ENGINEERING & COMPUTER SCIENCE ASSOCIATION
Concordia University

ENGLYMPICS CONCORDIA 2020

Programming Competition Statement

Presented by:

SnapX

Competition Background:

SnapX is an online platform to help accountants and businesses automate their accounting processes.

With SnapX, users can simply take a picture or upload a receipt (image or PDF) to store it in the cloud which eliminates the need to keep paper bills. The information on the receipt is then decoded and classified (taxes, supplier, payment method, purchase date, etc.) to be sent to the user's accounting software with the supporting document attached. SnapX then provides automatic bank reconciliation.

Decoding the information from each receipt is by far the most important and most complicated part of the SnapX process.

Your team has been tasked to create an algorithm that identifies suppliers on each receipt. To carry out the decoding process, you will be using OCR (Optical Character Recognition) and a list of approximately 1 million different registered companies in Canada and America to assure integrity of the data extracted from the receipt.

The provided files for testing during solution developpement:

- A set of 20 images of receipts
- The corresponding set of 20 OCR results in JSON
- Two supplier CSV files each containing half a million entries

Competition Challenge:

You will be given a folder(CompetitorPackage.rar) with a set of 20 images of test receipts (.jpg and .png). The images are for reference and for testing purposes only. CompetitorsPackage.rar will also contain a folder with the corresponding OCR results processed from the Google Vision API of each of the receipts. For example the j1.json file corresponds to the 1.jpg receipt image file.

In the folder you will also be given supplier list files (.csv) that contain all the possible suppliers the receipts could have.

The competition consists of automatically identifying the supplier (Selling company) on different forms of receipts and invoices. To ease the process, the 20 OCR JSON files (google vision) will already be provided for testing and final evaluation.

You must code an algorithm that will iterate through the 20 OCR JSON files in a specified folder which represent receipts. You must then identify the supplier from the OCR files using the supplierlist CSV files (1 & 2) for each of the OCR receipts.

All the suppliers can be found on the (shortened) list of Canadian and American suppliers provided.

The algorithm must return the following information about the supplier of each OCR JSON file:

- Supplier Name
- CSV file the Supplier is found in (1 or 2)
- Which line in the CSV file the supplier is found
- The SIC categories (found on the CSV) of the supplier.

The algorithm should be executable in a single command line which specifies the folder of JSON files to be analysed.

For example, say you create a command line application which uses the keyword 'analyze' followed by the path of the folder the command should be
'*analyze C:|Users|user|Desktop|Training Json*'

During the evaluation we will run your specified command with a different folder path containing 20 different OCR JSON result files than the ones you were given to test with.

To summarize, you are provided with a sample folder of 20 OCR JSON receipt files to test your algorithm with. The evaluation folder will contain 20 other OCR JSON receipt files that you have not seen but will be of the same format as the testing sample. Both the testing and evaluation samples suppliers are found on one of the two csv files provided. You must extract the Supplier information from all 20 OCR JSON files.

Deliverables

Deliverable 1:

At the halfway point of the competition, **at 2PM** all teams are expected to email competitions@ecaconcordia.ca , martin@snapx.com and alex.fl@snapx.com the programming language they will be using, and what stack needs to be installed on a WINDOWS computer to be able to execute their code entirely. A readme file should be attached with urls to all the required downloads for your solution. A penalty of 2% will be attributed for every 5 minutes late on the first deliverable.

Deliverable 2:

At the end of the competition, **at 6PM** all teams are expected to to email competitions@ecaconcordia.ca , martin@snapx.com and alex.fl@snapx.com a link to a github repo (invite KaylaCharky, MartinSpasov and Cyberunner23 to the repo). Please also include the powerpoint presentation you will be using for your presentation and a clear readme on how to run your solution. A penalty of 10% will be attributed for every 5 minutes late on the second deliverable.

Presentation

Solution Presentations

Competitors will have a maximum of seven (7) minutes to present their solutions. All team members must be present and participate in the presentation or be penalized by the judges. Judges then have a maximum of seven (7) minutes to ask questions. In order to ensure that all competitors cease to work on the case solutions once the design time has ended, the competitors cannot include any material in their oral presentation which is not included within their submitted written reports or presentation materials.

At the beginning of their presentation, one of the judges will run the competitors algorithm on their computer. At the end of the presentation the judge will inform the competitors of the amount of suppliers successfully identified and the time of execution of their algorithm. If the algorithm has not yet finished executing by the end of the question period. A score of 0 (zero) will be attributed for the performance criteria.

Evaluation Grid

Evaluation Criteria	Scoring
Performance*	**45%**
Amount of Suppliers successfully identified	30%
Speed of the algorithm	15%
Presentation	**25%**
Speaking ability of the participants (clarity, tone, speed)	15%
Visual aids (usefulness, clarity)	10%
Algorithm**	**30%**
Code Efficiency (Big-O) Code Reusability Code Readability Code Architecture/Design Code Modularity Code Testability Accuracy of the algorithm	30%

*Each supplier successfully identified is worth 1.5%. The points attributed to the speed of the algorithm are a multiplier on the amount of suppliers identified.
The fastest team will have 100% of the amount of points towards the speed criteria for a maximum of 15 points. The second fastest team 85%, third 70%, etc.

**The list of quality measures are there as suggested talking points for your presentation and what will be looked for in your code.