Text-Mining With iAnalyzer & R

Syllabus

REBO Skills Academy 2022/2023
Period 4 (2,5 ETCS)
Course code module (LEG.SA.11)
Neha Moopen (n.moopen@uu.nl)

1. Learning objectives

Text mining methods are quickly gaining popularity among researchers, including those from Law, Economics and Governance (LEG) disciplines. Text mining exploits sizeable collections of digitized texts for automatic analysis, using software such as R or Python. For example, newspaper articles have been to measure economic sentiments and digitized court records to analyze the evolution of jurisdiction.

In this course, we will get you up to speed with a simple workflow for data-driven research using digitized text collections, and learn to reflect on the pros and cons of using methods like these. We will start with selecting relevant texts in iAnalyzer and creating a dataset based on your own research question. After an introduction to R, you will gradually learn to analyze your dataset. This also allows you to develop general skills in R that are becoming ever more useful in the complex analysis of data. By the end of the course, you will be able to answer a simple research question based on your dataset, and to use an 'Open Science' approach to report on your workflow and the pros and cons of text-mining.

After completing this module, students will be able to:

- process substantial datasets containing textual information;
- select subsets of digitized text collections such as newspapers for analysis;
- import and harmonize textual data in R;
- visualize textual data using *n-grams* and relations between words;
- automatically identify themes and subjects within textual data (topic modeling);
- reflect on the benefits and drawbacks of using text-mining methods for research.

2. Structure of the course

The course will begin with a general introduction to text mining and a discussion of its relevance and applications to LEG disciplines. The meetings thereafter will include introductions to iAnalyzer and R and we will take you through the steps of a simple text mining pipeline using these tools. Through demonstrations and hands-on exercises, we will go from exploring corpora in iAnalyzer and creating a subset of a corpus (dataset) to importing that dataset in R and analyzing it with basic text mining techniques. Finally, we will touch upon how you can create a fully reproducible report of your workflow and results in line with open science practices.

During the course, we will focus on the Times Newspapers corpus* which covers a broad array of topics relevant to LEG disciplines. We will ask you to define a research question that can be applied to this corpus* and this will form the basis of your final assignment. You will already start building up your final assignment during the hands-on part of the meetings. In this way, you go through the entire pipeline based on a research question that you find to be interesting and relevant.

We will start working on steps of the pipeline during the meeting in which they are introduced and you will finish it in the time between meetings. Meetings may also require additional preparation, to ensure an effective use of the time in class. The last few course meetings can be utilized to revisit concepts where a refresher is required or simply to work on the final assignment.

Instructors will be available for support during the course meetings, as well as additional weekly Q&A sessions.

** it may be possible to work with another corpus from iAnalyzer, after consultation with the instructors*

3. Teachers and guest speakers

All instructors are part of the Research Data Management Support team at Utrecht University Library.

1. Neha Moopen, Research Data Manager (n.moopen@uu.nl, course coordinator and contactperson)
2.
3. Jacques Flores, Research Data Consultant (j.p.flores@uu.nl)

4. Andreas Franzke, Faculty Liaison for Science (a.w.franzke@uu.nl)

5. Lena Thole, Research Data Consultant (l.m.thole@uu.nl)

6. Stefano Rapisarda, Research Data Consultant (s.rapisarda@uu.nl)

4. Schedule

#	Time	Location	Subject	Preparation
1	Tuesday, 25th April, 17:00-19:00	Digital Humanities Lab, University Library	Introduction to Text-Mining	- Review recommended literature showing examples of text-mining in LEG disciplines. - Search for one additional article on text-mining in a LEG disciplines and bring a summary to the meeting.
2	Tuesday, 2nd May, 17:00-19:00	Digital Humanities Lab, University Library	Introduction to iAnalyzer	- Think about research questions/search queries that can be applied to the Times Newspapers data.
3	Tuesday, 9th May, 17:00-19:00	Digital Humanities Lab, University Library	Introduction to R	- Install R & RStudio - Install R Packages - Watch a video tour of RStudio
4	Tuesday, 16th May, 17:00-19:00	Digital Humanities Lab, University Library	Text-Mining with R	- Review/Complete exercises from *Introduction to R* exercises - Refine research questions based on the introduction from Day 1.
5	Tuesday, 23rd May, 17:00-19:00	Digital Humanities Lab, University Library	Generating Reproducible Reports with R Markdown	- Watch a video on R Markdown - Review/Complete exercises from *Text-Mining in R*.

6	Tuesday, 30th May, 17:00-19:00	Digital Humanities Lab, University Library	Refresher / Q&A / coworking	- Work on writing up the final report.
7	Tuesday, 6th June, 17:00-19:00	Digital Humanities Lab, University Library	Refresher / Q&A / coworking	- Work on writing up the final report.

5. Lecture overview and time investment

For this module 100% attendance is mandatory. Only under special circumstances and in consultation with the course coordinator an exception can be made. Even with a valid reason, if the student misses more than one meeting, they can no longer take part in assessment of the module. This is because modules that are part of the REBO Skills Academy are based on learning through experiencing.
For this reason, it is also not possible to join a meeting online.

Meeting 1

We will begin the course with presentation introducing computational text-analysis. The information covered will include:

- What is computational text-analysis?
- Why is it worth learning computational text-analysis?
- What kind of data is needed?
- What are the methods typically used in computational text-analysis?
  - *What are these texts about?* Word Frequency, Collocation, Topic Analysis, Significant Terms
  - *How are these texts connected?* Concordance, Network Analysis
  - *What emotions are expressed?* Sentiment Analysis
  - *What key names can I find?* Named Entity Recognition
  - *Which of these texts are similar?* Clustering, Supervised Machine Learning, Authorship Attribution

Literature & Resources:

LAW:

1. Dyevre, Arthur, Text-mining for Lawyers: How Machine Learning Techniques Can Advance our Understanding of Legal Discourse (November 5, 2021). Erasmus Law Review, Vol. 14, No. 1, 2021, Available at SSRN: https://ssrn.com/abstract=3957098
2. Wyner, A., Mochales-Palau, R., Moens, MF., Milward, D. (2010). Approaches to Text Mining Arguments from Legal Cases. In: Francesconi, E., Montemagni, S., Peters, W., Tiscornia, D. (eds) Semantic Processing of Legal Texts. Lecture Notes in Computer Science(), vol 6036. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-12837-0_4
3. Data Science for Lawyers: https://www.datascienceforlawyers.org/

ECONOMICS/FINANCE:

4. Siegel, M. (2018). Text Mining in Economics. In: Hoppe, T., Humm, B., Reibold, A. (eds) Semantic Applications. Springer Vieweg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-55433-3_5
5. Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as Data." *Journal of Economic Literature*, 57 (3): 535-74.DOI: https://doi.org/10.1257/jel.20181020

6. Gupta, A., Dengre, V., Kheruwala, H.A. *et al.* Comprehensive review of text-mining applications in finance. *Financ Innov* **6**, 39 (2020). https://doi.org/10.1186/s40854-020-00205-1

GOVERNANCE/POLICY:

7. Gyódi, K., Nawaro, Ł., Paliński, M. *et al.* Informing policy with text mining: technological change and social challenges. *Qual Quant* **57**, 933–954 (2023). https://doi.org/10.1007/s11135-022-01378-w
8. Abu-Shanab, E., & Harb, Y. (2019). E-government research insights: Text mining analysis. *Electronic Commerce Research and Applications*, *38*, 100892. https://doi.org/10.1016/j.elerap.2019.100892
9. 2020. "Big Data Analytics and Text Mining in Internet Governance Research: Computational Analysis of Transcripts from 12 Years of the Internet Governance Forum", Researching Internet Governance: **Methods, Frameworks, Futures**, Laura DeNardis, Derrick Cogburn, Nanette S. Levinson, Francesca Musiani

Meeting 2

In this meeting, we will introduce students to I-Analyzer. I-Analyzer is an online text and data mining application developed by the Digital Humanities Lab at Utrecht University. We will work on the Times Newspapers corpus in iAnalyzer.

We will then proceed with exploring how you can process the corpora in I-Analyzer. This will include:

- Searching and filtering the corpus, as well as visualizing the results.
- Creating subsets of the data (corpus) and exporting it for further analysis in R.

Meeting 2

The day of the course will be an introduction to R. This will include:

- R Syntax & Data Types
- Vectors in R
- Data Structures
- Missing Data
- Indexing Vectors & Lists
- Indexing a Data Frame

Note that we do not cover programming techniques such as if statements, functions, loops. The aim of this session is to familiarize students with R and present some basic data wrangling operations.

Meeting 3

The meeting is where we will dive into text-mining with R. This will include:

- importing and tidying textual data in R (tidy text format)
- sentiment analysis

- analyzing word and document frequency (tf-idf)
- calculating and visualizing relationships between words (n-grams and correlations)
- identifying themes and subjects within textual data (topic modeling)

Meeting 4

In this meeting, we will touch upon how to generate reproducible reports with R Markdown.

Students will have already been working in R Markdown documents for Day 2 & Day 3, but we will take a step further to add prose/text between the code and 'knit' or render the R Markdown file to pdf or HTML. This is a fully reproducible workflow which weaves together text, code, results into one output file.

The resulting text-mining report can be eventually be submitted for grading, along with the reflection report.

Meetings 6 & 7

These meetings are reserved for Q&A, refreshers on concepts if required, and coworking. Students can continue working on their reports during this time, with the instructors at hand for guidance.

6. Assignments and grading

Assignment

Students will already start building up their assignment during the course meetings. A full assignment will be a result of completing the exercises presented during each course meeting. If students are unable to finish the exercises during the course meeting itself, it is expected that they will attempt to complete it before the next meeting.

The criteria/checklist for a full assignment includes:

- Completion of all coding exercises from the *Introduction to R* session within the R Markdown script/file that will be provided by instructors. The completed script/file must be submitted.

- Completion of all coding exercises from the *Text-Mining in R* session within the R Markdown script/file that will be provided by instructors. The completed script/file must be submitted.

- A LEG-related research question applicable to a corpus from iAnalyzer.

- A dataset exported from I-Analyzer. The workflow to export this dataset (search strategy) must be documented in the final report as part of a methods section.

- A reproducible report addressing the research question selected by the student. The R Markdown script/file from the *Text-Mining in R* can be used as a basis for this report. In addition to the code and results, students will have to incorporate sections for: introduction/background, method, discussion, conclusion. The completed script/file must be submitted along with a rendered pdf version.

In addition to the course meetings, instructors will be available for Q&A and support during the Walk-In Hours for Data & Software Support offered by Research Data Management Support every Monday from 15:00 to 17:00.

Reflection assignment for REBO Skills Academy modules

During the past weeks, you have developed a new skill. You have done this as part of a group of students from different educational programmes: in different disciplines and both bachelor and master programmes. The final assessment of the module consists of two parts, that both need to be completed sufficiently. The first is an assignment in which you show the skill that you have acquired. This assignment is tailored to the module's skill. The second assignment is a reflection assignment where you link what you have learned about this skill to your own programme. This assignment is the same for all modules.

The core elements that should come to the fore in this assignment is how the skill relates to your own programme.

1. To do this, you start the assignment with a discussion of what you find to be the most important things that you have learned about the skill that you have developed. Please be sure to use the literature that you have read during the module to do this, and link this with the moments where you have actually practiced the skills. What did you experience? And what have you learned in these experiences?

2. Next, discuss how these most important lessons relate to what you have previously learned about the discipline of your education programme. In what way is the approach to questions similar or different? So: what are the most important differences between how you would approach a question or issue that arises in society from the discipline of your education programme, and how you would approach such a question based on the skill that you have developed? Think about styles of analysis, possible solutions, form vs. substance, etc.

3. And lastly, reflect on what might be the added value of this skill for professionals with your academic background?

Use about 500-700 words for your reflection (excluding references). References can be made in the style that you are used to in your own programme, but do not need to include specific pages except when direct citations are concerned. References should focus on the literature that is part of the module, so it is not necessary to reference all statements made about the discipline of your programme.

The assessment of your reflection will be based on whether you have done the three elements elaborated on above in an insightful and reflective manner, showing that you have a good understanding of the skill and the ability to link what you have learned in the module with what you have learned in the rest of your programme on a conceptual level.

The grading within this module will be either pass or fail. This means that both the mastery of the skill, as well as the reflection needs to be sufficient.

Assignments can be uploaded to a repository on the Utrecht University GitHub environment. This will make it easier for instructors to review. All written assignments can be additionally handed in through Blackboard; this will make sure the assignment is checked for plagiarism by Ouriginal (the plagiarism software). All the assignments handed in need to be authentic and written by the student themself.

The Education and Examination Regulation (Onderwijs- en examenregeling (OER)) of the bachelor program of the school of governance applies to this module.