

Machine Learning Based GDPR Database Sanitization Recommendation System

18 Fall CMPE-272 Project Report

Yuwen Li

*Department of Computer Engineering
San Jose State University*

San Jose, USA

yuwen.li01@sjsu.edu

Gaochao Wang

*Department of Computer Engineering
San Jose State University*

San Jose, USA

gaochao.wang@sjsu.edu

Abstract - Due to GDPR and other privacy regulations, there is increasing need for organization to identify and destroy data. This is something will have high impact/value on our platform. The report discusses and reflects on one method for data sanitization recommendation we implemented. Our system provides GUI and can simply search the relevant data and give user to option to delete or not. we use NLP tools and TF-IDF algorithm to do the recommendation. The source code is available at <https://github.com/SJSU272LabF18/Project-Team-29>

1. Introduction

Data breach is happening more frequently than ever in recent years. Society are paying attention to personal information privacy and protection. Data privacy and protection became a hot topic and was widely disguisedly, which affect many stakeholders such as business operators, privacy organization, scholars, government regulators, and individual customers. It is reported that privacy is one of the largest concerns for consumers in Public Opinion polls[1].

Information privacy control is about the data collection and dissemination in terms of technology, public expectation, legal and political issues. The two major category of private information are personally identifiable information and sensitive information, like age, occupation, salary, medical records. These data should be securely collected, stored, used, and finally destroyed or deleted based on clear user privacy agreement.

GDPR, General Data Protection Regulation, is a data protection and privacy regulation in EU law for all

individuals within the European Union (EU) and the European Economic Area (EEA), as well as the export of personal data outside the EU and EEA areas. The goal of GDPR is mainly to give the control to people over their personal data and to simplify the regulatory environment for international business by unifying the regulation within the EU.[2] The regulation states that personal data collection must be clearly disclosed and declared on the lawful basis to state the purpose, time period of data being retained and whether it is being shared with any third parties or outside of the EEA. The person of the data have the right to request a portable copy of the data, and the right to have their data deleted under certain situations.[3]

To solve these data privacy issue, multiple fields have involved, including computer security, data security, and information security. Among them, one of the striking problem is the existing data stored in the database of business premises. To search and identify the targeted data, a powerful and broadly suitable tool is strongly needed. This project proposed a solution and designed a software for this purpose. We used a NLP(nature language processing) machine learning algorithm on user input keywords to search all the related data in a given database, and enabled the deletion function for data sanitization.

2. Background

Data sanitization is to inspect the data on a memory device to remove or destroy target data permanently on purpose. Nowadays there are several ways to process the sanitization, such as software, hardware device and physical mechanism that destroys the device and the data in it so its data cannot be recovered. Some popular software tools are available for file and folder sanitization and disk Sanitization[4][5]. However, not a few software tool

is found for database sanitization. This is the motivation for us to build a system to target database sanitization.

3. Project Analysis

The project was aimed to solve the real life problem of GDPR data privacy in business premises. To build a useful and successful software tool, two aspects of a data sanitization recommendation system are important to be thoroughly considered and designed. One is user easiness operating functionality, the other is a complete and accurate data searching method for a database. In our project, we choose to use NLP machine learning method to process user input, which gives user the convenience of entering keywords only and the system will be able to find all related data in the target database.

Today, many factors that affect the quality of search engines, in addition to the user's click data, can be summarized into the following four aspects:
(1) Complete index. (2) A measure of the quality of a web page, such as PageRank. (3) User preferences. (4) A method of determining the relevance of a web page to a query. As a Sanitization Recommendation System, it has to determine the most relevant data for user for future sanitization.

4. Technique approach

Our system has two main part: front-end, which provides GUI functionalities to user, back-end, which process the data and do the sanitization. Multi-threading is applied for Frontend and Backend to achieve functionality and performance.

4.1 Front-end

In this project, we use python built in GUI library, tkinter and ttk to implement the UI functionalities. There are total 4 pages to interact with users, as shown below.

4.1.1 Home Page

(1) Provide user to enter database connection information(IP, Port, Database Name, Username and Password. (2) Check database connection. (3) Keywords entry and database scanning start button (4) Optional function to output result file



Fig. 1. System Home Page

4.1.2 Load Page

GUI waiting animation is presenting to users while backend is processing data sanitization with keywords.



Fig. 2. System Load Page

4.1.3 Result Page

(1) A summary of total scanned database tables and data rows in two scales of clean data quantities and suspected data quantities with arrow moving animation. (2) Total scan time. (3) Detail result button

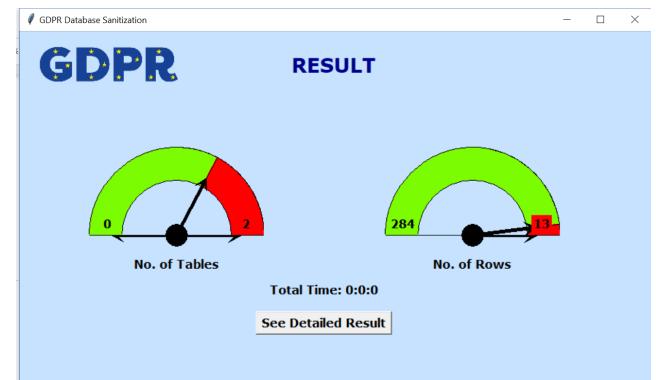


Fig. 3. System Result Page

4.1.4 List Page

(1) List related data rows of a table in GUI. (2) Allow user to browse all data in multiple tables, (3) Allow user to select multiple data row and remove them in database. (4) Allow user to re-arrange data row position in GUI

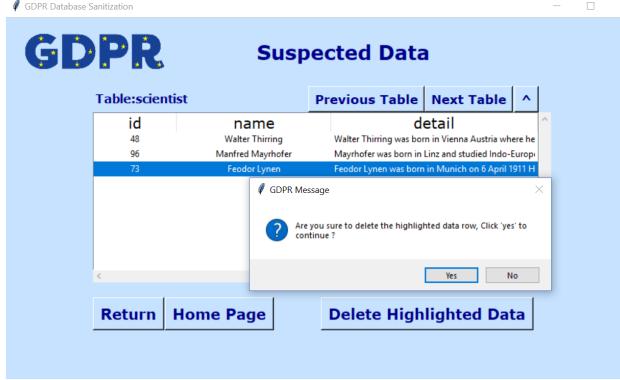


Fig. 4. System List Page

4.2 Back-end

4.2.1 Database

We use MySQL as target database, which has many library of API to communicate with our program.

4.2.2 Algorithm

We use nltk library to process nature language, in [7] TF-IDF to process the dataset and user's input. TF-IDF is an abbreviation of Term Frequency - Inverse Document Frequency, which is "word frequency - inverse text frequency". It consists of two parts, TF and IDF. The previous TF is also the word frequency we mentioned earlier. The vectorization we have done before is also the frequency of occurrence of each word in the text, and as a text feature, this is well understood. The key is how to understand this IDF, the "inverse text frequency". In the previous section, we talked about the "to" of almost all texts. Although the word frequency is high, the importance should be lower than the "China" and "Travel" with low frequency. Our IDF is to help us to reflect the importance of the word, and then to correct the word eigenvalues expressed only by word frequency. In summary, IDF reflects the frequency with which a word appears in all text. If a word appears in a lot of text, its IDF value should be low, such as "to" above. And conversely if a word appears in less text, its IDF value should be high. For example, some professional terms such as "Machine Learning". Such a word IDF value should be high. In an extreme case, if a word appears in all texts, its IDF value should be zero. The above is the role of the IDF described qualitatively. How to quantify the IDF of a word?

The basic formula for directly giving the IDF of a word x is as follows:

$$IDF(x) = \log \frac{N}{N(x)}$$

Where N represents the total number of texts in the corpus, and N(x) represents the total number of texts in the corpus containing the word x. The reason why is the basic formula of IDF supposed to be the same as above instead of N/N(x) is this involves some knowledge related to information theory. The above IDF formula is already available, but there are some minor problems in some special cases. For example, a certain uncommon word is not in the corpus, so our denominator is 0, and IDF has no meaning. So the commonly used IDF we need to do some smoothing, so that the words that do not appear in the corpus can also get a suitable IDF value. There are many ways to smooth, and one of the most common IDF smoothing formulas is:

$$IDF(x) = \log \frac{N + 1}{N(x) + 1} + 1$$

With the definition of IDF, we can calculate the TF-IDF value of a word:

$$TF - IDF(x) = TF(x) * IDF(*)$$

Where TF(x) refers to the word frequency of the word x in the current text.

After we got the original dataset TF-IDF value and user input words value, we simply Multiplied them, the result is the score that input matched the dataset.

4.2.3 Sanitization method

After the system finished matching, the GUI would show the result and the user is able to select the data he/she would delete. As soon as the back-end module get the feedback from the user, it can analysis the data and find the exact data in database, then delete the data.

4.3 Dataset

Scientist biography from Wikipedia

The table has 3 columns: id, name, detail

id	name	detail
66	In 1952	he was called to Vienna University as professor of Greek history, ancient history and epigraphy.
65	He acted as professor...	Heidelberg (1936) and Graz (1940).
21	Carl Friedrich von W...	A member of the prominent Weiszäcker family, he was son of the diplomat Ernst von Weiszäcker, elder brother of the former German Foreign Minister Hans-Dietrich Genscher.
21	Horst Dreier	A native of Hainichen, Lower Saxony, Dreier studied the violin at the Hochschule für Musik und Theater in Hannover, where he received his Staatsexamen in 1981. He played as a soloist in the orchestra of the Hannover State Opera.
35	Wolfgang Widerlik	After earning a doctorate in engineering at the Technical University in Graz, Hans List was appointed to the Tongji University in Shanghai, China, in 1982.
87	Hans List	After earning a doctorate in engineering at the Technical University in Graz, Hans List was appointed to the Tongji University in Shanghai, China, in 1982.
51	Bruno Snell	After studying law and economics at University of Edinburgh and University of Oxford, Snell gained interest in classical studies and philosophy.
61	Alfred Kubin	Alfred Leopold Iseid Kubin (April 1877 – 20 August 1959) was an Austrian printmaker, illustrator and occasional writer. Kubin was born in Prague, then part of the Austro-Hungarian Empire, and died in Vienna.
9	Nobuyoshi Araki	Araki was born in Tokyo on May 25, 1940. [4] He studied film cinematography at Chiba University from 1969, receiving a degree in 1973. After graduation, he worked as a cameraman for the Japanese television network NHK. From 1974 to 1976, he worked as a cameraman for the Japanese magazine "Animage".
36	H. C. Albrecht	Albrecht was born in 1900 in Vienna, Austria. He studied at the University of Vienna and the University of Graz. In 1924, he moved to Berlin, Germany, where he worked as a cameraman for the studios of Ufa.
74	Anton Farkas	Farkas has sung with leading orchestras including the Philharmonics of Vienna and London, Gewandhaus Leipzig, Radio-FM, and the Vienna State Opera.
77	Anny Falbemayr	Born Anna Maria Falbemayr-Székely in Vienna;[1] a family of craftsmen, she attended a Handelschule. She studied piano and singing.
94	Ernst Krenek	Born Ernst Heinrich Krenek in Vienna (then in Austria-Hungary), he was the son of a Czech soldier in the Austro-Hungarian army. ♀
109	Henryk Wolfson	Born in 1934 in Vienna, Henryk Wolfson remained intimately connected to the city of his birth. Between 1952–1957, he studied histology at the University of Vienna, under the supervision of Dr. Franz Schmidt.
44	Roland Rainer	Born in Klagenfurt, Roland Rainer dedicated himself to architecture from an early age. When he was 18, he so studied at the Vienna University of Technology.
49	Adolf Hitler	Born in Braunau am Inn, Austria, Hitler was the son of Alois Hitler, a customs official.
95	Wessem Franz	Born in Schärding, Austria, on January 21, 1900.[2] Franz studied mechanical engineering at the Graz University of Technology a year later.
75	Jerry Kurylowicz	Born in Stanislau, Austria-Hungary (now Ivano-Frankivsk, Ukraine), he is considered[who?] the most outstanding contemporary Polish scientist.
72	Carl Emil Schorle	Born in The Bronx, New York City, to Theodore Schorle and Gertrude Goldsmith. Schorle received his B.A. from Columbia in 1916.
2	György Ligeti	Born in Transylvania, Romania, he lived in Communist Hungary before immigrating to Austria in 1956. He became an Austrian citizen in 1961.
17	Ian Hacking	Born in Vancouver, British Columbia, Canada, he earned undergraduate degrees from the University of British Columbia (1956) an
81	Georg Kastner	Born in Vienna, he studied at the University of Vienna (1924–1928).
10	Marcus Rubin	Born in Vienna, where he studied with Richard Robert[1] and Franz Schmidt, he later emigrated to Paris, where he pursued further studies.
79	Bruno Gironcoli	Born in Villach, Gironcoli began training as a goldsmith in 1951 in Innsbruck, completing his apprenticeship in 1956. Between 1958 and 1960, he studied at the University of Applied Arts in Vienna.
67	Lorenz Böhler	Böhler is most notable as one of the creators—or even the creator—of modern accident surgery. He was the head of the AUVA-Hospital in Vienna.
4	Elias Canetti	Canetti moved to England in 1938 after the Anschluss to escape Nazi persecution. He became a British citizen in 1952. He is known as the son of Carl Leopold Cori [d] (1865, Brüx (Czech: Most), Bohemia, Imp.Austria-1954, Vienna), a zoologist, and Maria
43	Carl Ferdinand Cori	Carl was the son of Carl Leopold Cori [d] (1865, Brüx (Czech: Most), Bohemia, Imp.Austria-1954, Vienna), a zoologist, and Maria

Fig. 5. Table1 in database

Restaurant review from Yelp

The table has 6 columns: id, review_id, user_id, star, date, text

id	reviewerID	asin	reviewerName	helpfulness	star	text
1	xHD0BDSxEPQH0Dzw	mu2u7zJXZug0gb0Jsg		0.0	2011-02-25	The pizza was slow. Not the best I had. I prefer Saggio's on Flamingo / Fort Apache. The chef then left and the quality went down.
10	LWUtgPnHmMh0dLw	mrfa1z7Z_gxRqbg0Jsg		0.0	2012-11-19	Food is pretty good, no grime. But you have to make same effort to choose it. It like eating at home.
11	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-19	I've been coming here for years and town and city have changed. The food is still great. The service was GREAT. I am a CYCLES fan.
12	Fv1zqpl_1LSbsWytvA	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2014-07-05	I've been coming here for a while and I love how the food tastes the same every time. Delish!
13	AeHd0000000000000000	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-04-09	of 2 my girlfriends recommended this place as 3 of us went... ...made use to have an apheta we were all so happy to eat there again.
14	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-13	the service was great
15	sk9xV7yQgY7N9yMaWtA	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-13	The help was good! I got the mango toast and was a bit disappointed. The hate was very plain i
16	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-04-10	I have samples of the ice cream which are also very good, but I always go back to the Forno Red Rooster. I am not good at explaining why I want my hair since I always change my mind.
17	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
18	mc9yC5dPwBqJQ0DzTQ	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	not good at explaining why I want my hair since I always change my mind.
19	mc9yC5dPwBqJQ0DzTQ	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
20	ST1FM22w1z1FpBqJQ0DzTQ	TivuJChCzJxQwVg	TivuJChCzJxQwVg	0.0	2011-07-10	the service was good
21	D3TmNqQ0DzTQ	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
22	D3TmNqQ0DzTQ	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
23	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
24	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
25	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
26	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
27	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
28	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
29	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
30	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
31	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
32	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
33	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
34	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
35	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
36	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
37	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
38	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
39	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
40	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
41	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
42	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
43	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
44	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
45	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
46	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
47	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
48	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
49	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
50	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
51	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
52	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
53	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
54	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
55	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
56	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
57	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
58	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
59	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
60	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
61	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
62	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
63	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
64	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
65	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
66	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
67	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
68	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
69	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
70	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
71	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
72	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
73	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
74	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
75	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
76	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
77	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
78	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
79	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
80	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
81	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
82	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
83	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
84	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
85	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
86	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
87	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
88	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
89	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
90	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
91	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
92	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
93	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
94	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
95	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
96	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
97	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
98	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
99	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
100	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
101	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
102	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
103	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
104	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
105	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
106	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
107	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
108	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
109	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
110	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
111	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
112	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
113	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
114	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
115	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
116	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
117	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
118	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
119	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
120	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
121	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
122	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
123	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
124	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
125	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
126	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0	2011-07-10	the service was good
127	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	YJLjGKfCgkV0yv	0.0</td		

Fig. 6. Table2 in database

5. Result and Analysis

Input keyword(s): salary

Result: matched table: 0, matched raw:0

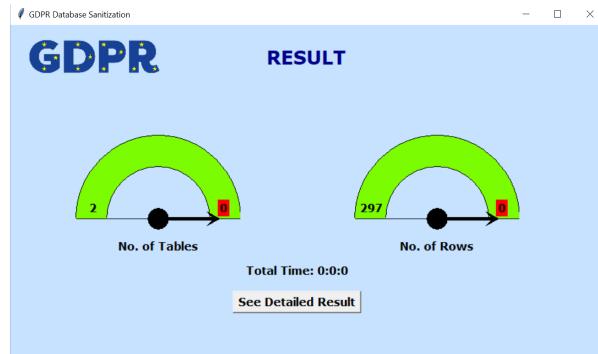


Fig. 7. Test result1

Input keyword(s): Senior, old, age
matched table: 2, matched raw:26

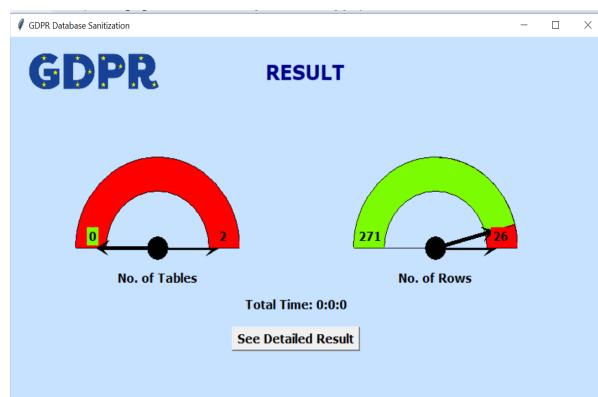


Fig. 8. Test result 2

Suspected Data			
Table:restaurant_review		Previous Table	Next Table
id	user_id	date	text
70	ogPvUe4_1p	2016-05-08	Comfortable seats happy they are not separated in clusters, groups like other th
84	ogPvUe4_1p	2016-05-08	Love the old school ambience Friendly staff Have been here several times ove
173	YHwLsB5gZz	2010-11-17	Huge Apple Danishes! Unfortunately huge doesn't mean better \nSo what does
114	YHwLsB5gZz	2010-11-22	Sometimes you ' ve ever driven by High Park on Bloor street seen the world-sign that
135	YHwLsB5gZz	2012-01-28	Sometimes I ride the 501 streetcar just for something to do it's a fun way to pet
147	YHwLsB5gZz	2011-04-20	Cupcakes have reached a point of utmost popularity in the world of baked good
133	YHwLsB5gZz	2011-03-07	For those that do n't know ONIers is that unique place where you dine entirely in
134	YHwLsB5gZz	2011-01-03	Having grown up enjoying pubs far and wide it does not take much to cajole me
166	YHwLsB5gZz	2011-10-29	I know St Lawrence market is full of great food but I never would 've thought th

Fig. 9. Test result 3

Table:scientist	Previous Table	Next Table	^
name	detail		
Walter Th.	Walter Thirring was born in Vienna Austria where he earned his Doctor of Physics degree in 1949 at the Günther W. Wilke's own area of interest focused on homogeneous catalysis by nickel complexes His group disco		
Manfred N.	Manfred N. Mayrhofer was born in Linz and studied Indo-European and Semitic linguistics and philosophy at the		
Feodor Lynen	Feodor Lynen was born in Munich on 6 April 1911 He started his studies at the chemistry department Valie Export Educated in a convent until the age of 14 Export studied painting drawing and design at the National Jerzy Kuryl Born in Stanislaw Austria-Hungary (now Ivano-Frankivsk Ukraine) he is considered [who] the most i		
Return	Home Page		Delete Highlighted Data

Fig. 10. Test result 4

Input keyword(s): george
matched table: 1, matched raw: 1

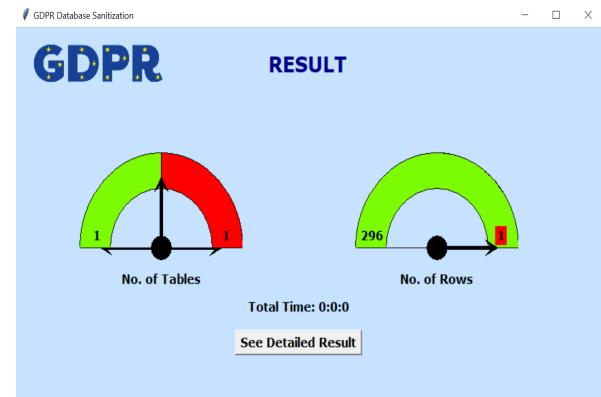


Fig. 11. Test result 5



Fig. 12. Test result 6

The result is showed to the user by the rank of the match score as a recommendation, since our algorithm is based on a statistic method, the key words would be 100% found and scored.

6. Conclusions and Further Research

Our system is easy to use as it provides GUI and has a good performance with matching the user's keywords and data in database, with using TF-IDF as the method are simple, fast, and easy to understand. However, the disadvantage is that sometimes the use of word frequency to measure the importance of a word in an article is not comprehensive enough. Sometimes important words may not appear enough, and such calculations cannot reflect positional information and cannot reflect the importance of words in context. For further research, we would try to embody the context structure of the word, then you may need to use the word2vec algorithm to support it. To meet the requirement of industry, we also need to increase our dataset into a large scale for validate the accuracy.

References

- [1] H. J. Smith, T. Dinev, H. Xu. *Information Privacy Research: An Interdisciplinary Review*. MN: University of Minnesota,2011
- [2] L. Thornton, N. Wallace. The EU General Data Protection Regulation: Implications for Research. 2018
- [3] NIBUSINESS, "General Data Protection Regulation " [Online]. Available: <https://www.nibusinessinfo.co.uk/content/data-subject-rights-under-gdpr> [Accessed Dec. 5, 2018]
- [4] Carnegie Mellon University, "Data Sanitization and Disposal Tools" [Online]. Available: <https://www.cmu.edu/iso/tools/data-sanitization-tools.html> [Accessed Dec. 5, 2018].

- [5] Lifewire, "40 Free Data Destruction Software Programs" [Online]. Available: <https://www.lifewire.com/free-data-destruction-software-program-s-2626174> [Accessed Dec. 5, 2018].
- [6] Scikit learn, "Feature extraction" [Online]. Available: https://scikit-learn.org/stable/modules/feature_extraction.html [Accessed Dec. 5, 2018].