Team Assignment 2: Building a Text Mining Pipeline

Objectives of this assignment

- Answer a research question using open data
  - Find and refine a simple, specific, and feasible research question that you can answer with at least one of the data sets gathered for Team Assignment 1 ("open data"). The links to these datasets (and readmes) are available here.
  - Understand the specifics of a data collection by reading the data documentation.
  - Analyze your research question using a simple research design (e.g., comparison of means, plots over time, regression analysis, logistic regression)
- Implement the text mining workflow described in Berger et al. (2020)[1]
  - Parse semi-structured JSON data into structured CSV files, using Python/Jupyter Notebook
  - Use Python's textblob package for conducting text analysis (see also the text mining tutorial in Jupyter Notebook on Canvas)
  - Use Python's pandas package to read in a CSV file back into Python, and add data to it (e.g., text mining metrics; cheat sheet here)
- Manage your project's infrastructure
  - Learn to use GitHub templates to manage your project infrastructure (https://github.com/hannesdatta/textmining-workflow)
  - Submit your work as a reproducible workflow in line with our tutorial (http://tilburgsciencehub.com/tutorial)

Prerequisites

- Please install all necessary software and packages (see our detailed installation instructions, and also the comments in the Notebooks we may provide to you).
- Familiarity with "common data operations"
- Follow our tutorials on…
  - …text mining using TextBlob (posted on Canvas)
  - …implementation of an efficient and reproducible text mining workflow (http://tilburgsciencehub.com/tutorial/)

Introduction

The academic community increasingly makes use of "open data" to research digital and social media. For example, Julian McAuley at UCSD has published his (web-scraped) Amazon product and review data at http://jmcauley.ucsd.edu/data/amazon/, which inspired a wealth of research on online

[1] Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., & Schweidel, D. A. (2020). Uniting the tribes: Using text for marketing insight. *Journal of Marketing*, 84(1), 1-25.

ratings and recommender systems. Similarly, we (your RSM profs) are structuring our data to release them along with our papers once they become published.

We can all benefit from each other's work. And while some data may just have been collected as (relatively minor) covariates for a regression analysis (like this open data set from Hannes), that data may become a key dependent variable in somebody else's research (though that hasn't happened yet – hopefully, there will be a link here in the future…).

The steps in this assignment are:

1) Define a research question
2) Answer the research question using available data from Team Assignment 1 (can be ANY data set, not necessarily the one you collected with your team)
3) Prepare a two-stage workflow
    - Data preparation [needs to be fully automized and reproducible]
        - Download the data set from our course server,
        - Parse the data,
        - Enrich the data using text mining,
        - Wrangle the data, if needed, and
        - Write a data description for the CSV data set which is ready for analysis now (describing variables and their operationalizations)
    - Analyze the data to answer your research question
        - The default is to implement an analysis using RMarkdown (like in the template). You can also use Python to conduct your analysis. We'll collect a bunch of additional links for you to explore additional functionality in R and Python.
        - If you feel that working in R is beyond the scope for your team, that's fine. In that case, you have the option to
            - use your *final data set from the data-preparation pipeline stage* (recall, that stage needs to be fully automated), and
            - analyze the data using some visualization/infographic tools, Excel, PowerBI, or any other software you may be comfortable with – as long as it can answer your RQ.

Tips

- Research question and data set choice
    - Although the question may be different, please incorporate our suggestions from the quiz & the feedback lecture for module 1!
    - Make your research question *simple, specific,* and *feasible*: simple to understand, specific enough to measure, and feasible to answer given your data and method skills.
    - You do not have to use your own data set, but you can use any data set collected for Team Assignment 1
    - Look for inspiration in the readme files for potential research questions/contributions (e.g., sections "motivation" and "use"), before downloading the data
    - Pick "rich" data sets – 200 tweets will probably be not providing you with sufficient data to conduct text analysis
    - Check for data quality (e.g., were there any problems when collecting the data?)

- Prototype your idea
    - Manually inspect a few tweets first to get an idea of whether answering your research question is feasible – if not, adjust your RQ!
    - Check whether students have included initial parsing scripts with their data – use those to investigate the data
- Clone our template from GitHub.com and implement a workflow
    - We have developed a template for you to start from! Fork this to your GitHub repository in developing your code and putting in your data. We explain this in detail in our tutorial at http://tilburgsciencehub.com/tutorial.
    - The template "works" but is very rudimentary – so try running it first and fix any issues you may encounter!
    - Adjust the template to meet your needs (e.g., other text mining metrics)
    - Test your workflow on different computers and operating systems (Mac, Windows) to ensure stability and portability.
- Analyze the data
    - Please try to stay with plots, comparing means by groups, or by running linear regression or logistic regression analysis. The main goal is to keep the analysis simple and clean.
- Conclusion
    - Summarize your research question, results of your analysis, and provide implications for relevant stakeholders (also check module 1 for our discussion on the various stakeholders that you could address).
- Please check the grading rubric for the implications of your choice for grading.

Submission

- Submit your workflow (i.e., your directory/file structure)
    - One zip file per team; named teamXX_assignment2.zip
        - Inside the zip, have only the files required to run your workflow.
        - Check the template again – source code and a readme file are sufficient to reproduce your work.
        - Email the zip file to rsm@tilburguniversity.edu exclusively using https://filesender.surf.nl. We do not accept submissions via Google Drive/WeTransfer, etc.
    - Alternatively, you can host your files on a GitHub repository [bonus; see rubric]
        - In that case, make sure to version only the files that need to be versioned (source code, readme; NO generated files, and NO raw data).
        - Email the repository URL to rsm@uvt.nl – do not change your repository before you receive your final grade.

- Submit a report
    - One document, rendered as a PDF – upload to Canvas
        - If you've worked in RMarkdown, please PDF your HTML file.
        - If you've worked in another tool, please produce a PDF with an overview of your analyses (e.g., you can use screenshots, etc.)
        - Keep it concise but provide sufficient detail to understand your work; 4-10 pages, excluding references, but including tables/figures. Put your team

number in the header of the document ("Team X"). Number your pages. Do not make use of a title page.

- Sections in your report
  - Research question
    - Name the research question
    - Add a brief motivation
  - Data collection
    - Briefly describe the data set and how you prepared it (e.g., what was parsed, what text mining metric used/added & why)
    - Any other data sets used? Briefly describe those, too.
  - Analysis
    - Short description on how the data was precleaned for analysis (e.g., if aggregation, merging or any other common data operations were used)
    - Short description of your analysis
    - Tables and plots with results
      - Each table or plot is numbered, has a title, and proper axis labels/column headers.
  - Conclusion
    - A brief summary of your RQ and your answers

Opportunity for Feedback / Q&A

- We will use the feedback livestream for module 4 and 5 to discuss the various issues you may encounter when working on this assignment.
- Quick questions can also be answered in our WhatsApp chat.

Good luck! / Succes!

Grading rubric

You can use the evaluation rubric below to understand how you are evaluated in this group assignment, and how we judge an element to be e.g., insufficient, adequate, or excellent. Further, you can use it to get detailed feedback on your performance once results are known and learn about improvements you could make. Your final grade is computed as the sum of all points.

	Ratings			
	Exceeds Expectations (10)	Meets Expectations (6.5)	Below Expectations (3)	Points
1. Research question	The RQ is **simple, specific, and feasible.** The formulation of this question is understandable, straightforward, and demarcates a specific topic. The RQ is feasible (can be answered in a valid way by the data that is available). Further, the RQ is highly attractive and important.	The RQ is sufficiently **simple, specific, and feasible.** The formulation is understandable, but could have been formulated in a more straightforward and better demarcated way. It is feasible to answer the suggested RQ with the available data in a broad sense, while some uncertainties remain.	The RQ is not sufficiently **simple, specific, and feasible.** This means that (for instance) the RQ is not formulated in an understandable way, is too vague or too broad, or cannot be answered with the data that is available.	1.5
2. Data preparation (workflow)	Automation is excellent. **Everything runs automatically**, without any errors, by typing 'make'. All directories are created correctly, and the workflow assigns all files to the **correct directories** (e.g., temp, output, audit – if required; directory structure).	Automation is adequate. **Make runs with some minor errors** (e.g., some directories are not created automatically), but can be fixed by minor corrections. Generally, workflow assigns files in the **correct directories, potentially with a few exceptions** (directory structure).	Automation was not successful. **Make does not run at all, contains obvious mistakes** or is not fixable by making a few corrections. Workflow does not assign files in the correct directories, **violating the directory structure**.	2.5
3. Data preparation (parsing and text mining)	Parsing of tweets is **complete** (i.e., either all or a motivated subsection of the tweets are parsed). The data is **enriched with relevant text mining metrics** that go **beyond merely adding a sentiment score at the tweet level** (i.e., there is some **creativity and uniqueness** with regard to the measurement). **No irrelevant metrics** to answer your RQ remain in the script. The code is written in an **efficient way** (e.g., no superfluous use of error handling, all variable and file names clear).	Parsing of tweets is **complete** (i.e., either all or a motivated subsection of the tweets are parsed). The data is enriched to some extent with **relevant text mining metrics** to answer the research question, but **several unnecessary metrics remain in the script**. Alternatively, the script **may be partly inefficient** (e.g. superfluous use of error handling, inaccessible variable or file names).	Parsing of tweets is **incomplete** (i.e., the script crashes, or only parses an (unmotivated) subsection of relevant tweets). The data is **not enriched with relevant text mining metrics** to answer the research question, beyond what was in the template. Alternatively, the metrics are **irrelevant** to answering the research question. **Many inefficiencies are still present in the code** (e.g., superfluous use of error handling, inaccessible variable or file names).	2

4. Analysis (workflow)	Automation is excellent. **Everything runs automatically**, without any errors, by typing 'make'. All directories are created correctly, and the workflow assigns all files to the **correct directories** (e.g., temp, output, audit – if required; directory structure). Common data operations (e.g., aggregating over time or other variables, merging data, etc.) have been applied to a significant extent, and are correct to prepare the data for analysis.	Automation is adequate. **Make runs with some minor errors** (e.g., some directories are not created automatically), but can be fixed by minor corrections. Generally, workflow assigns files in the **correct directories, potentially with a few exceptions** (directory structure). Common data operations (e.g., aggregating over time or other variables, merging data, etc.) have generally been applied correctly, but minor inaccuracies may remain. **Alternatively**, if students chose to submit an analysis which is *not automated,* it needs to adhere to the general standards of conducting statistical analysis. Further, common data operations to "preclean" the data are described well.	Automation was not successful. **Make does not run at all, contains obvious mistakes** or is not fixable by making a few corrections. Workflow does not assign files in the correct directories, **violating the directory structure**. **Alternatively**, if students chose to submit an analysis which is *not automated*, it may not adhere to the general standards of conducting statistical analysis. Alternatively, common data operations to further "preclean" the data are not documented well.	1
5. Analysis (adequacy to answer RQ)	**Statistics** in tables and figures are provided to answer the RQ, **which are correct, relevant, and complete**. The statistics provide **novel and/or interesting insights**, are presented **logically in an accessible layout**, and are clearly labeled. The analysis methods are **executed well**. The description is **well written**.	**Statistics** in tables and figures are provided such that the RQ can **largely be answered**, but **not all statistics may be relevant** or **fully correct**. For instance, graphs or tables that hold no additional value are included, or insights may not be complete. Alternatively, statistics are correct and complete but **presentation/layout is not accessible**. The description is generally understandable.	**Statistics** in tables and figures are **not provided** in a **sufficient way** to be able to answer the RQ, or are provided but **incorrectly calculated**. As such, **vital information is missing to answer the RQ**. In addition, layout may be very poor (i.e. layout is not accessible, not informative or not clearly labeled). The description is highly unclear.	2
6. Conclusion and overall report	The interpretation of the analyses is correct, and matches the provided statistics. This conclusion offers a solid, data-driven answer to the RQ, leading to **novel and interesting insights**. The PDF report adheres to the guidelines.	The interpretation of the analyses is largely correct (**while minor mistakes are made**), and matches the provided statistics. This conclusion offers a **sufficient answer** to the RQ, that is based on data. **Minor interpretation mistakes occur**, or details are overlooked which would	The interpretation is wrong (resulting in an **incorrect answer** to the RQ), does not match the provided statistics (as such, not data-driven), does not offer any useful insights. The PDF report does not adhere to the guidelines.	1

		add to the answer to the RQ or insights in general. The PDF report adheres to the guidelines.		
7. Bonus – submission via GitHub	**Submitted as a GitHub link via email.** The project builds without errors. Submission only contains the necessary files to build the project (i.e., it excludes any generated or raw data files).	Submitted as a zip file via filesender.surf.nl; the submission **only contains the necessary files to build the project** (i.e., it excludes any generated or raw data files).	Submitted as a zip file via filesender.surf.nl; submission **contains unnecessary files** that are not central to reproducing the workflow (i.e., raw data, or generated files).	1