Developing a text analytics algorithm in survey research

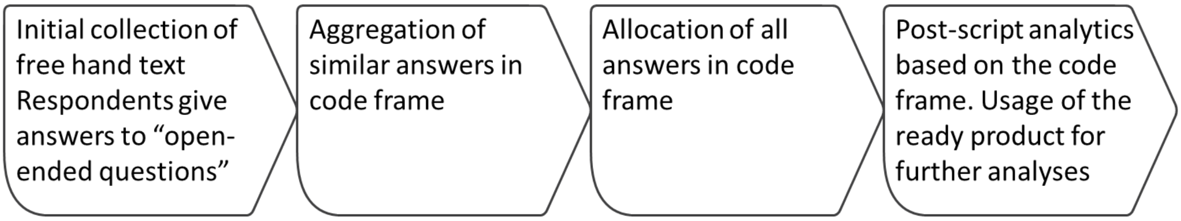
Company's overview:

GemSeek (www.gemseek.com) is a market intelligence and consulting company. It helps business leaders with decision support analytics that have a direct impact on bottom line and competition. Company's services are organized around Data science and predictive analytics, Market & Industry Intelligence, Customer Insight & Brand Analytics, Advanced Visualization Solutions and Competitive Intelligence.

Definition of the problem:

One of Gemseek's core activities is developing and implementing marketing survey research among different target groups across the world. The results serve as basic foundation for further analysis on customer perceptions, behavior and brand affiliation. The correct and appropriate data collection, on-time delivery and coordination of different levels is pivotal for obtaining correct and unbiased results, high customer satisfaction and enhance the industry image.

Apart from quantitative measurement provided by series of scales and variables the survey design also includes a separate set of data collection of qualitative type. These fields actually consist of freehand text, sometimes directly inputted from the respondent himself, where he/she gives more detailed information, opinions, statements in free text format. One of the biggest challenges for company was dealing with this type of data i.e. how to quantify and give numerical notation to that specific qualitative inputs. The process could be summarized by the following basic steps


Initial collection of free hand text Respondents give answers to "open-ended questions"
Aggregation of similar answers in code frame
Allocation of all answers in code frame
Post-script analytics based on the code frame. Usage of the ready product for further analyses

Business implication:

Currently a big portion of the work regarding text analytics and further aggregation is done manually, which costs a lot of human resources, additional budget and time. Since text analytics is one of the core domains of the industry an automatization of the whole process would optimize this process and innovate the way qualitative information is treated and utilized.

Task description:

Gemseek is looking at potential way to optimize the process of allocation of different opinions and statements into predefined categories which logically describe their meaning and sentiment. Thus, the amount of manual work will be reduced significantly.

A successful solution will be in the form of a **machine learning algorithm,** that learns from a set of previously classified comments and uses the information to assign new free text statements to the existing categories.

The participants will be provided with a dataset with unstructured text, containing various opinions, statements and sentiments expressed by respondents in one of the surveys administered by Gemseek. The dataset will also contain an individual line ID (respondent number). Additionally, the set of predefined categories will also be available, containing labels and sentiment.

The main task would be to develop an algorithm that reads through the text, takes into consideration the logic and the sense of both labels and sentiment, and suggest possible allocation of unstructured text within category levels. Participants will also have to align with the following:

- Some free-hand text could contain parts that could be affiliated to multiple levels of the categories. The algorithm should also take into consideration this option.
- If certain opinion or statement could not be allocated to any one of the categories, it could be marked as "Other" (such cases exist in the training data – variable "Not.Useful"). This option should contain no more than 10-15% of all unstructured opinions.
- The algorithm should take time to run within the reasonable boundaries.

Materials provided:

- Complete data set of unstructured text in CSV format with manually filled-in categories (train set)
- Evaluation set with no category information (only the original comments and country/person identifiers)
- Data dictionary
- An additional dataset from a different survey with classified comments (can be **optionally** used by the teams but some category labels may be different from the training/evaluation data).

Other information and resources and consultations will be readily available from Gemseek team on request.

Expected results:

- Brainstorm on various methods of solving the task.
- Presentation of different algorithms, stating pros and cons for each one.
- Used variables, predictors, distance measures, parameter estimates etc.;
- Suggestions of appropriate software and tools.
- Discussion of the results with bigger audience.
- <u>Final deliverables:</u>
    - A dataset containing all comments from the evaluation set with assigned categories (following the format of the train data).
    - Complete scripts and developer codes for completing the task.

Assessment criteria:

All suggestions for algorithms will be closely reviewed and assessed by Gemseek team. The following criteria will be used when choosing the most effective method:

- $\text{Accuracy} = \frac{\text{tp+tn}}{\text{tp+tn+fp+fn}}$
- $\text{Precision} = \frac{\text{tp}}{\text{tp+fp}}$
- $\text{Recall} = \frac{\text{tp}}{\text{tp+fn}}$

In order to calculate final model score, the following formula is going to be used:

$$Model\ Score = 60\% \times \text{Accuracy} + 25\% \times \text{Precision} + 15\% \times \text{Recall}$$

Gemseek will thoroughly review all suggested approaches and methods. The most effective will be considered the one achieved the highest model score.