MG­GY­9753 Business Analytics

Assignment 5

Submitted by Sunil Nandan N12934971

**PART 1**

**Case:**

The CEO has charged your analytics groups with a task: “What are the characteristics of an organization that adapts well to data analytics?” The CEO’s intention is to restructure the company to foster adoption and advancement of data analytics capabilities across units.

**Steps:**

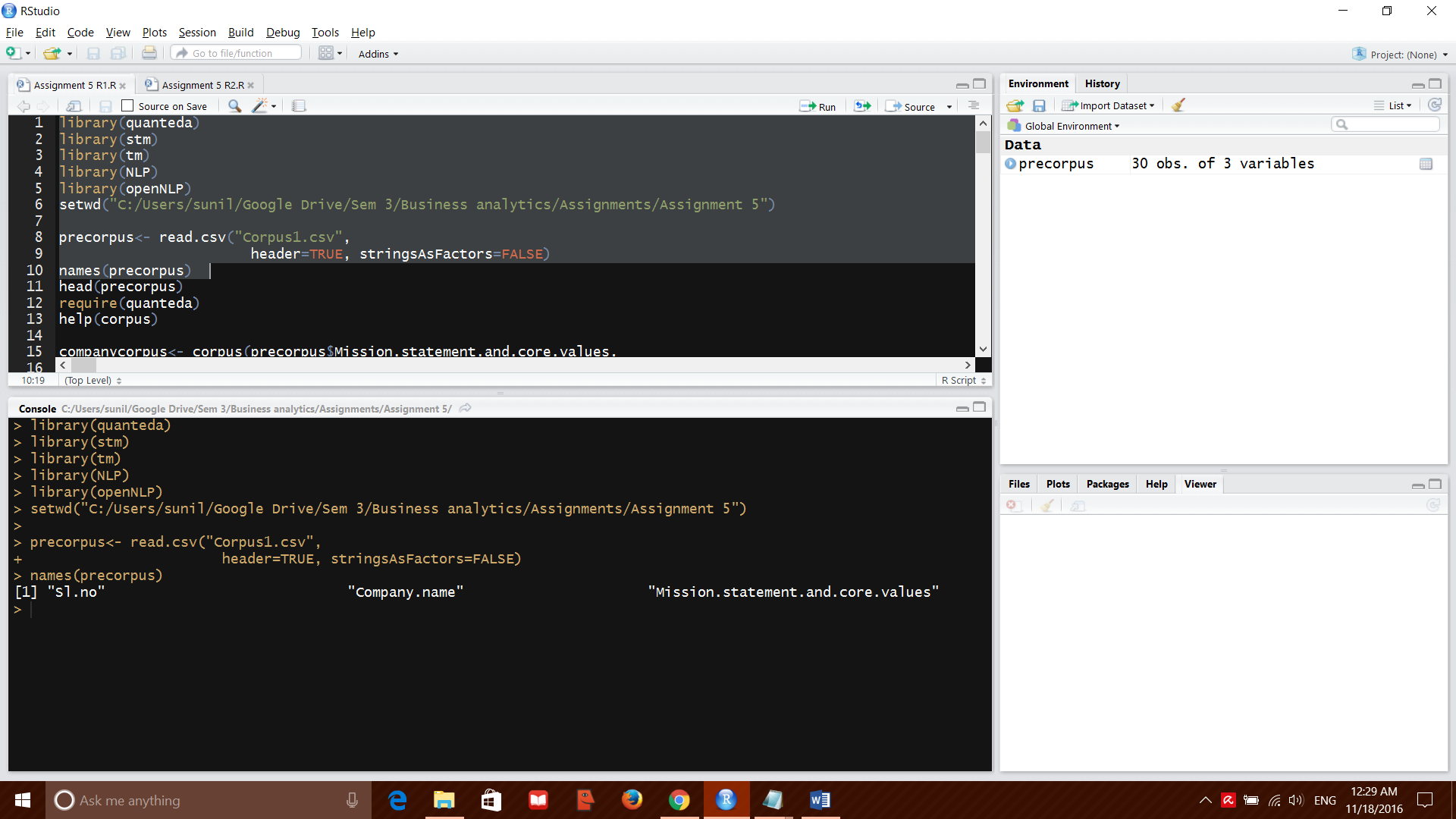
* Find a list of successful data-driven companies (See this [Forbes](http://www.forbes.com/sites/louiscolumbus/2015/05/09/the-best-big-data-and-business-analytics-companies-to-work-for-in-2015/#5569d331286f) article for inspiration) and select 30 organizations.

My pick of 30companies from the list

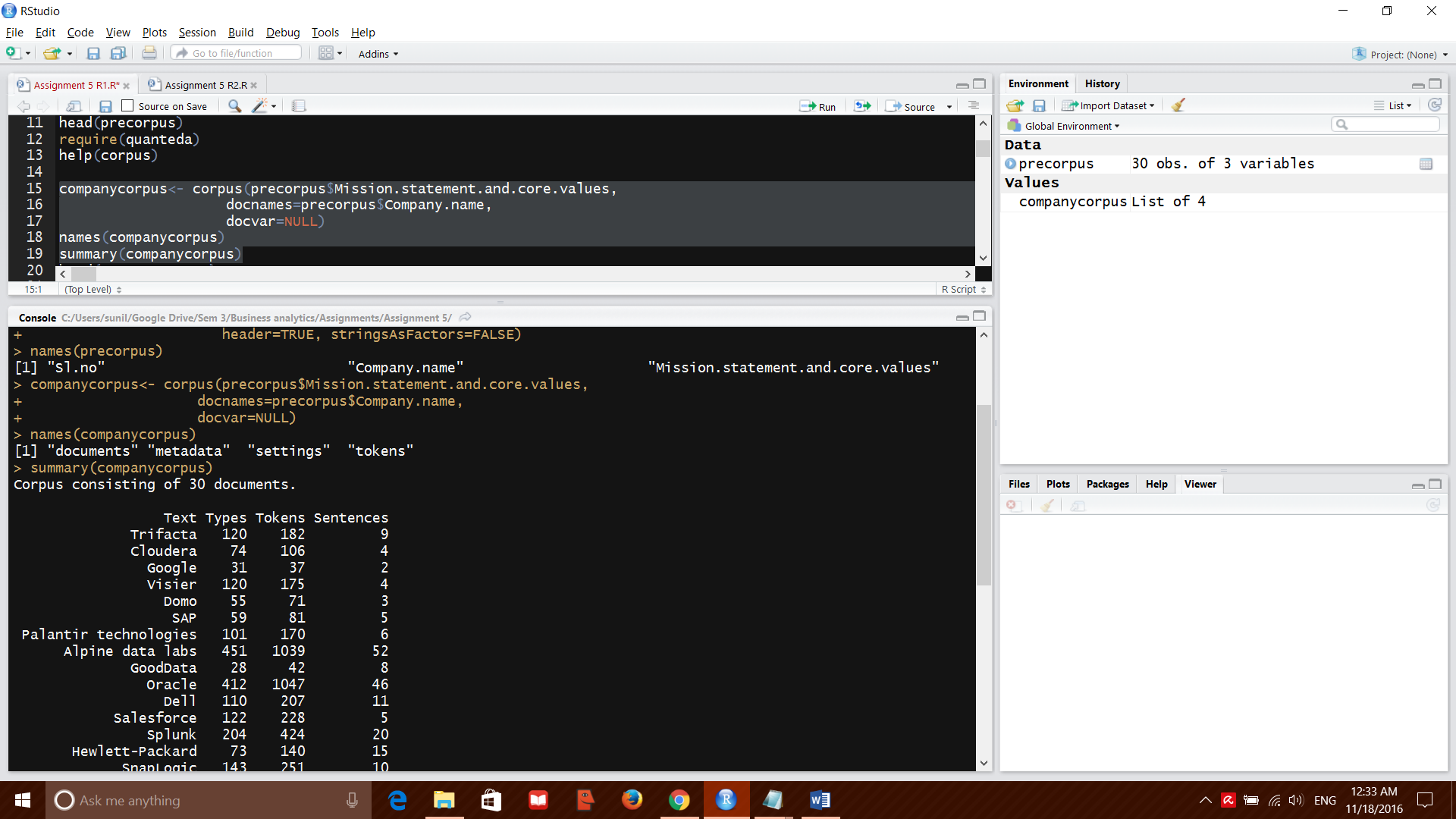
1. Trifacta
2. Cloudera
3. Google
4. Visier
5. Domo
6. SAP
7. Palantir technologies
8. Alpine data labs
9. GoodData
10. Oracle
11. Dell
12. Salesforce
13. Splunk
14. Hewlett-Packard
15. SnapLogic
16. Numerify
17. Attivio
18. Pivotal
19. IBM
20. Syncsort
21. Basho Technologies
22. Recommind
23. Aerospike
24. Cask
25. Confluent
26. Couchbase
27. DataRPM
28. Information builders
29. Qlik
30. Informatica

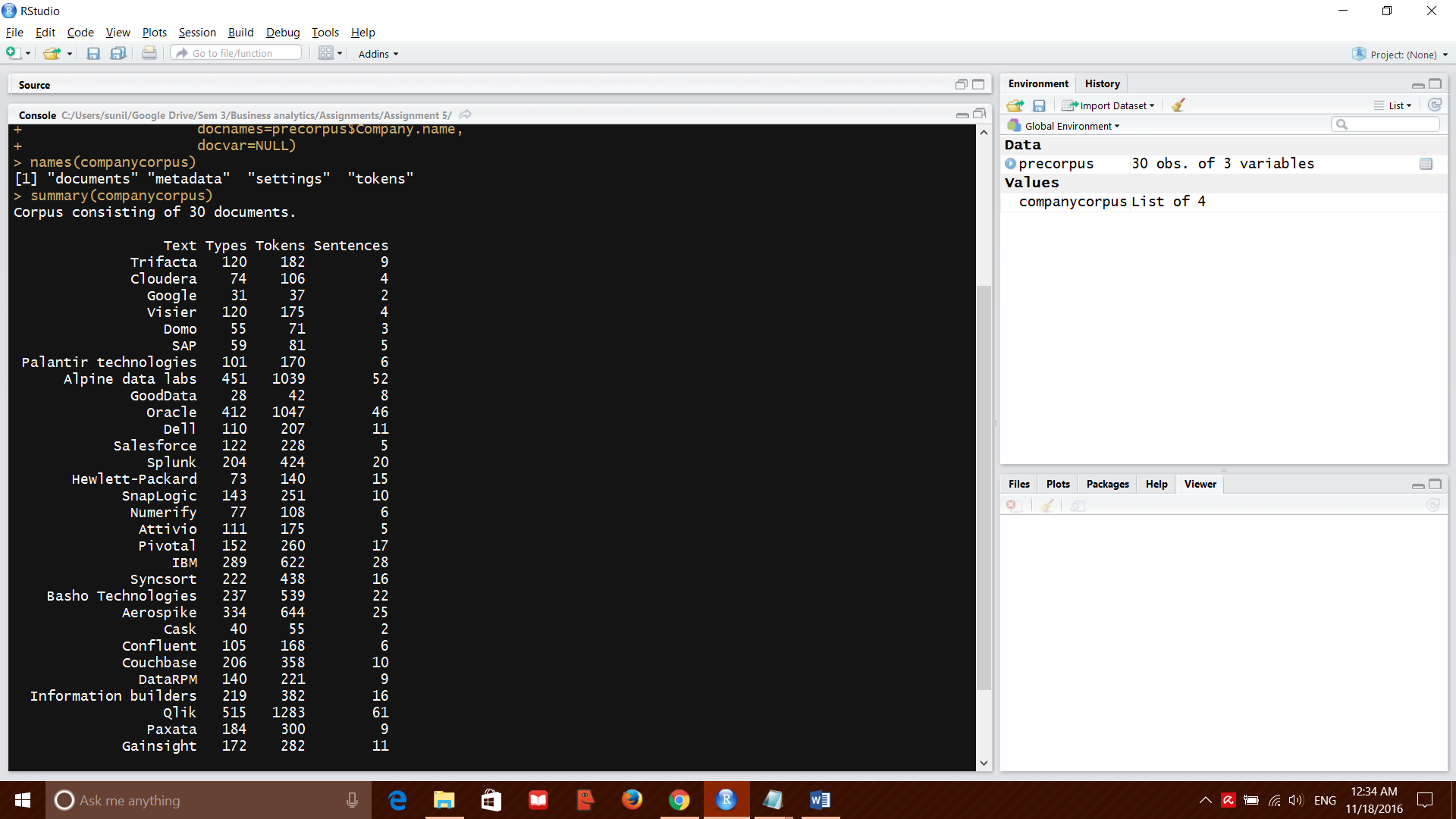
* Create a corpus of their mission statements
* Create a corpus of their core values
* Analyze the corpus and provide insight on how to structure a firm for data-analysis readiness
* Are there any other data-driven approaches you would recommend the CEO to implement?

Import the list of companies, their mission statements and goals from a .csv file into a data frame.

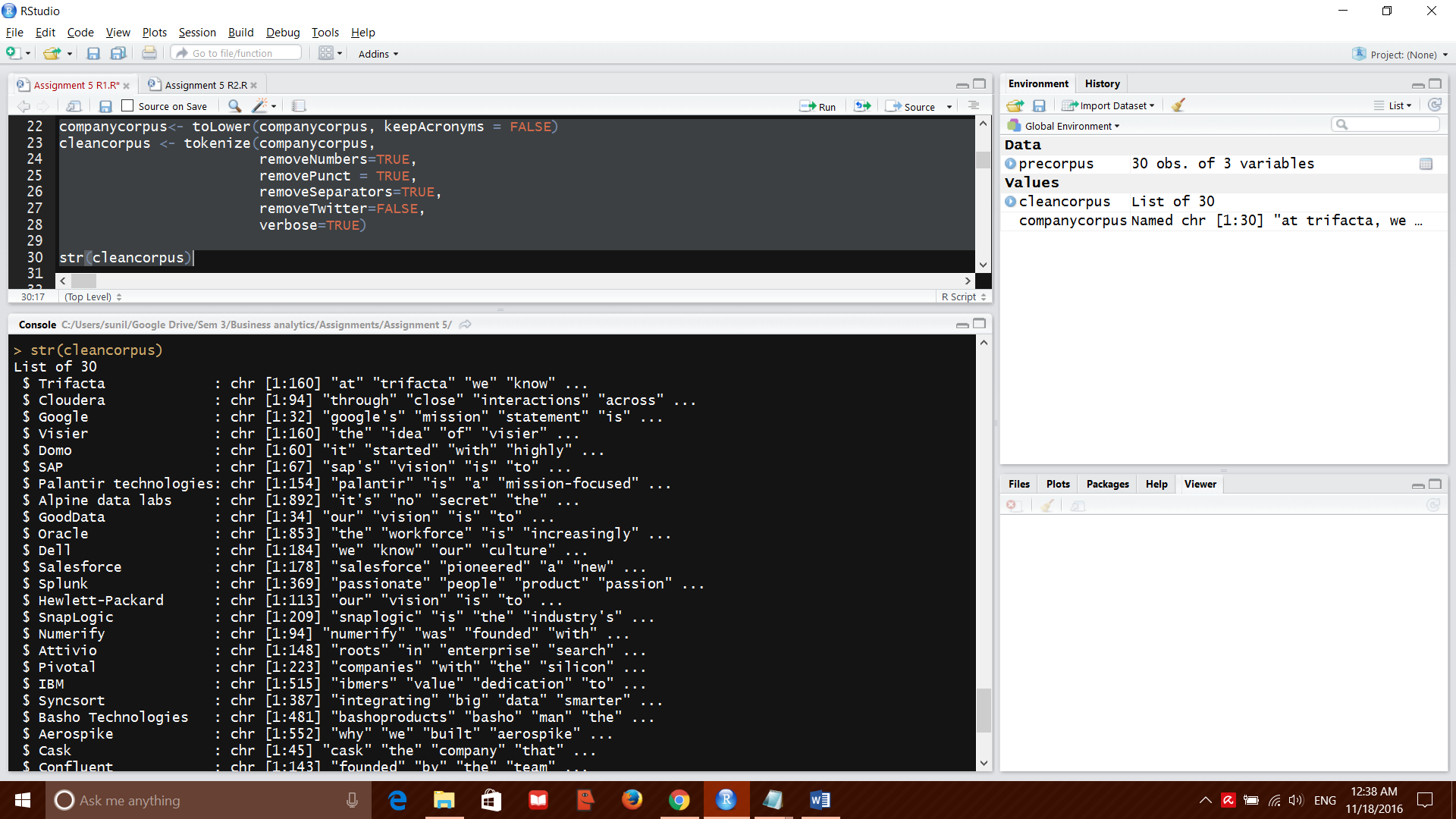


Create a corpus from the data frame , to be used for further analysis





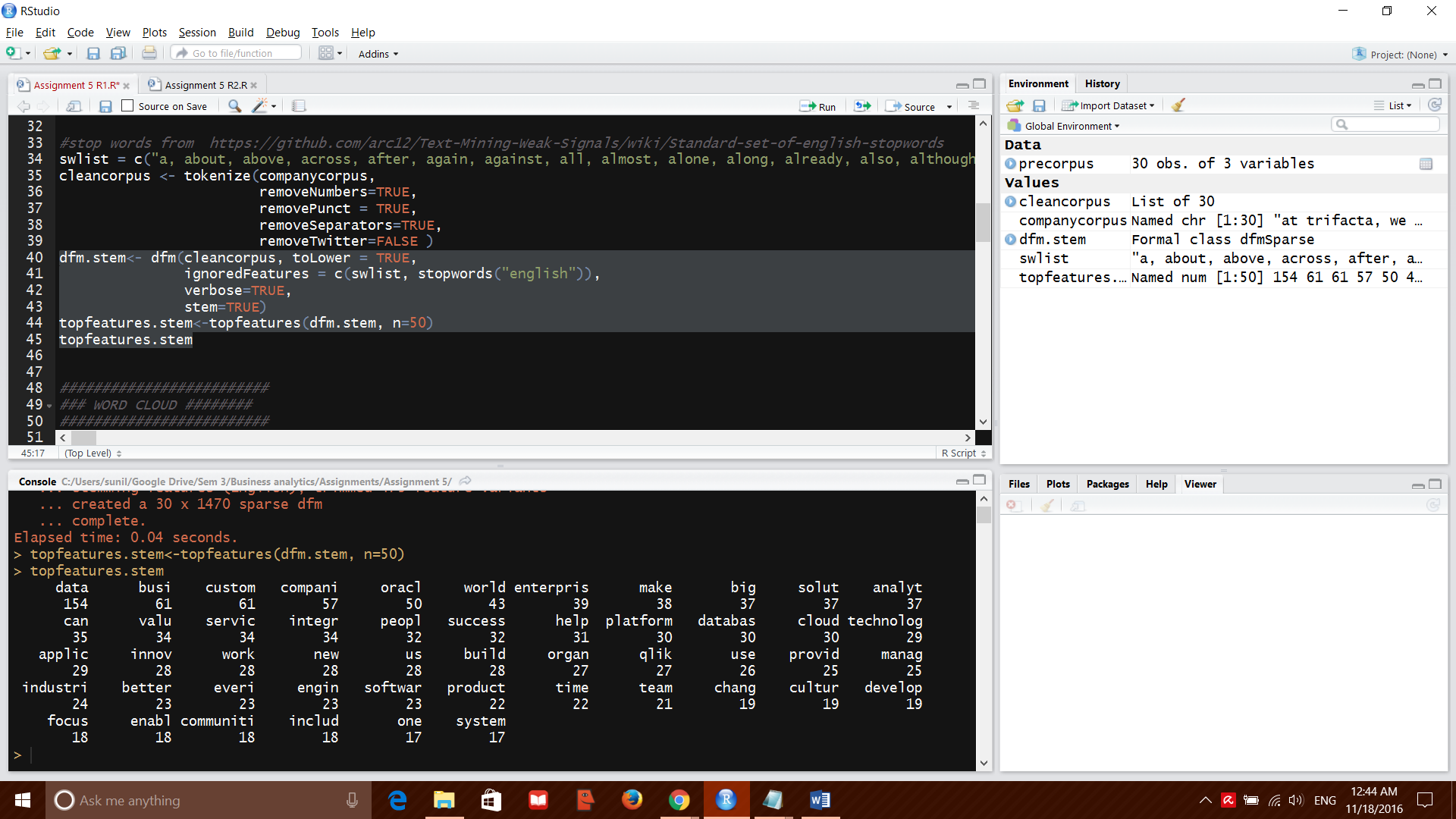
‘Clean’ the corpus before further operation



Stop words need to be eliminated from the text data before we can proceed. A list of common stop words were taken from the below url and the corpus was filtered using them.

<https://github.com/arc12/Text-Mining-Weak-Signals/wiki/Standard-set-of-english-stopwords>

top 50 features from corpus were listed.

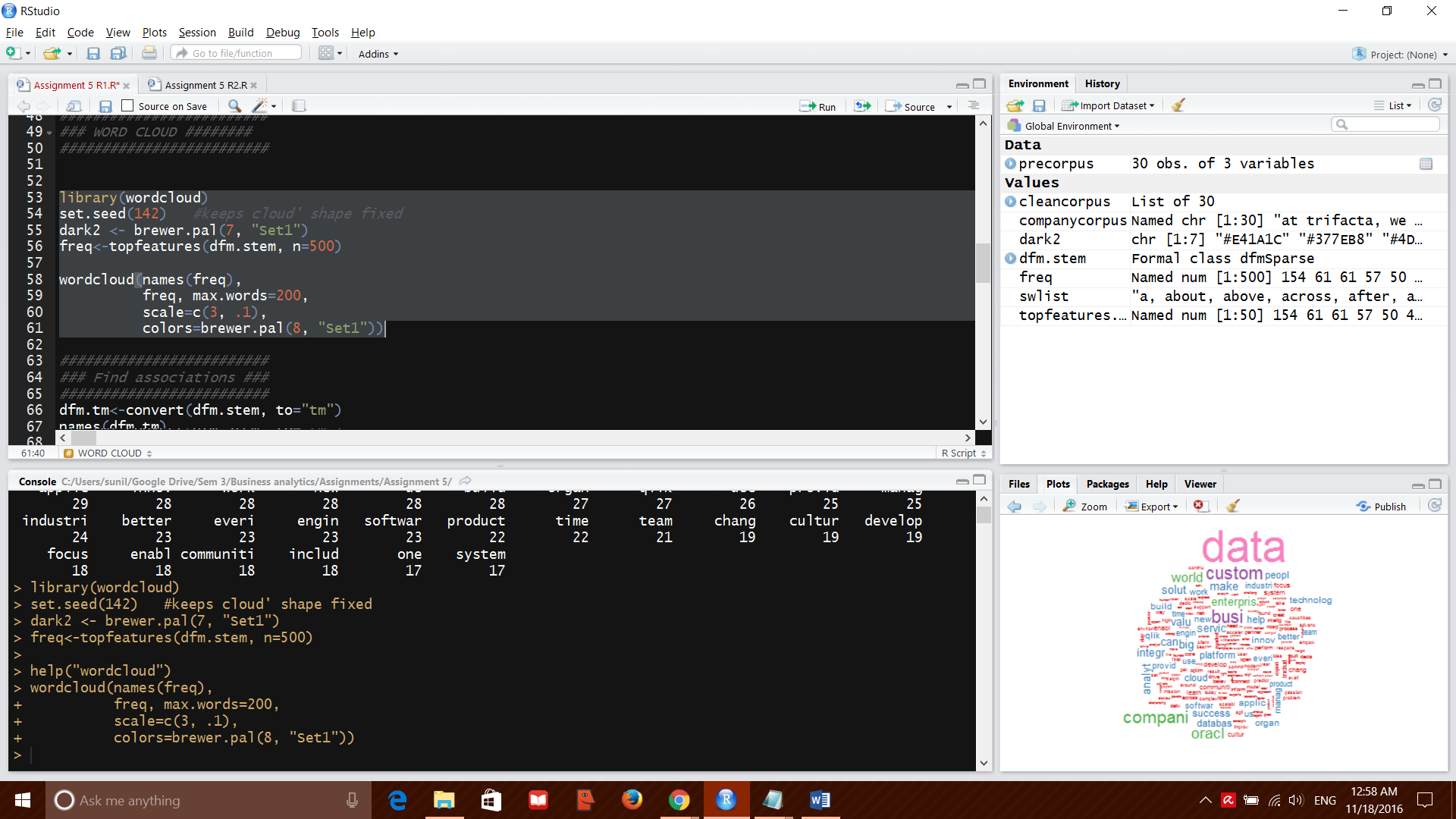


Inference:

From the above hierarchical list of word occurrences, we conclude that the top companies use data analytics to provide innovative solutions to customers in the global market. The focus of the companies is on providing value added services to their customers.

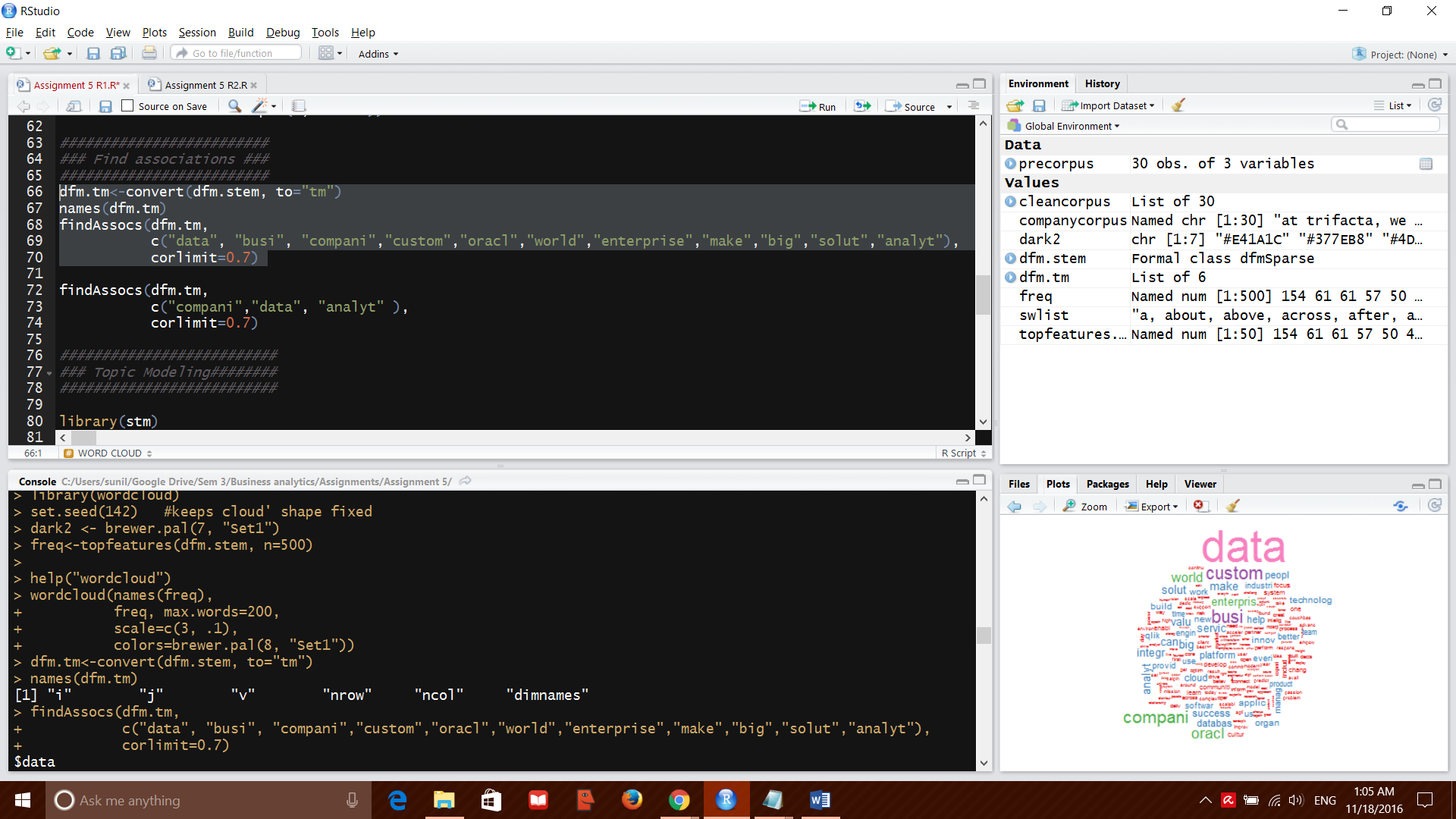
This is only one of the possible inference from the word occurrence matrix.

We form a word cloud next for visual representation of the word occurrence table.

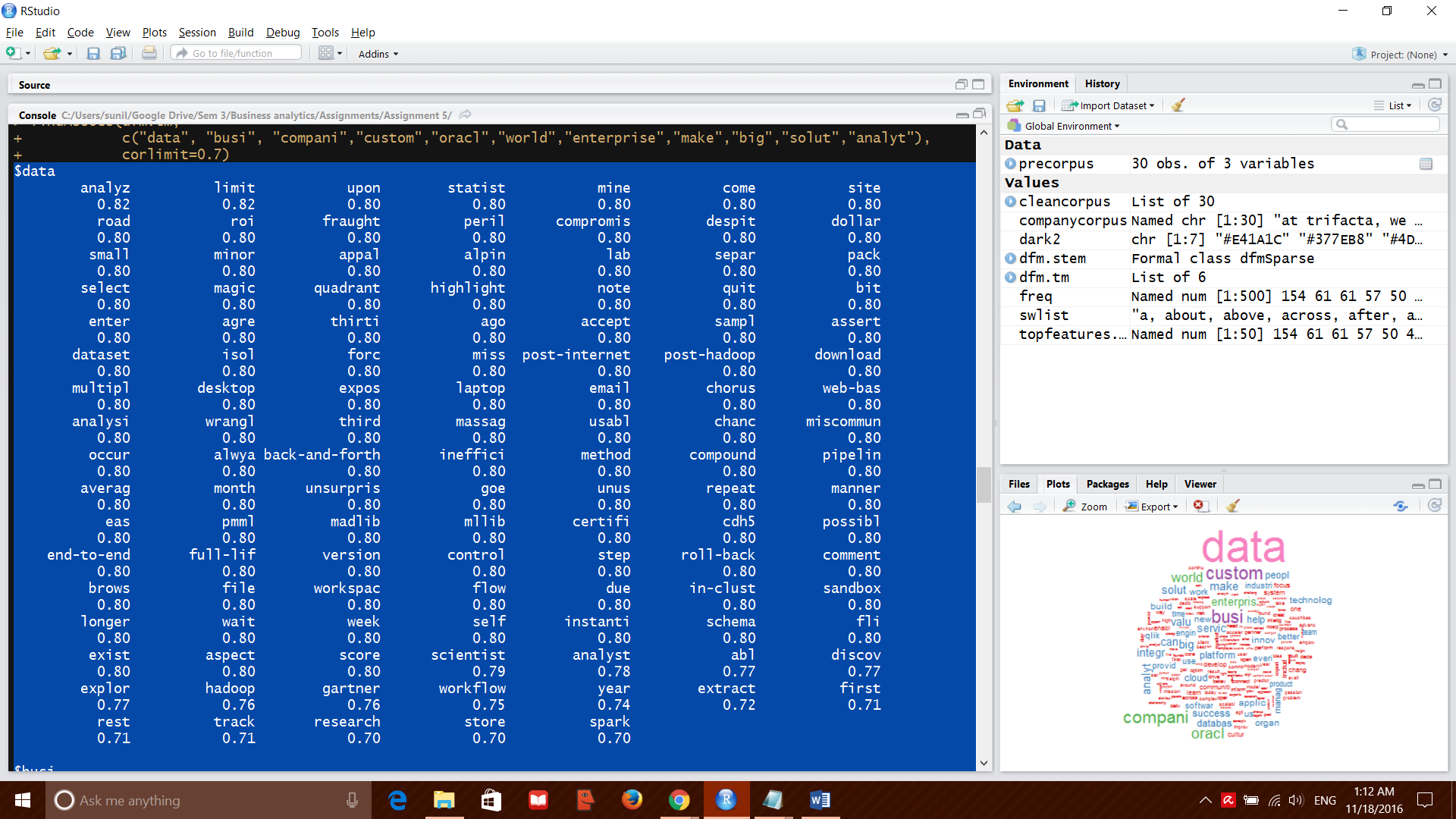


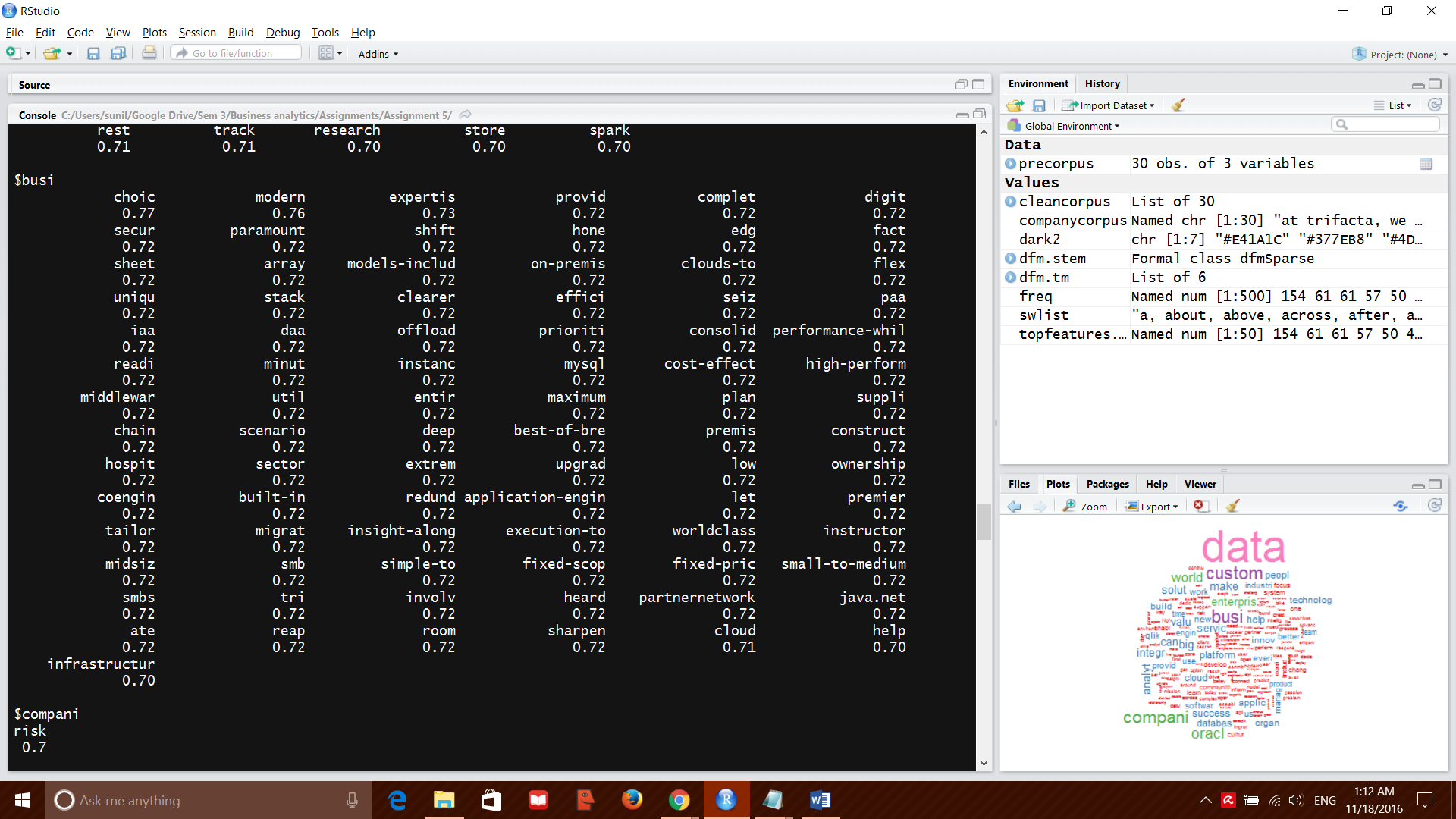


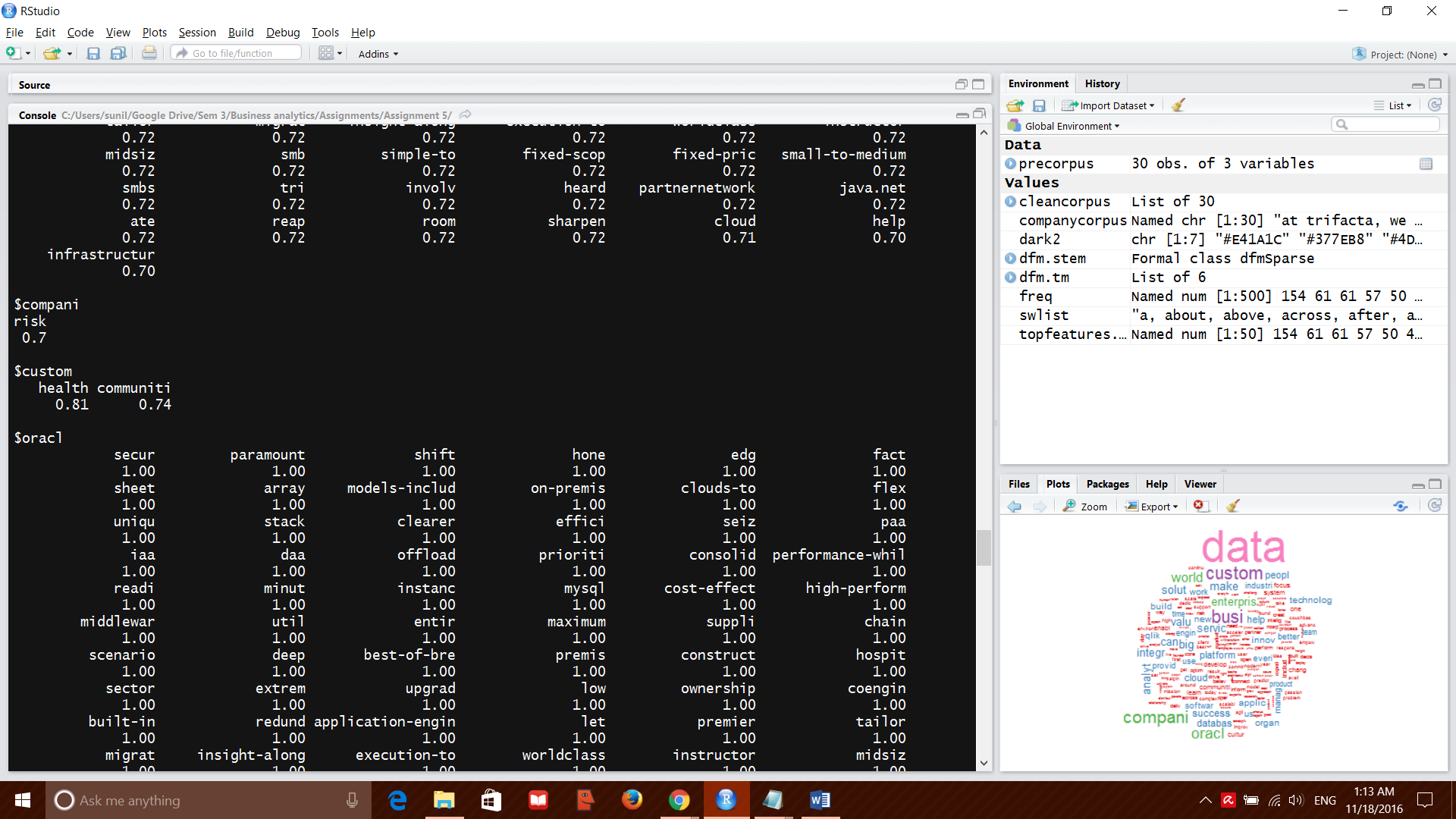
Next we move forward to find associations between the most common words in the corpus. We take the top 10 common words and define a correlation cap of 0.7 for our analysis

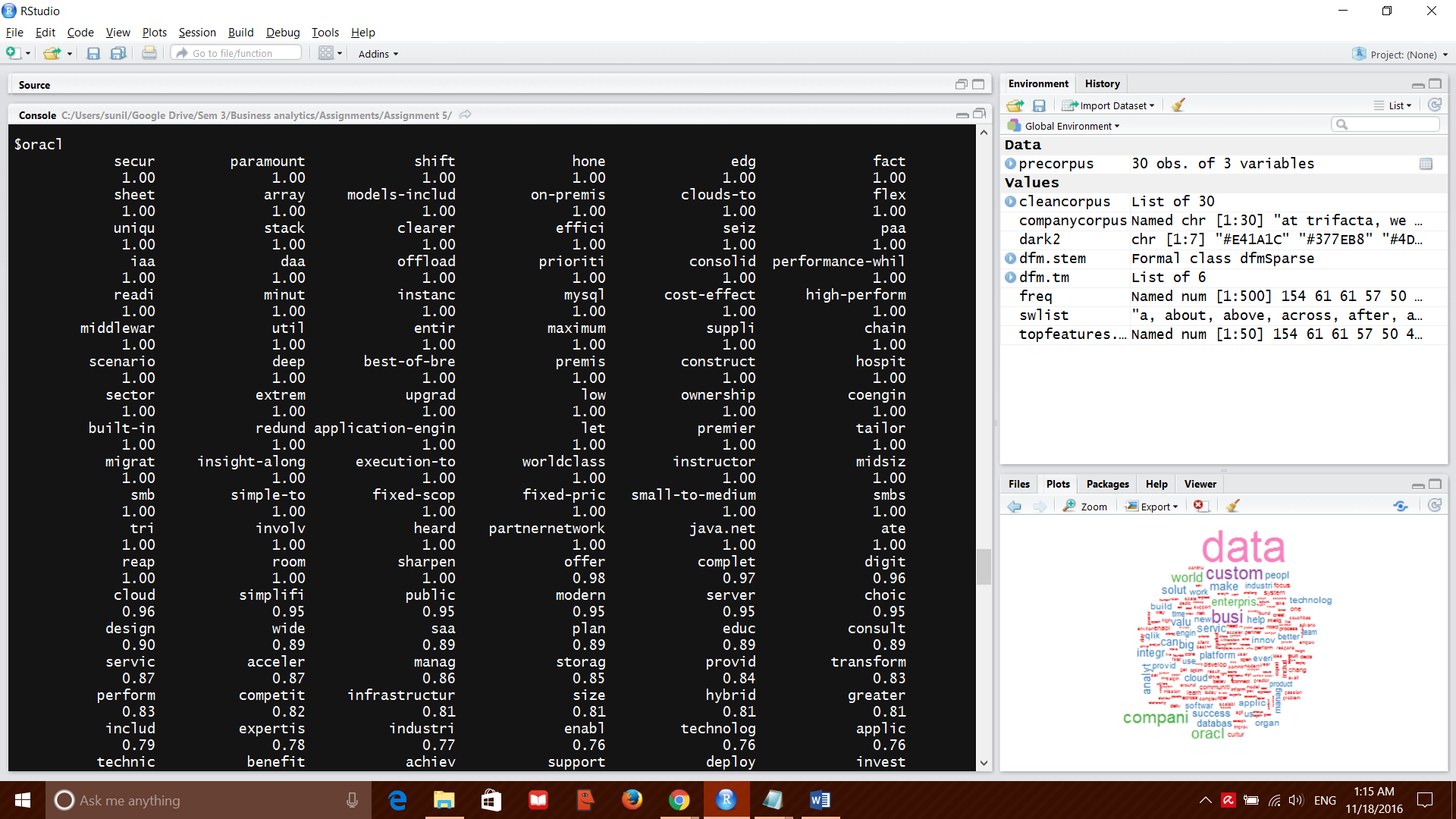


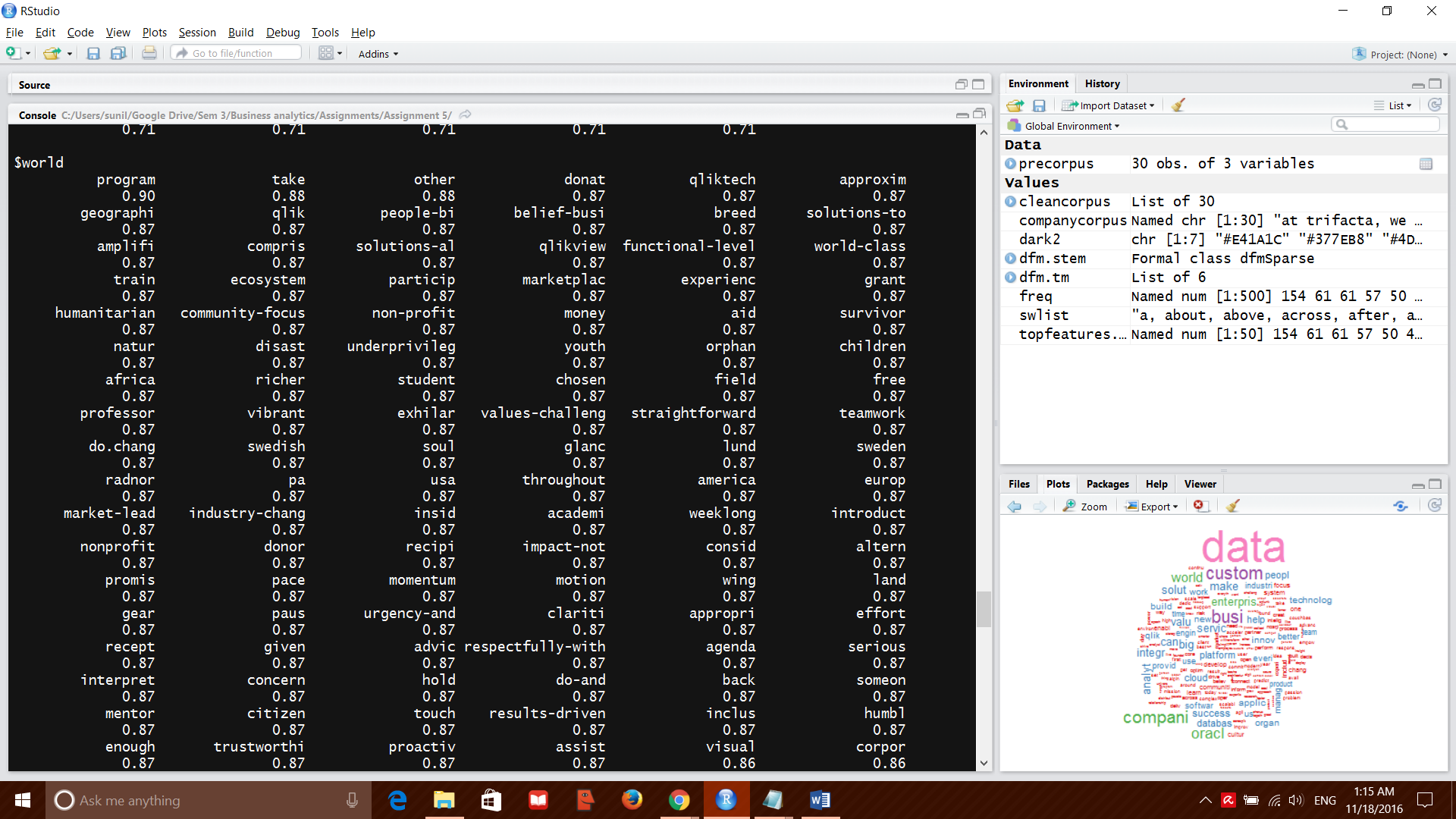
Below are the top results for the words that were queried

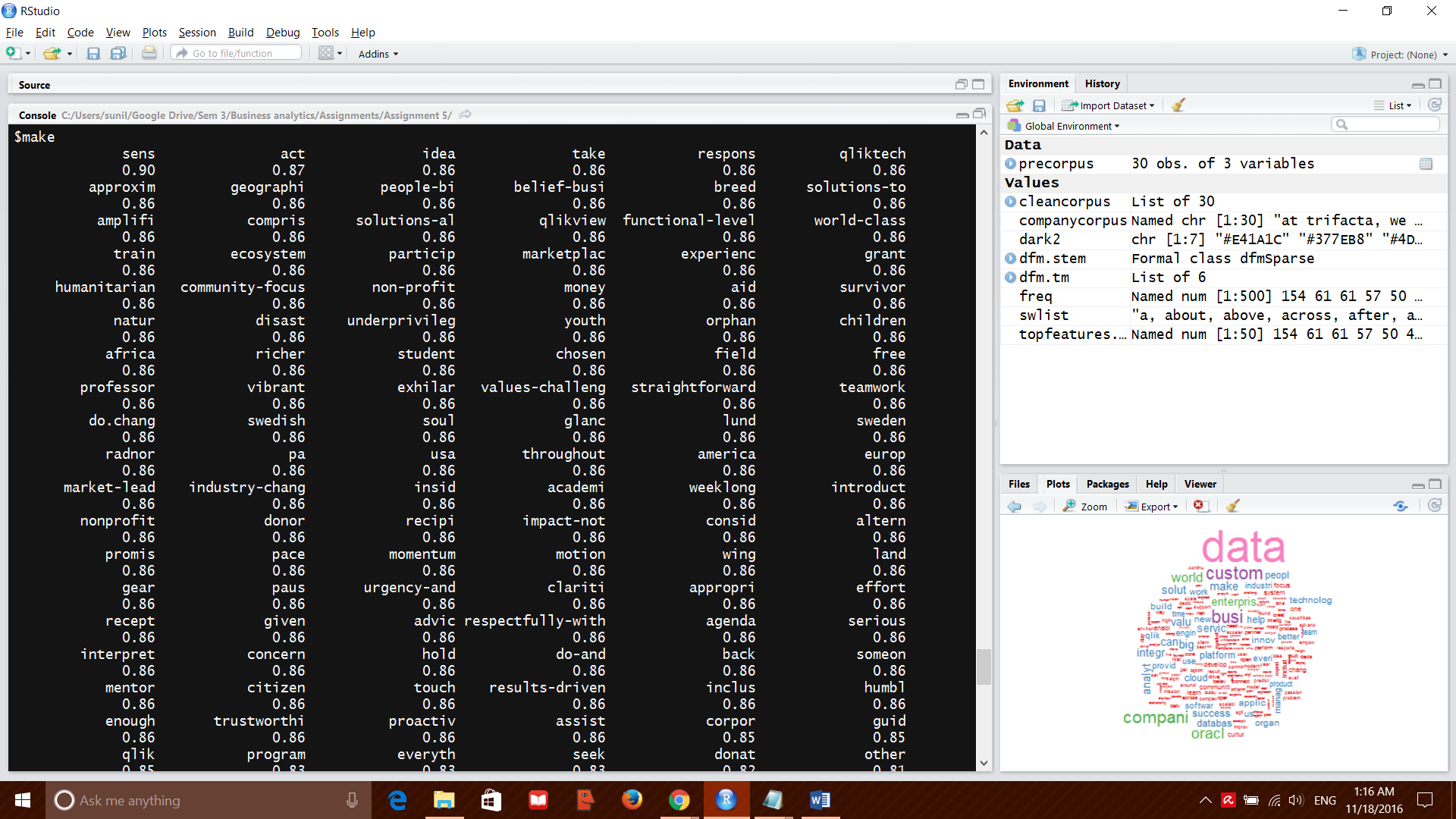


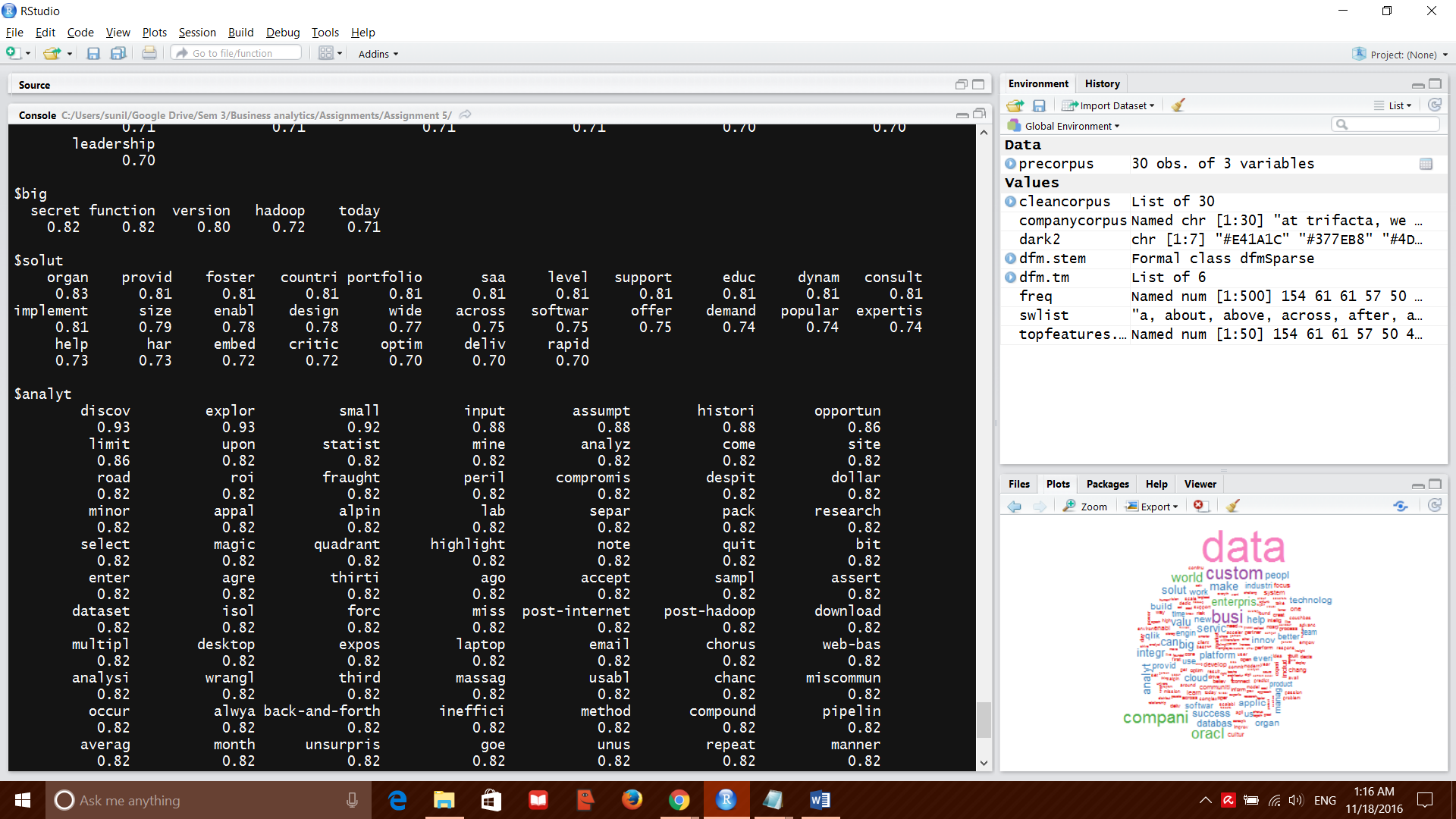


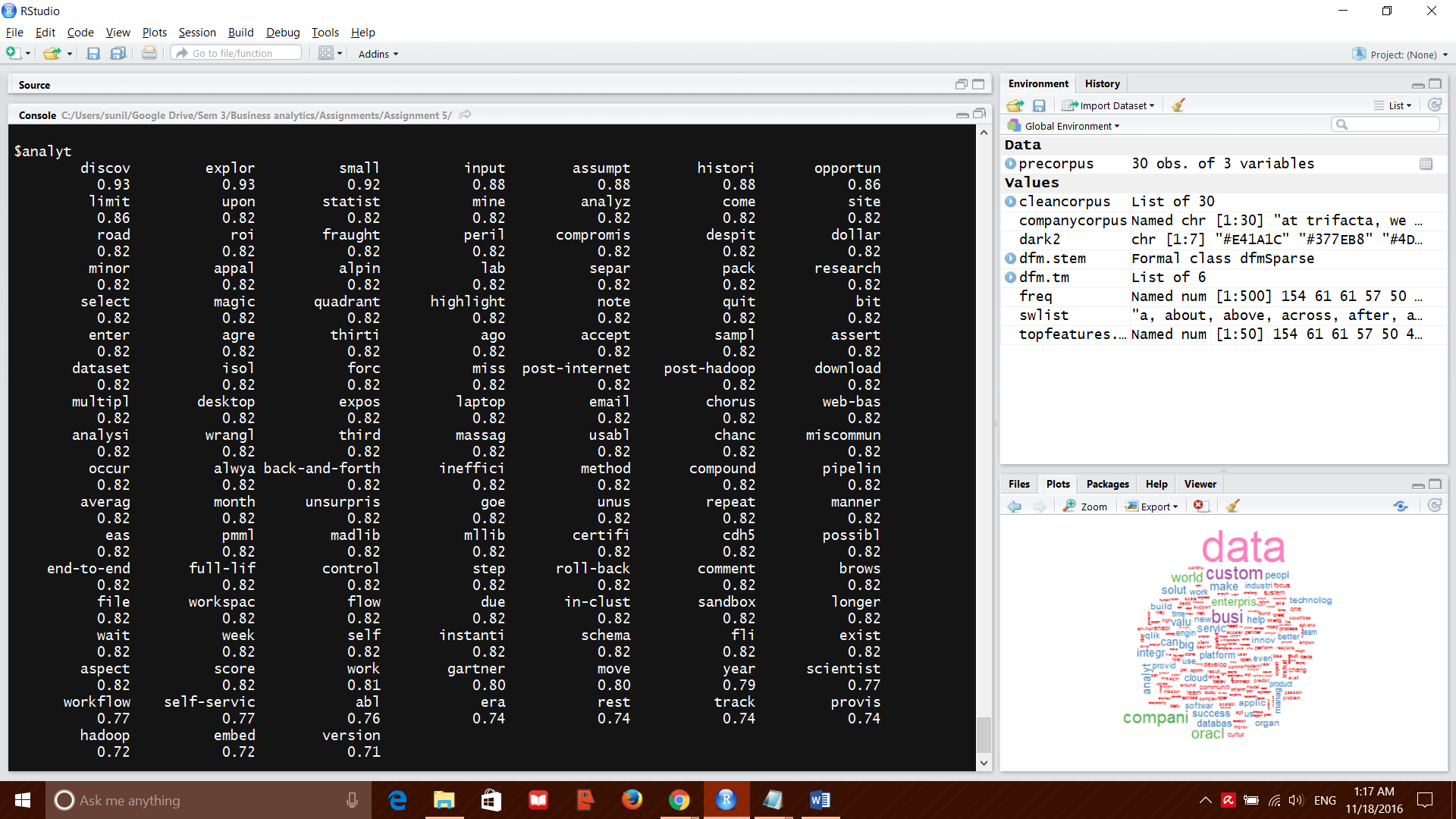












Each of these can be used to arrive at conclusions on what the mission statements and core values talk about.

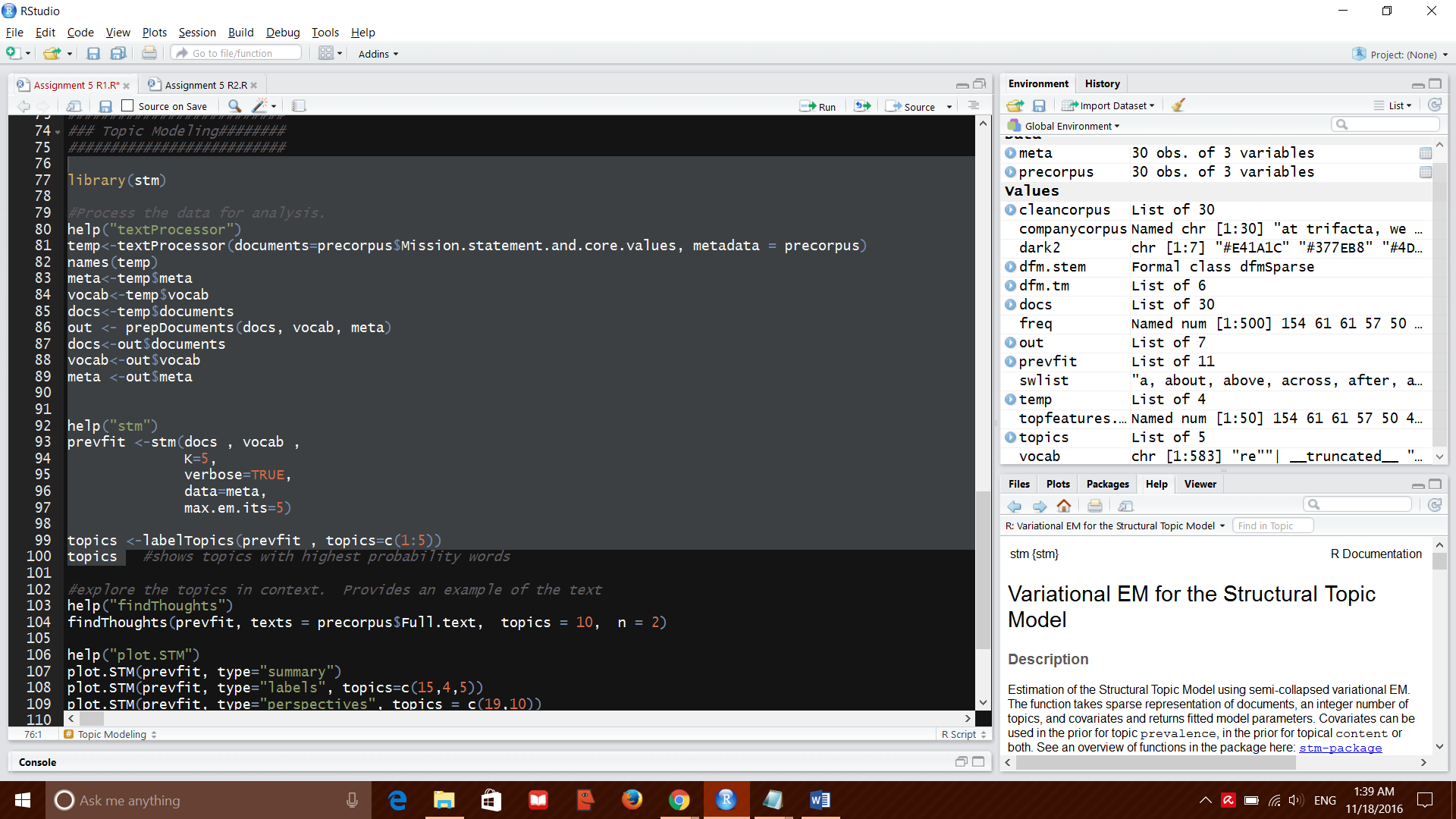
The word ‘data’ is used in the context of analyzing it statistically and mining the data to use the insights learnt from this study to make business decisions.

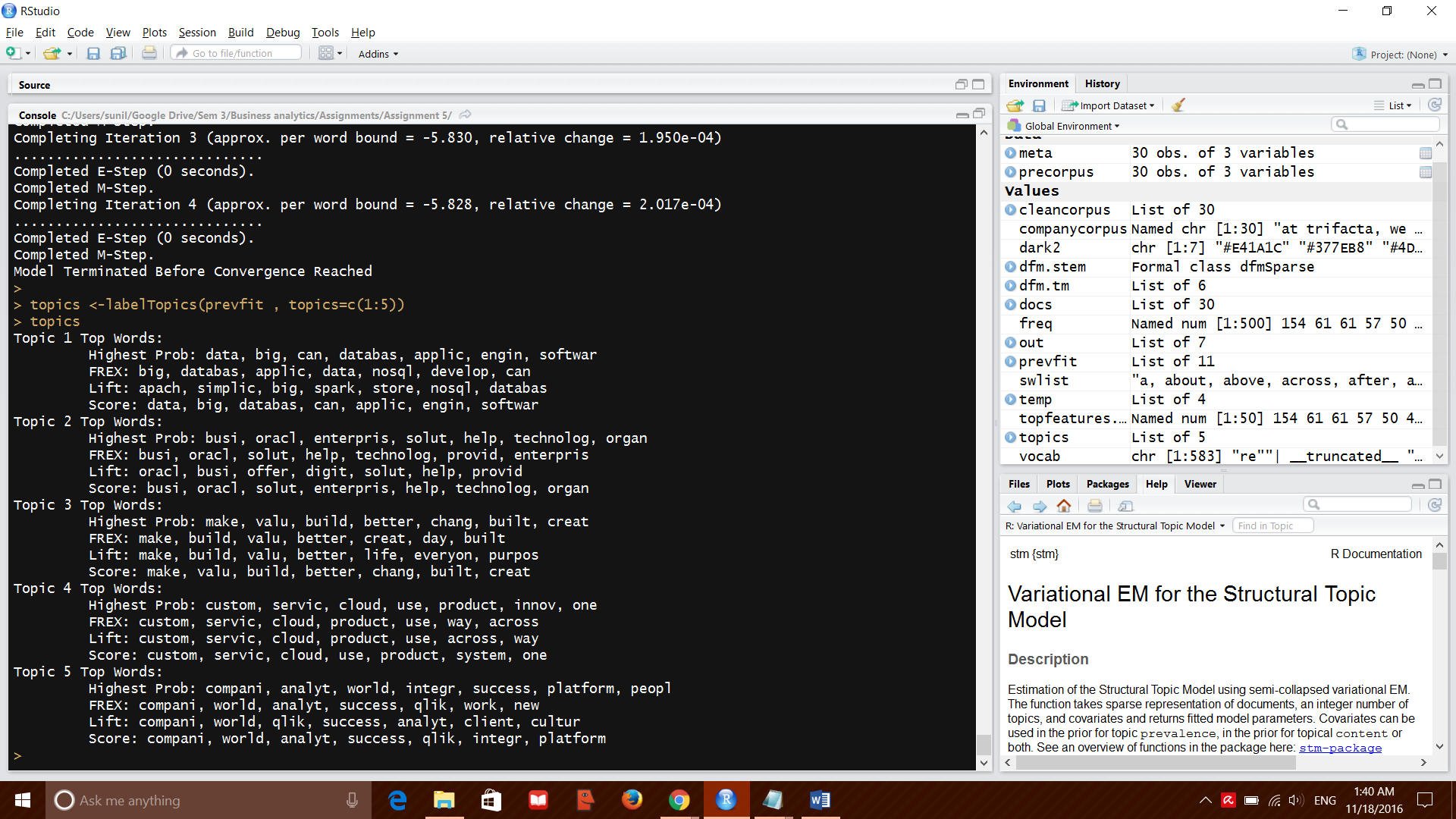
The word ‘business’ is used with choice, modern, expertise, complete and provide. This means that the mission statements discuss about making business choices and providing expertise to make business decisions.

The word ‘compani’ is used along with risk, may mean that the companies aim at reducing risks for companies.

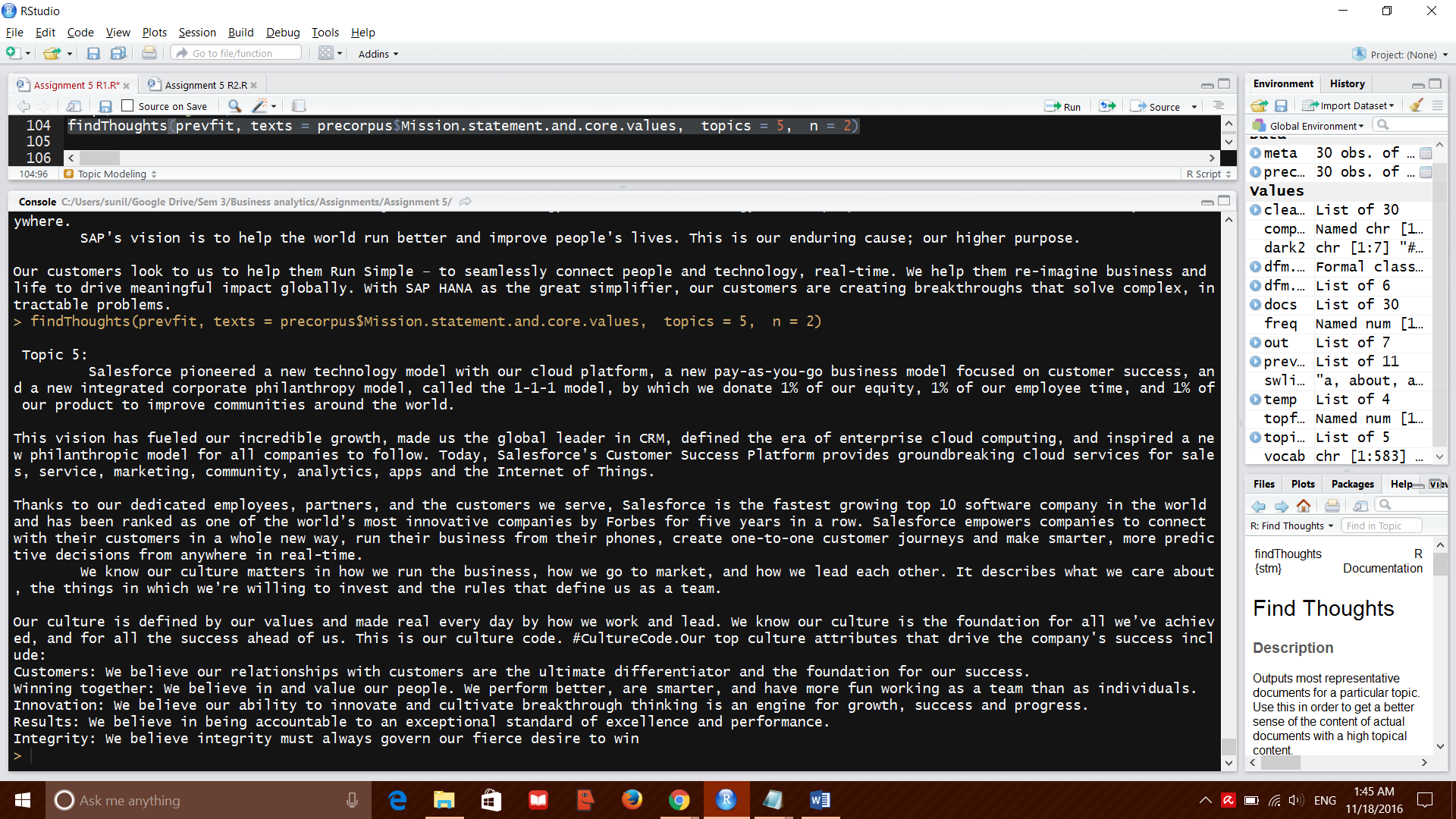
Another relevant inference is from analyt. It is used along with explor and discove. Might mean using analytics to explore and discover new business opportunities for development.

We could also try topic modelling to understand the important topics that the mission statements talk/discuss about.

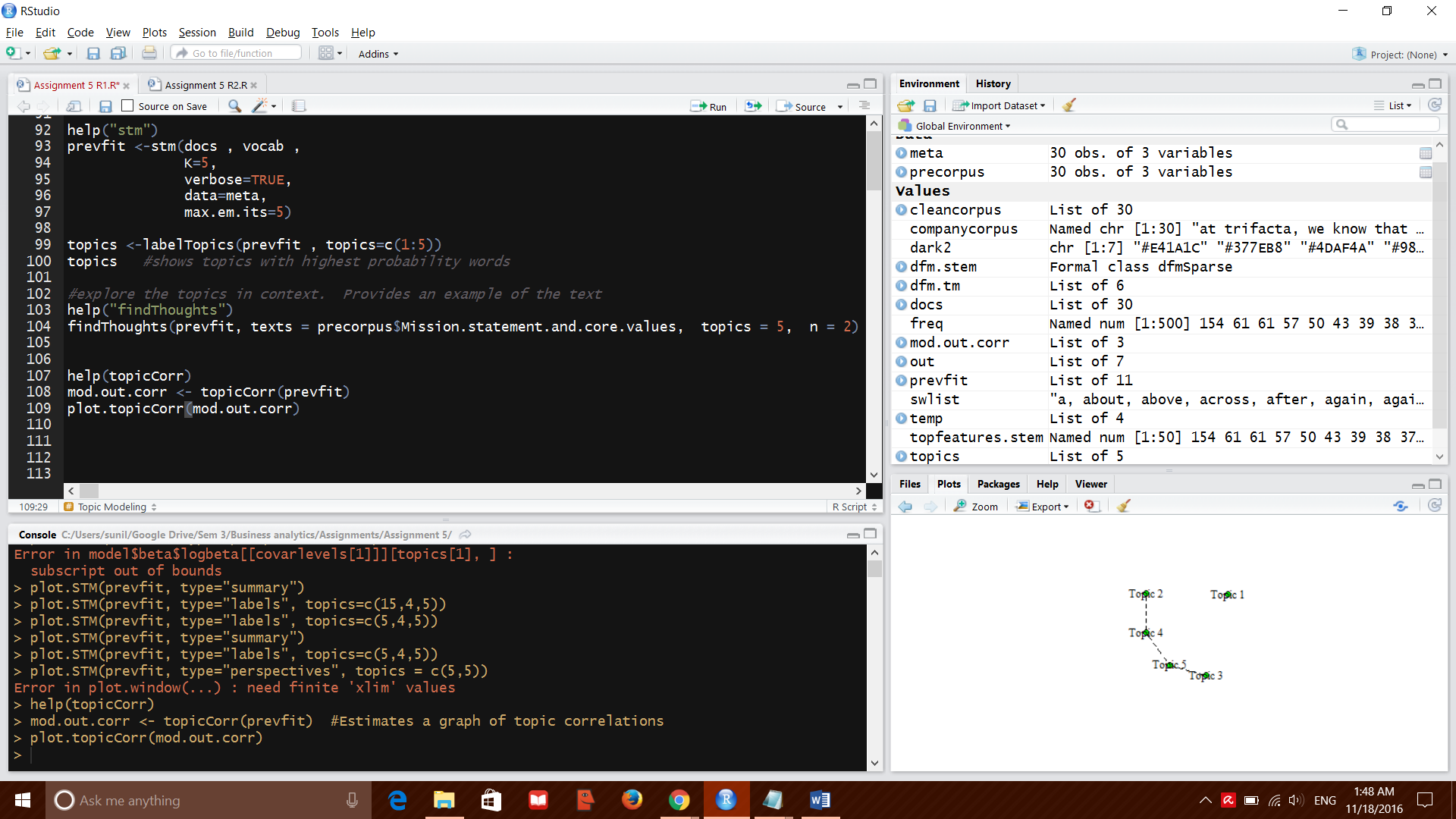


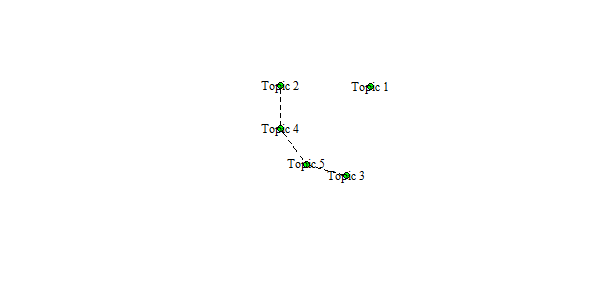


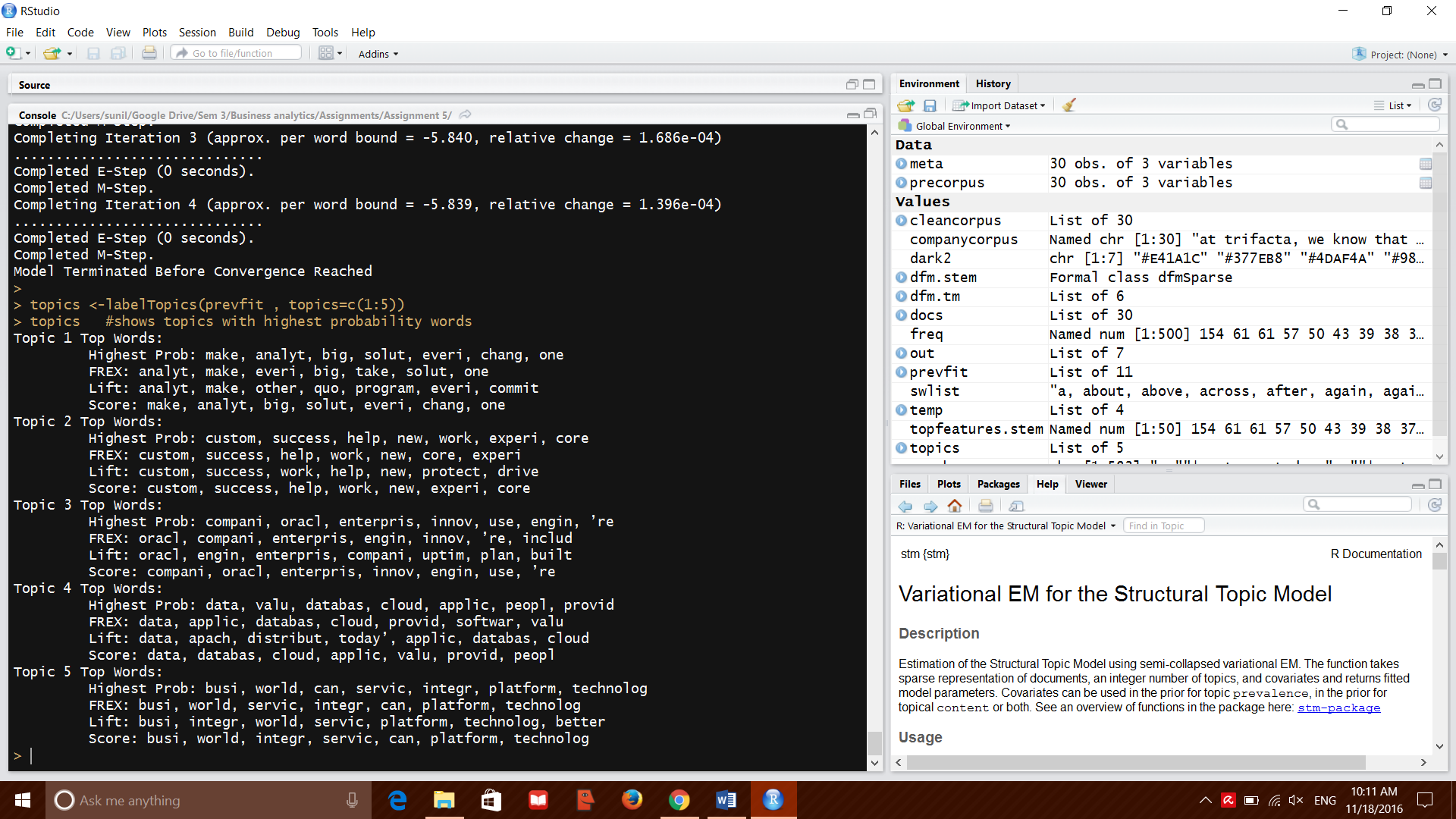
Next we use ‘find thoughts to find relations between the words.



Then we plot the topic correlation to find the related topics



we can infer from the plot that the topics listed under 3,5,4 and 2 are related while topic 1 is by itself. Therefore we must delve to incorporate the features under topics 3,5,4 and 2 in the company.



**PART 2**

**Case:**

This [video](http://digg.com/video/donald-trump-linguistics-answer-question) discusses Donald Trump’s linguistic style. Sometimes when he speaks he seems erratic and unfocused, yet many people like him and connect with him.  Assume you are providing “intelligence” to Trump’s campaign manager on what is making this candidate effective with people

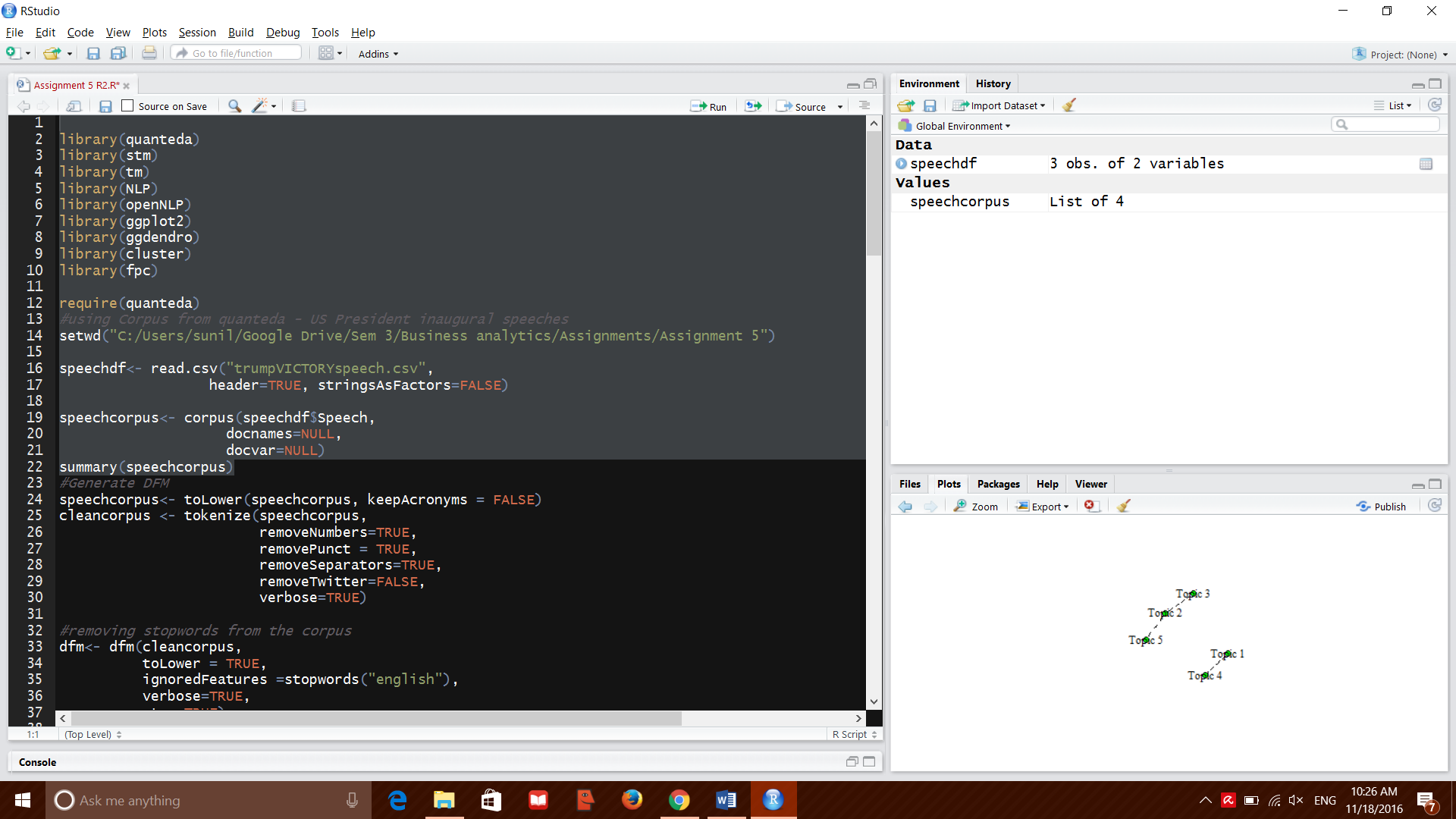
**Steps:**

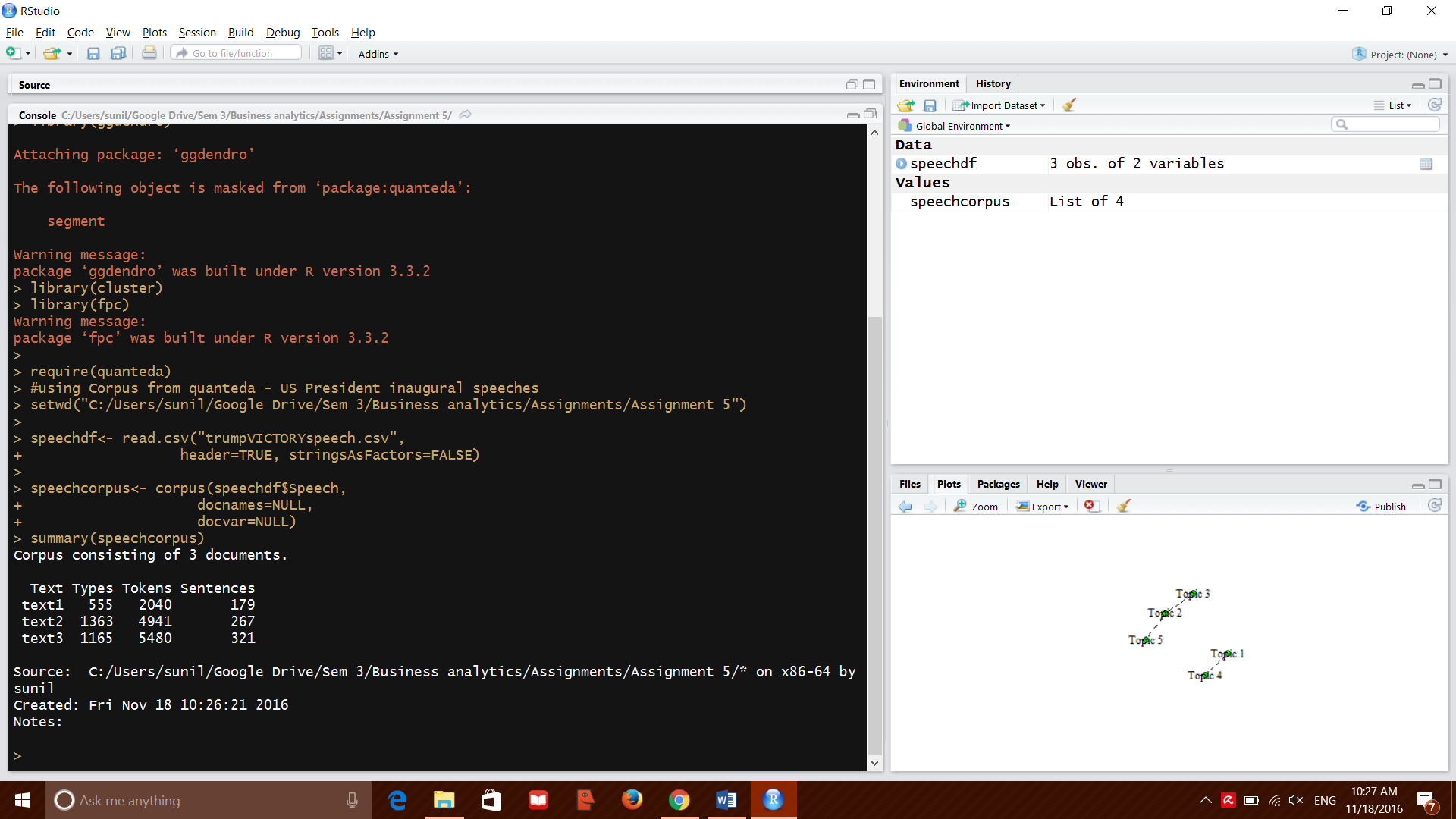
* Search for “Donald Trump speech transcript” and select 3 speeches of your choice
* Create a corpus for the speeches
* Complete a frequency analysis of word usage
* Complete a sentiment analysis
* What are the common topics in the corpus
* Write a memo style report summarizing Trump’s linguistic effectiveness

I have picked 3 of his iconic speeches for the comparison

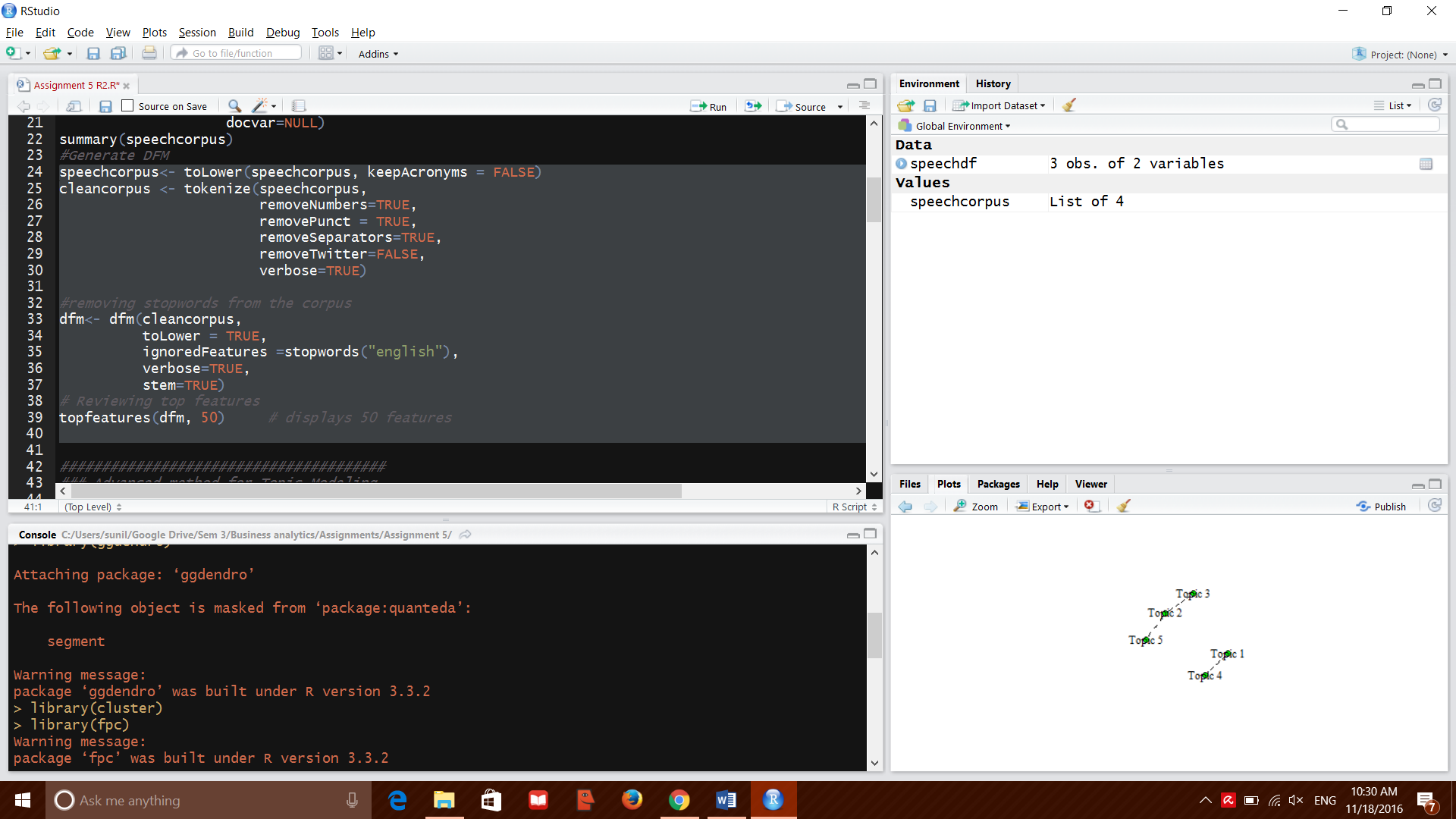
1. His victory speech after the results of the 2016 presidential election was announced
2. His presidential nomination acceptance speech
3. His speech in Pennsylvanian in June 2016

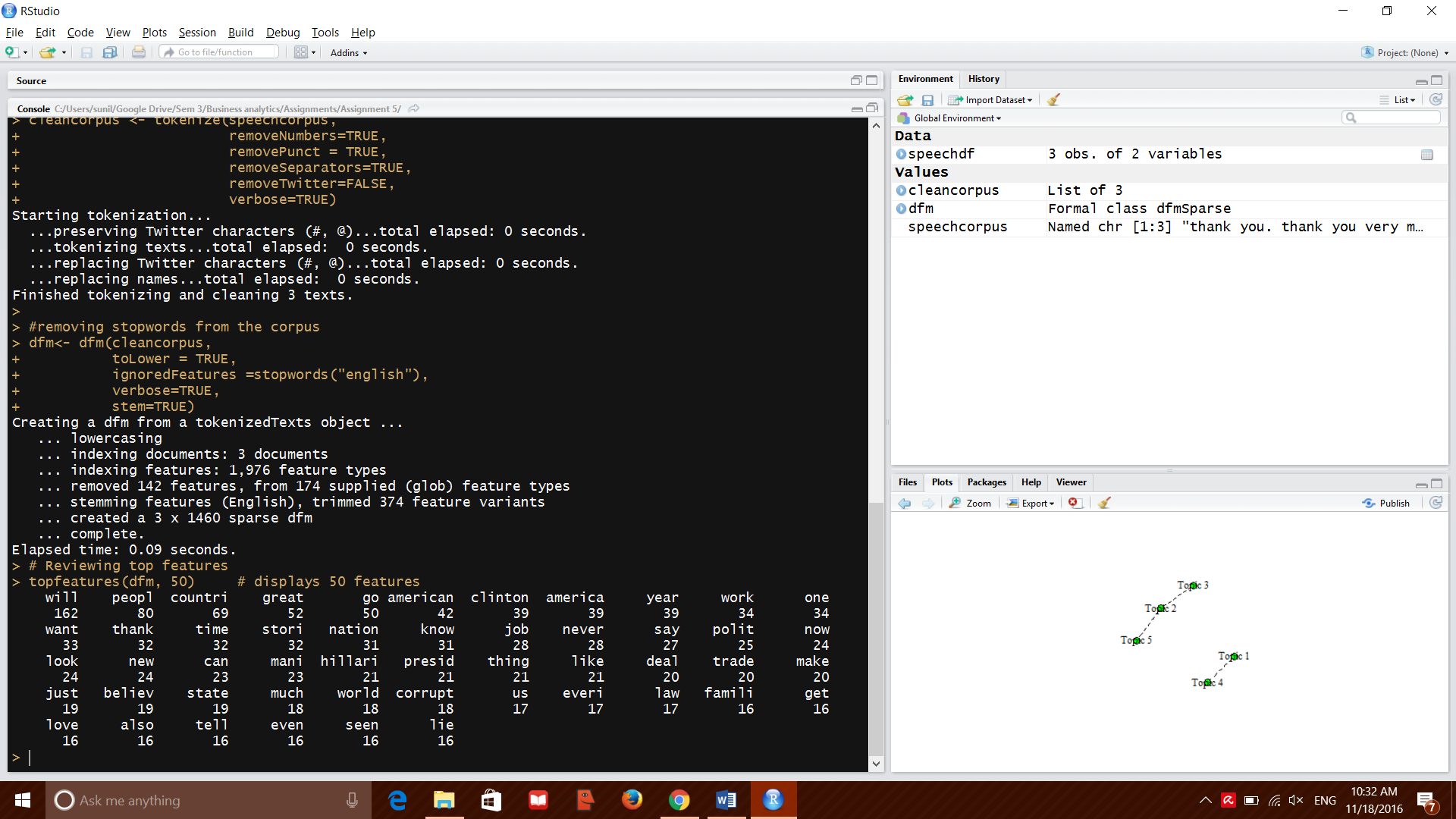
Corpus creation





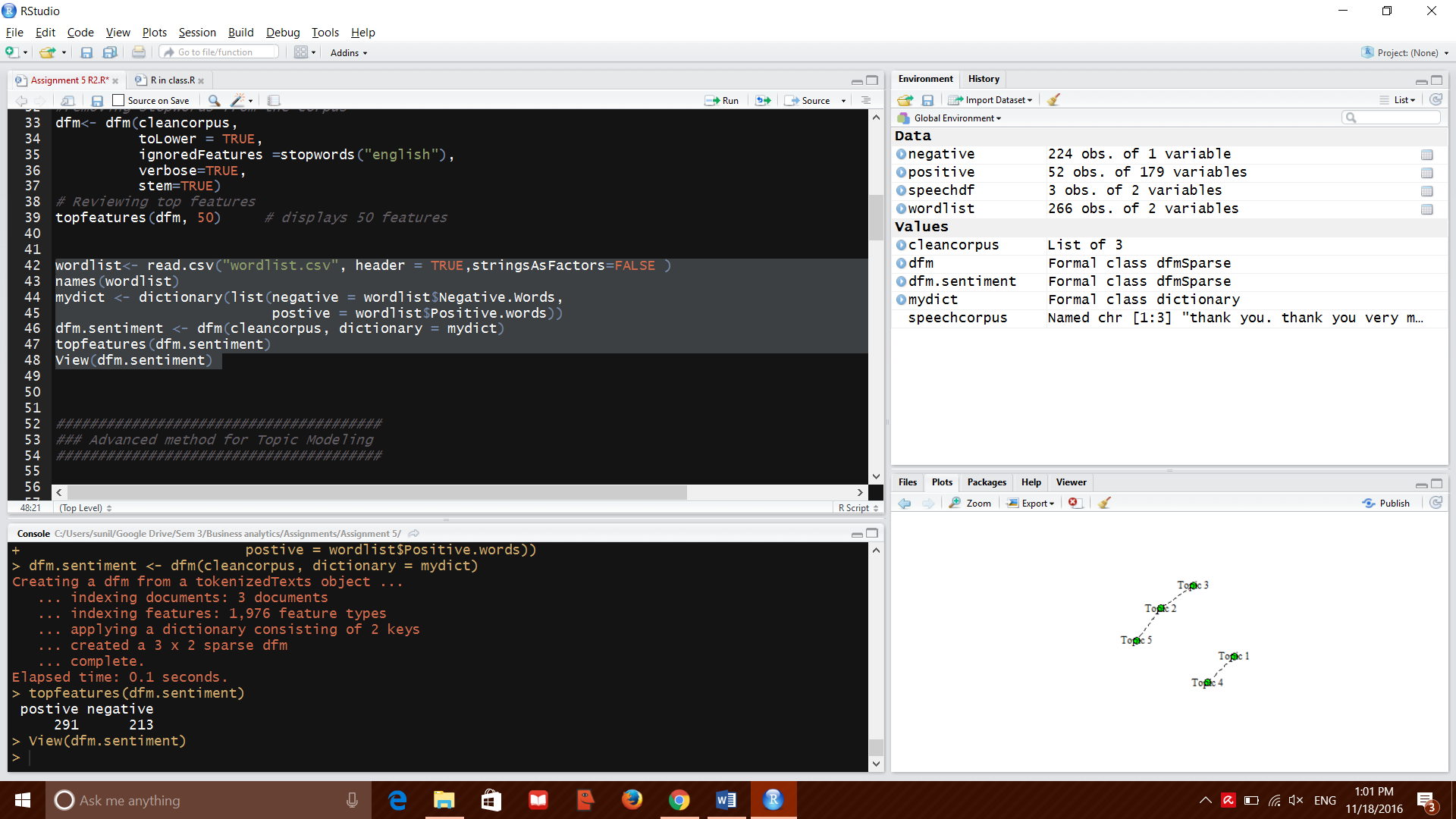
Clean up and display top 50 features in the corpus



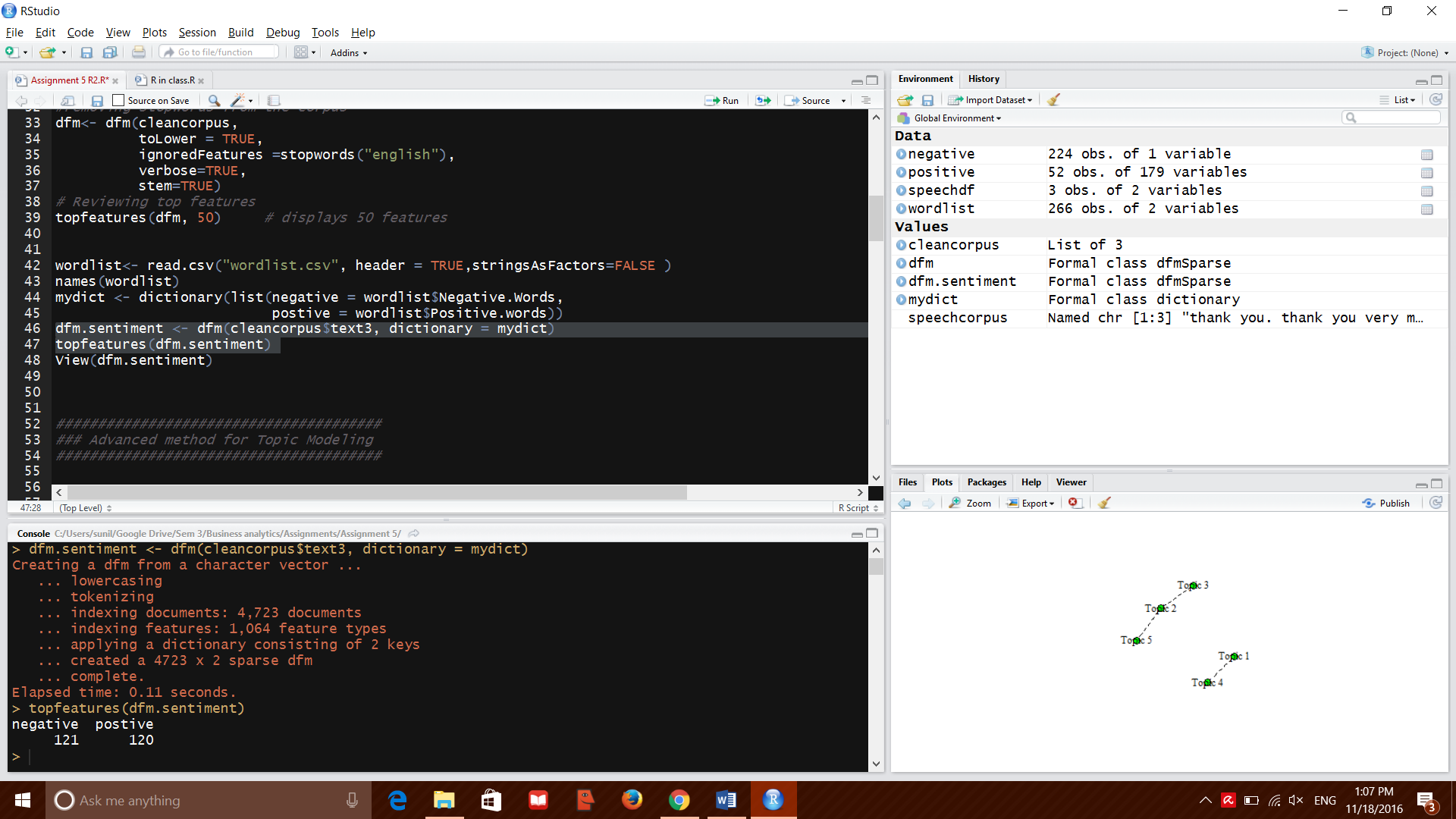


Sentiment analysis:

Creation of negative and positive word dictionaries and sentiment analysis of dfm created earlier



The sentiment analysis shows that the corpus has more positive words than negative. Maybe its because we are analyzing his victory speech and nomination acceptance speech. If we analyse his speech from Pennsylvania alone, we see that the negative words and positive words are interspersed equally.



Topic modelling output is at

<http://127.0.0.1:4321/#topic=0&lambda=1&term>=

at a glance, we can see that the people, country and Clinton feature prominently in all his speeches.