topics.



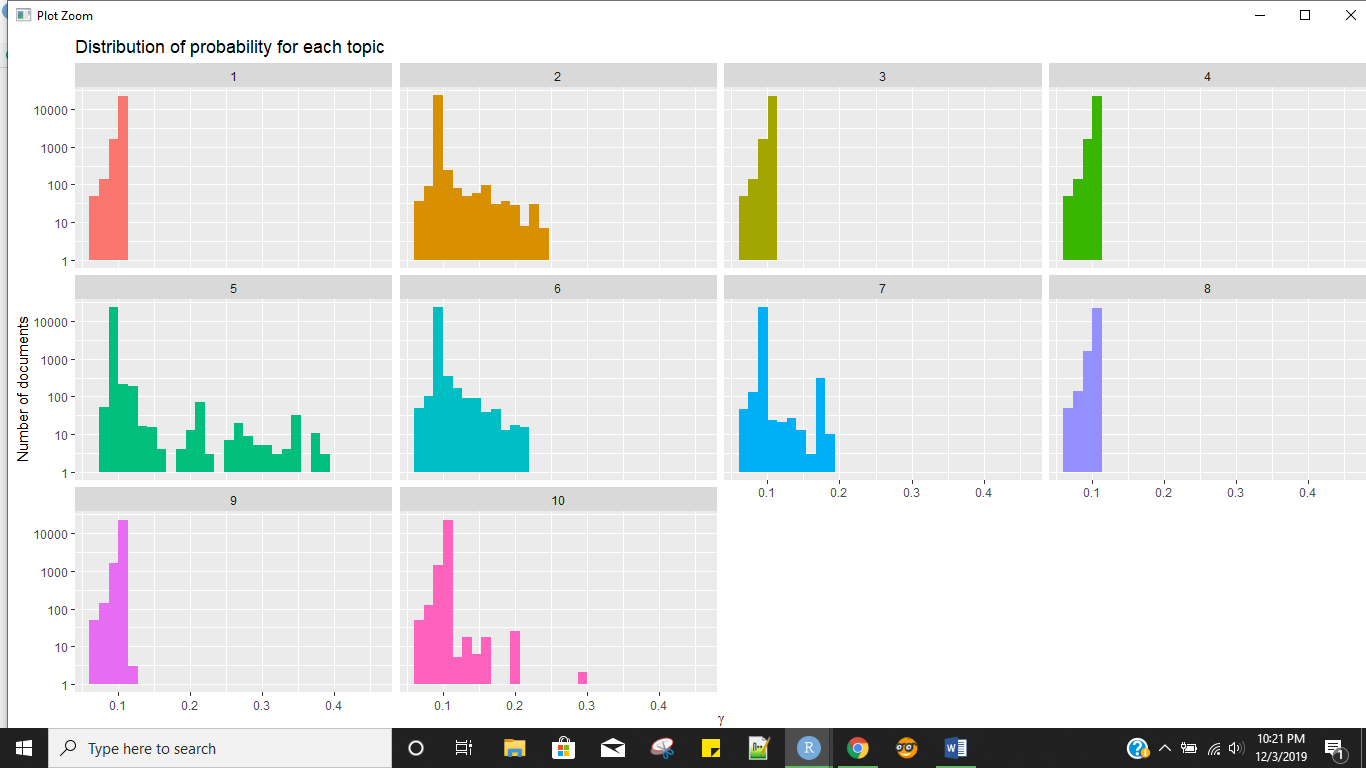
There are huge words which do not belong to the topics compared to the words which belong 100% to the topic, but on an overall the distributing from 0 to 1 seems good.



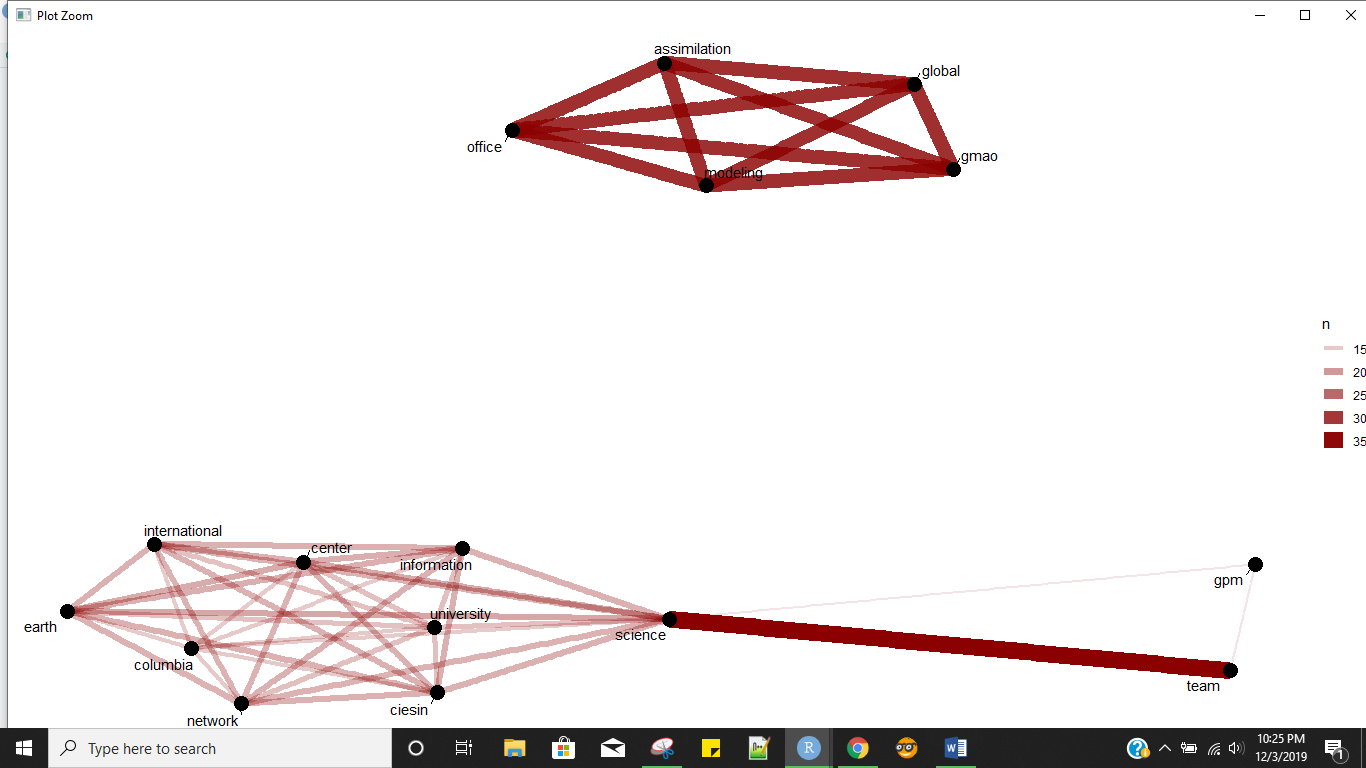
Completed, different research center names and ngda are very common and top words in most of the topics.

Extra:

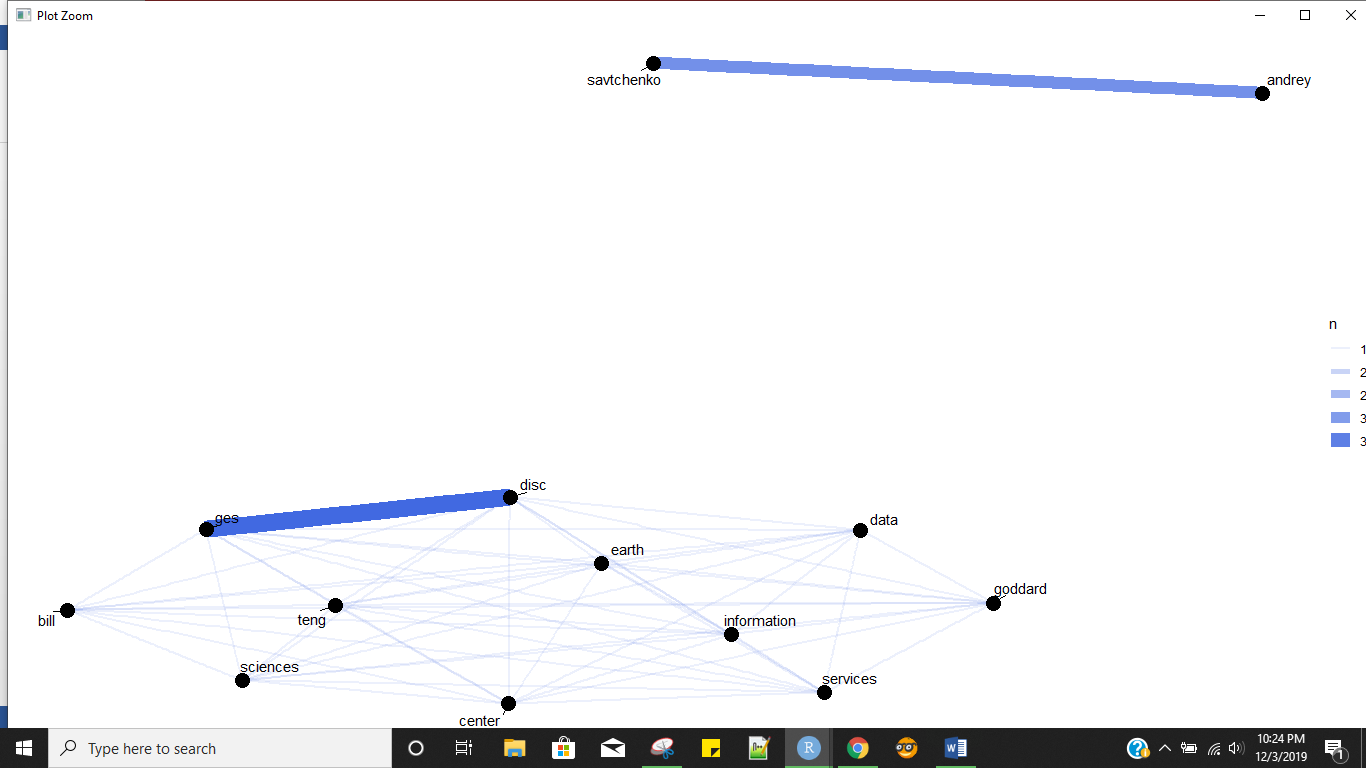
The same process was also tried to understand the relationship between title, creator and editor. But the distribution probability was very low ɤ <0.5.



The top creators being



And editors were



The colunm editors has many NA hence it was not a very useful choice of selection for analysis.

