Project Proposal

Data-Aid – Budget Management and Record Keeping Tool

David Sheehy, 16440304, x16440304@student.ncirl.ie

BSc (Hons) in Computing

Data Analytics

23/09/2019

# Objectives (What you want to build)

Data-Aid will be an application that will allow a user to record, analyse, predict and retrieve information on their spending over any specified period of time. The user will be able to record the information of any invoice/receipt they receive. This information may include the total cost on the receipt, the amount for individual items, the company who issued the receipt, the date of the receipt and the type (e.g. health insurance invoice, weekly shopping receipt, email receipt from online shopping etc).

The Project will be built as a Python Web Application. Python will be used to build the analytical tools of the application. The project will utilise open source libraries such as OpenCV and Pytersseract, Numpy, Matplotlib and Panda. These libraries will allow the application to process the data of images taken by the user. The main challenge will be ensuring the accuracy of the data, which is why the application will prompt the user to confirm the data is correct when an image is taken.

The user will be able to record these images for their future reference. The application will be able to organise and sort the images based on the company and date of the receipt (e.g. Aviva Car Insurance May 2019). In future the user will then be allowed to grab any recorded images of receipts over any defined time period, (VHI Health invoices January 2016- December 2018).

The financial data extrapolated from these images will be used by the application to analyse a user's spending and provide custom dashboards for the user to view. These dashboards will be built using the Python libraries outlined above and will be a fully configurable by the user. The default will be a user’s spending over a 30-day period and the spending data will categorised by types (e.g. Food, Recreation, Health, Utilities) and these categories will be presented in a charts and graphs, illustrating the user’s spending over the last 30 days. The user may configure any desired time period and any categories to view their spending data. They may additionally configure what way they wish the defined data displayed (bar chart, pie-chart, time graphs etc).

The application may also identify if a user could save money on certain items. A user may spend 2.10 on milk in a Shop A. The application may identify that in shop B in the local area may be selling the same or similar item for 1.50 in Shop B. This data is gathered by the activity of other users over time and over different shops, vendors, providers etc. The data of an individual's receipts will be stored on a user’s local machine, the cost of individual items may be stored in the cloud, however. This will enable the comparison functionality as discussed above. The user will be informed of this in the terms of conditions and of the application, which will be GDPR complaint.

# Background (Why You want to build this)

(Max. 2 Pages)

This idea of this project was conceived as a potential solution to numerous problems. Current budget management applications currently available have numerous flaws which this project hopes to improve upon. It is often the case at this low level of personal budget management, that data is usually loss which significantly reduces and degrades the accuracy of existing applications in their analysis of spending trends as well as the ability to intelligently provide potential optimizations on a user’s spending. In this section the proposal will:

* outline and explore the challenges current tools of similar features face
* Identify how Data-aid will be a solution to the existing challenges outlined above
* Illustrate how the project can incorporate existing methods and ideas on budget management, image recognition and machine learning can provide an innovative application which can provide the best user experience as well as a robust application.

Paragrpah One: YNAB, Quicken/Mint

Learning Workshops, Secure Data, “Eye Candy” graphs charts, award winning, goal setting, budget collaboration, monthly reports, Web/App/Apple Watch Alexa

Paragraph Two: Revolut

Bank Cards, Use with Google pay etc, easy recording of card transacations and catagoristaion, money transfers inter accounts

Paragraph Three: Data-Aid

Add Receipt Recording, removes tedium of inputting manual values, Record keeping, image data stored locally so less security concern

Paragraph Four: Innovative Approach

Use of old ideas that work, incorporate the receipt image recognition into the project, machine learning on a recommendation system, incorporation of Web APIs input of data

# Technical Approach

The approach to this project will be as follows. The project will examine existing implementations of Optical Character Recognition (OCR) and Computer Vision. This is crucial to the success to the project. The project will aim to implement separate existing technologies as outlined in the technical details section below.

The initial research involved in the project will be the applications of OCR and the accuracy of the image to text functionality. It will be critical to consult academic journals and papers on the subject in order to attain the best overview of the existing challenges and limitations of OCR and Computer Vision software. Research into these technologies will need to include study of common applications of the software, such as self-driving cars etc. This will be crucial for the critical evaluation of the technology for use in the project.

In terms of the machine learning implementation, there will be extensive prior research. The research will focus on the characteristics of common data sets used by machine learning algorithms to train models. The project will need to identify the scope of receipt data required to train an effective, robust and accurate model which can make intelligent recommendations to the user. The research will need to reference recent and relevant academic papers on the training of models as well as the common applications of those models. The recommending software on eCommerce sites such as Amazon and eBay would be an interesting application to examine as there are similarities between those models and the desired recommendation model which this project intends to create.

The project will additionally need to examine the best hosting software for the application. This application will be a web application so the hosting and data storage requirements will be crucial to its success. It will be necessary to examine factors such as the storage capacity of a hosting package, the uptime of a package, the reliability as well as cost. There are numerous options available such as AWS suggested below, or Blacknight or other hosting packages. In order to choose the most suitable hosting environment, it will be necessary to gain a more thorough understanding of the project requirements.

# Special resources required

If applicable, e.g., books, hardware, etc.

The only special resources the project will utilize are a phone camera and a scanner for paper receipts.

# Project Plan

Gantt chart using Microsoft Project with details on implementation steps and timelines

# Technical Details

Implementation language and principal libraries

The business logic for the application will use several languages to implement. The section of the project that will examine receipt images, as well as receipt data from the API data inputted into the application. The data will be extracted as plain text.

The libraries used for this functionality will be a combination of OpenCV and Pytesseract. OpenCV is a computer vision python library which is an open source library which is optimised for computer vision tasks. Python-tesseract is a Python wrapper for Google's Optical Character Recognition software. This will allow the application to extract and transform the data from the text of the receipt/invoice such as:

* Total Amount of Receipt/Invoice
* Costs of Individual Items
* Business Name, Address and Contact Details
* The time and date of the Invoice/Receipt

Python is appropriate as it is an ideal programming language for the data analytics involved. The project will retain a record of the image of the invoice/receipt inputted by the user for their future reference. These images will be automatically categorised by the application based on factors such as the business type (e.g. Aviva Car Insurance as Car Insurance expense), as well as the date of the receipt issue. This will allow the user to query the application for invoices of certain types and timeframes. For example, the user may request all grocery shopping receipts in the last 28 days, or all utility invoices in the last 12 months.

The application will be able to use the extracted and categorised data to produce custom dashboards for the user to vie w their expenses over any given time period. This will be implemented via Python data visualisation libraries such as Matplotlib and ggplot2, which are libraries designed to visualise data in a way which is easy to understand for a user.

The recommendation system, which will implement a basic machine learning algorithm, will allow the application to create basic recommendations to the user on optimise spending. This will involve the use of the data extracted from the receipts such as the cost of an item/service and the location of the vendor of that item/service. The ML algorithm should be able to recommend to a user for example if they can save on going for a haircut in a barber’s/hairdresser's that is nearby their area (e.g. Dublin 1). The TensorFlow library will allow the application to build a basic machine learning model to create basic recommendations such as outlined above by using the extracted data from the invoices.

The user interface of this application will be implemented by creating a web application using HTML and CSS. The application will be hosted on a service such as Amazon Web Service (AWS) web hosting, which will mean that the application will be hosted on a reliable and secure server. This user interface will allow the user to use the business logic of the application developed in Python as outlined above.

# Evaluation

Describe how you will evaluate the system with real technical data using system tests, integration tests etc. In addition, where possible describe how you will evaluate the system with an **end user. (be careful here re Ethics etc)**

The testing of the software will be undertaken through a phased approach.

# Invention Disclosure Form (Optional – remove if not filled)

***Please fill in the following sections, if you think your idea is innovative***:

1. Title of Invention

2. Inventors

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | School/Research Institute | Affiliation with Institute (i.e. department, student, staff, visitor) | Address, contact phone no., e-mail | % Contribution to the Invention |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

3. Contribution to the Invention

Each contributor/potential inventor should write a paragraph relating to his/her contribution and include a signature and date at the end of the paragraph.

4. Description of Invention

(Please highlight the novelty/patentable aspect. Attach extra sheets if necessary including diagrams where appropriate). What is novel, the ‘inventive step’? For more information on patents, please look at <http://www.patentsoffice.ie/en/patents.aspx>

5. Why is this invention more advantageous than present technology?

What is its novel or unusual features? What problems does it solve? What are the problems associated with these technologies, products or processes? Explain how this invention overcomes these problems (*i.e*. what are its advantages).

6. What is the current stage of development / testing of the invention?

7. List the names of companies which you think would be interested in using, developing or marketing this invention

8. Funding Partner(s)

|  |  |
| --- | --- |
| Government Agency & Department |  |
| % Support |  |
| Contract/Grant No. |  |
| Contact Name |  |
| Phone No. |  |
| Address |  |

|  |  |
| --- | --- |
| Industry or other Sponsor |  |
| % Support |  |
| Contract/Grant No. |  |
| Contact Name |  |
| Phone No. |  |
| Address |  |

9. Where was the research carried out?

10. What is the potential commercial application of this invention?

11. Was there transfer of any materials/information to or from other institutions regarding this invention?

If so please give details and provide signed agreements where relevant.

12. Have any third parties any rights to this invention?

If yes, give names and addresses and a brief explanation of involvement.

13. Are there any existing or planned disclosures regarding this invention?

Please give details.

14. Has any patent application been made? Yes/No

If yes, give date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Application No.: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name of patent agent: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Please supply copy of specification.

15. Is a model or prototype available? Has the invention been demonstrated practically?

**I/we acknowledge that I/we have read, understood and agree with this form and the Institute’s *Intellectual Property and Procedures* and that all the information provided in this disclosure is complete and correct.**

**I/we shall take all reasonable precautions to protect the integrity and confidentiality of the IP in question.**

Inventor: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature Date

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature of student and date