ITECH2303 Data Analytics

Assignment 1: Data wrangling

Overview

This assignment will explore issues around data preparation/manipulation and simple visualisation. Data from different sources will be required to be merged and several transformations applied, for instance dealing with missing values, deriving new features. Simple visualisation techniques will be applied on the data (for instance line charts, density plots). The assignment will be implemented through Jupyter notebooks (or equivalent). Submission will include the Jupyter notebooks (or equivalent) and the data sources used (or links/URLs to sources).

Domain

You will use the “Amazon product data” compiled by Julian McAuley. This dataset contains product reviews and metadata from Amazon, including 142.8 million reviews spanning May 1996 - July 2014. This dataset includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs).

Timelines and Expectations

Percentage Value of Task: 30%
Learning Outcomes Assessed: S1, S2, A1, A2
Due: Week 9 11.59 Friday 13th May

ITECH2303 Data Analytics

Assessment Details

Data preparation [10 marks]

You are to download any 5 of the datasets available from here:
http://jmcauley.ucsd.edu/data/amazon/index_2014.html
Scroll down to the "*Per-category files*" which are for individual product categories, which have already had duplicate item reviews removed. Please be aware that some of the datasets are very large, so you are advised to only download those that contain less than 300,000 reviews.
Load the data into a Jupyter Notebook and combine the 5 datasets into appropriate data structures (for instance *pandas* dataframe), with an identifier attribute that indicates their source (dataset identifier).

Initial analysis [20 marks]

You will have already realised that the format of the data is JSON when you initially examined it to load it into your program - you will need to make decisions about which features to include in your dataframe, and how to deal with missing values.
For instance, by putting the most frequently occurring features into your dataframe and excluding some others.
Pay attention to whether there are products that occur in several of the datasets.
Which are the most frequently occurring features and how will you represent them in your dataframe.
Use appropriate *pandas* functions to initially analyse the data, for instance descriptive statistics of each attribute.

GroupBy analysis [20 marks]

Use the *GroupBy* function in *pandas* to analyse the data, for instance against the average reviews for the various products.
The field "*asin*" is the ID of the product (e.g. 0000013714)
Implement various aggregate functions that will provide interesting insights into the data.

Linechart display [10 marks]

Find the products that have the most reviews – choose the top 5.
Plot the Star rating against time for each of these, and display in a single linechart.

ITECH2303 Data Analytics

Histogram display [20 marks]

Draw a histogram for the 5 datasets based on the date fields.
You will need to chunk the date fields into periods of time like weeks or months.
Combine these into a single stacked histogram.
Use whichever you prefer from either *matplotlib* (matplotlib.pyplot.hist), *pandas* (pandas.DataFrame.plot) or *seaborn* (seaborn.histplot).

Some URL's you might find useful:

- https://matplotlib.org/3.1.1/gallery/statistics/histogram_multihist.html
- https://stackoverflow.com/questions/18449602/matplotlib-creating-stacked-histogram-from-three-unequal-length-arrays
- https://towardsdatascience.com/histograms-and-density-plots-in-python-f6bda88f5ac0
- https://www.forbes.com/sites/naomirobbins/2012/01/04/a-histogram-is-not-a-bar-chart/?sh=57ac8be76d77
- https://seaborn.pydata.org/examples/histogram_stacked.html

Sentiment analysis [10 marks]

Discuss how you would approach the subject of *sentiment analysis* to make use of the text review section.
Where possible, illustrate your answer with code that works using your data.

Web scraping [10 marks]

Discuss methods of *web scraping* the product details from the URL's provided in the "asin" or ID of the product. How could you use this additional data to augment your analysis?
Where possible, illustrate your answer with code that works using your data.

ITECH2303 Data Analytics

Documentation requirements

1. Prepare a Jupyter Notebook which contains the following:

   a). Details of the data files used

   b). Code for all the tasks you have attempted

   c). Any extensive documentation provided as Mark Down text, for instance explanations of algorithms or functions used.

   d). Statement of any resources used as Mark Down text. These includes full disclosure of assistance from all sources including tutors and other students. Full APA referencing of any resources used.

2. Save your Jupyter Notebook as an ipynb file. You need to also export your results as a PDF file. Zip these files together in a single archive:

   Assignment file name:
   ITECH2303_Assignment1_Report_*yourname_studentID.zip*

3. Upload your Zip file through Moodle.

Submission

The assignment is to be submitted via the Assignment submission box in Moodle.
This can be found in the Assessments section of the course Moodle shell.

Feedback

Feedback and marks will be provided in Moodle. Marks will also be available in FDL Marks.

Plagiarism

Plagiarism is the presentation of the expressed thought or work of another person as though it is one's own without properly acknowledging that person. You must not allow other students to copy your work and must take care to safeguard against this happening. More information about the plagiarism policy and procedure for the university can be found at http://federation.edu.au/students/learning-and-study/online-help-with/plagiarism
Please refer to the *Course Description* for information regarding late assignments, extensions, and special consideration. A reminder all academic regulations can be accessed via the university's website, see: http://federation.edu.au/staff/governance/legal/feduni-legislation

ITECH2303 Data Analytics

Marking Criteria/Rubric

Task	Notes	Marks
Data preparation		/10
Initial analysis		/20
GroupBy analysis		/20
Linechart display		/10
Histogram display		/20
Sentiment analysis		/10
Web scraping		/10
	Total:	**/100**