

2021-11-6

# MESSENGER BOOKKEEPING CHATBOT PROJECT REPORT

CHEN ZEFANG  
ZENG TENG YUE  
ZHOU XINYI  
ZHOU YIYANG

## Content

Executive Summary .....	1
Business justification .....	1
System Design .....	7
System Development and Implement.....	11
Future Improvements .....	15
Appendix 1 Proposal.....	16
Appendix 2 Installation and User Guide .....	17
Appendix 3 Individual Report .....	21
Appendix 4 Mapped System Functionalities against knowledge.....	27

## **1. Executive Summary**

As consumers continue to struggle with the financial uncertainty caused by more than a year of pandemic economic disruptions, budgeting has become a top priority for many people. While there is no shame in using a handwritten bookkeeping, bookkeeping apps or intelligent bookkeeping chatbot have become a popular way to track spending and savings habits.

The project has developed a Messenger bookkeeping chatbot based on natural language processing techniques. Since transactions for modern people are, nowadays, very frequent and mostly digital, and since recording transactions are still important for an individual's financial management, an auto-bookkeeping tool is in demand. The system developed in the project has two main functions, transaction history query and recording transactions. The model applies Naive Bayes algorithm to classify different actions of input sentences. The input sentences are transformed into word vectors using Jieba functions. Moreover, the project has a potential to be further developed to be a data-mining tool.

Keywords: Naive Bayes, bookkeeping, expense, income

## **2. Business justification**

### **2.1 Business Problem Background**

With close to 1.3 billion downloads in the second quarter of 2020 and finance apps being used over 1 trillion times on Android devices throughout 2019, it paints a decent picture in terms of growth of adoption of personal finance apps.

In fact, according to research conducted by Google, a smartphone user on average has close to 3 personal finance apps installed on their mobile devices with 4 in 10 using their devices for managing finance, checking account history, tracking investments, paying bills among other financial activities.

This creates an opportunity for intelligent bookkeeping chatbot, because covid triggers a boom in personal finance app market, and users need a simple and

convenient bookkeeping solution to integrate their expenses and income on all platforms, including online and offline.

## 2.2 Market Research

To evaluate competitive position of our solution and to develop strategic planning. We use SWOT analysis to assess internal and external factors, as well as current and future potential.

### 2.2.1 Strength

1. One-step bookkeeping and one-step query.



2021年11月3日的账单

	ID	描述	金额
0	6181fa7e903a8b5143b16a27	今天吃早餐花了20块钱	-20.0
1	6181fb0e503a8b5143b16a28	工资进账200	200.0
2	6181fb0e503a8b5143b16a29	淘宝购物500	-500.0
3	6181fb45903a8b5143b16a2a	淘宝购物了200.	-200.0
4	汇总	总支出:720.0总收入:200.0	-520.0

**Fig. 2-1:** Example of one query

2. Support text, audio and video as input.
3. Fast opening speed, no advertisements.
4. Automatically classify and extract time, value and type through algorithms and complex rules for bookkeeping.

5. Supports multiple expressions, like “ The day before yesterday I paid 127.85 for 20 pieces of candy ”

### **2.2.2 Weakness**

1. No graphical analysis interface for now.
2. No further classification of income and expenses for now.

### **2.2.3 Opportunity**




1. More and more users are choosing digital ways to keep track of expenses, manage, and grow wealth. There are boundless scaling opportunities for personal finance management solution.
2. The existing bookkeeping apps are too complicated and difficult for the elderly to use.
3. There are many existing speech recognition systems for various languages, so the project is highly scalable.

### **2.2.4 Threat**

With so many bookkeeping apps available, and a wide range of features among all of them, it can be daunting to sort through the numerous options. Here we selected the highest ranked bookkeeping app “Shark bookkeeping” and another bookkeeping chatbot app “DaoDao bookkeeping” from app store China. Figure 2-1 shows the user interface of our EZcharge and our competitors DaoDao bookkeeping and Shark bookkeeping, and Table 2-1 compares the functionality of the three bookkeeping solutions.



Fig. 2-2: User interface of EZcharge, DaoDao bookkeeping and Shark bookkeeping

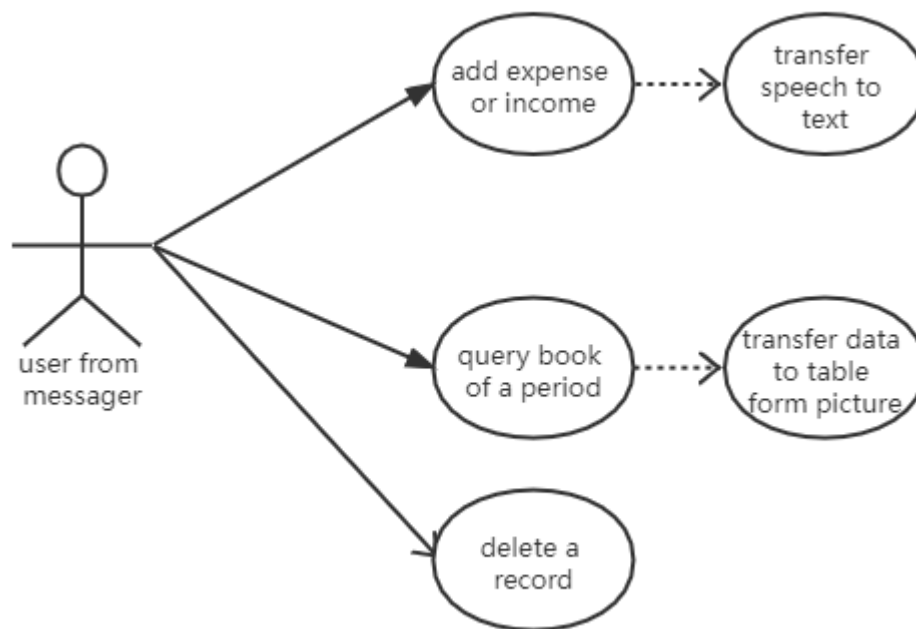
APP	Bookkeeping Steps	Start time	Strength	Weakness
EZcharge 	1 step input of text, audio and video	1s	1. Clear user interface with no ads.	1. No graphical analysis.
Competitor- DaoDao 	1. Choose income/expenses 2. Select specific income/expenses type 3. Enter the amount and comments	3s with start-up ad	1. Support graphical analysis. 2. Attract young female users through user-defined chat partner.	1. Complex user registration procedure. 2. Too many ads on the user interface.
Competitor- Shark 	1. Choose income/expenses 2. Select specific income/expenses type 3. Enter the amount and comments	3s with start-up ad	1. Support graphical analysis.	1. The user interface is too complicated for elderly user.

**Table 2-1:** Function comparison of EZcharge, DaoDao bookkeeping and Shark bookkeeping

In addition to the mentioned limitations, these bookkeeping app competitors all have unforeseen security issues. Therefore, users need a bookkeeping solution that offers a high level of data encryption and multiple ways to verify their identity. This is how we can improve our intelligent bookkeeping chatbot in the future.

### 3. System Design

#### 3.1. Use case and Business Flow



**Fig. 3-1:** Use case diagram

Our system aims to simplify the process of bookkeeping, as the machine learning model can help to recognize the type of query or add record. User can easily send a text message or audio/video message to our system through Facebook Messenger chat window. Then the backend will process these data into text form by calling Baidu speech transfer API, if the message is in audio or video form. The system will check whether this message contains a special command “delete” and followed by an record ID, which is used to delete this certain record. If not, a naive Bayesian classifier will figure out the type of this message (record or query).

If the message aims to add a record, then another Bayesian classifier will be used to identify the type of this record (income or expense). Then a bunch of matching rules, mainly based on regular matching rules, will process the message into explicit number, time and other related information. These data will be fed into our Mongo DB database together with the raw text data. And the success saved message will be sent to user through the chat window. (Shown as below)

If the message aims to query the records during a certain period of time, a bunch of extracting rules, mainly based on regular matching rules, will process the message into time period information. By now, our system supports different time period in days, month and years. Then the query will be sent to our Mongo Db to



get all the records in this period, and the records will be converted into more readable and understandable image form to send back to user in the Facebook Messenger chat window. (Shown as below)

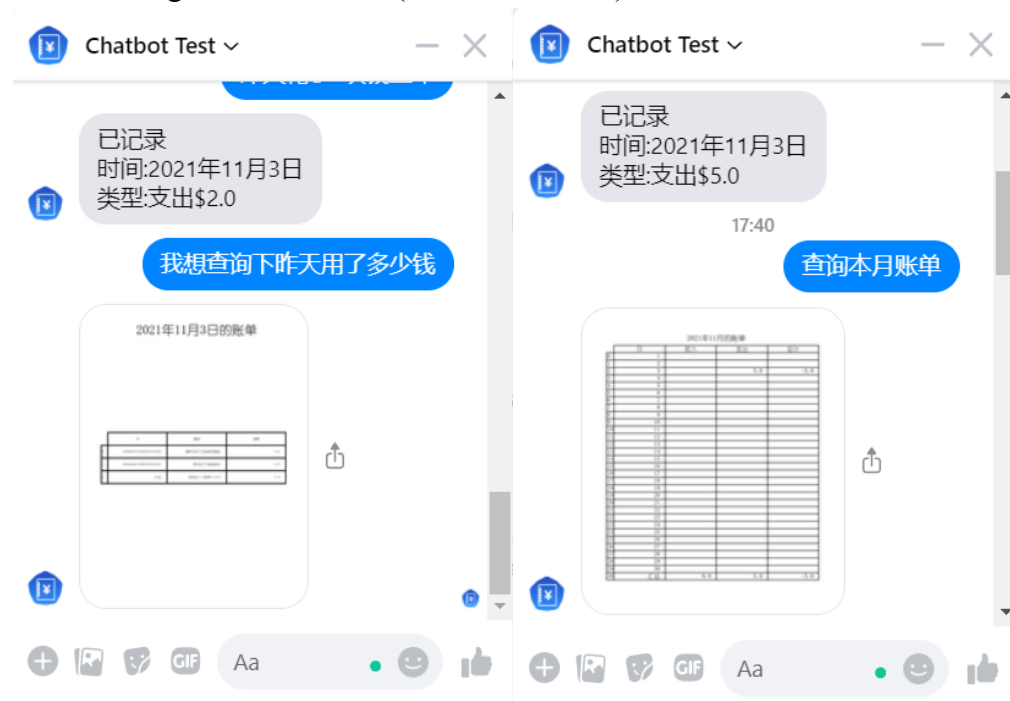


Fig. 3-2: Sample chat

## 3.2. Reasoning Models

We use different machine learning technics to classify the type of message and to extract the useful information in the message.

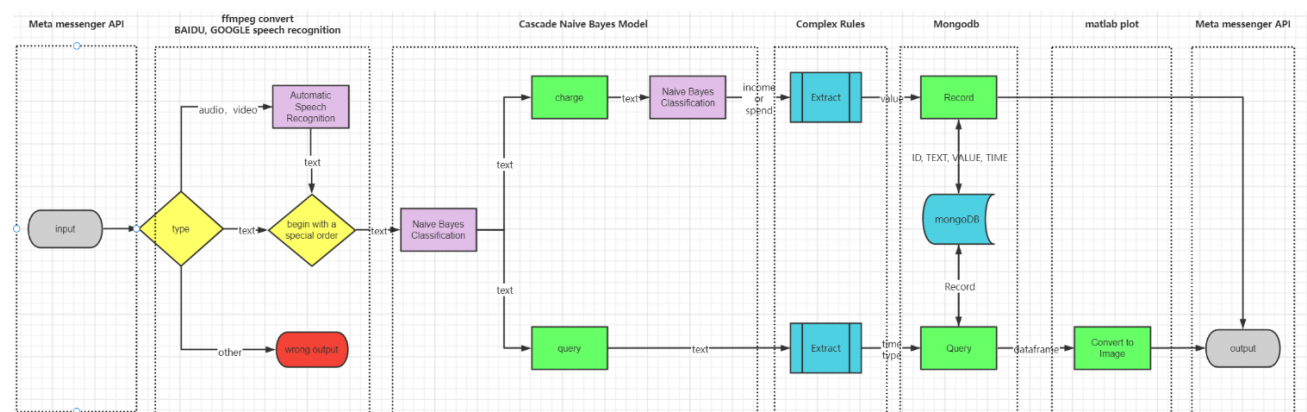


Fig. 3-3: Technical stacks

### 3.2.1. Naïve Bayesian classifier

First, we use Jieba to do word segmentation for better accuracy in classifying Chinese text. Then we use one Naïve Bayesian classifier to confirm that the user wants to look up the accounting book or add more records of expense and income.

One more Naïve Bayesian classifier is added to distinguish the record type, income or expense. The details of this classifier will be explained later in the implement part.

### 3.2.2. Regular Match

After classification, we use a bunch of regular rules to extract the accurate amount of money and date, including matching the different expression of spend and earn in Chinese and number in words in Chinese.

Expense related verb in Chinese	‘花了’, ‘用了’, ‘给了’, ‘付了’, ‘转了’, ‘亏了’, ‘刷了’, ‘借了’, ‘交了’, ‘用掉’, ‘掉了’
Income related verb in Chinese	‘赚了’, ‘得了’, ‘收入’
Number in Chinese words	‘零’, ‘一’, ‘两’, ‘二’, ‘三’, ‘四’, ‘五’, ‘六’, ‘七’, ‘八’, ‘九’

**Table 3-1:** examples of matching words

## 3.3. Database Structure – MongoDB

Considering the storage of raw data and flexible format to save all the records from all users all time, we use MongoDB as our database. The document-based model is unstructured database avoiding potential schema changing in future development. Besides, it offers powerful searching with index, which give us more possible searching based on index of different attributes.

```

564 // -----
565 {
566   "_id" : ObjectId("6181fb45503a8b5143b16a2a"),
567   "client_id" : "6758070074210966",
568   "year" : NumberInt(2021),
569   "month" : NumberInt(11),
570   "day" : NumberInt(3),
571   "description" : "海底捞吃了200。",
572   "type" : NumberInt(-1),
573   "value" : 200.0,
574   "add_time" : ISODate("2021-11-03T03:00:21.045+0000")
575 }
576 // -----
577 {
578   "_id" : ObjectId("6181fc46503a8b5143b16a2b"),
579   "client_id" : "6758070074210966",
580   "year" : NumberInt(2021),
581   "month" : NumberInt(11),
582   "day" : NumberInt(3),
583   "description" : "今天花了三块两毛嗯。",
584   "type" : NumberInt(-1),
585   "value" : 3.2,
586   "add_time" : ISODate("2021-11-03T03:04:38.571+0000")
587 }
588 // -----
589 {
590   "_id" : ObjectId("61838e07f130492fb321e622"),
591   "client_id" : "4406925109425092",
592   "year" : NumberInt(2021),
593   "month" : NumberInt(11),
594   "day" : NumberInt(3),
595   "description" : "我昨天花了五块钱买面包",
596   "type" : NumberInt(-1),
597   "value" : 5.0,
598   "add_time" : ISODate("2021-11-04T07:38:47.811+0000")
599 }

```

**Fig. 3-4:** Structure of database

_id	Auto-generate distinctive record id
client_id	User id
year	Year of the date
month	Month of the date
day	Day of the date
description	Raw text data of the message
type	Expense or income
value	value of money
add_time	The time of inserting the record

**Table 3-1:** Explanation of the attributes

## 4. System Development and Implement

### 4.1. System Modules

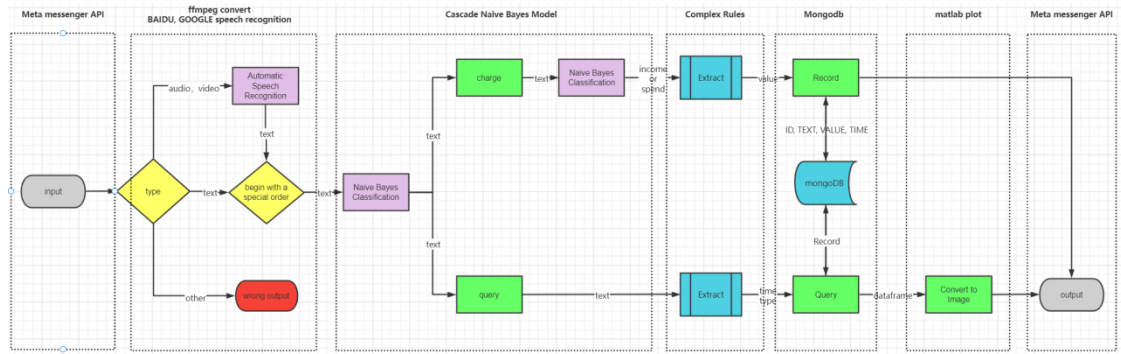


Fig. 3-3: The technical stacks of the system

The whole process flow can be divided into 7 main technical stacks, of which the input and output are both finished by open-source API of Facebook Messenger.

- First up after input, the module is to convert all forms of input data to raw text with special starting words.
- Then it is the Naive Bayes classifier to tell whether the text is to extract a bookkeeping record or to add a new bookkeeping record. Additionally, for the latter, apply another Naive Bayes classifier to distinguish income or expense. In all, there are three classes, retrieving record, recording expense, and recording income.
- The Complex Rules module is to extract values from the raw text. After acquiring the class of the text (mentioned above, one of the three classes), the system execute actions respectively. If the class is query, the system extracts the date and time value. If the class is income or expense, the system extracts the cost value (positive/negative).
- MongoDB module is to adjust the database. After receiving the class type and value, the system does executions respectively. When taking notes in bookkeeping, it adds the current data into the database (transaction ID, transaction value, transaction time). When extracting notes in bookkeeping, it retrieves the sub-dataset of the given time interval.
- For query, there is a image generation module. The module uses MATLAB plot to generate an image of a sheet based on the data received.

### 4.2. Naive Bayes Model Construction:

The Naive Bayes classifier figures out whether a text sentence is to ask for note taking or looking up history. Also, within the bookkeeping section, there is another classifier to label out income or expense. In this cascade classifier process, the system inputs new sentence and outputs the result based on the trained model. The model is trained by the dataset below:

	A	B	C	D
1	text	label	cascade	lb2
2	查账	0		
3	查询	0		
4	历史记录	0		
5	历史账单	0		
6	查询一下	0		
7	查询账单	0		
8	查看记录	0		
9	查一下账	0		
10	想查账	0		
11	想看账单	0		
12	想查下账	0		
13	买	1	1	
14	买了	1	1	
15	给了	1	1	
16	给	1	1	
17	花了	1	1	
18	借了	1	1	
19	借给	1	1	
20	用了	1	1	
21	退了	1	0	
22	退	1	0	
23	找到	1	0	
24	发了	1	0	
25	收入	1	0	
26	收	1	0	
27	收了	1	0	
28	回款	1	0	
29				

**Fig. 4-1:** This is the trimmed dataset, not the whole

Then, apply Jieba functions to convert those keywords into word vectors. The matrix is as the following:

[illegible]

**Fig. 4-2:** Example of word vectors of the two classifiers

When a sentence is transformed to an array using Jieba functions, sentences with the same label share some similarities. We can the label to be the output and those arrays to be the input, thus, construct a 3-class Naive Bayes classifier. When the system encounters a new sentence, it scores the sentence with 3 possible labels and rule out the one with the highest score. This is the procedure of Naive Bayes classifier.

### 4.3. System features

### 1. Language features:

The language of input is Mandarin. Mandarin is a Sino-Tibetan language that there is no punctuation (spaces) between words. Usually, one word is composed

of one or more characters. And a character may contain multiple meanings with similar frequencies. The meaning of a character is determined only when put into a sentence, that is, the meaning of a character is decided by the combination with other characters.

## 2. User-friendly:

The target of this software is everyone using bookkeeping, so the process shall be easy, user-friendly, and interactive. We are not supposed to develop a console window that only programmers know how to operate. This software develops a chatbot that can be interactive to users, and the interface is the same as Messenger.

## 4.4. System scope

Since the project faces bookkeeping usage cases, the input can be done through simple text. While Messenger provides multiple input types, including image, video, etc. Therefore, we define the input data type within text and audio only.



Fig. 4-3: Text and audio can be both recognized by the system

## 4.5. System Measurement

We have tried with different speech sentences and text sentences. The result is accurate that when we used different implicit Mandarin verbs to express expenses (not saying anything related to costing), it can detect expenses with the correct value. The income and query functions performed well too.

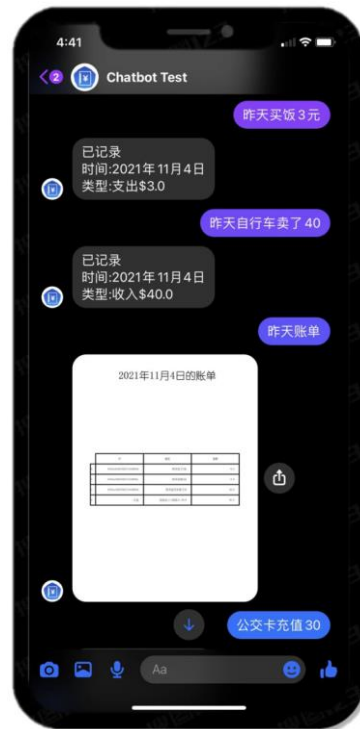


Fig. 4-4: Example of testing

## 5. Future Improvements

### 5.2. Categories of cost classification:

Until now, the system can only detect whether the transaction is income or expense. However, it does not classify the category of costs, such as food, education, entertainment, etc. Also, there are sub-categories that can be applicable.

打麻将赢了##	Gained ## playing Mojang	+ Entertainment
违章交了##	Fined with ## for traffic violation	- Fines
买菜花了##	Costed ## for VegeFresh®	- Necessities (sub food)
Steam氪了##	[Term for topping up] ## for Steam	- Entertainment

Fig. 5-1: Samples of input and English translations and converted data (gain or expense and category)

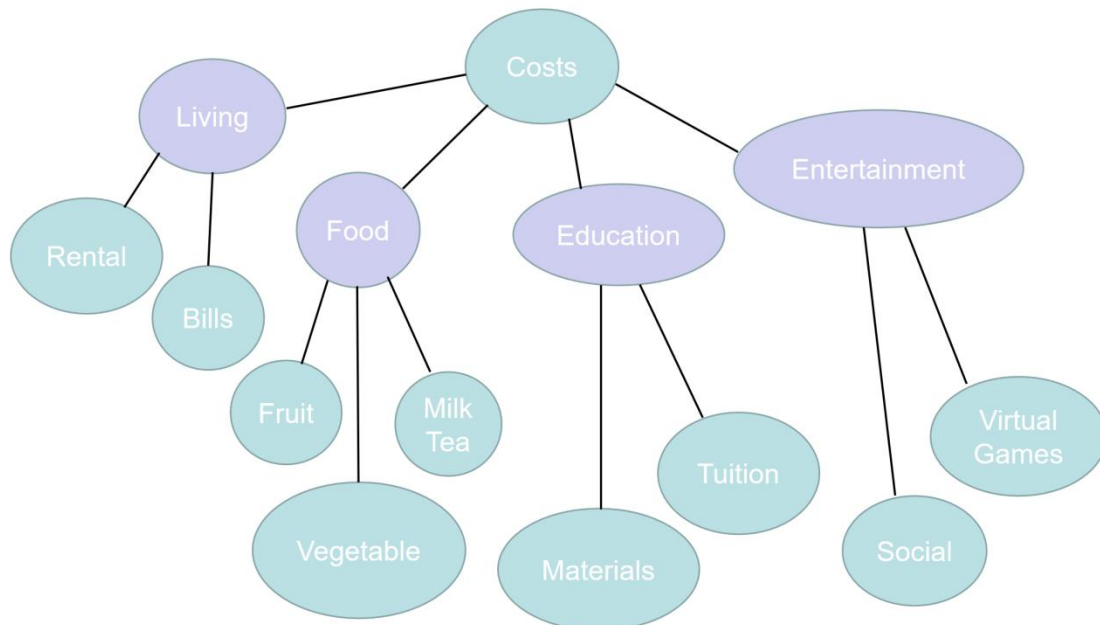


Fig. 5-2: Example of cost category tree-classification

### 5.3. Data mining:

Since we have developed a language database including a variety of transaction-related words, there is a potential to develop a data-mining tool. Generally, if a user wants to acquire the bill or take notes of transactions, the system can require the user to input beginning with only one special word, but we have taken plenty of special words into the database. Thus, the database can do more than taking notes, but data mining. When it is installed into a device, it can mine the chat history of the user and thus estimate the personal preference.



## Appendix 1: Proposal

<b>Date of proposal:</b> Aug 08, 2021
<b>Project Title:</b> Messenger Bookkeeping Chatbot
<b>Group ID (As Enrolled in LumiNUS Class Groups):</b> 14 <b>Group Members (Name , Student ID):</b> Zeng Tengyue, A0231549A Zhou Xinyi, A0231538H Zhou Yiyang, A0231545L Chen Zefang, A0231380R
<b>Background/Aims/Objectives:</b> In the modern society, people have high frequency of small-amount transactions. Also, digital transactions are taking more and more proportion. This becomes a problem of bookkeeping in financial management. Thus, a new type of auto bookkeeping method is demanded. We, thus, propose a chatbot to do the task, which is user-friendly, convenient, and functionable.
<b>Project Descriptions:</b>  Structure: When a user speaks or texts to this the chatbot, it makes decisions to do specific actions, recording income, recording expense, and looking up transaction history. There will be an natural language processing module to convert language information to data.  Method: We plan to use Naive Bayes method to do the classification, and we plan to use NLP algorithms to convert text to word vectors as the input.  Detailed Objectives (sorted by difficulty): <ol style="list-style-type: none"> <li>1) Build an interactive chatbot in Messenger</li> <li>2) Convert sentences to word vectors</li> <li>3) Construct Bayesian classifiers with those vectors and validate them by testing</li> </ol>

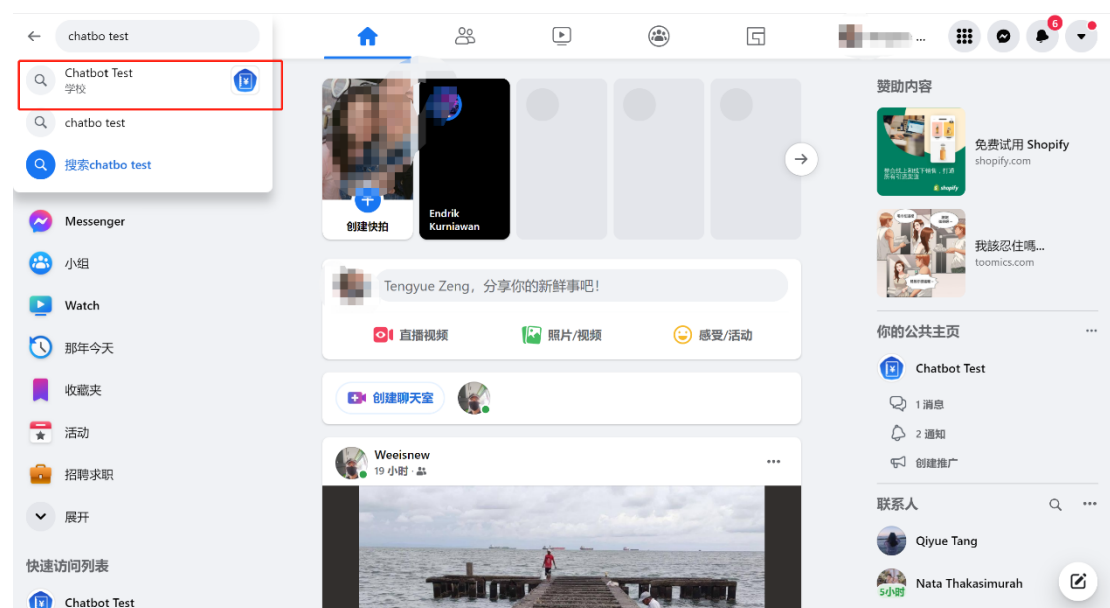
## Appendix 2: Installation and User Guide

### Installation:

Our system is running on our server, so users can easily get the robot's help just by chat with it through Facebook Messenger.

You can first log into your Facebook account on <https://www.messenger.com/> or your already download Messenger APP.

Then you tap "Chatbot Test" in the search bar. Like below:



You will find the proper one with this profile picture.



Chatbot Test

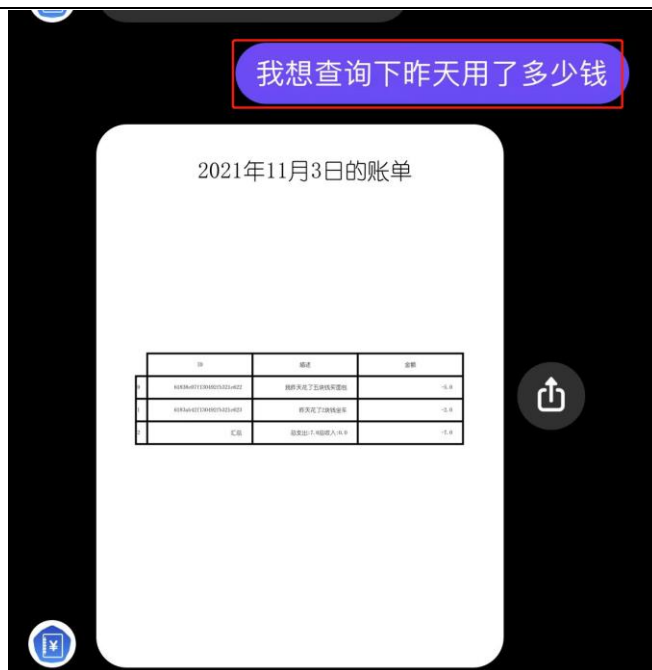
Customise chat ▼

Privacy and support ▼

Shared media ▼

Click it to start the conversation and enjoy your convenient bookkeeping with it!





Then it will return you a table like below:

## 2021年11月3日的账单

	ID	描述	金额
0	61838e07f130492fb321e622	我昨天花了五块钱买面包	-5.0
1	6183ab42f130492fb321e623	昨天花了2块钱坐车	-2.0
2	汇总	总支出:7.0总收入:0.0	-7.0

2021年11月的账单

	日	收入	支出	总计
0	1			
1	2			
2	3		7.0	-7.0
3	4			
4	5			
5	6			
6	7			
7	8			
8	9			
9	10			
10	11			
11	12			
12	13			
13	14			
14	15			
15	16			
16	17			
17	18			
18	19			
19	20			
20	21			
21	22			
22	23			
23	24			
24	25			
25	26			
26	27			
27	28			
28	29			
29	30			
30	汇总	0.0	7.0	-7.0

- If you add a record by mistake and need to delete it, please ask to check the daily records of the wrong record. Use the ID to delete the record in this format:

delete@61838e07f130492fb321e622

	ID	描述	金额
0	61838e07f130492fb321e622	我昨天花了五块钱买面包	-5.0
1	6183ab42f130492fb321e623	昨天花了2块钱坐车	-2.0
2	汇总	总支出:7.0总收入:0.0	-7.0

## Appendix 3: Individual Report – Zhou Xinyi

ISS\_IRS\_PM Messenger Bookkeeping Chatbot

**Zhou Xinyi A0231538H**

### 1. Personal Contribution

Throughout the project, I participated in model choosing and testing. And I also contributed to the project management, which includes project planning, deciding the scope of our application, scheduling the events and discussion in our group.

At first we decide to use deep neural network to identify the message type and the user's intent, including causal chat, add records, lookup accounting book, etc. But later we found out that we don't have enough data to train this neural network to produce precise result. Considering we only have hundreds level data, we move on to more direct and manual controllable model. This is when I suggest we should try out naïve bayes classifier after we do pre-process of the raw Chinese message. Naïve bayes classifier learn much faster than neural network and it performs well for our project.

After all parts were developed, we step into testing stage. I organized the test cases and distributed them to team members to carry out and check the results. It's lucky to find that our system can perform well to do bookkeeping and look-up.

### 2. Learning Outcome

During this project, I tried out how to build a classifier model to meet a certain demand in real life and I learnt some developing skills to quickly develop a minimal runnable program.

The pre-processing of raw data is really important and can help a lot to model training. Our system is based on natural language, in terms of this, the pre-processing becomes more indispensable. It needs sentence segmentation and stop words removal before it is fed into our bayes classifier.

In addition, developing a software connected to other company's API is another important lesson I learnt. We develop our chatbot based on Facebook's service, so I got to know the process of getting permission from such big company.

### 3. Knowledge and Skill Application

The knowledge and skills gained through this project will be very useful in future jobs. As mentioned above, I gained more experience in building practical models to meet the requirements and how to improve the model with limited data.

In addition, getting more familiar with Python and the use of Pandas will also be of great help to me. They are indispensable tools today to processing data and analyze data.

## Individual Report – Zeng Tengyue

ISS\_IRS\_PM Messenger Bookkeeping Chatbot

**Zeng Tengyue A0231549A**

1. I learned to how to use ngrok to map an Intranet port to an Extranet. This will help me learn how to turn my computer into a server for everyone to access.
2. I learned how to use Flask to build a back end and listen for port requests. I never knew how to build a back end using Python, but now I can build a back end using Flask and register a function as an activation function that runs the set calculation steps when listening events are fired. This allows me to create more powerful Python code.
3. I learned how to call Baidu and Google mature API port to development my procedures. In reality, there are many powerful and accurate apis for us to use, in the actual system development, we do not need to develop each function alone, the proper use of these API interfaces can not only simplify our development process, but also often make our programs more powerful.
4. I am familiar with the use of Naive Bayes model and understand its shortcomings. In this development, the naive Bayes algorithm is not the most suitable algorithm, because a very big disadvantage of the naive Bayes algorithm is that its default words are independent between words, so it cannot identify the short text well, while the algorithm like RNN is more suitable. However, we can still use the naive Bayes model, but I have slightly improved it. Through the method of cascading model, we can judge one type first and then another type, so as to improve the accuracy of the Bayesian model. This can give me a lot of inspiration in the future.
5. Learned how to use mongoDB. Learning how to use mongoDB helped me a lot. It didn't require loading content into memory in real time and learning how to add databases to the system made my application more like an application, not a demo.



## Individual Report – Chen Zefang

ISS\_IRS\_PM Messenger Bookkeeping Chatbot  
**Chen Zefang A0231380R**

### 1. Personal Contribution

We participated in the project together, but my main part was focusing on collecting training data and doing the test of the classifier. I have tried many cases of the usage to enlarge the diversity of the training set. There are a lot expressions about incoming and expensing that are implicit, not saying sensitive words. I worked on this situation with many use cases.

Also, I had tried and evaluated different methods. At first, we planned to use machine learning methods to identify the message data type and message meanings. But afterwards we realized that the dataset size was hard to be acquired enough. Then I suggested to apply something simpler in training process, since the product we were to develop did not have to meet such requirements as to be that robust. Then we discussed to step onto Naïve bayes classifier to do such tasks.

### 2. What I have learned

In this project, I have learned some practical techniques.

I have learned to apply NLP functions to analyze sentences, to convert a sentence to word vectors, to extract features of a sentence. Also, I have got to know how to build a Naïve Bayesian classifier and had a further understanding of the theory of NB classifier.

This project also helped me to get to know open-source API. Messenger has open-source API, and I can modify functions there to build my customized chatbots. Also, during the project process, I learned to use Jieba, mangoDB, etc. This could be a training of such skill, since the skill of applying API will be used in multiple projects and multiple fields.

## Individual Report – Zhou Yiyang

ISS\_IRS\_PM Messenger Bookkeeping Chatbot  
Zhou Yiyang A0231545L

### 1. Personal Contribution

In this project, I participated in the test of Naïve bayes classifier and complex rule module. I considered different age and geographical language habits to construct the dataset of keywords. So there are various referring expressions for income and expenses, for example, win represents income and a recharge represents expense. Besides, I created some test case for our complex rule module to test whether the model accurately finds the matching amount when the income and expense keywords appear, like I bought \$20 fruit for \$15 today.

Apart from that, I analyzed potential project development opportunities and competitors for our projects based on external market factors and our own attributes. Then conducted project test to ensure that it meets the business requirements.

At the end of the project, I finished the video recording of business case and demo, then completed the project report with other team members together.

### 2. Learning Outcome

This was my first complete involvement in the development of a project. My teammates shared a lot of technical details with me and I learned a lot from them.

Firstly, from our complex rule module I learned that basic method is not necessarily worse than the latest models. For example, the accuracy of our complex rules is higher than TextCNN. In addition, I learned how to use some open source tools to help us better complete our projects, such as Baidu and Google mature API. After that, through the business case study and market analysis, I learned how to analyze the value of a product rationally from the perspective of a developer to that of a product manager.

During the project development, I better understand a point mentioned in our courses - there are more gray areas in real life. In the process of doing a project we have to learn to make trade-offs and learn to choose the most suitable one from several mediocre ideas and then do our best to optimize it.

### 3. Knowledge and Skill Application

Through this project I became more familiar with the use of python and some open source tools, which will be very important in my future study and work.

In addition, this bookkeeping chatbot also made me realize the extensibility of chatbot, we can also add graphical analysis or memo function to improve our bookkeeping chatbot. In the future, we can use the chatbot to design apps for senior users, as they can use Wechat or Whatsapp, I believe an intelligent chatbot will help them a lot.

## Appendix 4: Mapped System Functionalities Against Knowledge

Business rules and process	<p>1) Collect dataset of various (long/short, implicit/explicit, common/uncommon) expressions meaning query or bookkeeping</p> <p>2) Label the data with cascaded labels: expense, income and query (based on daily life experience)</p> <p>3) extract key information based on regular rules which represent the most common used words to describe money flows.</p>
Business resource optimization	We use evolutionary computing techniques to optimize our extract rules generation by generation and finally come out with a bunch of powerful regular rules.
System designed with cognitive techniques	The cognition process is detection of languages. It aims to convert human natural language to specific demands. We used Jieba to do sentence segmentation (to convert Chinese sentences to word vectors) and Baidu speech translator to recognize audio and video information.