brottslighet



Mohamad almazrli

EC Utbildning

projektkurs

2024-10

# 

# 

# Abstract

This project automates the collection of crime data from the Swedish Police API, storing it in a structured SQL database for comprehensive analysis. The system explores automatic news text generation from crime reports, using natural language processing to keep the public informed on recent incidents. Additionally, data visualization in Power BI provides insights into crime patterns, highlighting trends such as frequently targeted locations and common crime types. By integrating data automation, text generation, and visual analytics, this project aims to enhance crime analysis and support proactive safety efforts.

Innehållsförteckning

[Abstract 2](#_Toc181012644)

[1 Inledning 1](#_Toc181012645)

[2 Teori 2](#_Toc181012646)

[2.1 Automatiserad Datainsamling 2](#_Toc181012647)

[2.2 SQL och Datahantering 2](#_Toc181012648)

[2.3 Maskininlärning för Textgenerering 2](#_Toc181012649)

[2.4 Visualisering och Analys i Power BI 3](#_Toc181012650)

[2.5 Sammanfattning av Teori 3](#_Toc181012651)

[3 Metod 4](#_Toc181012652)

[3.1 Datainsamling från Polisens API 4](#_Toc181012653)

[3.2 Lagring i SQL-databas 4](#_Toc181012654)

[3.3 Modellering för Textgenerering 4](#_Toc181012655)

[3.4 Analys och Visualisering i Power BI 4](#_Toc181012656)

[3.5 Modellutvärdering 4](#_Toc181012657)

[3.6 Agil Arbetsmetodik 4](#_Toc181012658)

[4 Resultat 5](#_Toc181012659)

[4.1 Resultat av Datainsamling och Lagring 5](#_Toc181012660)

[4.2 Resultat av Modellering för Textgenerering 5](#_Toc181012661)

[4.3 Visualisering och Analys i Power BI 5](#_Toc181012662)

[5 Diskussion 5](#_Toc181012663)

[6 Slutsatser 6](#_Toc181012664)

[6.1 Hur kan automatiserad insamling och lagring av brottsdata förbättra datakvaliteten och ge en mer korrekt bild av brottstrender? 6](#_Toc181012665)

[6.2 Vilka fördelar och begränsningar finns med att använda maskininlärningsmodeller för att generera text om brottshändelser? 6](#_Toc181012666)

[7 Sammanfattning 6](#_Toc181012667)

[8 Självutvärdering 7](#_Toc181012668)

# Inledning

I en alltmer digitaliserad värld blir effektiv datahantering och analys avgörande inom många områden, särskilt för säkerhet och brottsbekämpning. Att kunna samla in, strukturera och analysera stora mängder data systematiskt är en nyckel till att snabbt upptäcka mönster och fatta välgrundade beslut. Denna rapport beskriver ett projekt som automatiserar insamlingen av brottsdata från polisens API. Den insamlade informationen lagras i en SQL-databas för vidare analys och visualisering. Projektet undersöker även användningen av maskininlärning för att generera automatiska nyhetstexter utifrån brottsrapporter.

Projektet fokuserar på följande frågor:

1. Hur kan automatiserad insamling och lagring av brottsdata förbättra datakvaliteten och ge en mer korrekt bild av brottstrender?
2. Vilka fördelar och begränsningar finns med att använda maskininlärningsmodeller för att generera text om brottshändelser?

Målet är att skapa ett system som effektiviserar hela processen från datainsamling till analys och presentation, vilket i sin tur möjliggör snabbare och mer precisa insikter för beslutsfattande.

# Teori

För att utveckla ett automatiserat system för brottsanalys är en gedigen teoretisk grund inom flera områden avgörande. Projektet bygger på principer inom automatiserad datainsamling, databashantering med SQL, maskininlärning för textgenerering och visualiseringstekniker. Dessa områden utgör basen för att skapa ett system som kan hantera stora datamängder och omvandla dem till användbara insikter.

## 2.1 Automatiserad Datainsamling

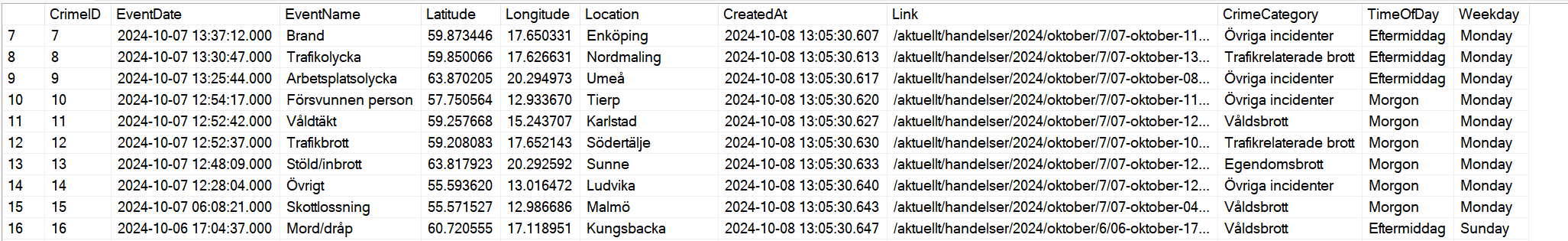
Automatiserad datainsamling innebär att data kontinuerligt hämtas utan behov av manuell input. I detta projekt används polisens öppna API för att samla in data om brottshändelser i realtid. Python-skript används för att interagera med API och schemaläggs med hjälp av Windows Task Scheduler för att hämta nya data dagligen. Automatiseringen säkerställer att brottsdata är uppdaterade och omedelbart tillgängliga för analys. Genom att använda Python-bibliotek som requests kan data i JSON-format enkelt omvandlas till strukturerade data som sedan lagras i en SQL-databas.

A screen shot of a computer program

Description automatically generated

## 2.2 SQL och Datahantering

SQL används i projektet för att effektivt lagra, bearbeta och analysera brottsdata. SQL-databaser erbjuder ett flexibelt och strukturerat sätt att hantera stora mängder data, vilket gör det möjligt att utföra komplexa frågor för att filtrera, aggregera och analysera informationen. SQL-funktioner möjliggör snabb och exakt hantering av data, vilket är avgörande för att extrahera relevanta insikter om brottstrender och vanliga brottstyper i olika områden.



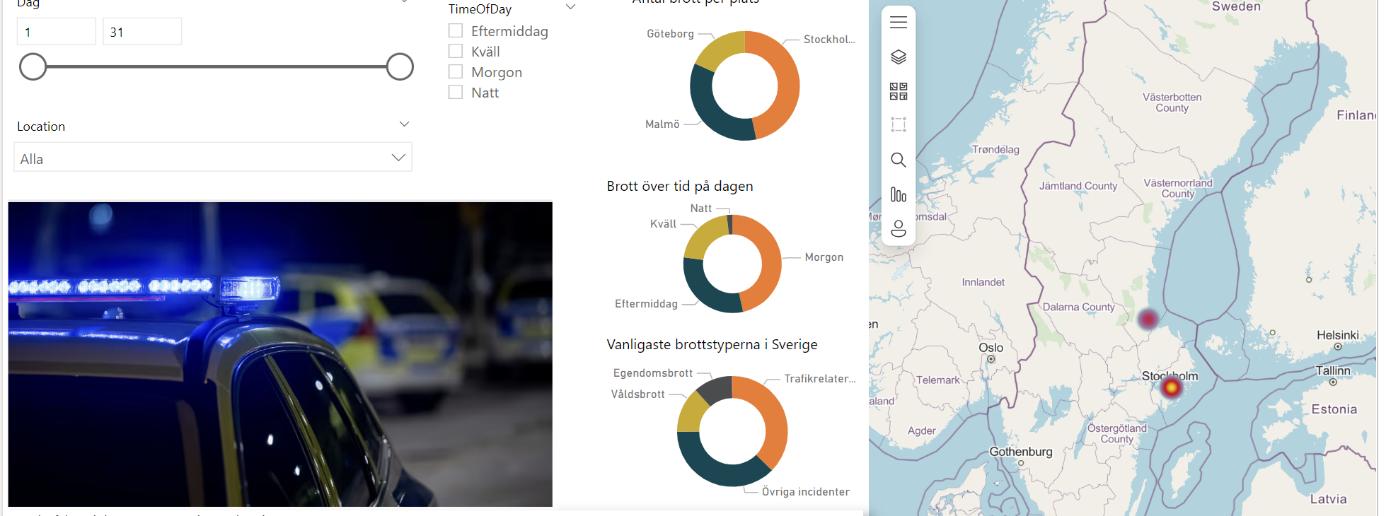
## 2.3 Maskininlärning för Textgenerering

Maskininlärning spelar en viktig roll i projektet genom att möjliggöra automatiserad textgenerering baserat på brottsdata. Målet är att utveckla en modell som kan skapa korta nyhetstexter baserade på insamlade brottsrapporter. Modeller som GPT och T5, som är transformerbaserade neurala nätverk, används här för att generera text baserat på en kort sammanfattning av händelsedata (som typ av brott, tid och plats). Trots att tekniken är kraftfull krävs betydande mängd träningsdata för att uppnå hög kvalitet i de genererade texterna, och modellen har stött på utmaningar relaterade till datamängd och variation.



## 2.4 Visualisering och Analys i Power BI

När data har samlats in och bearbetats i SQL används Power BI för att skapa interaktiva visualiseringar som belyser mönster i brottsligheten. Power BI är ett omfattande verktyg för att utveckla dashboards och analysvyer som ger användarna insikter i brottstrender över tid, olika geografiska områden och för specifika brottstyper. Visualiseringarna inkluderar bland annat stapeldiagram, linjediagram och kartor som visuellt representerar var och när brott är mest koncentrerade. Genom att länka Power BI till SQL-databasen kan visualiseringarna enkelt uppdateras, vilket skapar en dynamisk och användbar lösning för brottsanalys och rapportering.



## 2.5 Sammanfattning av Teori

Projektet är uppbyggt kring en rad tekniska områden, från automatiserad datainsamling och SQL-baserad datahantering till maskininlärning och avancerade visualiseringar i Power BI. Genom att kombinera dessa teknologier skapas ett effektivt system som möjliggör snabb och automatiserad insamling, analys och rapportering av brottsdata. Det automatiserade arbetsflödet gör det möjligt för beslutsfattare att enkelt få tillgång till aktuella insikter och information som kan stödja effektiva åtgärder mot brottslighet.

# Metod

För att uppnå projektmålet om att automatisera insamling, lagring, analys och visualisering av brottsdata från polisens API, delades projektet in i följande huvudsakliga steg.

## 3.1 Datainsamling från Polisens API

Det första steget i projektet var att samla in brottsdata från polisens API på ett automatiserat sätt. Genom att använda ett Python-skript med biblioteket requests kan API, anropas för att hämta brottsdata i JSON-format. Python-skriptet schemaläggs med Windows Task Scheduler för att köras dagligen, vilket säkerställer att datan alltid är aktuell. Rådata rensas och formateras för att underlätta lagring och framtida analyser.

## 3.2 Lagring i SQL-databas

Efter insamling lagras data i en SQL-databas för effektiv hantering och analys. Python-biblioteket SQLAlchemy används för att skapa en koppling mellan Python-skriptet och SQL-databasen. Den insamlade informationen lagras i en tabell som innehåller detaljer som brottstyp, plats, tidpunkt och en beskrivning av händelsen. Genom att strukturera data på detta sätt blir det enkelt att köra SQL-frågor för att utforska och analysera olika aspekter av brottsligheten.

## 3.3 Modellering för Textgenerering

Ett centralt steg i projektet är att bygga en maskininlärningsmodell som kan generera automatiserade nyhetsartiklar baserade på brottsdata. En förtränad modell, som T5 (Text-to-Text Transfer Transformer), har använts och finjusterats med den insamlade brottsdata för att skapa en modell som genererar sammanhängande och informativa texter. Modellen tränas med input från brottsrapporter som brottstyp, plats och sammanfattning och genererar nyhetsliknande text som output. Denna del av projektet stötte dock på vissa utmaningar med att uppnå en önskvärd textkvalitet, vilket kan bero på begränsningar i träningsdatan.

## 3.4 Analys och Visualisering i Power BI

För att visualisera brottstrender och möjliggöra en djupare analys importeras SQL-data till Power BI. SQL-frågor körs för att extrahera specifika mönster i brottsdata, som tidpunkter och platser för olika brottstyper. Dessa resultat importeras sedan till Power BI för att skapa interaktiva dashboards. Genom Power BIs olika visualiseringsverktyg, såsom stapeldiagram och kartvisualiseringar, kan användaren se tydliga trender i brottsligheten och få insikter om vilka områden och tider som är mest utsatta för olika typer av brott.

## 3.5 Modellutvärdering

Den generativa modellen för text utvärderas genom att jämföra modellens genererade texter med faktiska nyhetsartiklar om brott. Kriterier för utvärdering inkluderar textens flyt, relevans och korrekthet i förhållande till brottsdata. Denna utvärdering görs för att avgöra om modellen producerar artiklar som är informativa och kan användas som en del av den automatiserade rapporteringen.

## 3.6 Agil Arbetsmetodik

Projektet följer en agil arbetsmetodik där vi håller dagliga möten för att diskutera framsteg, identifiera utmaningar och göra eventuella justeringar. Detta arbetssätt främjar effektiv kommunikation och flexibilitet, vilket är avgörande när nya insikter och behov uppstår under projektets gång. Genom att dela upp projektet i mindre steg kunde vi fokusera på varje del och säkerställa att vi kontinuerligt levererade fungerande resultat. Efter varje delmoment reflekterade teamet över sitt arbetssätt för att optimera processer och hantera utmaningar som exempelvis komplexiteten i textgenereringsmodellen.

# Resultat

Detta projekt visar på effektiviteten i att automatisera processen från datainsamling till visualisering av brottsdata. Genom att implementera Python-skript för daglig insamling, SQL för databashantering och Power BI för visualisering, kunde ett automatiserat arbetsflöde upprättas. Nedan beskrivs resultaten från varje delsteg, samt diskussion kring framgångar och utmaningar i projektet.

## 4.1 Resultat av Datainsamling och Lagring

Datainsamlingen från polisens API fungerade som planerat, och schemaläggningen via Windows Task Scheduler säkerställde att datan uppdaterades dagligen utan manuell inblandning. Data rensades och lagrades i SQL-databasen med detaljer såsom brottstyp, plats och tidpunkt. Genom denna metod erhölls en välstrukturerad databas som möjliggjorde flexibel analys. En begränsning är att API-datan kan variera i struktur, vilket kräver fortsatt övervakning av skriptets funktionalitet.

## 4.2 Resultat av Modellering för Textgenerering

Trots flera iterationer av modellträning och finjustering visade modellen på utmaningar när det gällde att generera koherenta och informativa texter. Modellen kunde skapa artiklar baserade på enkla brottsbeskrivningar men hade svårigheter att hantera mer komplexa händelser eller varierande brottstyper. Detta kan till stor del tillskrivas begränsningar i träningsdatan, vilket påverkade modellens förmåga att förstå och korrekt tolka olika typer av brottsdata. Resultaten visar att mer omfattande och varierad träningsdata skulle kunna förbättra modellens prestanda.

## 4.3 Visualisering och Analys i Power BI

Power BI-visualiseringarna gav tydliga och användbara insikter i brottstrender. Data från SQL-databasen importerades och användes för att skapa interaktiva dashboards med stapeldiagram, linjediagram och kartor som illustrerade olika brottstyper i specifika geografiska områden och under olika tidpunkter. Exempelvis kunde visualiseringarna visa att vissa brott var vanligare i specifika områden och vid vissa tider på dygnet, vilket kan vara användbart för polisiära resurser och preventionsstrategier.

# Diskussion

Projektet visar att automatiserad insamling och analys av brottsdata kan effektivisera arbetet och generera meningsfulla insikter. Automatiseringen av dataflödet gör att analysen kan hållas aktuell och anpassad till förändringar i brottstrender. Dock avslöjade textgenereringsmodellen vissa begränsningar, särskilt i fråga om datakvalitet och variation. Att modellens genererade artiklar inte alltid uppfyllde kraven på koherens och relevans belyser vikten av datamängd och kvalitet för maskininlärningsprojekt.

Projektet har även visat värdet av visualiseringar i brottsanalys. Genom Power BI kunde data presenteras på ett överskådligt och interaktivt sätt, vilket hjälper beslutsfattare att lättare identifiera viktiga mönster och trender. En tydlig lärdom är att framtida förbättringar skulle kunna göras genom att samla in en större datamängd för modellträningen samt vidareutveckla visualiseringsmöjligheterna för att möta specifika användarbehov.

Sammanfattningsvis demonstrerar projektet potentialen i ett automatiserat arbetsflöde för brottsanalys och pekar på framtida förbättringsområden som skulle kunna höja analysens värde och precision.

# Slutsatser

Projektet visar att automatisering av insamling och lagring av brottsdata från polisens API kan ge både en mer konsekvent datakvalitet och en bättre överblick av brottstrender. Nedan sammanfattas svaren på de ursprungliga frågeställningarna.

## Hur kan automatiserad insamling och lagring av brottsdata förbättra datakvaliteten och ge en mer korrekt bild av brottstrender?

Automatiserad insamling och lagring av data via Python-skript och SQL-databaser erbjuder en konsekvent och strukturerad metod för att hantera stora mängder brottsdata. Genom att schemalägga insamling och lagring av brottsdata minskas risken för manuell felhantering och inkonsekvenser, vilket säkerställer en hög och jämn datakvalitet. Dessutom gör lagringen i SQL-databasen det möjligt att utföra detaljerade analyser av brottstrender över tid och geografiska områden, vilket ger en mer korrekt och omfattande bild av brottsmönster och frekvenser.

## Vilka fördelar och begränsningar finns med att använda maskininlärningsmodeller för att generera text om brottshändelser?

Maskininlärningsmodeller som T5 och GPT kan generera automatiserade texter om brottshändelser och därmed minska behovet av manuellt arbete inom rapportering. Modellerna är särskilt användbara för att snabbt producera sammanfattningar baserade på strukturerade brottsdata, vilket kan förbättra informationsspridningen. Dock visade projektet att begränsad träningsdata och varierande brottsbeskrivningar påverkar modellens prestanda negativt, vilket ibland resulterade i texter som saknade sammanhang eller tydlighet. För att förbättra dessa modeller krävs en större mängd varierade data samt ytterligare finjustering, särskilt för att hantera komplexa brottshändelser och skapa texter med högre språklig kvalitet och relevans.

# Sammanfattning

Projektet visar på styrkorna i att automatisera insamling, analys och visualisering av brottsdata, vilket underlättar snabbare och mer informerade beslut. Med förbättrad modellering och tillgång till mer omfattande data kan systemet utökas för att ge ännu mer detaljerade och användbara insikter inom brottsanalys. Automatiseringen av dataprocessen, i kombination med maskininlärning och visualisering, har potential att bli ett kraftfullt verktyg för framtida brottsrapportering och förebyggande åtgärder.

# Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.   
   tillgänglig av data och modellering
2. Vilket betyg du anser att du skall ha och varför.   
   VG
3. Något du vill lyfta fram till Antonio?

nej