**(Business Data Analytics Project title in 24 pt font)**

**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](#_Toc196159428)

[Literature review 6](#_Toc196159429)

[Individuals that are protected by the regulation 6](#_Toc196159430)

[Introducing the general principles of GDPR 6](#_Toc196159431)

[Summary of related laws and regulations 6](#_Toc196159432)

[Defining data breaches and data protection incidents 6](#_Toc196159433)

[Research methodology 7](#_Toc196159434)

[Introducing the Python libraries (Pandas, chosen NLP model, etc.) 7](#_Toc196159435)

[Exploratory data analysis 8](#_Toc196159436)

[Algorithms and models 9](#_Toc196159437)

[Result analysis 10](#_Toc196159438)

[Recommendations 11](#_Toc196159439)

[Conclusion 12](#_Toc196159440)

[List of references 13](#_Toc196159441)

[Appendix 14](#_Toc196159442)

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 have had to step in 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 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.

From a technological standpoint this project introduces a low-cost alternative to existing LLM based text processing. Many businesses cannot afford commercial licenses for

Literature review

## Individuals that are protected by the regulation

## 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