Series 4

1. Collecting data: Data would be collected as follows (categorical. semantic differential, rank-order, multiple-choice):

    a. You want to evaluate the preferred characteristics of a new product (different forms, different colours, and so on)

    Solution: Categorical. One wants to collect the choices or opinions about different categories e.g. colour in white, black, grey, red or different forms of a product.

    b. You want to test the analytical thinking of job candidates by a test during the hiring process.

    Solution: multiple-choice. In test during a hiring process there are typically presented 3 or 4 different answers to the candidate, thus, it is a multiple-choice test. The tests are also typically online / computer-based such that with multiple-choice an automated evaluation can be performed.

    c. You want to rate the importance of the opening hours in the evening of a local store.

    Solution:

    Rank-order. Either one is asking different opening hours option and is asking the customers to order them according preference (evaluation of different opening hours) or one is asking different characteristics of a good store e.g. variety of products, brands of products, opening hours, individual services and so on and let them rank against each other.

    Categorical: One is as giving different characteristics of a store e.g. opening hours and asking if this is not important at all, somewhat important, important, very important, not sure / not applicable.

    d. You want to find out how many foreign languages people are speaking.

    Solution: categorical. One want to know the number of languages, thus, 1, 2, 3, 4, ….

    e. You want to find out the personal personality traits.

    Solution: semantic differential. Personality traits are typically measured in a range of converse characteristics like concrete thinking – conceptual thinking, seriousness – sanguineness, …

    f. You want to know the 5 most important benefits of a new mobile app.

    Solution: rank-order. One gives a choice of e.g. 10 possible benefits and the customer should rank them (ranking of all given benefits) or the customer should rank only the 5 most important benefits out of the 10 (partial rank-order) or the customer should only mark the 5 most important benefits without giving any further order. It depends on the purpose of the data collection.

g. You want to know the least useful and most useful features of a new vacuum cleaner.

Solution: rank-order. One is giving a choice of e.g. 5 or 10 possible benefits and the customer should rank them (ranking of all given benefits) or

h. You want to find out what people are thinking about the political agenda of a party.

Solution: categorical. One can give different topics out of that agenda and ask the people what they are thinking about these topics, by giving them the possible answers of strongly disagree, disagree, neutral, agree, strongly agree.

2. A bank is asking you in supporting them in rogue trading analytics in the front office (i.e. the business department where the traders are working and performing). The management of the bank is expecting that you are performing data analytics and with advanced analytics procedures you would then detect patterns of a possible rogue trading. You know that there are several departments with traders. The back office does the work for all trading departments and are booking and reporting all trades.

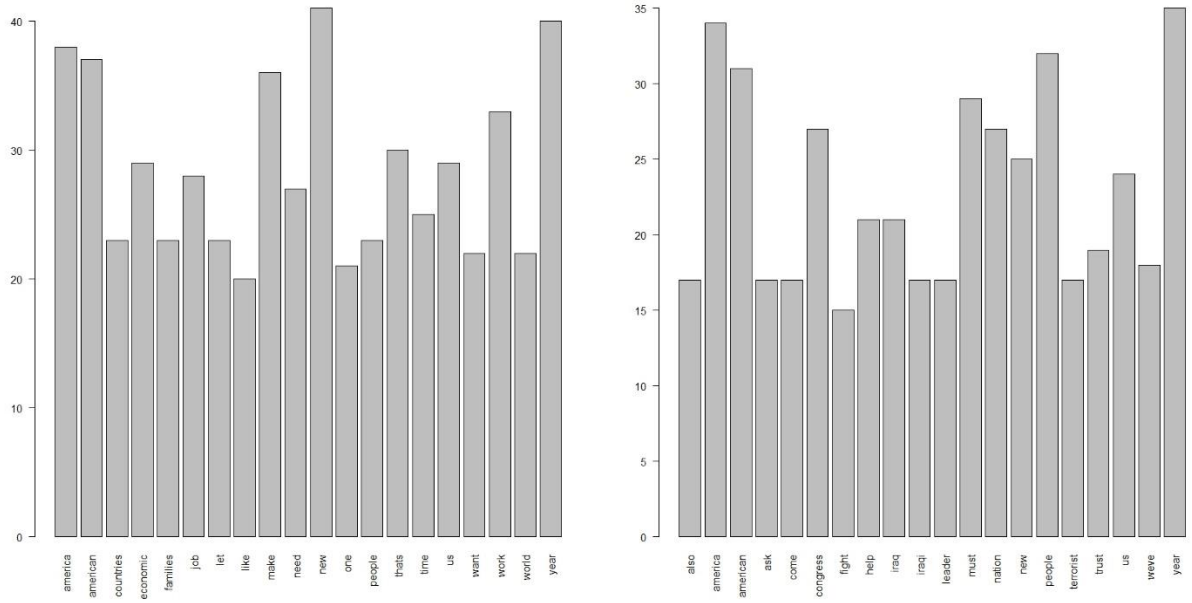
Before you can start with applying methods, you need data. What are all the questions (a bit concretised to this example) you would ask for getting the required data?

Possible questions to ask:

- How many trading department / trading desks are in total?
- What data are collected per desk?
- Are only the trading data per desk collected or also the communications of the traders (e.g. Bloomberg chat)?
- Who is entering the data in the front office? Which are entered by the traders and which are collected / generated automatically?
- Do all departments / desks have the same systems / infrastructure or what are the different systems?
- Are all systems collect the same data or are there differences?
- What are other data available which could be linked to the traders? E.g. the limit system, the for trading approved instruments, and so on.
- Who has access rights to these recorded data?
- How we can have access to these data?
- Can the data be downloaded?
- In which format the data can be downloaded?
- What is the volume of the data?
- What are other fields recorded?
- Can we have a description of the data fields of the recorded data?
- Who has the rights to change data in the front office system afterwards?
- How are the data transferred from the front office systems to the back office systems?
- What are the data available in the back office?
- Is there only one back office for all trading desks / instrument types or are there different back offices?
- Which back office is handling which type of instruments?
- What is the delay between the front office bookings and the bookings in the back office? (X minutes, Y days, Z months)
- Who has access rights to the back office systems?

- Again for back office data:
  - How we can have access to these data?
  - Can the data be downloaded?
  - In which format the data can be downloaded?
  - What is the volume of the data?
  - What are other fields recorded?
- Are there identifiers how one can merge the front and back office data?

3. Running the script "TextMiner.R" with the two given files yields the following charts:



Some frequent vocabulary is common for State of the Union Addresses, such as for instance "America" and "American".
In the first speech terms like "new", but also society related terms like "family", "job", "economic" and "work" are dominant. Further, the verb "make" dominates. These are actually the topics of president Barack Obama.
In the second speech, terms related to the US state dominate like "congress", "nation" and "people". Also the predominant verb is "must". Further, the speech contains terms related to Iraq exceptionally often. This is a strong indication for a speech of George W. Bush, as the situation in Iraq used to be a hot topic during his presidency.

It is also interesting to see how our brain can analyse text only based on the information of terms and the frequency of terms. Thus, if the information is somehow structured and not too high-dimensional our brain is a very efficient data analytics "tool".