SIT220/731 2022.T3: Task 8HD

Data Cleansing and Text Analysis Challenge

Last updated: 2022-10-30

Contents

1 Task

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/.

Create a single Jupyter/IPython notebook (see the *Artefacts* section below for all the requirements), 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.).

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.

The PDF version of the report must be at least 10 pages long (not including the data conversion/import part). Make it aesthetic and interesting to read.

> *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.

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

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

3 Artefacts

The solution to the task must be included in a single Jupyter/IPython notebook (an .ipynb file) running against a Python 3 kernel.

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.

Make sure that your notebook has a **readable structure**; in particular, that it is divided into sections. Use rich Markdown formatting.

Imagine it is a report 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.

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

Submit one file via OnTrack:

1. the version of the Jupyter/IPython notebook converted to a PDF file (e.g., via *File → Export Notebook As → PDF* or convert to HTML and from that to PDF with your web browser; any method will do).

You do not need to submit the .ipynb file via OnTrack, but you must store it for further reference – a marking tutor might ask for it later, e.g., at the end of the trimester.

4 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