BAN432 fall 2022 - First group assignment

Goal of this group assignment

This assignment is based on the insight of Hoberg, G. and Phillips, G. (2016) "Text-based network industries and endogenous product differentiation". The task focuses on the implementation of a clustering analysis in a portfolio choice problem. This assignment also evaluates different representation of the text in the n-gram space (uni versus bi-grams) and term limits based on Log-Likelihood.

Please submit your assignment even though you were not able to find a solution to all tasks.

Hint: Before starting to code, try to separate the task into smaller pieces.

Formalities

This assignment will be handed out on 19th of October, 2022 at 14:00 and has to be submitted no later than the 26th of October, 2012 at 12:00. Submit your commented coding file and a pdf including the numerical results as well as answers to questions posted. You do not need to describe your coding in the pdf file. Please comment your code in the .R file shortly so that the grader can reconstruct your thinking. You do not need to explain the used functions.

Work together in groups of four and submit the assignment via Canvas.

Constructing a portfolio tracking oil firms

It is the 31.12.2013 and you are a portfolio manager constructing a portfolio for 2014. You can allocate 1 million NOK across the 500 companies available to you. You are convinced that oil companies will perform very well in 2014. Unfortunately, regulation prevents you to invest a single NOK in this sector, in your data defined as (`industry.fama.french.49 == 30 Oil` in the `raw.data` data frame). Luckily, from Hoberg G. and Phillips G. (2016) you know that industry assignment is not perfect and that there might be firms which business is actually in the oil sector while not classified as such.

You develop the following strategy using insights from textual analysis: You identify a subset of firms that – based on their business description in their annual report – sound like oil firms even though they are not classified as such. You invest an equal amount in each of those firms.

Use the `firm_dataset.Rdata` as introduced in class. In the data file you find:

- `raw.data`: a data frame with meta information about the firms, such as CIK, industry classification, and monthly returns
- `section.1.business`: a data frame that contains all business descriptions of the firms. Please use this text for your analysis.

The basic procedure for your analysis is as follows:

- Clean the data: remove punctuation, remove numbers, remove stopwords, make lower case, only consider terms with 3 to 20 letters, and delete excess white spaces.
- Transform the data into bigrams. One easy implementation using the `tokenizers` package:

```
require(tokenizers)
temp <- "this is a random sentence"
temp.bigram <- tokenize_ngrams(temp, n = 2, ngram_delim = "_")
print(temp.bigram)
```

```
## [[1]]
## [1] "this_is"         "is_a"            "a_random"        "random_sentence"
```

```
print(paste(temp.bigram[[1]], collapse = " "))
```

```
## [1] "this_is is_a a_random random_sentence"
```

You can also use any other implementation. However, make sure that you work with bigrams from now on.

- Make a document term matrix only including tokens that appear in more than 5 but less than 100 documents.
- Identify firms in the Oil sector and firms not in the Oil sector.
- For each token (bi-grams) in the corpora compute the Log-Likelihood of being in the subset of oil firms and the remaining. Display the top tokens. Are these tokens associated with the oil sector? For the remainder of the analysis, only use the 500 tokens with the highest Log-Likelihood. Log-Likelihood was mentioned in one of the guest lectures and you calculate it this way:

```
calculate.ll <- function(a, b, c, d){
e1 <- c*(a+b)/(c+d)
e2 <- d*(a+b)/(c+d)
ll <- 2*((a*log(a/e1)) + (b*log(b/e2)))
return(ll)
}

# a = freq. of a word in the oil corpus
# b = freq. of a word in the non-oil corpus
# c = sum of all words in the oil corpus
# d = sum of all words in the non-oil corpus
```

- For each oil firm, find 5 peer firms using cosine similarity. In particular,
  - For each oil firm compute the cosine similarity with all other firms
  - Select the five closest non-oil firms
  - Note, allow for overlapping identification. E.g. the same firm being a peer for several oil firms.
- Compute the average return for the oil portfolio and the peer group for the year 2014 (use variables `return.monthly.NY.m01-12`). Note, if a peer firm is selected several times, it is included several times in the portfolio (e.g. more money allocated to that firm).
- Evaluate the performance of your algorithm. Compute the Root Mean Squared Error (just google the definition) between both portfolios.
- Evaluate whether using uni-grams (versus bi-grams) performs better/worse? For the remainder of steps, use exactly the same approach. An easy way to implement this analysis is by recycling your current code skipping the step in which you transform the corpus into bigrams.
- [Optional] Does this approach of constructing tracking portfolios work? Knowing the value of the RMSE is not necessarily enough to assess performance against alternative portfolio selections. Construct 10.000 random portfolios of similar number of firms and compute the corresponding RMSE. Display it as a histogram. Does the text approach do a better/worse job?