Instruction Manual for Jaktent Chatbot Project

Client:

Mr. Goldy T Wijaya

Team Member:

AUT ID	Name
20108405	Tianyang Li
20108728	Shijie Ma
20108736	Hanpeng Jiang
20108511	Jiacheng Lv

Mentor Name:

Mr. Petteri Kaskenpalo

Date: 15/5/2022

Version 1.1

Executive Summary

The main purpose of this project instruction manual is to guide clients and developers to carry out subsequent project deployment and use, and at the same time provide some important technical details for further development and use of the project. The project instruction manual mainly includes the following two parts:

①**Project Run and Deployment:** This part is mainly to guide developers how to deploy and use the project, please follow the steps of the manual to execute step by step.

②**Important Project Technical Details Instruction:** We have further explained some important technical details. If developers encounter problems in the follow-up development process, this part can be used as a reference.

Table of Contents

Project Run and Deployment

1. Operation Environment:

Anaconda3.7(Recommend to download 2020.2 version), **Pycharm Professional Edition**(Don't install Community Edition as it not supports to use database)

2. Programming Language

①Front-end: HTML+CSS+Javascript, JQuery, etc.

②Back-end: Python3.7

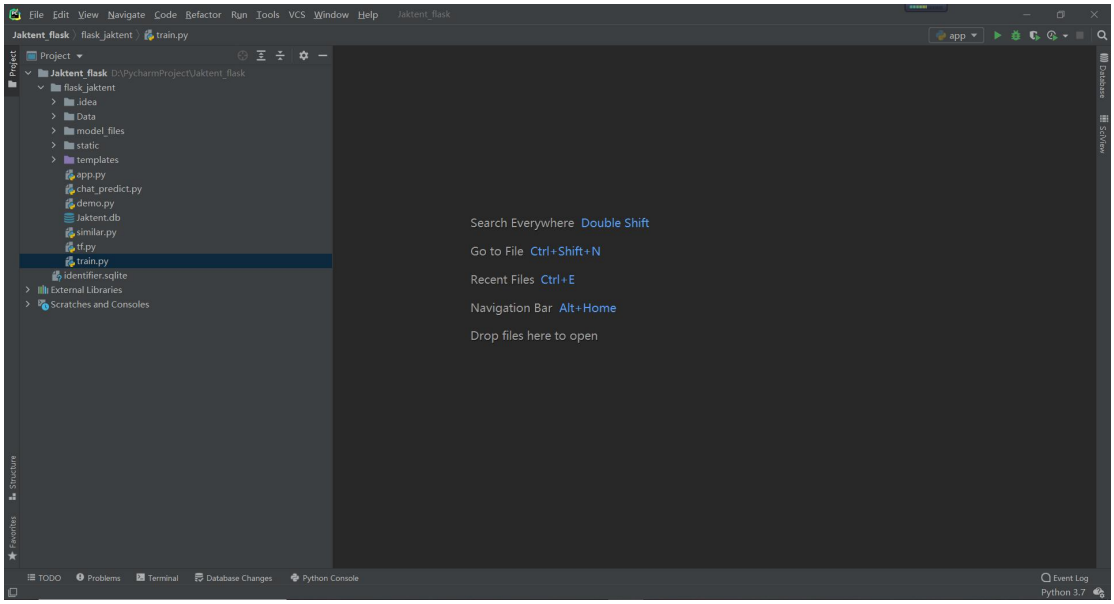
3. Python packages and plugins that require additional installation

①Python Package: Tensorflow 2.8.0**(Must be Tensorflow 2.0 version)**, flask 2.1.0, nltk 3.4.5, flask_cors 3.0.10, requests 2.22.0

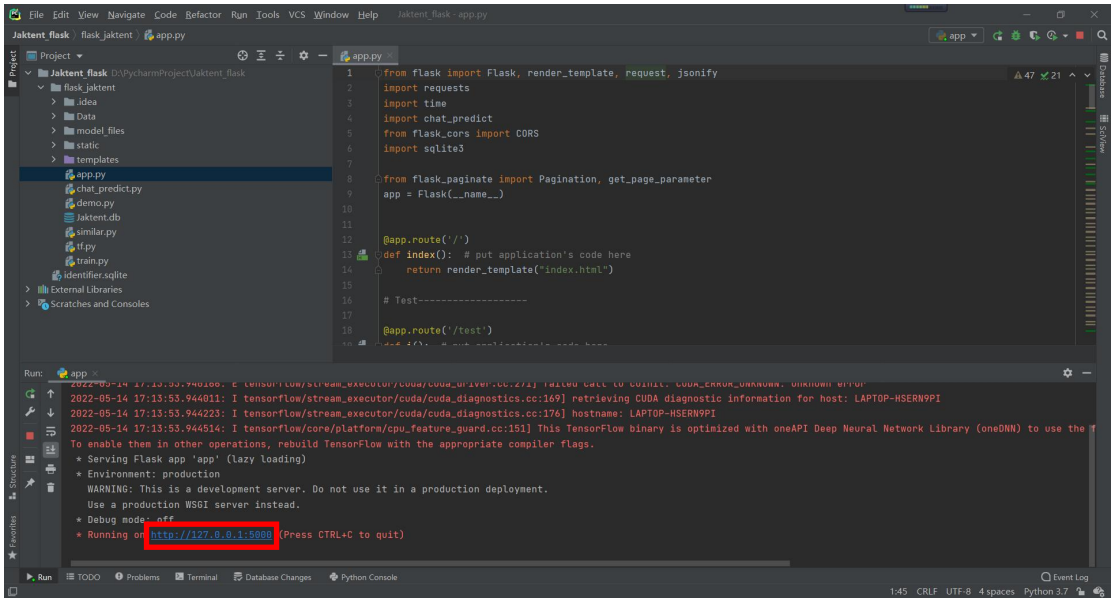
② flask_paginate 2022.1.8, sqlite3 3.330, xlwt 1.30, json 5, wordcloud 1.8.1, PIL 7.0.0

4. Project Run and Deployment

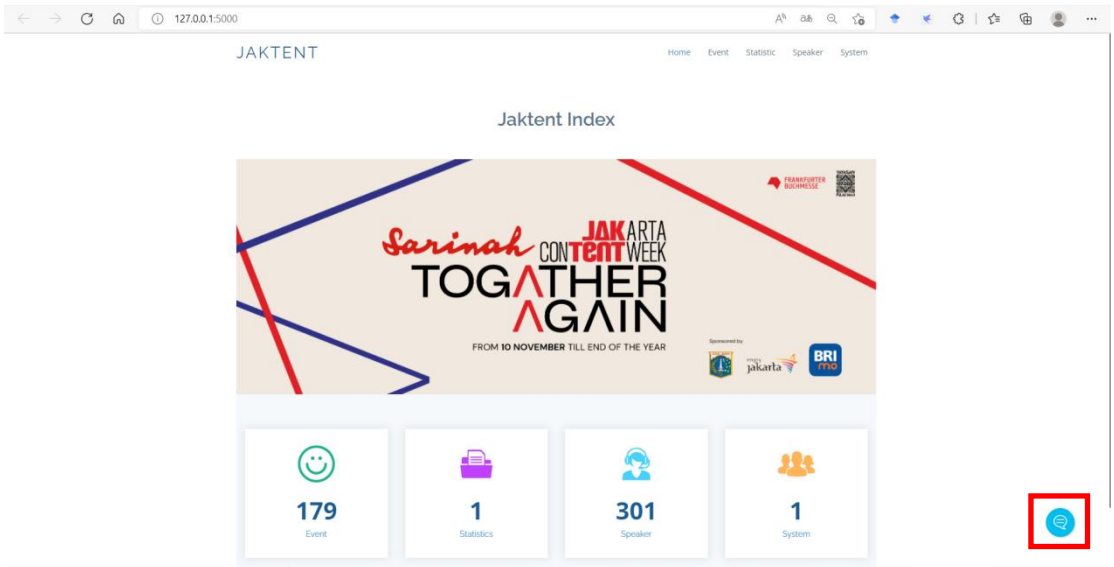
①In Pycharm, Click "File" → Click "Open" → Open the folder "Jaktent_flask". At the same time, deploy the pycharm operation environment.



②Run the "app.py" file



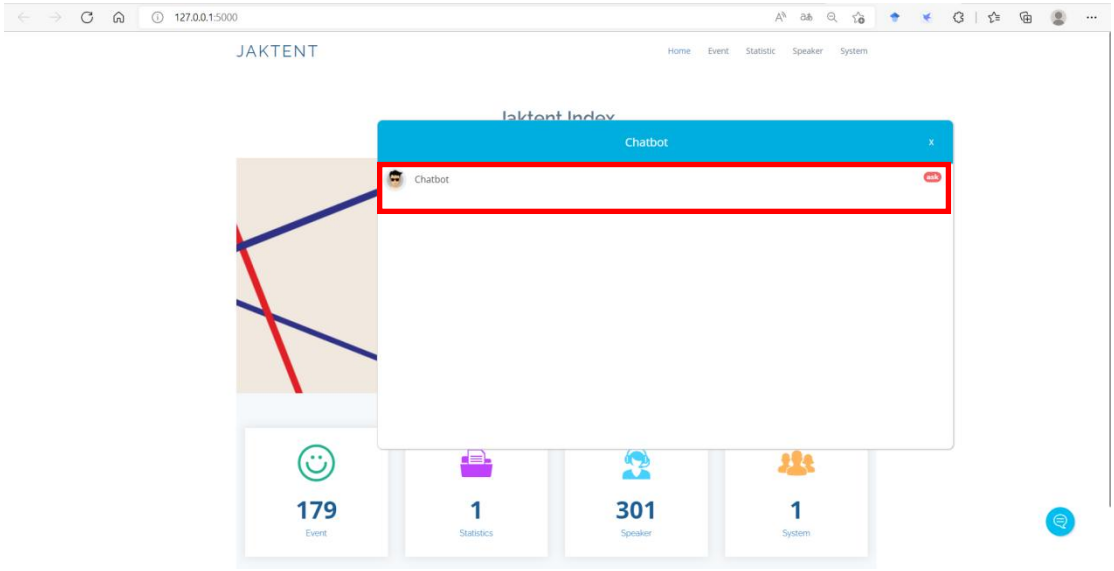
③Click on the IP address in the red box above to enter the Jaktent Index website interface(**Attention:** Clicking the IP address to enter the webpage may sometimes appear stuck, some icons cannot be displayed, some links cannot be clicked, etc. These are normal phenomena and are related to the browser configuration. If you encounter the above problem, you can solve it by refreshing the webpage.)

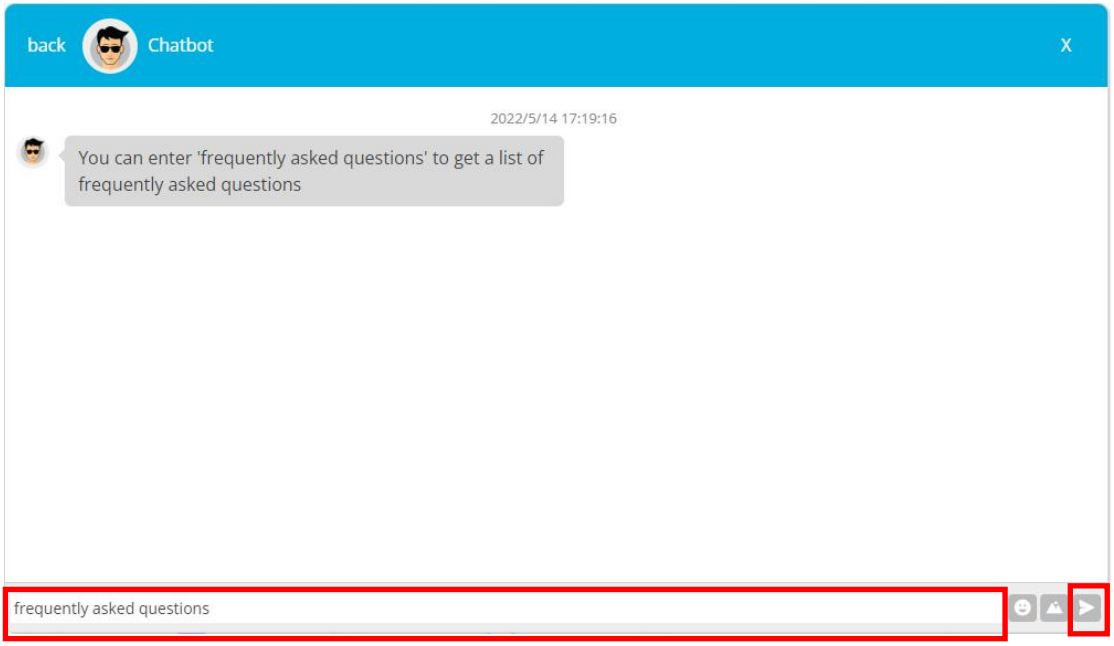


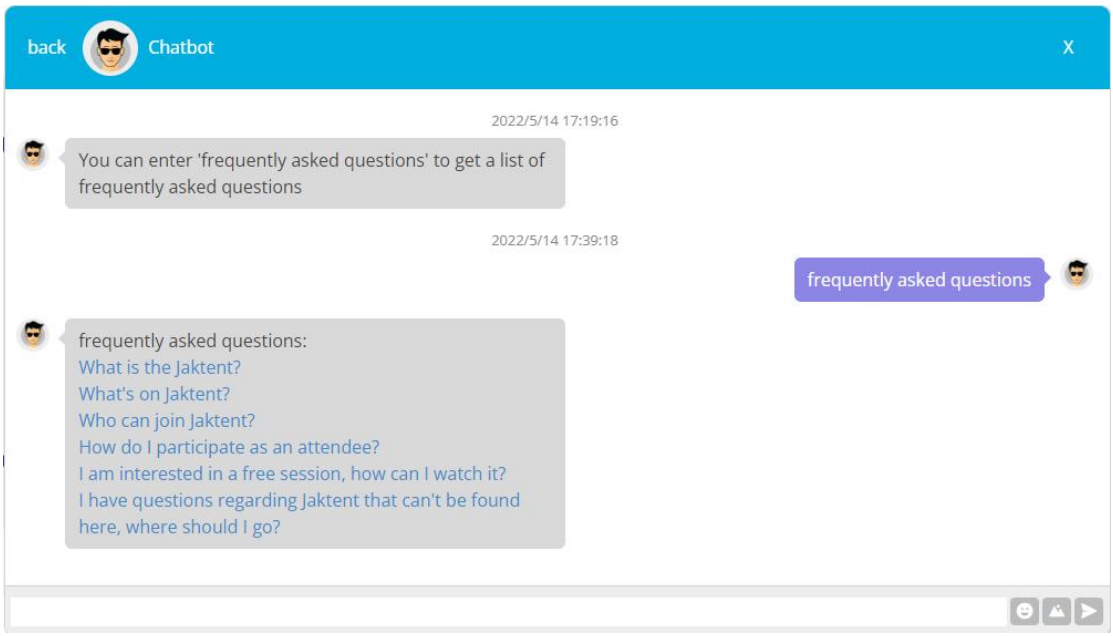
The website is temporarily connected to the local address. The default is 127.0.0.1:5000. If you need to deploy the website and chatbot online in the future, you can change the domain name.

④For the Jaktent Chatbot part, click the icon in the lower right corner of the image above to jump out of the chatbot window. Then click "ask" to enter the session

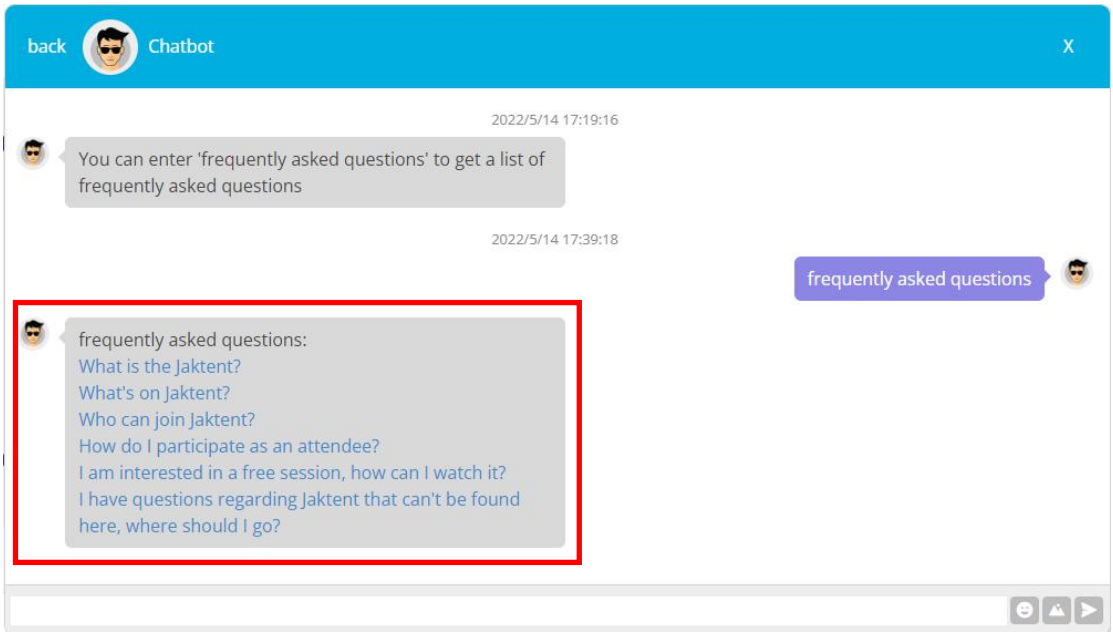
interface. Type "frequently asked questions" in the input box to start a conversation with the chatbot.

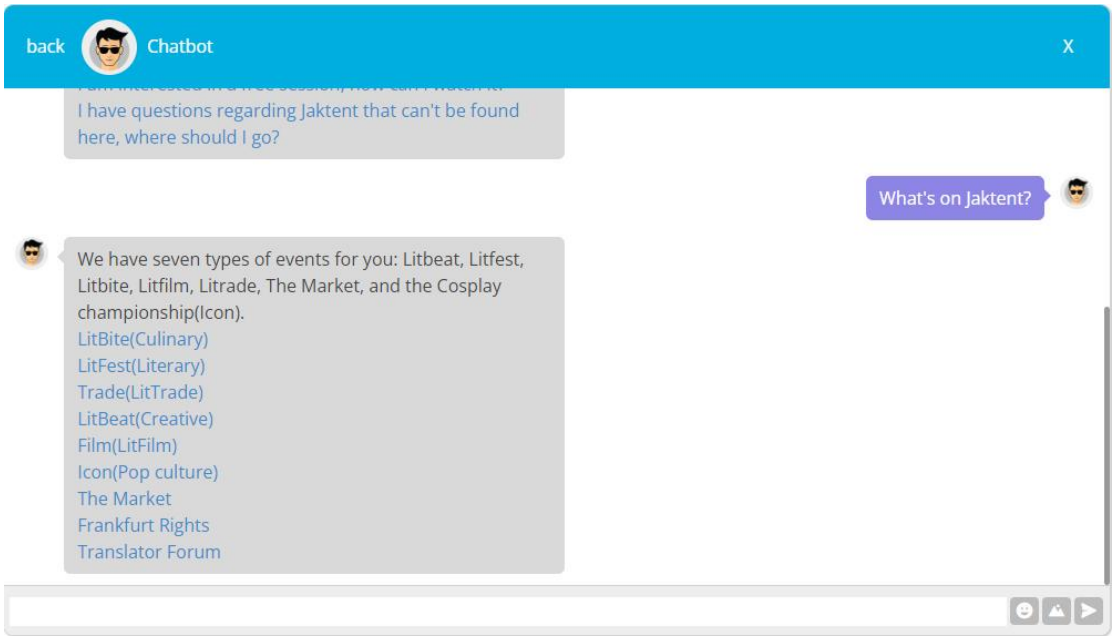


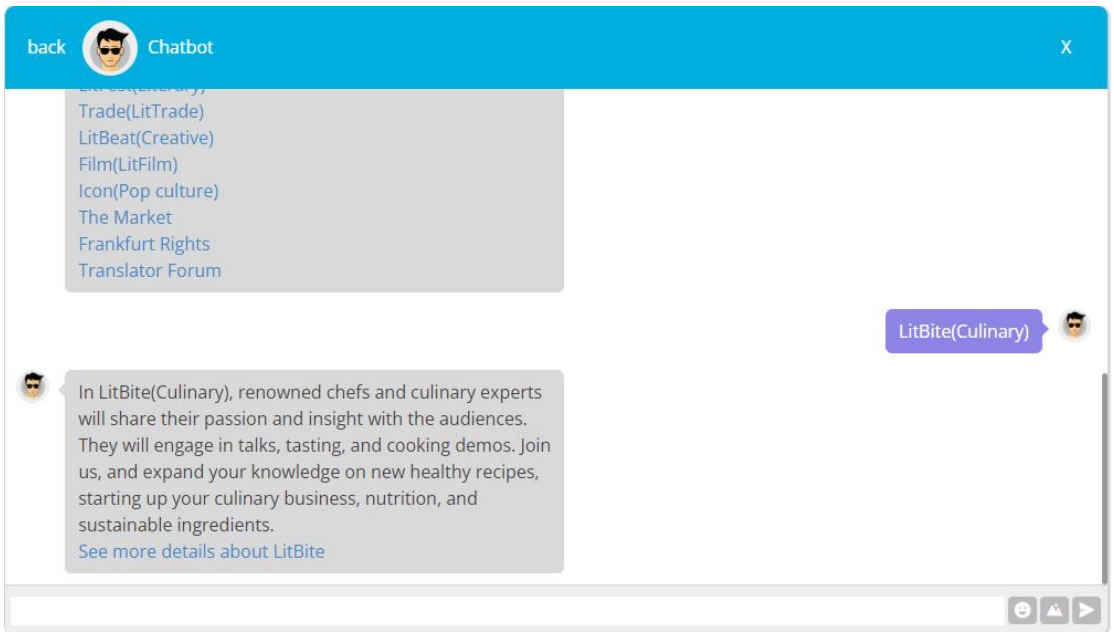




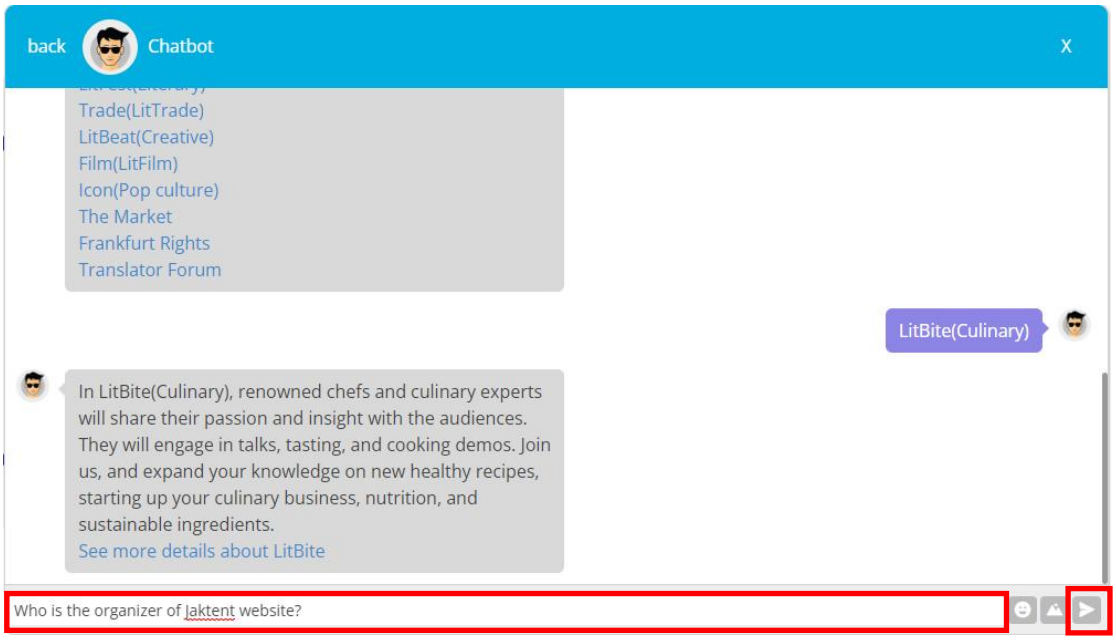
⑤For some frequently asked questions, you can directly click the hyperlinks in the dialog box to get answers, and some answers will be followed by other questions that may be asked, and you can also get answers after clicking.

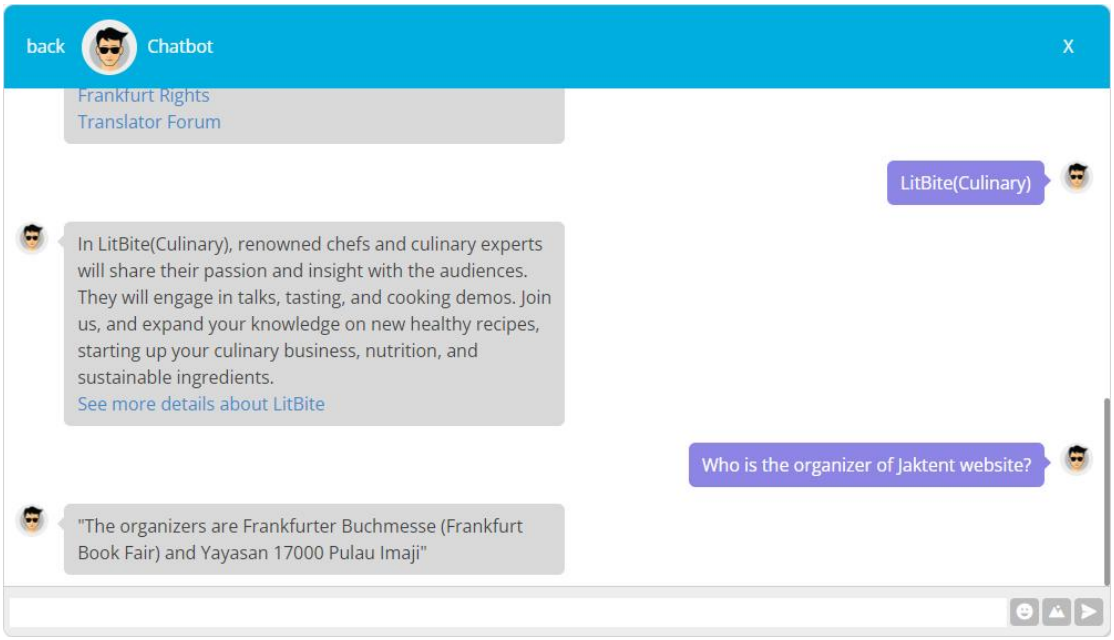


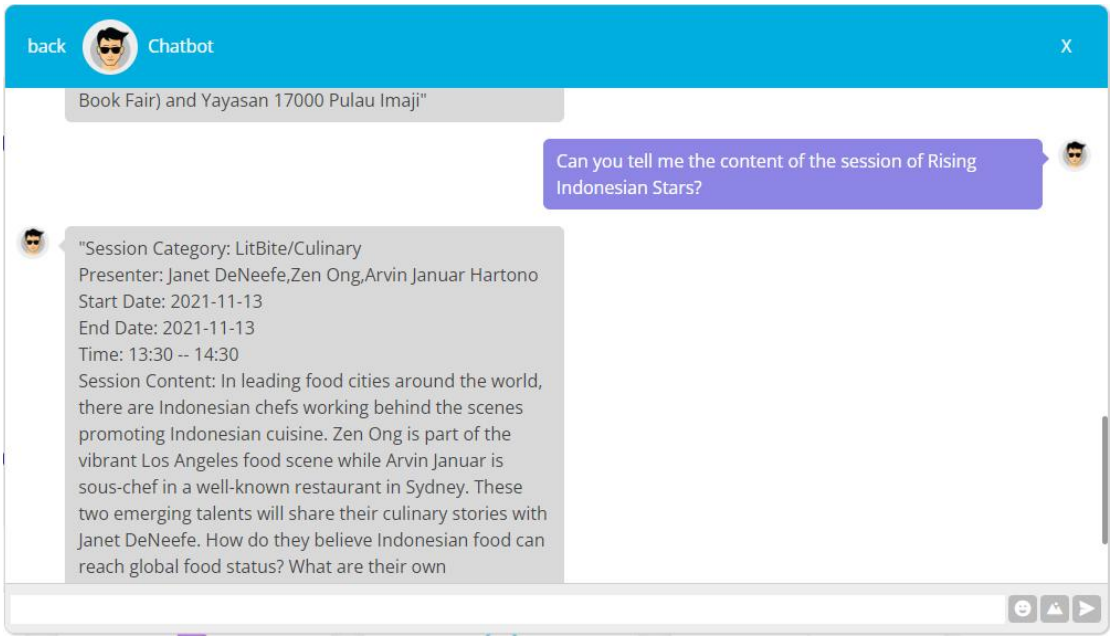




⑥If you have other questions that want to ask, you can directly enter the questions you want to ask in the input box. After clicking send, the chatbot will automatically give a reply.


back
Chatbot
X
Trade(LitTrade)
LitBeat(Creative)
Film(LitFilm)
Icon(Pop culture)
The Market
Frankfurt Rights
Translator Forum
LitBite(Culinary)
In LitBite(Culinary), renowned chefs and culinary experts will share their passion and insight with the audiences. They will engage in talks, tasting, and cooking demos. Join us, and expand your knowledge on new healthy recipes, starting up your culinary business, nutrition, and sustainable ingredients.
See more details about LitBite
Who is the organizer of Jaktent website?

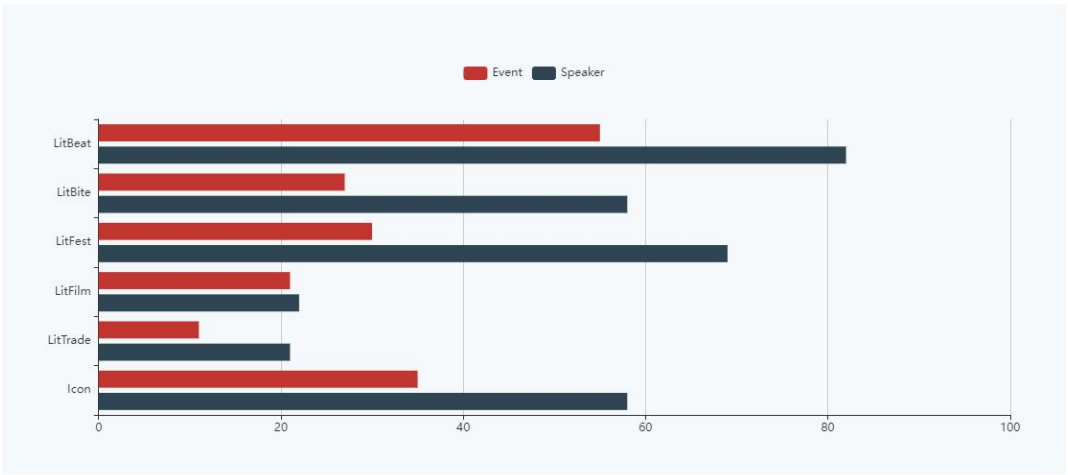

back
Chatbot
X
Frankfurt Rights
Translator Forum
LitBite(Culinary)
In LitBite(Culinary), renowned chefs and culinary experts will share their passion and insight with the audiences. They will engage in talks, tasting, and cooking demos. Join us, and expand your knowledge on new healthy recipes, starting up your culinary business, nutrition, and sustainable ingredients.
See more details about LitBite
Who is the organizer of Jaktent website?
"The organizers are Frankfurter Buchmesse (Frankfurt Book Fair) and Yayasan 17000 Pulau Imaji"



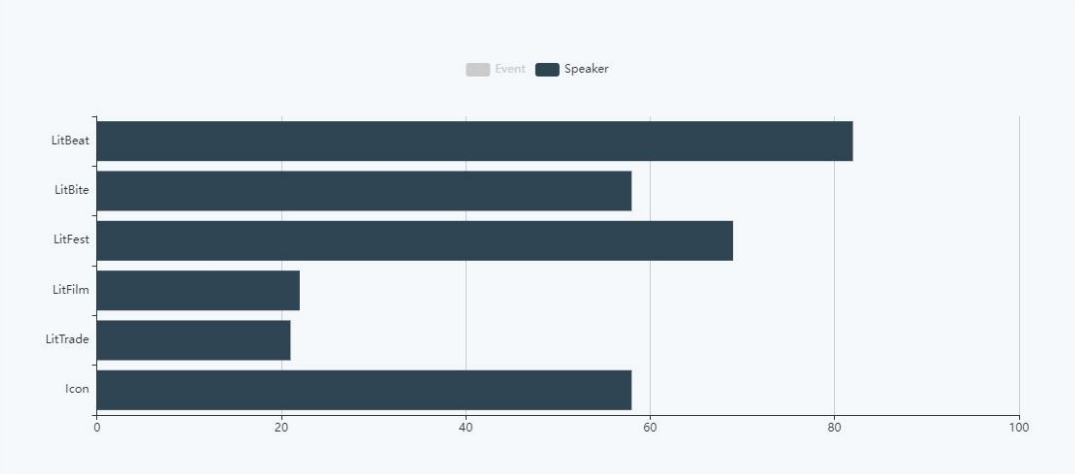
⑦For the Jaktent Index Website part, the website consists of five pages, the first is the main page, the second is the conference search page, the third is the conference and the number of speakers visual page, the fourth is the conference speaker search page, the last is the conference recommendation system page. Click the content in the red box or the word in the upper right corner to jump to the page.

⑧The following figure is the conference search box of Jaktent. You can enter the conference name or keyword in the red box. After entering, click the Search button, the corresponding conference name, type, date, start and end time and introduction will be displayed below. The same goes for the speaker interface.

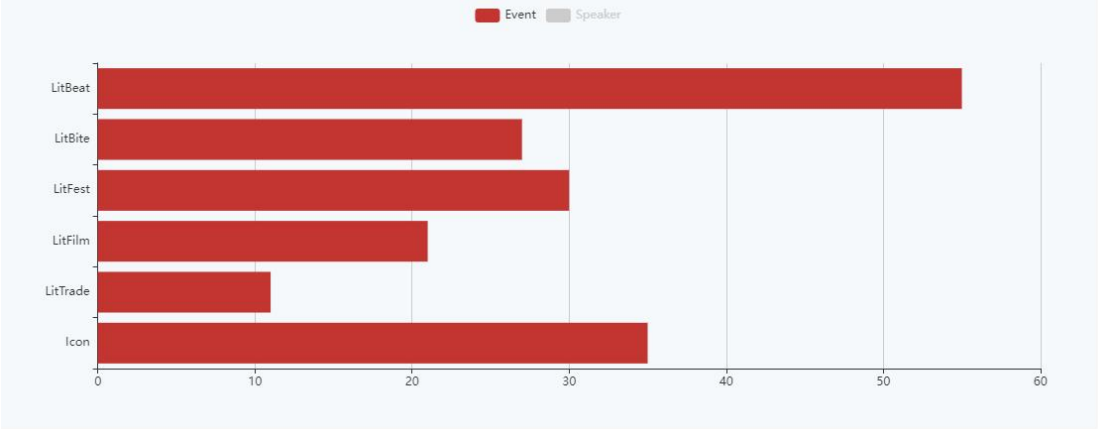
⑨The following interface is a visual interface of the number of sessions and speakers. Click the two interfaces in the red box to display the number of sessions or speakers separately, which shows the number of each session type.



The Number of Events and Speakers



The Number of Events and Speakers



⑩

JAKTENT

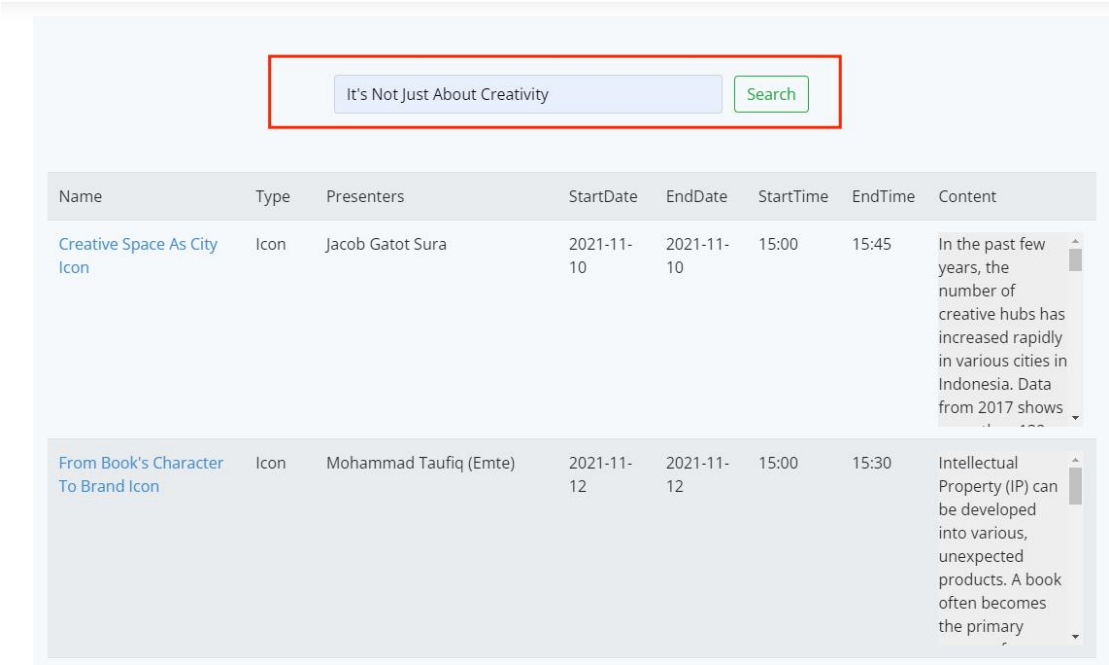
Home Event Statistic Speaker System

Guess You Like

The interface displays the top ten hot frequency words in the meeting tag.

- Filmmaking
- Literature
- Digital-technology
- Food-industry
- Pandemic
- Social-media
- Children
- Book-industry
- Virtual-film-class
- Feminism

The ten words shown in the figure are the ten tags with the highest frequency obtained through word frequency analysis. If you click the words in the red box, the page will jump to the session presentation page with similar tags.



Enter the name of the meeting you are interested in in the red box, and the meeting with similar content will be displayed below.

Important Project Technical Details Instruction

1. Jaktent Chatbot Part

1.1. Chatbot Front-end Interface and Rule-based Part

(1) If you want to integrate Jaktent Chatbot Interface on other web pages, the files we need are :

A. static/iconfont.css

B. static/chat.css

C. static/jquery-3.4.1.min.js

D. Templates/kefu.html

By locating or moving **"kefu.html"**, you can place the chatbot on the web page you want.

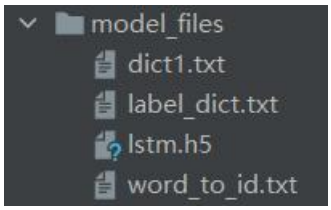
(2) For the FAQ(Frequently Asked Question) Settings, You can find in **"kefu.html"** and you need to find the following modules as shown:

```
if (textContent.indexOf('frequently asked questions') != '-1') {
    var re =
        'frequently asked questions:' +
        '<br/>' +
        "<a href='javascript:;' >What is  the Jaktent?</a>" +
        '<br/>' +
        "<a href='javascript:;' >What's on Jaktent?</a>" +
        '<br/>' +
        "<a href='javascript:;' >Who can join Jaktent?</a>" +
        '<br/>' +
        "<a href='javascript:;' >How do I participate as an attendee?</a>" +
        '<br/>' +
        "<a href='javascript:;' >I am interested in a free session, how can I watch it?</a>"
        '<br/>' +
        "<a href='javascript:;' >I have questions regarding Jaktent that can't be found here,
    reply(re)
} else if (textContent.indexOf('What is  the Jaktent?') != '-1') {
    var re =
        'Jakarta Content Week (Jaktent) is an international event comprising conferences, wor
    reply(re)
} else if (textContent.indexOf("What's on Jaktent?") != '-1') {
    var re =
        'We have seven types of events for you: Litbeat, Litfest, Litbite, Litfilm, Litrade,
        '<br/>' +
        "<a href='javascript:;' >LitBite(Culinary)</a>" +
        '<br/>' +
        "<a href='javascript:;' >LitFest(Literary)</a>" +
        '<br/>' +
        "<a href='javascript:;' >Trade(LitTrade)</a>" +
        '<br/>' +
```

If you want to add or modify the frequently asked questions, you can set the basic questions you want to set by modifying the value of XXX in **"textContent.indexof('XXX')"**

1.2. Custom Chatbot(Back-end)

In the whole project, the part related to the custom chatbot mainly involves two files: **"train.py"** and **"chat_predict.py"**. **“train.py”** is used to train the chatbot model. The trained model is saved in the **"model_files"** folder. The main function of **"chat_predict.py"** is to load the trained model and predict the question answer. When the user enters a question in the input box at the front-end, the data will automatically return to the back-end. The back-end calls the model to give the answer to the corresponding problem input by users. The functions related to front-end and back-end interaction are uniformly written in the **"app.py"** file, the part related to the custom chatbot is the **"update_task()"** function.

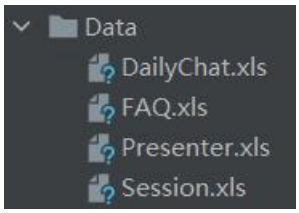


“model_files” Folder

```
@app.route('/predict1', methods=['POST'])
def update_task():
    # Solve the cross-domain problem of data transmission when the front-end and back-end are separated and developed
    CORS(app, supports_credentials=True,resources=r'/*')
    input = request.form.get('input')
    result = chat_predict.predict([input])
    result = result.replace('\n', '<br/>')

    # connect the database
    conn = sqlite3.connect("Jaktent.db")
    cur = conn.cursor()
    sql = "insert into recording(question,recording) values(?,?) "

    cur.execute(sql,(input,result))
    conn.commit()
    cur.close()
    conn.close()

    return jsonify({'result':result})
```

“update_task()” Function in “app.py” Files

Because it is a custom chatbot, we constructed a dataset based on the actual situation of the Jaktent conference website. The dataset contains the following four Excel files: **"DailyChat.xls"** mainly involves some dialogue materials commonly used in daily life; **"FAQ.xls"** mainly involves some corpus of Frequently Asked Questions related to Jaktent; "Presenter.xls" mainly involves the corpus related to the session presenter; "Session.xls" mainly involves corpus related to each session.



Because the performance of the custom chatbot is closely related to the corpus dataset, only by constantly improving the corpus dataset can the performance of the chatbot be better. Since future client and developer will definitely involve the problem of modifying the corpus dataset. Let's take "FAQ.xls" as an example to introduce how we constructed the dataset.

As shown in the following two figures, when you open the "FAQ.xls" file, you will find that there are two sheets, each Excel file has the same structure.

Looking at **Figure 1**, the main column fields involved in this table are **"Standard Question"** and **"Standard Question Answer"**. **"Standard Question"** is mainly some common standard questions that we have sorted out, and there are also corresponding answers. Then we focus on **Figure 2**, this table involves **"Standard Question"** and **"Non-standard Question"**, **"Non-standard Question"** means for one question, users may use different expressions to ask. There are countless **"Non-standard Questions"** for each **"Standard Question"**, because you can't estimate which expression users will use. This is also the biggest problem with chatbots, so all we can do is to continuously improve the corpus dataset as much as possible.

	A	B	C	D
1	Standard Question ID	Standard Question	Standard Question Answer	URL
2	1	What is Jaktent?	Jakarta Content Week (Jaktent) is an international event comprising conferences, workshops, seminars, festivals, and exhibitions. Jaktent will bring together and showcase various sectors in the creative industries, with a focus on the Asia Pacific region. Jakarta Content Week aims to be the new hub for creative content and rights trading in Asia Pacific, focusing on license for publishing products and intellectual property rights from cultural and creative sectors in the thriving region. Our tent is open for global and local players, from established companies to small and medium enterprises.	
3	2	What's on Jaktent?	We have seven types of events for you: Litbeat, Litfest, Litbite, Litfilm, Litrade, The Market, and the Cosplay championship(Icon).	
4	3	When and where can I visit?	Due to the COVID-19 pandemic, 2020 Jaktent will be an online event that you can stream from the comfort of your couch. Register at jaktent.com to enjoy our sessions. Our week-long event will be held on November 11-15, 2020.	
5	4	Who are the organizers?	The organizers are Frankfurter Buchmesse (Frankfurt Book Fair) and Yayasan 17000 Pulau Imaji	
6	5	Where can I see Jaktent's complete programming and event schedule?	You can check out our website, jaktent.com, and our social media accounts.	
7	6	What language will Jaktent be held in?	Some of our events will be held in English, while some others will be held in Bahasa Indonesia.	
8	7	Is Jaktent a book fair or literary festival?	Partly. We have LitBeat and LitFest for literature enthusiasts, but we are not limiting ourselves to this. Our main focus is content and all things that are based on creativity and have intellectual property. Check out our programming and we guarantee that there's always something for everyone.	See All Programs
9	8	Who can join Jaktent?	Everyone who is anyone in the business of content and intellectual property can join us. Our tent is open for global and local players, from established companies to small and medium enterprises.	Register as an exhibitor here
10	9	How do I participate as an attendee?	Go to our registration page and choose “attendee”. You will be directed to a page where you can select the programs that you are interested in, or to buy the day pass for Jaktent. We strongly encourage you to buy the day pass as you will get to enjoy more programs with better pricing!	
11	10	Do I need a ticket to attend?	Yes. We want to make sure that you'll have the best experience during 2020 Jaktent, so please register first on the website and go buy a ticket.	
12	11	I am interested in a free session, how can I	Go to our registration page and choose “attendee”. Free sessions will be marked as free and you don’t need to pay for the tickets. Please note that you still have to complete the whole order process.	

Sheet1 Sheet2

Figure 1

	A	B	C
1	**Standard Question ID**	**Standard Question**	**Non Standard Question**
2	1	What is Jaktent?	Jaktent
3	1	What is Jaktent?	Jakarta Content Week
4	1	What is Jaktent?	What's Jaktent?
5	1	What is Jaktent?	What's Jakarta Content Week?
6	1	What is Jaktent?	What is Jakarta Content Week?
7	1	What is Jaktent?	I want to know Jaktent.
8	1	What is Jaktent?	I want to know something about Jaktent.
9	1	What is Jaktent?	I want to know something about the website.
10	1	What is Jaktent?	I want to know something about website.
11	1	What is Jaktent?	I want to know more about Jaktent.
12	1	What is Jaktent?	I want to know more about the website.
13	1	What is Jaktent?	I want to know more about website.
14	1	What is Jaktent?	Learn something about website
15	1	What is Jaktent?	Learn something about the website
16	1	What is Jaktent?	Learn more about website
17	1	What is Jaktent?	Learn more about the website
18	1	What is Jaktent?	Learn about website
19	1	What is Jaktent?	Learn about the website
20	1	What is Jaktent?	Learn something about Jaktent
21	1	What is Jaktent?	Learn about Jaktent
22	1	What is Jaktent?	Learn more about Jaktent
23	1	What is Jaktent?	I want to learn more about website.
24	1	What is Jaktent?	I want to learn more about the website.
25	1	What is Jaktent?	I want to learn something about website.
26	1	What is Jaktent?	I want to learn something about the website.
27	1	What is Jaktent?	I want to learn about website.
28	1	What is Jaktent?	I want to learn about the website.
29	1	What is Jaktent?	I want to learn something about Jaktent.
30	1	What is Jaktent?	I want to learn more about Jaktent.
31	1	What is Jaktent?	I want to learn about Jaktent.
32	1	What is Jaktent?	Give some information about the website.

Sheet1 Sheet2

Figure 2

If you want to further improve the corpus dataset in the future, you can continue to add data in each sheet. Other Excel files are the same. There are two points that need to pay attention:

(1) When adding a **“Standard Question”**, several **"Non-standard Questions"** corresponding to the **"Standard Question"** must also be added.

(2) The **"Standard Question ID"** column field cannot be empty, this is to ensure that the final questions can be one-to-one correspondence.

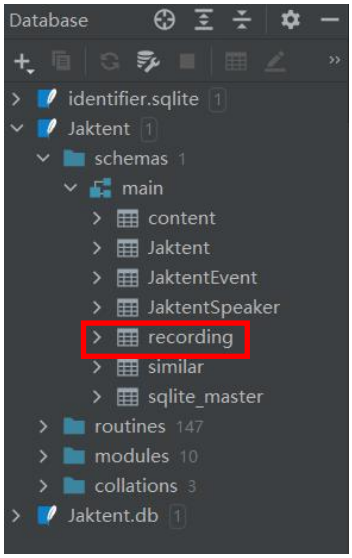
So if you want to modify and improve the dataset, the steps are as follows:

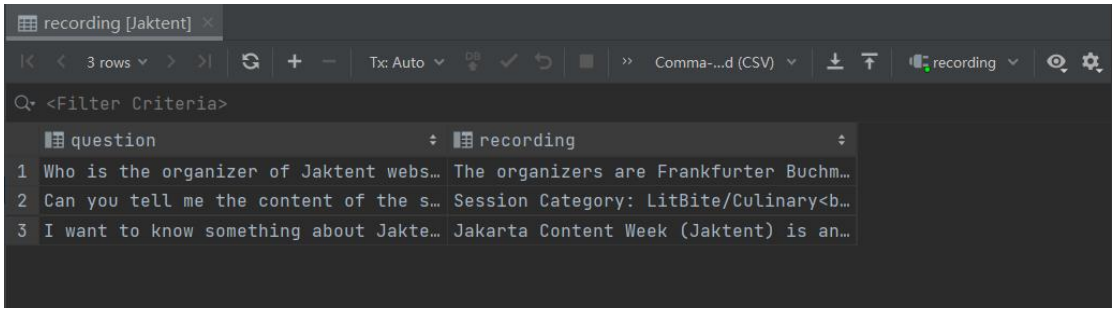
Modify and improve each dataset in Excel format → Run “train.py” file → Run “app.py” file → Use Jaktent Chatbot and Index Website

The figure below shows a sample of the training results

```
Epoch 1/10
625/625 [==============================] - 48s 75ms/step - loss: 6.3532 - acc: 0.0058 - val_loss: 6.3160 - val_acc: 0.0046
Epoch 2/10
625/625 [==============================] - 45s 72ms/step - loss: 6.3079 - acc: 0.0055 - val_loss: 6.3061 - val_acc: 0.0056
Epoch 3/10
625/625 [==============================] - 46s 73ms/step - loss: 5.3118 - acc: 0.0812 - val_loss: 3.4074 - val_acc: 0.3863
Epoch 4/10
625/625 [==============================] - 47s 75ms/step - loss: 1.7307 - acc: 0.6654 - val_loss: 0.7531 - val_acc: 0.8564
Epoch 5/10
625/625 [==============================] - 45s 73ms/step - loss: 0.5589 - acc: 0.8866 - val_loss: 0.3366 - val_acc: 0.9277
Epoch 6/10
625/625 [==============================] - 45s 72ms/step - loss: 0.2889 - acc: 0.9365 - val_loss: 0.2123 - val_acc: 0.9530
Epoch 7/10
625/625 [==============================] - 44s 70ms/step - loss: 0.2059 - acc: 0.9529 - val_loss: 0.1619 - val_acc: 0.9546
Epoch 8/10
625/625 [==============================] - 44s 70ms/step - loss: 0.1791 - acc: 0.9559 - val_loss: 0.1473 - val_acc: 0.9629
Epoch 9/10
625/625 [==============================] - 44s 70ms/step - loss: 0.1296 - acc: 0.9686 - val_loss: 0.1107 - val_acc: 0.9692
Epoch 10/10
625/625 [==============================] - 43s 69ms/step - loss: 0.0873 - acc: 0.9779 - val_loss: 0.0752 - val_acc: 0.9825

Process finished with exit code 0
```

In addition, we consider that after the project is set up online in the future, the background needs to have a function that can record the user's input questions and the corresponding answers, so that developers can find some wrong answers in time by checking these data regularly. And at the same time, they can modify and enhance corpus datasets. Therefore, we also set up a database that can record the questions input by users and the corresponding answers in real time. When users input a question on the front-end, the database can automatically record the question and the corresponding answer. The main database involved is the **"recording"** database, as shown below.





2. Jaktent Index Website Part

①

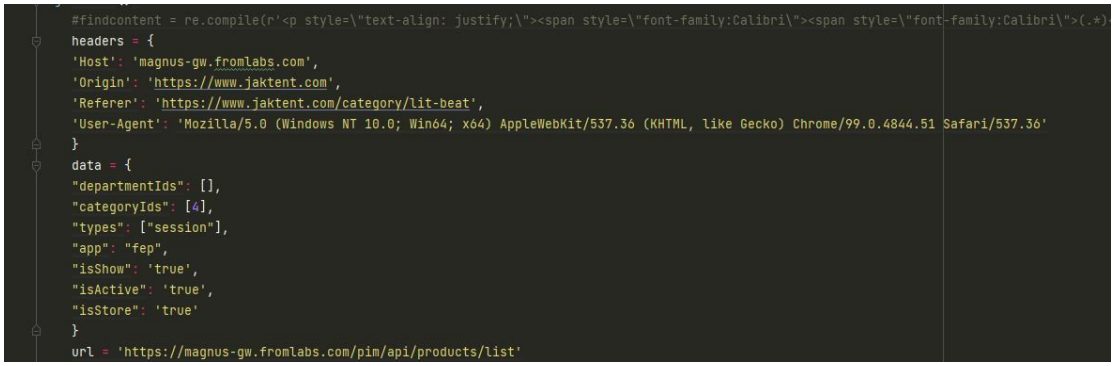
Libraries that need to be prepared in the early stage:

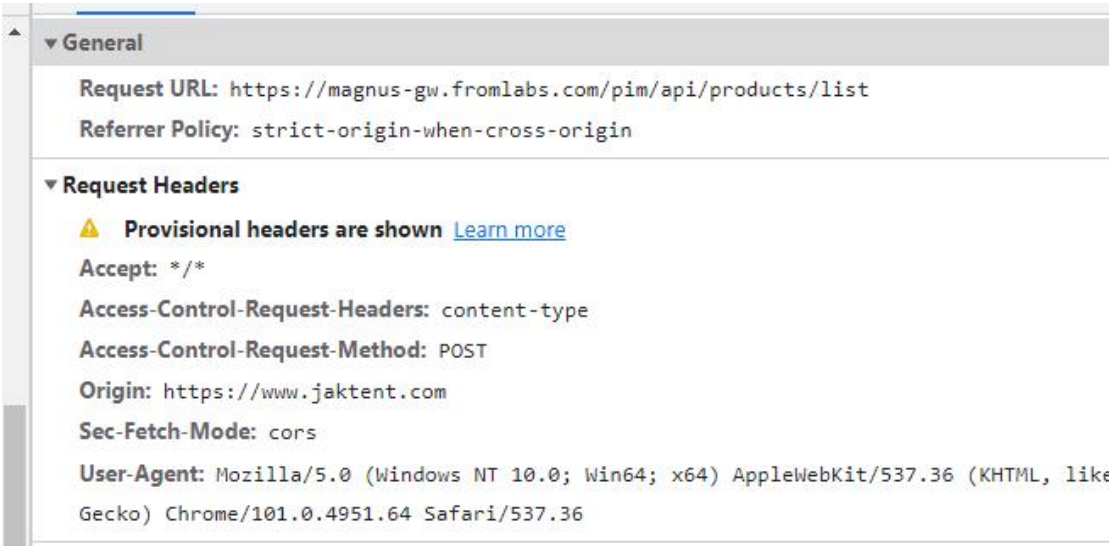
Requests: The Requests library is a Python third-party library that can be called to help us automatically crawl HTML web pages and automatically submit web Requests that simulate human access to the server.

Json: Mainly used to encode Python objects as JSON for output or storage, and to decode JSON objects into Python objects.

Xlwt: A common library for manipulating Excel in Python.

How do user change the content of a page crawled:





(1) Press F12 on the required web site to obtain the detailed HTTP request for the page to determine whether to use the POST or GET method to obtain the information of the page.

(2) Change user-agent, host, refer and other related information according to the actual information. In particular, if the request mode is POST, the corresponding data data needs to be passed in.

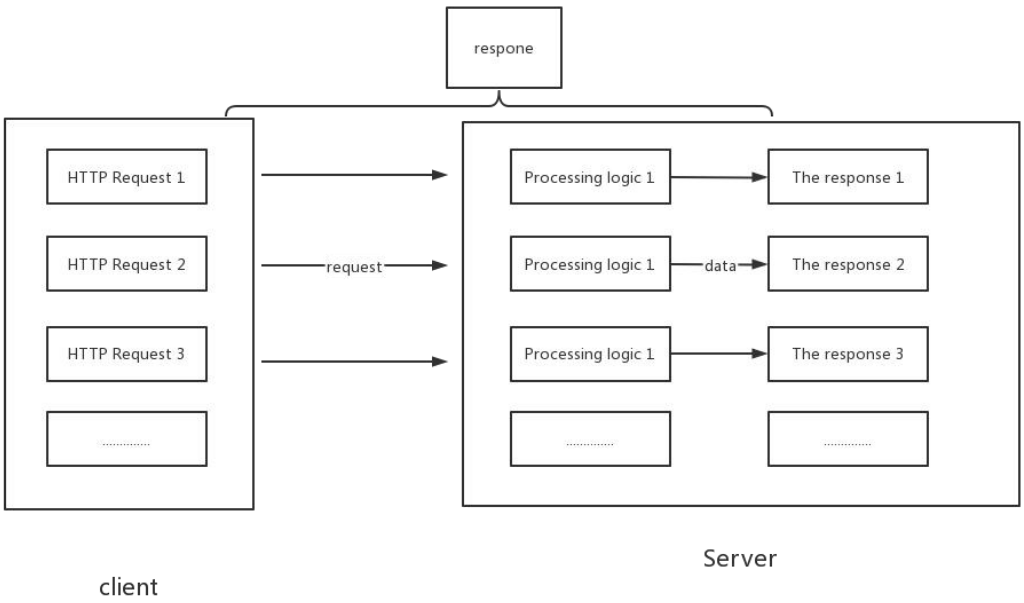
```
r = requests.post(url,json = data,headers = headers )
jss = json.loads(r.text)
book = xlwt.Workbook(encoding="utf-8", style_compression=0)
sheet = book.add_sheet('LitBeat', cell_overwrite_ok=True)
col = ("Name", "Presenters", "startDate", "endDate", "startTime", "endTime","content","URL")
for i in range(0, 8):
    sheet.write(0, i, col[i])
```

Pass the data into an Excel spreadsheet:

Loads are used to convert data from JSON to Python objects using the loads method in the JSON library, and to extract the required data information using the special properties of the Python dictionary format.

②

Flask framework use



All actions are initiated from the client.Flask's basic principles:

- All Flask programs must create an instance of the program
- When a client wants to obtain resources, it usually initiates an HTTP request through a browser
- At this point, the Web Server uses a WSGI (Web Server Gateway Interface) called the Web Server Gateway Interface.The Flask application instance receives requests from the client.
- Flask uses Werkzeug for route distribution (mapping between URL requests and view functions). Based on each URL request,Find the specific view function.

```python
@app.route('/movie')
def movie():

    datalist = []
    conn = sqlite3.connect("Jaktent.db")
    cur = conn.cursor()
    sql = "select * from JaktentEvent"
    data =cur.execute(sql)
    #print(data)
    for item in data :
        datalist.append(item)

    cur.close()
    conn.close()
    page = request.args.get(get_page_parameter(), type=int, default=1)
    limit = 10
    start = (page - 1) * limit
    end = start + limit
    res = datalist[start:end]
    pagination = Pagination(page=page, total=len(datalist), per_page=limit)
    return render_template("movie.html", movies=res, pagination=pagination)
```

In this project, the built-in SQLite database of PyCharm was used for data operation, and the data was extracted from the database and stored in the list. The contents of the list are transferred to the front HTML interface by the interaction of flask's front and back parameters.

At the same time, due to the large number of meetings, the flask_Paginate package was used in the project to paginate the front-end pages. If you need to change the number of presentation meetings on the same page, you can change limit =? (? The number of meetings displayed on the delegate page).

```
</div>
  <table class = "table table-striped">

      <tr>
          <td>Name</td>
          <td>Type</td>
          <td>Presenters</td>
          <td>StartDate</td>
          <td>EndDate</td>
          <td>StartTime</td>
          <td>EndTime</td>
          <td>Content</td>
      </tr>

      {% for movie in movies %}
      <tr>
          <td>
              <a  href="{{ movie[8] }}" target = "_blank" >
                  {{ movie[0] }}
              </a>
          </td>
          <td>{{ movie[1] }}</td>
          <td>{{ movie[2] }}</td>
          <td>{{ movie[3] }}</td>
          <td>{{ movie[4] }}</td>
          <td>{{ movie[5] }}</td>
          <td>{{ movie[6] }}</td>
          <td>
          <div id = d1>
              {{ movie[7] }}
          </div>
          </td>
      </tr>

      {% endfor %}

  </table>
  {{ pagination.links }}
```

This is the code for the front end, which is presented in the form of a table showing the name of the meeting, the time, the speaker, and a brief introduction of the content. If you need to make changes later, contact the Python code at the back end to make changes together.

③

JAKTENT

Home Event Statistic Speaker System

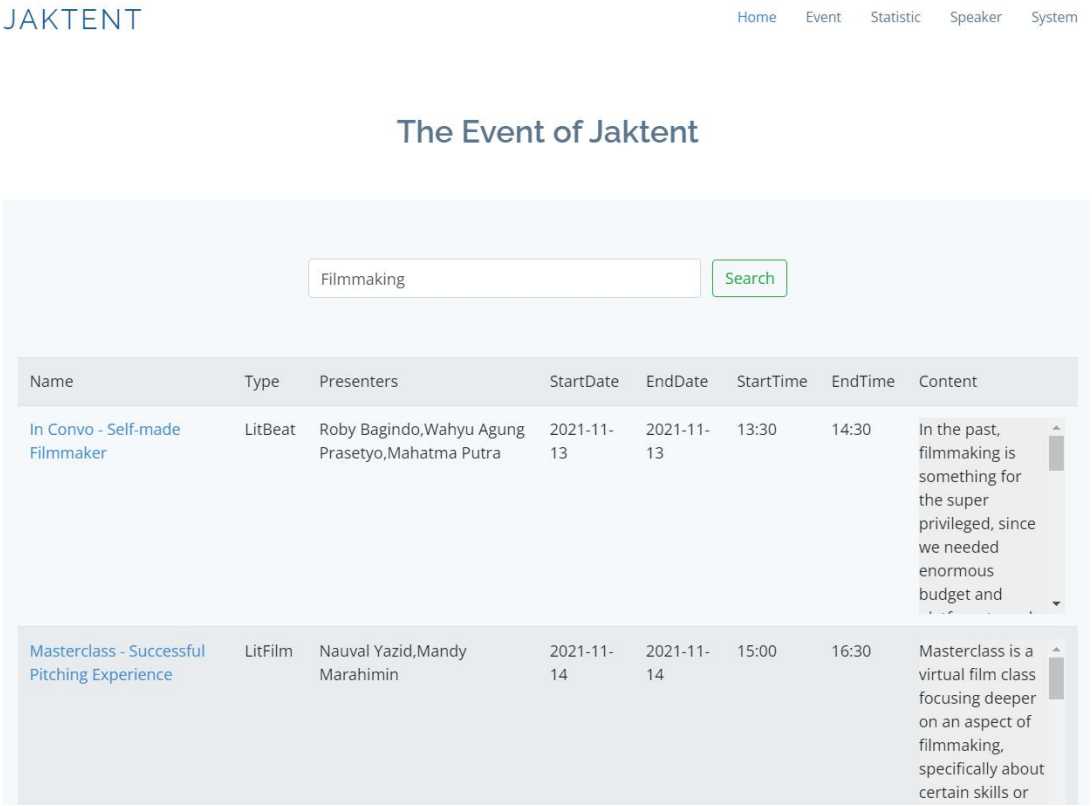
Guess You Like

Magnam dolores commodi suscipit. Necessitatibus eius consequatur ex aliquid fuga eum quidem. Sit sint consectetur velit. Quisquam quos quisquam cupiditate. Et nemo qui impedit suscipit alias ea.

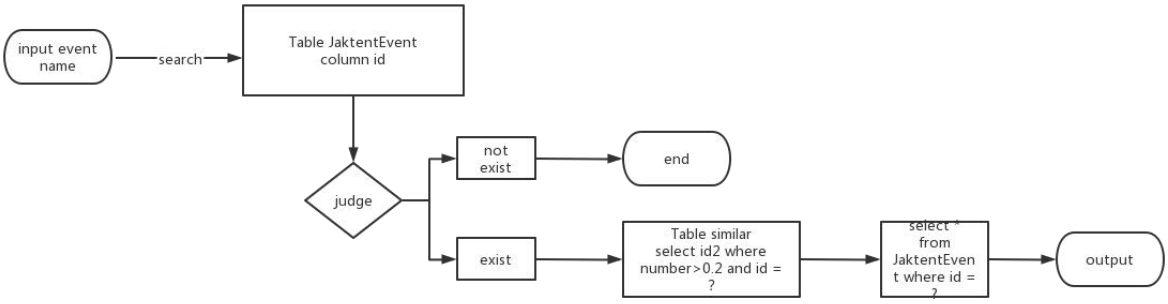
- Filmmaking
- Literature
- Digital-technology
- Food-industry
- Pandemic
- Social-media
- Children
- Book-industry
- Virtual-film-class
- Feminism

It's Not Just About Creativity Search

Name	Type	Presenters	StartDate	EndDate	StartTime	EndTime	Content
Creative Space As City Icon	Icon	Jacob Gatot Sura	2021-11-10	2021-11-10	15:00	15:45	In the past few years, the number of creative hubs has increased rapidly in various cities in Indonesia. Data



The recommendation system page is mainly divided into two sections. The first section is recommendation section, and the second section is hot words section. We first conduct natural language processing on the content of the meeting, and then get the key words of each meeting. After the keyword is obtained, word frequency analysis is carried out on the keyword. After analysis, ten keywords with the highest frequency were obtained, which were made into click form on the web page, and the conferences containing the keywords were selected for display. In addition, the word cloud technology is used to make the word cloud map of the meeting on the right. Make the user enter the interface after more intuitive experience. As for the recommendation section, due to the lack of users' browsing records, number of likes and other information, we can only make it through TF and Wordbag. The similarity between each two meetings was analyzed and stored in a database. At the same time, set the search box, and enter the meeting you are interested in. Similar meeting descriptions are displayed.




```python
@app.route('/re')
def re():
    a = request.args.get("b")
    print(a)
    datalist = []
    rdata = []
    pdata = []

    conn = sqlite3.connect("Jaktent.db")
    cur = conn.cursor()

    sql2 = "select * from JaktentEvent"
    data2 = cur.execute(sql2)
    for item in data2:
        if a in item[0]:
            rdata.append(item)

    num = rdata[0][9]


    sql1 = "select id2 from similar where number > 0.2 and id = ?"
    v = (num,)
    data1 = cur.execute(sql1,v)
    for i in data1:
        datalist.append(i[0])
    print(datalist)

    for item in datalist:
        sql3 = "select * from JaktentEvent where id = ? "
        v = (item,)
        data3 = cur.execute(sql3,v)
        for a in data3:
            pdata.append(a)

    cur.close()
    conn.close()

    return render_template("wordcloud.html",movies = pdata )
```

```
def bow_cosine(s1, s2):
    vectorizer = CountVectorizer()
    vectorizer.fit([s1, s2])
    X = vectorizer.transform([s1, s2])  # Get the vector of s1 and s2 represente
    #print(X.toarray())
    a = cosine_similarity(X[0], X[1])
    b = a[0].tolist()
    print(b[0])
    #print(type(b[0]))

    #print(cosine_similarity(X[0], X[1]))

def way():
    datalist = []
    conn = sqlite3.connect("Jaktent.db")
    cur = conn.cursor()
    sql = "select content from content"
    data = cur.execute(sql)
    for item in data:
        datalist.append(item[0])
    cur.close()
    conn.close()
    # print(datalist)
    # for item in datalist :
    #     print(item)
    data = []
    for i in range(len(datalist)):
        for j in range(i+1,len(datalist)):
            bow_cosine(datalist[i],datalist[j])

way()
```

● Similarity calculation method based on text feature

(1) Convert the text to Feature Vectors. Tf-idf is used to obtain feature vectors. The vector dimension is the dictionary size, and each dimension of the vector is the TF-IDF value calculated in the text of the word at this position in the dictionary. If it does not appear in the text, it is 0.

(2) Feature vectors are obtained by bag of Words. The vector dimension is the dictionary size, and each dimension of the vector is the frequency of occurrence of words at this position in the dictionary in the text, and 0 if they do not appear in the text.

(3) Feature Vectors were used to calculate the similarity between texts. Cosine similarity can be used to calculate the similarity of two texts based on their feature vectors.

Get the vector of s1 and s2 represented by bag of words.

Using cosine similarity, the similarity of two texts is calculated based on their feature vectors.

Connect to the database, obtain the label name of each meeting from the database, compare the label values of each two meetings, and save the similarity values in the

database.

At present, the similarity is directly calculated and stored in the database to speed up the operation of the project

Word frequency analysis part：

```
article
article.replace(",","").replace(".","").replace(":","").replace(";","").repl
ace("?","")
```

Replace punctuation with Spaces

```
exchange = article.lower();print(exchange)
```

Uppercase letters are converted to lowercase letters

```
list = exchange.split()print(list)
```

Generating word lists

```
dic = {}for i in list:
    count = list.count(i)
    dic[i] = countprint(dic)
```

Generate word frequency statistics

```
word = {'and','the','with','in','by','its','for','of','an','to'}for i in word:
    del(dic[i])print(dic)
```

Eliminate specific words

```
dic1= sorted(dic.items(),key=lambda d:d[1],reverse= True)print(dic1)
```

Sort

```
for i in range(10):
    print(dic1[i])
```

Output the top ten words with the highest frequency