SIT220/731 2024.T1: Task 8HD

Data Cleansing and Text Analysis Challenge

14th April 2024

Contents

1 Introduction

> Tasks 5–8 are not obligatory; you can submit them in any order (or decide not to tackle them at all). C/D/HD is merely a subjective estimate of their difficulty level. For each task that you successfully complete, you score 10 points (and for those that are not 100% correct, no points will be given).

The discussion due day with the teaching team in OnTrack is **Week 11 (Thursday)**. After this due day, the teaching team will not provide any feedback for your work in OnTrack. The submission due day is **Week 11 (Sunday)**. Start tackling this task as early as possible. If we find your first solution incomplete or otherwise incorrect, you will still be able to amend it based on the generous feedback we will give you (allow 3–5 working days). In case of any problems/questions, do hot hesitate to attend our on-campus/online classes or use the Discussion Board on the unit site.

Submitting after the aforementioned due date might incur a late penalty. The **cut-off date is Week 12 (Friday)**. There will be **no extensions** and no solutions will be accepted thereafter. At that time, if your submission is not 100% complete, it will be marked as FAIL, without the possibility of correcting and resubmitting. To ensure a fair environment for all, we are always very strict about deadlines.

2 Task

First, research/gather the data:

1. Choose one StackExchange site dealing with topics that you find interesting; see https://stackexchange.com/sites?view=list#traffic for a list. The site cannot be too small, but also avoid selecting any of the largest ones (especially *StackOverflow*, *Mathematics*) unless you *really* want to challenge yourself. As a rule of thumb, let's say that the site must have at least 10,000 questions *and* 10,000 answers.

2. Download the site's most recent data dump from https://archive.org/details/stackexchange/.
3. Read the description of all the data tables published at https://meta.stackexchange.com/questions/2677/.

Then, create a single Quarto .qmd file[1] that you will be rendering to a PDF report (how to do that you will have to learn yourself – this is part of this HD-level task), where you perform what follows.

1. Convert all the data tables (Badges, Comments, PostHistory, PostLinks, Posts, Tags, Users, Votes) from XML to CSV, using custom code that you write yourself. Ideally, you should write a Python function that takes a single input file name (.xml) and output file name (.csv) and performs the conversion of a single dataset.
2. Load the CSV files as **pandas** data frames.
3. Create at least five nontrivial data visualisations and/or tables, at least three of which are based on the extraction of information from text (e.g., tags, keywords, locations, etc.). You must demonstrate that you have learned how to write your own regular expressions (regexes).
4. Draw insightful and interesting conclusions. Do not forget to reflect on the potential data privacy and ethics issues that arise during the data analysis process.

   *This HD-level task is purposely under-defined – you will not be told precisely what to do. Your aim is to generate some **interesting** insights into data featuring lots of textual information.*

In the course of the report preparation, you should apply a wide range of data frame wrangling and text processing techniques. In particular, you must demonstrate that you mastered *regular expressions*.

Do not use pie charts (as we discussed during the lecture). Go beyond the basic plots that we have covered in this course. Draw at least one map (e.g., of the world) and a word cloud.

3 Additional Tasks for Postgraduate (SIT731) Students (*)

There are no specific additional tasks, because the whole exercise has an open-ended formulation.

4 Artefacts

Make sure that your notebook has a **readable structure**; in particular, that it is divided into sections. Use rich Markdown formatting (text in dedicated Markdown chunks – not just Python comments).

Do not include the questions/tasks from the task specification. Your notebook should read nicely and smoothly – like a report from data analysis that you designed yourself. Make the flow read natural (e.g., *First, let us load the data on… Then, let us determine… etc.*). Imagine it is a piece of work that you would like to show to your manager or clients — you certainly want to make a good impression. Check your spelling and grammar. Also, use formal language.

At the start of the notebook, you need to provide: the **title** of the report (e.g., *Task 42: How Much I Love This Unit*), your **name**, **student number**, **email address**, and whether you are an **undergraduate (SIT220) or postgraduate (SIT731)** student.

Then, add 1–2 introductory paragraphs (an introduction/abstract – what the task is about).

Before each nontrivial code chunk, briefly **explain** what its purpose is. After each code chunk, **summarise and discuss the obtained results** (in a few sentences).

[1] OnTrack does not accept files with the .qmd extension. Before submitting the file, rename it so that it has the .md extension.

Conclude the report with 1–2 paragraphs (summary/discussion/possible extensions of the analysis etc.).

Limitations of the OnTrack ipynb-to-pdf renderer:

Ensure that your report as seen in OnTrack is aesthetic (see *Download submision PDF* after uploading the .ipynb file). The OnTrack ipynb-to-pdf renderer is imperfect. We work with what we have. Here are the most common Markdown-related errors.

- Do not include any externally loaded images (via the `![label](href)` Markdown command), for they lead to upload errors.
- Do not input HTML code in Markdown.
- Make sure you leave one blank line before and after each paragraph and bullet list. Do not use backslashes at the end of the line.
- Currently, also *LaTeXformulae* and Markdown tables are not recognised. However, they do not lead to any errors.

Checklist:

1. Header, introduction, conclusion (Markdown chunks).
2. Text divided into sections, all major code chunks commented and discussed in your own words (Markdown chunks).
3. Every subtask addressed/solved. In particular, all reference results that are part of the task specification have been reproduced (plots, computed aggregates, etc.).
4. The report is readable and neat. In particular:
    - all code lines are visible in their entirety (they are not too long),
    - code chunks use consecutive numbering (select *Kernel - Restart and Run All* from the Jupyter menu),
    - rich Markdown formatting is used (`# Section Title`, `* bullet list`, `1. enumerated list`, `| table |`, `*italic*`, etc.),
    - the printing of unnecessary/intermediate objects is minimised (focus on reporting the results specifically requested in the task specification).

Submissions which do not *fully* (100%) conform to the task specification *on* the cut-off date will be marked as FAIL.

Good luck!

5 Intended Learning Outcomes

ULO	Is Related?
ULO1 (Data Processing/Wrangling)	YES
ULO2 (Data Discovery/Extraction)	YES
ULO3 (Requirement Analysis/Data Sources)	YES
ULO4 (Exploratory Data Analysis)	YES
ULO5 (Data Privacy and Ethics)	YES