TStat

Residential SUMMER SCHOOL

TEXT ANALYSIS:
A QUALITATIVE AND QUANTITATIVE APPROACH

Florence, 14-17 September 2020

Today researchers across a wide variety of fields find themselves having to analyse an increasing amount of qualitative information. The objective of this summer school therefore, is to provide participants the requisite toolkit necessary for the successful planning, conducting and subsequent statistical analysis of qualitative text. To this end, an overview of the following methodologies: qualitative analysis, quantitative content analysis and text mining, to text analysis is provided. The opening sessions focus on the fundamental role of data preparation to the analysis, before moving on to identifying themes and correlations using both the text mining and content analysis approach. The final sessions address the more advanced topics of importing and exporting data, together with document classification.

In common with TStat's training philosophy, the summer school takes very much a hands-on approach to qualitative and quantitative text analysis. Each individual session is composed of both a theoretical component (in which the techniques and underlying principles behind them are explained), and an applied (hands-on) segment, during which participants have the opportunity to implement the techniques using real data under the watchful eye of the course tutor. Theoretical sessions are reinforced by case study examples, in which the course tutor discusses and highlights potential pitfalls and the advantages of individual techniques. The intuition behind the choice and implementation of a specific technique is of the utmost importance. In this manner, the course leader is able to bridge the "often difficult" gap between abstract theoretical methodologies, and the practical issues one encounters when conducting text analysis on real data. Throughout the course the applied sessions are carried out using Provalis Research's QDA Miner, WordStat and SimStat text analysis software. WordStat is a flexible text analysis software, offering both text mining tools for fast extraction of themes and trends and state-of-the-art quantitative content analysis tools, which in conjunction with SimStat (Provalis Research's statistical data analysis tool) and QDA Miner (for qualitative data analysis) offer users an extremely powerful and flexible integrated toolkit for qualitative and quantitative text analysis.

At the end of the summer school participants are expected to be in a position to autonomously implement, with the aid of the routines utilized during the sessions, the theories and methodologies discussed during the course of the week. In particular, participants should be in able to identify the type of data required for their specific research topic; evaluate which methodology is more appropriate for the analysis in hand; and finally test the appropriateness and sensitivity of their estimated model and the robustness of the results obtained.

SUMMER SCHOOL CODE

I-SS14

DATE AND LOCATION

Florence, 14-17 September 2020
CISL Studium Center
Via Della Piazzola, 71
I-50123 Florence
http://www.centrostudi.cisl.it

TEXT ANALYSIS: A QUALITATIVE AND QUANTITATIVE APPROACH

TARGET AUDIENCE

The summer school is aimed at:

i) academic researchers, evaluators, policy advisers, social workers, educators and students working in economics, public health, sociology, psychology and political science;

ii) data mining and market research analysts based in the automotive, market research, logistics or transportation, telecommunications sectors, needing to analyse comments from surveys, blogs, websites, social media platforms and other textual format sources;

iii) insurance analysts looking to analyse and categorize claims from customers;

iv) researchers based in pharmaceutical companies and medical research laboratories required to analyse healthcare reports, notes from medical doctors, interviews and/or focus groups with patients.

SETTING THE SCENE

SESSION I: THREE APPROACHES TO TEXT ANALYSIS

- Qualitative Analysis
- Quantitative Content Analysis
- Text Mining

SESSION II: QDA MINER AND WORDSTAT – A BRIEF OVERVIEW QDA MINER

QDA MINER
1. Introduction and project management
2. Codebook management and manual coding
3. Security features and text retrieval tools
4. Coding Frequency and Retrieval
5. Code co-occurrence and case similarity analysis
6. Assessing relationship between coding and variables
7. Using the Report Manager and the Command Log
8. Performing teamwork
9. Miscellaneous Functions

WORDSTAT
1. Content Analysis or Text Mining
2. Analyzing words without dictionaries – a text mining approach
3. Content Analysis – Principles of dictionary construction
4. Importing and exporting data
5. Introduction to automatic document classification

QDA MINER

SESSION I: INTRODUCTION AND PROJECT MANAGEMENT
1. Introduction to CAQDAS using QDA Miner
   - The CASE x VARIABLE file structure
   - The Mixed-Method approach
2. Quick overview of the work environment
   - The four windows – CASE, VARIABLES, CODES, and DOCUMENT
   - The menu system
3. Creating of a new project
   - Creating a new project from a list of documents

TAX   SENATOR   ENERGY   TStat

- Creating a new project from an existing data file
- Creating an empty project / defining structure
- Using the document conversion wizard

4. Customizing and personalizing the project
   - The PROJECT | PROPERTIES dialog
   - The PROJECT | NOTES command

5. Manipulating variables
   - Adding a variable
   - Deleting a variable
   - Changing the variable data type
   - Recoding the values of a variable
   - Reordering variables
   - Changing variable properties

6. Manipulating cases

7. Add a new case
   - Deleting cases
   - Importing new documents in new cases
   - Changing the case grouping and description

SESSION II:
CODEBOOK MANAGEMENT
AND MANUAL CODING

1. Creating codes and managing the codebook
   - Creating codes and categories
   - Modifying an existing code
   - Delete existing codes
   - Moving codes in the codebook
   - Merging codes in the codebook
   - Splitting codes in the codebook
   - Importing an existing codebook

2. Manual coding of documents (versus autocoding)
   - The four basic methods for assigning codes to text segments:
     - Highlight text segment then drag a code
     - Highlight text segment then double-click a code
     - Highlight text segment then select code and button (toolbar)
     - Drag and drop a code over a paragraph (or a sentence – press ALT)
   - Assignment of multiple codes to the same segment (press CTRL)

3. Modifying existing coding
   - Working with code marks
   - Viewing coding information
   - Adding a comment to a coding
   - Remove a coding
   - Change the code assigned to a text segment
   - Resizing a segment
   - Consolidating codes
   - Searching and replacing codes
   - Hiding code marks
   - Highlighting coded segments

SESSION III:
SECURITY FEATURES
AND TEXT RETRIEVAL
TOOLS

1. Using backup features
   - Creating a permanent backup
   - Restoring a backup
   - Using the temporary session backup

2. Text retrieval tools (4)
   - Searching for text
   - Performing a simple text search

https://www.tstattraining.eu/training/summer-school-text-analysis/

TAX   SENATOR   ENERGY   TStat

- Performing a complex text search (using Boolean and wildcard
- Performing a thesaurus search
- Using the "search hits" table
- Performing manual coding and autocoding
- Saving to disk or printing the table

3. Retrieving sections in structured documents
4. Performing a query by example
    - Finding text similar to a sample text segment
    - Providing relevance feedback to improve search results
    - Finding text similar to specific coded segments
    - Performing a "fuzzy string matching"
5. Performing a keyword search
    - Assigning keywords to codes
    - Performing a keyword retrieval on internal codes
    - Performing a keyword retrieval on WordStat dictionary files

SESSION IV:
CODING FREQUENCY
AND RETRIEVAL

1. Coding frequency
    - Creating a frequency list of all codes
    - Creating a barchart or a pie chart on selected codes
    - Customizing the chart
2. Coding Retrieval
    - Performing a simple coding retrieval
    - Performing a complex search
    - Creating a text report
    - Creating a new project from
    - A shortcut for simple coding retrieval
3. Saving and Retrieving Queries
4. Retrieving a list of comments

SESSION V:
CODE CO-OCCURRENCE
AND CASE SIMILARITY ANALYSIS

1. Analyzing codes co-occurrences
    - Hierarchical clustering of codes
    - 2D and 3D multidimensional scaling plots
    - Using the Proximity plots
    - Assessing similarity of cases
2. Analyzing code sequences
    - Choosing codes and setting minimum / maximum distances
    - Using the Sequence matrix
    - Searching and coding specific sequences

SESSION VI:
ASSESSING RELATIONSHIP BETWEEN
CODING AND VARIABLES

1. Analyzing coding by variables
    - Crosstabulating coding frequency by variables
    - Setting the content and format of the table
    - Computing correlation or comparison statistics
    - Comparing frequencies using barcharts or line charts
    - Creating and interpreting 2D and 3D correspondence plots
    - Creating and interpreting heatmaps
2. A quick overview of graphic coding features

SESSION VII:
USING THE REPORT MANAGER
AND THE COMMAND LOG

1. Using the Report Manager
    - Accessing the Report Manager
    - The Report Manager interface

https://www.tstattraining.eu/training/summer-school-text-analysis/

TAX   SENATOR   ENERGY   TStat

- Appending tables, graphics and quotes
- Moving and organizing items using the table of content
- Editing existing items / adding comments
- Adding empty documents or folders and deleting existing items
- Importing documents, images or tables
- Searching and replacing text
- Exporting results to HTML, Word or RTF files

2. Using the Command Log
   - Introduction to the command log – Filtering log entries
   - Adding comments to log entries
   - Undoing previously performed operations
   - Repeating previously performed operations
   - Exporting the log table to disk

**SESSION VIII:
PERFORMING TEAMWORK**

1. Preparing projects for teamwork
2. Creating user accounts and setting privileges
3. Creating new accounts
4. Defining users access rights
5. Forcing users to log in
6. Creating duplicate copies of a project
7. Sending a project by email
8. Merging projects and assessing coding reliability
9. Merging two or more projects
   - Planning teamwork for assessing coding agreement
   - Adjusting colors of code marks
   - Computing coding agreement
   - The codebook and segmentation problems
   - Four levels of agreement
   - Presence or absence (0 or 1)
   - Frequency (0, 1, 2, etc.)
   - Coding importance (% of words)
   - Coding overlap
   - Correcting (or not) for chance agreement
   - Identifying disagreements

WORDSTAT

**SESSION IX:
BASIC WORD STATISTICS
AND TEXT MINING**

1. Content Analysis or Text Mining
2. Running WordStat from QDA Miner or Simstat
3. Analyzing words without dictionaries – a text mining approach
4. Data preparation – misspelling and control characters
5. Basic word frequency analysis
   - Application of text pre-processing methods
   - Exclusion list – use with care
   - Lemmatization and stemming – limits and benefits
   - Setting upper and lower frequency criteria
   - A few additional options
   - Numeric and other non-alphabetic characters Braces and square brackets
   - Random sampling
   - Using disk or memory as the working space
6. Identifying themes using word co-occurrence analysis
   - Clustering words and measuring their proximity
   - Clustering documents based on the words they contains
7. Correlation and comparison analysis based on word usage
   - Performing crosstabs and computing statistics
   - Comparing words among the sources (document or text variables)
   - Correspondence analysis and heatmaps

https://www.tstattraining.eu/training/summer-school-text-analysis/

SESSION X:
CONTENT ANALYSIS
PRINCIPLES OF DICTIONARY
CONSTRUCTION

1. Introduction to WordStat categorization dictionary
   - Dictionary structure and functions
   - Opening, saving, and creating categorization dictionaries
   - Creating manually categories of words and phrases
   - Principles of dictionary construction – Extracting features
   - Identification of technical terms and proper names (persons, places, products)
   - Identification of common misspellings
   - Extracting phrases
   - Creating an initial dictionary – Phrases technical terms and proper nouns words
   - Adding words manually
   - Adding words from tables Using the drag and drop editor
   - Organizing the dictionary (drag and drop)
2. Applying the dictionary
   - Setting different levels
   - Mixing dictionaries with words
3. Validating the dictionary
   - Finding words or phrases with improper meanings using the KWIC list
   - WordStat evaluation order – how to use this at your advantage
   - Disambiguation methods
   - Manual disambiguation Disambiguation using phrases Disambiguation using rules
4. Improving categorization dictionaries
   - Creating comprehensive dictionaries using the Suggest button
   - Assessing coverage using the keyword retrieval feature

SESSION XI:
ADVANCED FEATURES

1. Importing and exporting data
   - Exportation of frequency data

TAX   SENATOR   ENERGY   TStat

TEXT ANALYSIS: A QUALITATIVE AND QUANTITATIVE APPROACH

REGISTRATION FEES

Full-Time Students*: € 1080.00
Academic: € 1760.00
Commercial: € 2600.00

*To be eligible for student prices, participants must provide proof of their full-time student status for the current academic year. Residential costs for full time students are completely sponsored by TStat Training through our **Investing in Young Researchers Programme**. Participation is however restricted to a maximum of 3 students.

Fees are subject to VAT (applied at the current Italian rate of 22%). Under current EU fiscal regulations, VAT will not however applied to companies, Institutions or Universities providing a valid tax registration number.

Please note that a *non-refundable deposit* of €100.00 for full-time students and €250.00 for Academic and Commercial participants, is required to secure a place and is payable upon registration. The number of participants is limited to 15. Places will be allocated on a first come, first serve basis.

Course fees cover: i) teaching materials (copies of lecture slides, databases and Stata routines used during the summer school; ii) a temporary licence of Stata valid for 30 days from the day before the beginning of the school; iii) half board accommodation (breakfast, lunch and coffee breaks) in a single room at the CISL Studium Centre or equivalent (4 nights). Participants requiring accommodation the night of the final day of the school, are requested to contact us as soon as possible.

To maximize the usefulness of this summer school, we strongly recommend that participants bring their own laptops with them, to enable them to actively participate in the empirical sessions.

Individuals interested in attending this summer school must return their completed registration forms by email (training@tstat.eu) to TStat by the 24th August 2020.

Further details regarding our registration procedures, including our commercial terms and conditions, can be found at https://www.tstattraining.eu/training/summer-school-text-analysis/.

CONTACTS

Monica Gianni
TStat Training | Kleebergstraße, 8
D-60322 Frankfurt am Main

TStat S.r.l. | Via Rettangolo, 12-14
I-67039 Sulmona (AQ)
T. +39 0864 210101

training@tstat.eu
www.tstattraining.eu - www.tstat.eu