**Application of text mining for understanding data protection incidents**

**Nguyen Quang Anh**

Dissertation submitted to International Business School

for the partial fulfillment of the requirement for the degree of

MASTER OF SCIENCE IN IT FOR BUSINESS DATA ANALYTICS

<May, 2025>

**DECLARATION**

This dissertation is a product of my own work and is the result of nothing done in collaboration.

*I consent to International Business School’s free use including/excluding online reproduction, including/excluding electronically, and including/excluding adaptation for teaching and education activities of any whole or part item of this dissertation.*

(Student signature)

Nguyen Quang Anh

Word length: ##,### words

Executive Summary

Table of Contents

[Introduction 5](#_Toc196253205)

[Literature review 7](#_Toc196253206)

[Individuals that are protected by the regulation 7](#_Toc196253207)

[Introducing the general principles of GDPR 7](#_Toc196253208)

[Summary of related laws and regulations 7](#_Toc196253209)

[Defining data breaches and data protection incidents 7](#_Toc196253210)

[Research methodology 8](#_Toc196253211)

[Introducing the Python libraries (Pandas, chosen NLP model, etc.) 8](#_Toc196253212)

[Exploratory data analysis 9](#_Toc196253213)

[Algorithms and models 10](#_Toc196253214)

[Result analysis 11](#_Toc196253215)

[Recommendations 12](#_Toc196253216)

[Conclusion 13](#_Toc196253217)

[List of references 14](#_Toc196253218)

[Appendix 15](#_Toc196253219)

Introduction

Data protection and regulations have been a hot topic in recent years due to the growing expansion of internet users and the rise of social media. As many tech companies are collecting data from their users, the governing authorities had acted to regulate the unlawful collection and processing of personal information. In recent years, the UK and the US banned TikTok from government devices, while India banned the app altogether from the country, citing national security concerns and espionage.

In the European Union, an initiative was started in 2016 called the General Data Protection Regulation, commonly known as the GDPR, to protect people's rights and freedoms. The collection of data privacy laws aimed to harmonize European countries and their data protection authorities, known as DPAs. The ruling became relevant in May 2018 and has been in effect since.

Even though the regulation was released in various forms, due to its difficult legal language and complex connection to different laws and articles, few people know its effects and success. For most internet users, the only noticeable change was a pop-up window asking to opt in to process cookies when browsing, however, the GDPR changed the practices of how companies can collect, store, and process personal data.

The project aims to facilitate the understanding of GDPR and its surrounding laws and definitions for individuals, startups, and small to mid-sized businesses without access to consulting services. As the articles can range from hundreds of words to thousands, it is very time-consuming to read, understand, and apply the rules written. There are estimates that over 90% of people do not read the terms and services conditions before accepting them. Based on this information, we can assume that even fewer people read the regulation on their own.

By utilizing automation software and text mining Python libraries, I am creating an approach to process legal documents and create a list of common mistakes that businesses make. Using the final rulings of penalized businesses for text mining, the expected output is the causes for the incidents, which can be investigated concerning the amount of the fine and the breached article(s). Examining this result can help define the severity of data protection incidents from the perspective of DPAs. If businesses can avoid following the same mistakes that are extracted from the documents, then the likelihood of incidents could drop significantly.

Even though the problem statement focuses more on the business perspective of the regulation, reading the project can also help regular people as well. Understanding our own data protection rights and how companies might misuse data could assist us in taking preventive measures. Knowing what personal data is being collected and what dangers it is exposed to changes our view on the internet and security. Even as an individual we can become data processor by collecting it without knowing. For example, when creating a survey for research or work we can inadvertently collect sensitive information without knowing. Or doing personal projects such as social media scraping for data analysis could be against the rules of GDPR, which is why grasping the concept of it is important.

From a technological standpoint this project introduces a low-cost alternative to existing LLM based text processing. Many businesses cannot afford commercial licenses for these services, therefore a script that can run in a Jupyter notebook should make it accessible. Even though this project focuses processing legal documents, the principles that will be presented can be easily applied to any other text-based research. To be as relevant as possible to the rapidly improving LLM and other text mining models the project will be applying modern text mining techniques where achievable.

However, this collection won’t cover all the possible causes as the rulings are in multiple European languages and due to limitations in time and processing power, only part of them will be added. Articles that haven’t been breached or fined yet also will be missing from the list as there is no input for them. These constraints should be kept in mind when drawing conclusions from the output of this research project.

The scope of the research includes two languages: English and Hungarian. The former is an obvious choice as most text-mining libraries are optimized for this language with many vocabularies. I have included Hungarian as a second language as this is the mother tongue of most of my professors and due to its complexity, it would be interesting to see what results we get compared to English.

The expected outcome of the project is a collection of words and phrases that are connected to the cause of incidents, with exploratory data analysis presenting the legal and technical context surrounding it. The person reading the contents of this paper should get better understanding of European data protection laws while getting a basic level of introduction to data mining methods using Python.

The project utilizes prior knowledge and literature from my bachelor’s degree thesis work. However, the technical application and methodology used for the project are different and have no relation to my prior work. During this project I will apply newly acquired skills gained during courses held by IBS. The research method can be considered both primary and secondary as I will be applying libraries and vocabularies prepared by other programmers, but the data collection, preprocessing and conclusion will be my own work.

Literature review

To understand where companies can fail data protection inspections, some legal context is needed, which will be briefly explained in the following sections. It is important to know why and how an individual is protected to recognize the breach of one’s rights. After a quick summary of the relevant articles of the GDPR the next section will introduce some of the basic text mining theories in order show the framework the project will be built upon.

## Individuals that are protected by the regulation

The first and foremost concept that need to be clarified is which individuals are protected by the articles of the regulation. In law a natural person is defined

## Introducing the general principles of GDPR

## Summary of related laws and regulations

## Defining data breaches and data protection incidents

Research methodology

The main objective of the research is to find the most common causes for data protection incidents that were investigated by a supervisory authority and were penalized for insufficient handling of personal data. By employing text mining and natural language processing methods, the expected outcome is a list of words that are present numerous times within these legal documents issued by said authorities.  
  
The main source of data comes from <https://www.enforcementtracker.com/>, which is a website that collects fines and penalties from multiple data protection authorities across Europe. The site tracks the ID of the case (ETid), country, date, the amount of fine, data controller or processor, the article(s) breached, and lastly, the type of issue summarized by the site. As we have access to much of the relevant data extracted already, instead of focusing on extracting this information from the files, I plan to focus on finding the connection between the most common.  
Causes found and the number of fines for articles breached.

While the page hosts many cases from various countries, I will be focusing on documents written in English. As one of the most spoken languages, many Python libraries and vocabularies are built upon it. Of course, the university program is also held in English, which would make it a requirement however, I will be including Hungarian cases as well, either by translating the document or by text mining it utilizing Hungarian vocabularies.

## Introducing the Python libraries (Pandas, chosen NLP model, etc.)

Exploratory data analysis

Algorithms and models

Result analysis

Recommendations

Conclusion

List of references

Appendix