Postgraduate coursework, Information School

INF6032 Big Data Analytics

1 Introduction

The assessment for "INF6032 Big Data Analytics" consists of a piece of **individual** coursework to assess your ability 1) to implement a solution which performs some simple Big Data Analytics, 2) describe and select the most appropriate big data infrastructure solutions, and 3) discuss the meaning of the results with respect to your client.

You should **write a 3,000 word structured report** (see Section 4) that presents and explains your solutions, justifies your choice of techniques and discusses the implications of your answers along with exploring any further investigations.

This assessment is worth 100% of the overall module mark for INF6032. A pass mark of 50 is required to pass the module. A complete submission is formed of two parts: your report, via Turnitin, and your code, via Blackboard (see Section 5 for more details). Both portions, code and report, need to be your own work. See Section 6 for more general information about Coursework Submission Requirements within the Information School.

Submission deadline: Wednesday, 31st May 2023.

2 The data

You will be using two datasets in this assignment. The first is a dataset of job descriptions (JDs) – a CSV file where each row contains a JD identifier, followed by the skills extracted from the JD with each skill appearing in a separate column. (Note that in order to achieve a fixed number of columns, this means that for a large number of JDs, there are many empty columns in the file.) The file does not have a header. The solutions you submit should be based on the `skill2vec_50K.csv.gz` file, available from `https://github.com/duyet/`

`skill2vec-dataset` – you can either download this file to your local system and manually upload to Databricks or your AWS storage, or access it directly from a Databricks notebook via the file's "permalink". When you save this to Databricks, you need to ensure that the full filepath is: `/FileStore/tables/skill2vec_50K.csv.gz`

The second file is part of the O*NET database. [1] This database contains descriptions of jobs and the skills they use, unifying these in various hierarchically ordered categories, such as "Hot Technologies" or "In Demand" skills. You will use the "Technology Skills" file, available from `https://www.onetcenter.org/dl_files/database/db_27_2_text/Technology%20Skills.txt` Note that only this version (27.2) will give you the expected results.

The O*NET technology skills file has a header, and can be treated as a tab separated file with Windows style endings ('\r\n') on each line. The column "Example" can be thought of as corresponding to a skill as listed in the first dataset. The full filepath on Databricks for this file should be: `/FileStore/tables/Technology_Skills.txt`

3 Questions

Implement a big data solution (i.e. you must use the tools taught in this course, pySpark, dataframes or sql, or AWS or a combination thereof). Note that solutions based on pandas, excel or similar non parallel approaches will NOT be awarded any implementation marks.

1. Programmatically confirm that the number of job descriptions is as expected (i.e. that there are 50,000 **distinct** job descriptions in the skill2vec_50K dataset).
2. Work out the frequencies with which distinct skills are mentioned in job descriptions, and present the top 10 (in order of decreasing frequency) skills in your report, alongside the frequency of each across the entire dataset. I.e. if your dataset consists of the following:

   JD1,python,databricks,amazon web services aws software,,
   JD2,python,deep learning,,,
   JD3,machine learning,python,amazon web services aws software,,

   Your output would be:

[1] More details available at: `https://www.onetonline.org/`

Skill	Freq
python	3
amazon web services aws software	2
databricks	1
deep learning	1
machine learning	1

3. Find the 5 most frequent numbers of skills in JDs across the dataset. I.e. given the example with JD1, JD2 and JD3 above, the expected result would be:

Num skills	Freq
3	2
2	1

   indicating that 2 JDs contain 3 skills while 1 JD contains only 2 skills.

4. So far, you've explored the dataset in its original form. Check how the distribution of the frequencies with which distinct skills are mentioned in JDs changes if you lower case all the skills. As in question 2, present the top 10 (in order of decreasing frequency) skills in your report.

5. To gain some additional information about the sought after skills, you'd like to join the (lower cased) skills from JDs with the skills listed in the Example column in the O*NET dataset (don't forget to lower case the example column!). Find the change in the number of skills before and after the join (i.e. report the number of original skills and the skills that are both in the JD dataset and the O*NET dataset – reporting two separate numbers).

6. The join you performed in Question 5 gives you access to the "Commodity Title" column. Find the 10 most frequent "Commodity Title"s across all the job descriptions. I.e. using the example from Question 2, the output should be:

Object or component oriented development software	3
Data base user interface and query software	2

   Note that more than one skill can map to the same "Commodity Title".

To allow you to check your code, the answers to the above questions will be provided to you alongside this assignment brief on two smaller JD subsets. If you use these as input, your code should give you the results supplied – however, it is the results on the 50,000 dataset that you should include in your report.

4 Report structure

A 3000 word report that documents your solution should be included with your submission. In this module, an abstract, background, or literature review are **not** required. The format of the report should be as follows:

1. Description of any setup required
2. Any data cleaning and preparation (including descriptions, justifications and screenshots of all code)
3. Problem answers
   (a) Question 1
      - Assumptions made
      - Implementation outline (description of main ideas in words and screenshot(s) of code)
      - Result on the submission (final) data – screenshots of the results must be included
      - Discussion of result, outline(s) and / or implementations of any further explorations the result suggests (implementation screenshots must be included in report)

   (b) Question 2
      - Assumptions made
      - Implementation outline (description of main ideas in words and screenshot(s) of code)
      - Result on the submission (final) data – screenshots of the results must be included
      - Discussion of result, outline(s) and / or implementations of any further explorations the result suggests (implementation screenshots must be included in report)

   (c) Same arrangement for remaining questions

5 Submission

Your submission consists of multiple parts:

1. A 3,000 word report to be submitted via Turnitin
2. Your code, to be submitted via Blackboard. This should be a single zip file containing the python notebook (i.e. the `.ipynb` file), **and** an `.html` export of the notebook which

has been executed (i.e. all results of your run are visible). If your solution is AWS based, you must ensure that the appendix of your report includes the steps taken (in the style provided in the Data Analytics Labs) and sufficient screenshots for us to follow all the steps of your solution.

6 Information School coursework submission requirements

It is the student's responsibility to ensure no aspect of their work is plagiarised or the result of other unfair means. The University's and Information School's Advice on unfair means can be found in your Student Handbook, available via `http://www.sheffield.ac.uk/is/current`

Your assignment has a word count limit. A deduction of 3 marks will be applied for coursework that is 10% or more above or below the word count as specified above or that does not state the word count.

It is your responsibility to ensure your coursework is correctly submitted before the deadline. It is highly recommended that you submit well before the deadline. Coursework submitted after 10am on the stated submission date will result in a deduction of 5% of the mark awarded for each working day after the submission date/time up to a maximum of 5 working days, where 'working day' includes Monday to Friday (excluding public holidays) and runs from 10am to 10am. Coursework submitted after the maximum period will receive zero marks.

Work submitted electronically, including through Turnitin, should be reviewed to ensure it appears as you intended. Before the submission deadline, you can submit coursework to Turnitin numerous times. Each submission will overwrite the previous submission. Only your most recent submission will be assessed. However, after the submission deadline, the coursework can only be submitted once.

Details about the submission of work via Turnitin can be found at `http://youtu.be/C_wO9vHHheo`

If you encounter any problems during the electronic submission of your coursework, you should immediately contact the module coordinator and one of the Information School Teaching Support Team on is-teaching-support@shef.ac.uk. This does not negate your responsibilities to submit your coursework on time and correctly.