

Contents

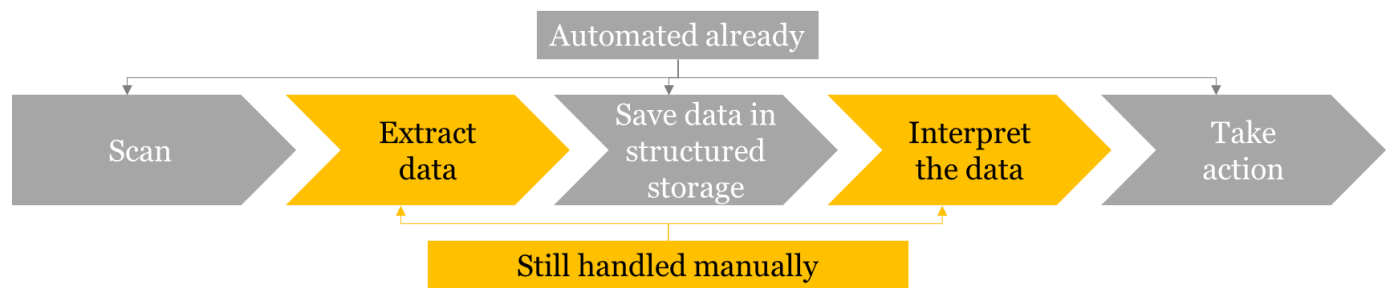
Problem statement	2
Problem – “Expenses Processing”.....	2
Guide for the Hackathon.....	3
Checklist (before the event).....	3
Who Am I?.....	3
Problems to address	4
Problem 1. Implementation of full process using existing tools.....	4
Problem 2. Design and architecture of full process mentioning existing tools	5
Problem 3. Data extraction (Computer Vision).....	6
Problem 4. Interpret extracted data (NLP+logic).....	7
Problem 5. Handwriting (Intelligent Character Recognition)	8
Problem 6. Scan cutter	9
Problem 7. Enhance existing OCR solutions	10
Problem 8. Dataset preparation	11
Problem 9. Market research (test existing solutions).....	12
Problem 10. Leverage Natural Language Processing on raw OCR data.....	13
Problem 11. Locate the receipt issuer	14
Problem 12. Multi-language receipt processing.	15
Problem 13. Microservices.....	16
Problem 14. UX for Expense Management process.....	17
Problem 15. Different idea	18
Prizes and evaluation.....	19
Tools and resources worth looking at in advance.....	20
Contacts.....	20

Problem statement

Problem – “Expenses Processing”

In many industries, companies reimburse business-related expenses (expenses incurred while conducting business on behalf of a company, e.g., during a business trip) to their employees. These include expenses incurred at restaurants, hotels, and transportation. However, to get the reimbursement, in many instances employees have to manually enter the information from an expense receipt into a company-mandated expense management system.

Once an expense is submitted, the finance department will review it. If the expense is within policy, it will be approved, otherwise it will be declined. In a large organization (imagine a heavy sales organization e.g., Pharma), reviewing receipts and approving them using simple rules is a full time job of multiple employees.



On the diagram - generic receipt processing work flow

There is no overarching market solution to this *problem*, which is broader than described above. With **yellow** we marked steps for which we could not find any industry-level tool. These are the steps that we would like to address during this Hackathon. If we count resources spent by employees and finance departments globally we won't be too far off if we estimated the dollar equivalent to be in billions. This makes it a very interesting problem!

For this Hackathon we have split the process described above into a *number of problems* (see the list below), solving any of which would allow us to get closer to the target solution. However, we realize this is not the full list and new problem statements are welcome.

Select a problem and 1-2 use-cases which would be interesting for you to work on throughout the day and try to put them together. It is better to complete one simple problem rather than leave a harder problem half-way done. Each problem is marked with team roles to support it, however, this is only our recommendation. The ultimate decision about the roles and problem to solve is yours.

Guide for the Hackathon

Checklist (before the event)

-Understand the main problem we are solving

-Think, which role you should pick for the day, do not worry if you have little experience in certain role – be ready to learn and motivation is more important

-Think of use-cases (see list of problems below) in background during the last days before the Hackathon

-You do not have to join with a team – we will facilitate strong teams formation for sole joiners

Who Am I?

We believe that both **Developer** and **Non-Developer** roles are required to address most of the challenges during this Hackathon.

The problem can be approached by solely developers/data-scientists, but for better results we recommend to have a mixed team. In fact linguists, accountants, auditors, developers, designers, researchers, consultants and many other specialists can contribute to the solution.

Data scientist

– focus on Computer Vision/NLP or other applications of ML.

Developer

– solution development, putting together components solving each problem, creating a functional prototype.

Designer

– thinking the solution through, creating a concept, not necessarily functional.

Entrepreneur

– coming up with new problems and ideas for solutions. Challenging designers, developers and data-scientists from business perspective.

Researcher

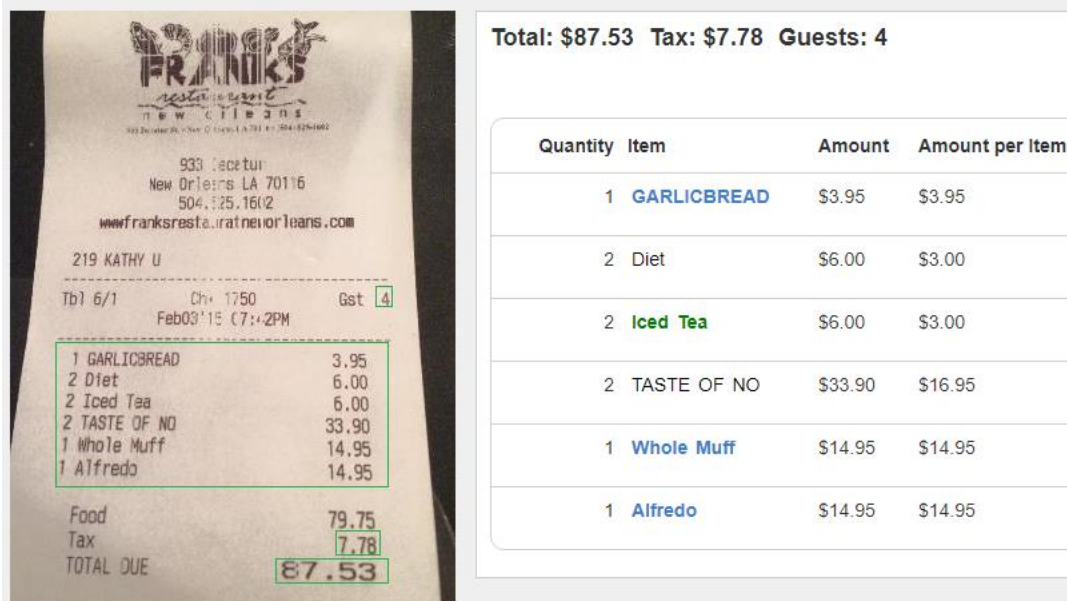
– cross-checking ideas and solutions with experts and internet. Testing the ideas, collecting data and supporting the development.

Problems to address

Problem 1. Implementation of full process using existing tools

You find a set of existing tools and build a solution which implements the process similar to stated above.

You demonstrate the solution in action to process receipts.



The image shows a receipt from Franks restaurant and a digital summary of the receipt data. The receipt is for a meal for 4 guests, totaling \$87.53. The items ordered are 1 Garlicbread, 2 Diet, 2 Iced Tea, 2 Taste of No, 1 Whole Muff, and 1 Alfredo. The total amount due is \$87.53, including a tax of \$7.78.

Quantity	Item	Amount	Amount per Item
1	GARLICBREAD	\$3.95	\$3.95
2	Diet	\$6.00	\$3.00
2	Iced Tea	\$6.00	\$3.00
2	TASTE OF NO	\$33.90	\$16.95
1	Whole Muff	\$14.95	\$14.95
1	Alfredo	\$14.95	\$14.95

Total: \$87.53 Tax: \$7.78 Guests: 4

Food 79.75
Tax 7.78
TOTAL DUE 87.53

Developer

Data scientist

Designer

Entrepreneur

Problem 2. Design and architecture of full process mentioning existing tools

You develop mockup design referencing existing tools, but do not implement the solution itself. You should explain why it will work very well. We will challenge it.

Designer

Entrepreneur

Researcher

Problem 3. Data extraction (Computer Vision)

In the end of the day you have a system which takes a scanned receipt and outputs a CSV file with receipt data, structured or unstructured. You explain your solution limitations and elaborate what tools did you reuse and what were your team's findings, what was something which you had to research.



row #	raw text
1	The Edison
2	Disney Springs
3	Orlando FL, 32830
4	375659 George C
5	Tbl 309/1 Chk 5551 Gst 5
6	Nov06'18 06:38PM
7	2 DNR Cruzan Light Mojito \$24.00 \$12.00
8	2 Dft Copper Tail \$16.00 \$8.00
9	1 Dft Buddha Pine \$8.00 \$8.00
10	2 Pineapple Juice \$10.00 \$5.00
11	1 Chips&Guac \$14.00 \$14.00
12	1 Mac&Chzz \$12.00 \$12.00
13	1 Salmon \$28.00 \$28.00
14	1 QueenCut \$34.00 \$34.00
15	1 Tacos \$23.00 \$23.00
16	1 TheEdison \$22.00 \$22.00
17	Subtotal 191.00
18	Tax 12.42
19	07:48PM Total Due 203.42
20	Suggested Gratuities
21	15% gratuity: \$30.51
22	18% gratuity: \$36.62
23	20% gratuity: \$40.68
24	Thank you for dining with us!

Developer

Researcher

Data scientist

Problem 4. Interpret extracted data (NLP+logic)

Output from Problem 3 solution is an input for Problem 4. You get a CSV file with the receipt data and you analyze each item in the receipt. For example, based on that information, identify a type of receipt – hotel/restaurant/shop/taxi/flight or identify a type of content in the receipt rows.

Or, given name of the restaurant dish – interpret – is it alcohol or not, is it luxury, is it overpriced for the area, does it contain fats or is it healthy? Scope a use-case interesting for you.

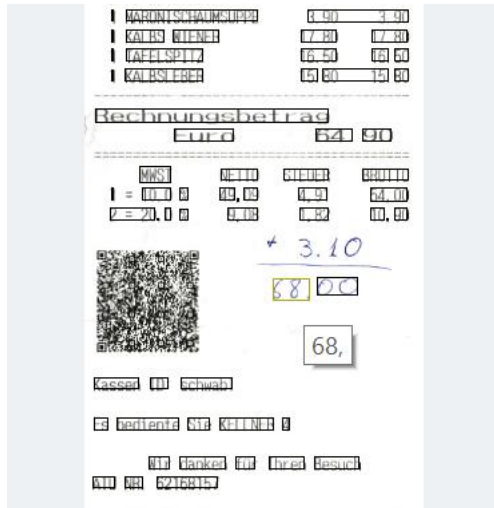
Input	Example output			Ordered items				
row #	raw text	Identified type of content	# of the table	# of guests	Quantity	Item	Amount	per item
1	The Edison	Name of the restaurant						
2	Dianey Springs	Address line 1						
3	Orlando FL 32830	Address line 2						
4	375659 George C	ambiguous						
5	Tbl 3091 Chk 5551 Get 5	Restaurant visit info	3091	5				
6	Nov06 18 00 38Pdt	Date/time						
7	2 DNR Cruzan Light Mojito \$24.00 \$12.00	Ordered items			2	DNR Cruzan Light Mojito	\$24.00	\$12.00
8	2 DR Copper Tail \$16.00 \$8.00	Ordered items			2	DR Copper Tail	\$16.00	\$8.00
9	1 DR Buddha Pine \$8.00 \$8.00	Ordered items			1	DR Buddha Pine	\$8.00	\$8.00
10	2 Pineapple Juice \$10.00 \$5.00	Ordered items			2	Pineapple Juice	\$10.00	\$5.00
11	1 Chips&Guac \$14.00 \$14.00	Ordered items			1	Chips&Guac	\$14.00	\$14.00
12	1 Mac&Chzz \$12.00 \$12.00	Ordered items			1	Mac&Chzz	\$12.00	\$12.00
13	1 Salmon \$28.00 \$28.00	Ordered items			1	Salmon	\$28.00	\$28.00
14	1 QueenCut \$34.00 \$34.00	Ordered items			1	QueenCut	\$34.00	\$34.00
15	1 Tacos \$23.00 \$23.00	Ordered items			1	Tacos	\$23.00	\$23.00
16	1 TheEdison \$22.00 \$22.00	Ordered items			1	TheEdison	\$22.00	\$22.00
17	Subtotal 191.00	Total without Tax						
18	Tax 12.42	Tax						
19	07:48PM Total Due 203.42	Date and Total						
20	Suggested Gratuites	ambiguous						
21	15% gratuity \$30.51	ambiguous						
22	18% gratuity \$36.62	ambiguous						
23	20% gratuity \$40.68	ambiguous						
24	Thank you for dining with us!	ambiguous						

Data scientist

Entrepreneur

Problem 5. Handwriting (Intelligent Character Recognition)

Some of the receipts contain handwritten tip or other comments. How this can be approached? Can we either extract the information using ICR tools or avoid corrupting the results of recognition of printed data?



Developer

Data scientist

Problem 7. Enhance existing OCR solutions

Can you make Google/ABBYY/Microsoft or other tools to return better results?

Processing the image prior to calling existing OCR tool will allow extracted data to be cleaner?

Which OCR tool would you recommend? Did you develop an OCR tool?

You will know when this problem is for your team.



Developer

Data scientist

Problem 8. Dataset preparation

You would like to work on something which will be definitely useful in future? Help Data scientists by preparing a dataset or cleaning and processing the data you find elsewhere. Without good dataset the problem will never be resolved. This is also very creative – of course you can use internet to get original receipts. But also you can visit restaurants and ask them to give you their receipt for last month. You may try to call to other countries restaurants and ask them to send you copies of their receipts. Organizing this is much more interesting than it seems, but also very challenging.

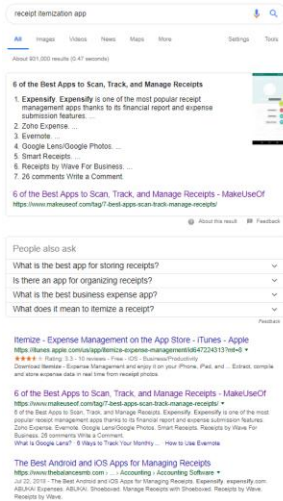
Automate scanning process for each receipt template and prepare a database and guidelines of how to grow it going forward.

Entrepreneur

Researcher

Problem 9. Market research (test existing solutions)

Deeply research the market – check for existing solutions and try them. Give us clear presentation on what was useful, what was not, maybe our hackathon is not needed at all and some startup solved all our use-cases (not). This challenge seemingly simple might require very advanced research skills to find something useful.



Entrepreneur

Designer

Researcher

Problem 10. Leverage Natural Language Processing on raw OCR data

You believe that with correctly applied NLP you can extract data and even restore a receipt structure. Use raw OCR output without symbol location to restore receipt structure.

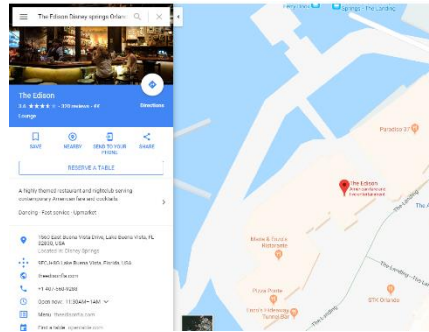
Similar to problem 5, but instead of nice and tidy CSV with recognized text try to use raw output from Google Cloud Vision or another tool.


Data scientist

Researcher

Problem 11. Locate the receipt issuer

You get a receipt scan as an input. Find the website/google location of the entity who issued it.



 theedisonfla.com

Researcher

Problem 12. Multi-language receipt processing.

Add Multilanguage support to any of the use-cases and double the set of problems:

品名	數量	單價	金額
茶	5	3.0	15.0
沙茶雞絲	1	30.0	30.0
紅燒鵝皮卷	1	22.0	22.0
紅燒鵝皮卷	1	22.0	22.0
古法炸蛋卷	1	14.0	14.0
香滑雞絲卷	1	15.0	15.0
香滑雞絲卷	1	15.0	15.0
蟹肉鮮魷卷	1	24.0	24.0
香芒奶皇卷	1	24.0	24.0
蛋黃叉燒包	1	13.0	13.0
椰菜肉叉燒包	1	16.0	16.0
博愛時菜卷	1	26.0	26.0
玫瑰紅薯爪	1	15.0	15.0
小計			251.0
數量:		17	
總共:		251.0	
現金:		251.0	

Developer

Researcher

Problem 13. Microservices

Microservices for receipts processing – elaborate approach and build services which would be limited to getting 1 piece of information from the receipt. E.g., get the Total, get the name of the restaurant, get number of guests.

Developer

Researcher

Problem 14. UX for Expense Management process

Design of the wire frames UI which can be used by accountants and other users to facilitate their interactions with expense management process.

Designer

Entrepreneur

Problem 15. Different idea

Use problems above as inspiration as you come up with your own way to solve the problem or to improve the receipt processing.

Developer

Data scientist

Designer

Researcher

Entrepreneur

Prizes and evaluation

Mentors and experts will evaluate your team's:

- ideas and creativity
- effort
- added value of your implementation
- fair play and collaborative attitude

Among the jury we have a professional in each of the areas involved – data science, business, accounting and IT. Any of the above mentioned use-cases can lead your team to success and place on the podium. The main goal of all of all participants is to contribute to solution of the ultimate problem.

The prizes will be granted to the notable teams by collectively selected by the jury.

Important. For each existing tool used to solve certain problem, the creators must be mentioned in the final presentation. Failing to do so will revoke your team's participation in the competition and trigger further actions by organizers. If you openly use 3rd party tools this will not degrade your achievements.

Tools and resources worth looking at in advance

You are encouraged to propose completely different approaches to address a business case and use tools of your choice. However, if you have not such experience we recommend to start with the following tools – as they are beginner friendly.

[Google Cloud Vision](#) (OCR)

[SpaCy.io](#) (NLP)

Contacts

Please, send your questions/suggestions to Ivan Glinka ivan.glinka@pwc.com