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.

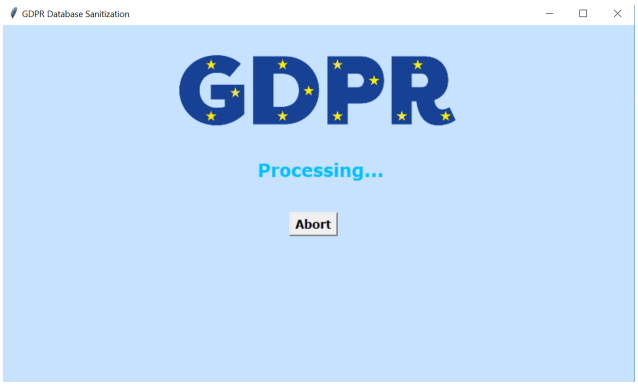


Fig. 2. System Load Page

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

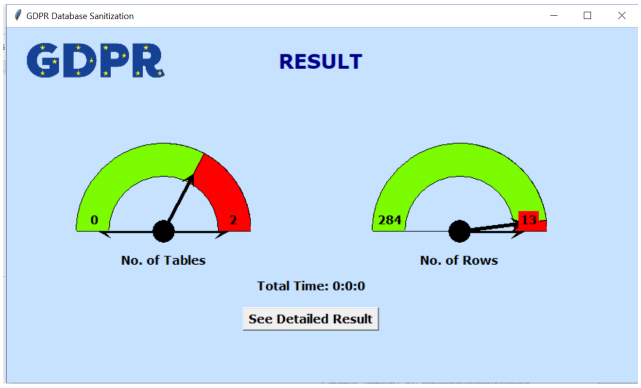


Fig. 3. System Result Page

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

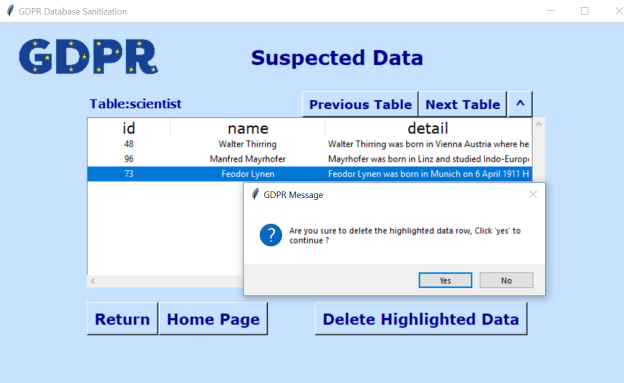


Fig. 4. System List Page

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-IDE 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.

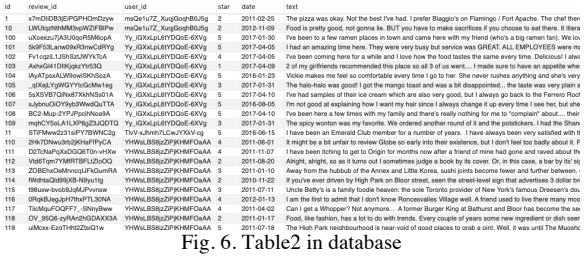
4.3 Dataset

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 professo…	Heidelberg (1936) and Graz (1940).
19	Carl Friedrich von W…	A member of the prominent Weizsäcker family, he was son of the diplomat Ernst von Weizsäcker, elder brother of the former Germ
21	Horst Dreier	A native of Hannover, Lower Saxony, Dreier attended the University of Hannover, where he received his Staatsexamen in 1981. He
35	Kurt Schwertsik	A pupil of Joseph Marx and Karl Schiske at the Academy of Music in Vienna, he later studied with Karlheinz Stockhausen in Cologr
87	Hans List	After earning a doctorate in engineering at the Technical University in Graz, Hans List was appointed to the Tongji University in Chi
51	Bruno Snell	After studying law and economics at University of Edinburgh and University of Oxford, Snell gained interest in classical studies and
61	Alfred Kubin	Alfred Leopold Isidor Kubin (10 April 1877 – 20 August 1959) was an Austrian printmaker, illustrator, and occasional writer. Kubin is
9	Nobuyoshi Araki	Araki was born in Tokyo on May 25, 1940.[4] He studied film and photography at Chiba University from 1959, receiving a degree in
36	H. C. Artmann	Artmann was born in Vienna as the son of shoe maker Johann Artmann and his wife Marie (née Schneider). After growing up in Vie
74	Bernarda Fink	Bernarda Fink has sung with leading orchestras including the Philharmonics of Vienna and London, Gewandhaus Leipzig, Radio-F
77	Anny Felbermayer	Born Anna Maria Felbermayer-Szekely in Vienna[1] to a family of craftsmen, she attended a Handelsschule. She studied piano and
94	Ernst Krenek	Born Ernst Heinrich Křenek in Vienna (then in Austria-Hungary), he was the son of a Czech soldier in the Austro-Hungarian army. H
109	Herwig Wolfram	Born in 1934 in Vienna, Herwig Wolfram remained intimately connected to the city of his birth. Between 1952–1957, he studied hist
44	Roland Rainer	Born in Klagenfurt, Roland Rainer decided to become an architect when he was 18, so he studied at the Vienna University of Techr
49	Adolf Butenandt	Born in Lehe, near Bremerhaven, he started his studies at the University of Marburg.
95	Anselm Franz	Born in Schladming, Austria, on January 21, 1900,[2] Franz studied mechanical engineering at the Graz University of Technology a
75	Jerzy Kuryłowicz	Born in Stanisław, Austria–Hungary (now Ivano-Frankivsk, Ukraine), he is considered[who?] the most outstanding contemporary Pc
72	Carl Emil Schorske	Born in The Bronx, New York City, to Theodore Schorske and Gertrude Goldsmith, Schorske received his B.A. from Columbia in 19
2	György Ligeti	Born in Transylvania, Romania, he lived in Communist Hungary before emigrating to Austria in 1956. He became an Austrian citize
17	Ian Hacking	Born in Vancouver, British Columbia, Canada, he earned undergraduate degrees from the University of British Columbia (1956) an
81	Josef Mikl	Born in Vienna, he received his first training at the Graphische Lehr- und Versuchsanstalt, studying at the prominent Viennese acac
10	Marcel Rubin	Born in Vienna, where he studied with Richard Robert[1] and Franz Schmidt, he later emigrated to Paris, where he pursued further
79	Bruno Gironcoli	Born in Villach, Gironcoli began training as a goldsmith in 1951 in Innsbruck, completing his apprenticeship in 1956. Between 1957
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-Ho
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 knov
43	Carl Ferdinand Cori	Carl was the son of Carl Isidor Cori [de] (1865, Brüx (Czech: Most), R.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	review_id	user_id	star	date	text
1	x7mDIDB3jEiPGPHOmDzyw	msQe1u7Z_XuqGoqhB0J5g	2	2011-02-25	The pizza was okay. Not the best I've had. I prefer Blaggio's on Flamingo / Fort Apache. The chef then
10	UWJtqyNrhMM3vpW2F8iPw	msQe1u7Z_XuqGoqhB0J5g	2	2012-11-09	Food is pretty good, not gonna lie. BUT you have to make sacrifices if you choose to eat there. It litera
100	uXeexzu7jA3U0qoR5M6opA	Yy_iGXxLpL6tYDQoE-6XVg	5	2017-01-30	I've been to a few ramen places in town and came here with my friend (who's a big ramen fan). We lov
101	5k9F53LamW0tkR3hwCdRYg	Yy_iGXxLpL6tYDQoE-6XVg	5	2017-04-05	I had an amazing time here. They were very busy but service was GREAT. ALL EMPLOYEES were m
102	Fv1cqzLLJSrSzUWYkTcA	Yy_iGXxLpL6tYDQoE-6XVg	4	2017-04-05	I've been coming here for a while and I love how the food tastes the same every time. Delicious! I alwa
103	AxhxGd41DlKjdxzYh53Q	Yy_iGXxLpL6tYDQoE-6XVg	1	2017-04-09	2 of my girlfriends recommended this place so all 3 of us went.... I made sure to have an appetite whe
104	IAyA7pxxALW8owISKhSozA	Yy_iGXxLpL6tYDQoE-6XVg	5	2016-01-23	Vickie makes me feel so comfortable every time I go to her. She never rushes anything and she's very
105	_qXadLYgWGYYlcGcMw1eg	Yy_iGXxLpL6tYDQoE-6XVg	3	2017-01-31	The halo-halo was good! I got the mango toast and was a bit disappointed... the taste was very plain a
106	5xX5VBTQRx87XkWNSs01A	Yy_iGXxLpL6tYDQoE-6XVg	5	2017-04-10	I've had samples of their ice cream which are also very good, but I always go back to the Ferrero Roch
107	sJybnuOiOY9yb3WwdQuTTA	Yy_iGXxLpL6tYDQoE-6XVg	5	2016-08-05	I'm not good at explaining how I want my hair since I always change it up every time I see her, but she
108	BC2-Mup-2YPJPpoPNoadA	Yy_iGXxLpL6tYDQoE-6XVg	5	2017-04-10	I've been here a few times with my family and there's really nothing for me to "complain" about.... their
109	mqhC75sLA1LXP6gZlJQDTQ	Yy_iGXxLpL6tYDQoE-6XVg	3	2017-01-31	The spicy wonton was my favorite. We ordered another round of it and the potstickers. I had the Shan
11	STiFMww2z31sIPY79WNC2g	TiVf-xJhnh7LCwJYXkV-cg	5	2016-06-15	I have been an Emerald Club member for a number of years. I have always been very satisfied with t
110	2Hk7DNwu3rb2jKHaFfPyCA	YHWsLBS8jzZIPJKHMFOaAA	4	2011-06-01	It might be a bit unfair to review Globe so early into their existence, but I don't feel too badly about it. F
111	DZToNaPpXxDGQETDn-xKXw	YHWsLBS8jzZIPJKHMFOaAA	4	2011-11-07	I have been itching to get to Origin for months now after a friend of mine had gone and raved about th
112	Vtd6Tqm7YMfRTBFLlZfoOQ	YHWsLBS8jzZIPJKHMFOaAA	2	2011-08-20	Alright, alright, so as it turns out I sometimes judge a book by its cover. Or, in this case, a bar by its' si
113	ZOBEhxOeMnnvqiJJFxGumRA	YHWsLBS8jzZIPJKHMFOaAA	3	2011-01-10	Away from the hubbub of the Annex and Little Korea, sushi joints become fewer and further between.
114	fWdHaQe8RjKlB-N9yu1tg	YHWsLBS8jzZIPJKHMFOaAA	2	2010-11-22	If you've ever driven by High Park on Bloor street, seen the street-level sign that advertises 3 dollar be
115	t9Euaw-bvcbJUqMJPvvnaw	YHWsLBS8jzZIPJKHMFOaAA	3	2011-07-11	Uncle Betty's is a family foodie heaven: the sole Toronto provider of New York's famous Dreesen's dou
116	0RqkBJeg-jpH7hxPTL30NA	YHWsLBS8jzZIPJKHMFOaAA	4	2012-01-13	I am the first to admit that I don't know Roncesvalles Village well. A friend used to live there many moo
117	TscMuyFOQFF7_-SNnyBww	YHWsLBS8jzZIPJKHMFOaAA	4	2011-04-02	Can I get a Whopper? Not anymore... A former Burger King at Bathurst and Bloor has become the sa
118	OV_9SQ6-zyRAn2hGDAX3A	YHWsLBS8jzZIPJKHMFOaAA	2	2011-01-17	Food, like fashion, has a lot to do with trends. Every couple of years some new ingredient or dish seer
119	uMcxx-EzoTHht2ZtaiQ1w	YHWsLBS8jzZIPJKHMFOaAA	5	2011-07-18	The High Park neighbourhood is near-void of good places to grab a pint. Well, it was until The Muoshc

Fig. 6. Table2 in database

5. Result and Analysis

Input keyword(s): salary
Result: matched table: 0, matched raw:0

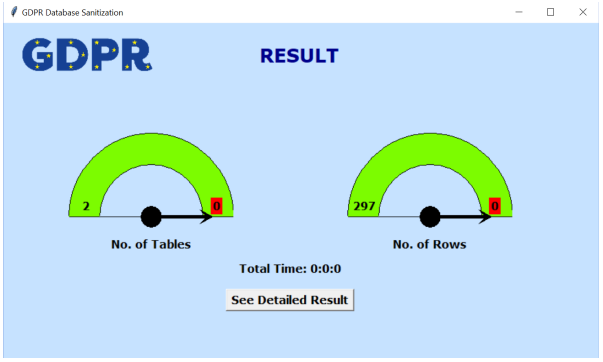


Fig. 7. Test result1

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

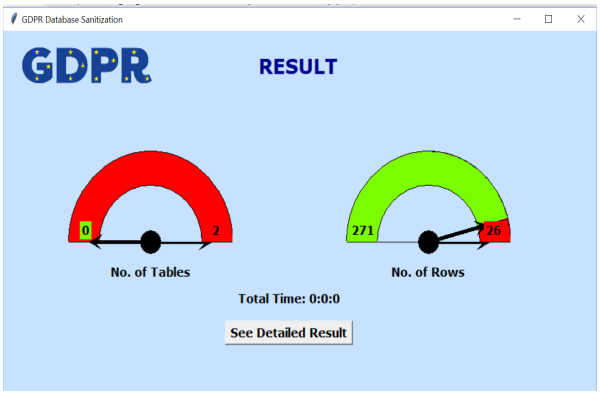


Fig. 8. Test result 2

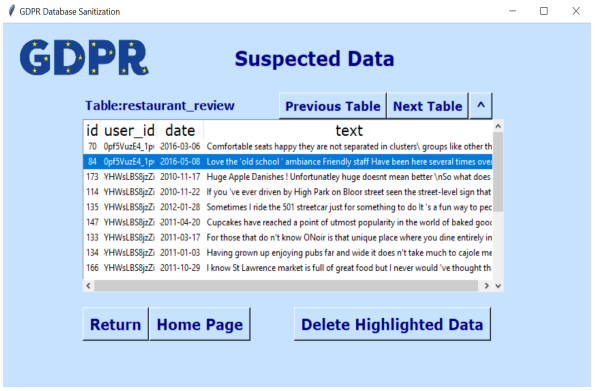


Fig. 9. Test result 3

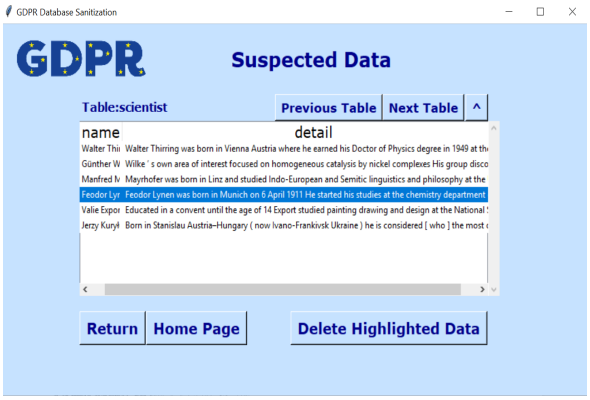


Fig. 10. Test result 4

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

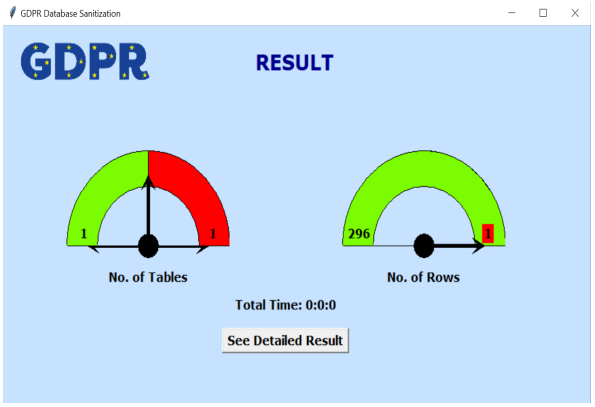


Fig. 11. Test result 5

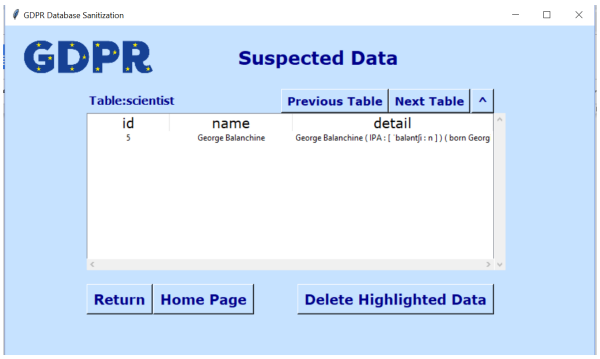


Fig. 12. Test result 6

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-programs-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].