Brott i Sverige



Mustafa almazerli

EC Utbildning

Projekt arbeta

202410

# Abstract

This project automates the daily retrieval of data and its storage in a SQL database using a Python script. The data is automatically collected, stored, and then analyzed through SQL queries to uncover patterns and trends. These insights are subsequently visualized in Power BI, providing a comprehensive view of the results. By automating the data collection, analysis, and visualization processes, the project significantly improves efficiency and ensures timely, data-driven insights for decision-making.

Innehållsförteckning

[Abstract 2](#_Toc179993153)

[1 Inledning 1](#_Toc179993154)

[2 Teori 2](#_Toc179993155)

[2.1 Automatiserad Datainsamling 2](#_Toc179993156)

[2.2 SQL och Datahantering 2](#_Toc179993157)

[2.3 Maskininlärning för Textgenerering 2](#_Toc179993159)

[2.4 Visualisering och Analys i Power BI 3](#_Toc179993160)

[2.5 Sammanfattning av Teori 3](#_Toc179993161)

[3 Metod 4](#_Toc179993162)

[**3.1** **Datainsamling från Polisens API** 4](#_Toc179993163)

[**3.2** **Lagring i SQL-databas** 4](#_Toc179993164)

[3.3 Modellering för Textgenerering 4](#_Toc179993165)

[3.4 3.4 Analys och Visualisering i Power BI 4](#_Toc179993166)

[3.5 3.5 Modellutvärdering 4](#_Toc179993167)

[3.6 Agilt 5](#_Toc179993168)

[4 Resultat och Diskussion 6](#_Toc179993169)

[**4.1 Resultat av Datainsamling** 6](#_Toc179993170)

[4.2 Modellering för Textgenerering 6](#_Toc179993171)

[4.3 Visualisering och Analys 6](#_Toc179993172)

[5 Slutsatser 7](#_Toc179993173)

[6 Självutvärdering 8](#_Toc179993174)

# Inledning

I dagens digitaliserade samhälle spelar automatiserad datahantering och analys en central roll inom många områden. Behovet av att kunna samla in, bearbeta och analysera stora mängder data är viktigt för att fatta informerade beslut, särskilt inom områden som säkerhet och brottsbekämpning. Genom att effektivisera dataflöden kan vi förbättra både precisionen och snabbheten i att analysera trender och göra förutsägelser.

Denna rapport fokuserar på att automatisera insamlingen av data från polisens API, lagra denna data i en SQL-databas, och sedan analysera den för att skapa en modell som kan generera automatiserade nyhetsartiklar baserade på brottshändelser. Vidare kommer data att användas i Power BI för att visualisera brottstrender och skapa insiktsfulla analyser. Detta projekt syftar till att förbättra effektiviteten i datainsamling och analys, samt utforska möjligheten att använda maskininlärning för att skapa automatiserade nyhetstexter baserade på brottsrapporter.

Syftet med denna rapport är att automatisera datainsamling, lagring och analys för att stödja prediktiv analys och automatiserad rapportgenerering. För att uppfylla syftet kommer följande frågeställningar att besvaras:

1. Hur kan data från polisens API hämtas och lagras automatiskt i en SQL-databas?
2. Hur kan en modell utvecklas för att generera automatiserade artiklar baserade på brottstyper och rapporter från polisen?
3. Hur kan data från SQL-databasen användas i Power BI för att visualisera brottstrender och skapa insiktsfulla analyser?

Vidare kommer en SQL-analys att genomföras, följt av arbete i Power BI för att visualisera och analysera den data som hämtats och lagrats.

# Teori

För att förstå och utveckla ett system för brottsanalys är det viktigt att ha en solid teoretisk grund inom flera nyckelområden, inklusive automatiserad datainsamling, datahantering, maskininlärning och visualiseringstekniker. Den här avdelningen presenterar de teoretiska principer som ligger till grund för projektet och deras tillämpningar i varje steg.

## Automatiserad Datainsamling

Automatiserad datainsamling innebär att data samlas in kontinuerligt och systematiskt utan mänsklig inblandning. I detta projekt används polisens öppna API för att hämta brottshändelser i realtid. Genom att använda Python-skript tillsammans med schemaläggningsverktyg som Windows Task Scheduler(schmaläggaren) kan datainsamlingen ske dagligen och lagras automatiskt i en SQL-databas. Automatiseringen säkerställer att de insamlade brottsdata alltid är aktuella och tillgängliga för vidare analys.

Ett vanligt sätt att hämta data från ett API är att använda requests-biblioteket i Python, som gör HTTP-förfrågningar och returnerar data i JSON-format. Detta JSON-format omvandlas sedan till en strukturerad form som kan lagras i SQL-databasen för framtida bearbetning.

A computer screen with text and images

Description automatically generated

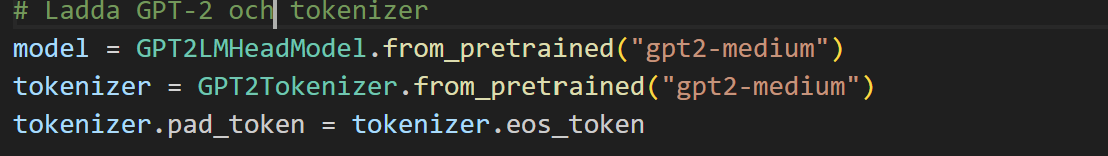
## SQL och Datahantering

SQL är ett verktyg som används för att lagra, hämta och bearbeta stora mängder data i en relationsdatabas. I detta projekt används SQL både för att lagra brottshändelser som hämtats från polisens API och för att bearbeta och analysera dessa data. SQL möjliggör kraftfull hantering av datan, inklusive filtrering, aggregering och sortering, vilket gör det lätt att extrahera meningsfulla insikter ur datan.

## A screenshot of a computer Description automatically generated

## Maskininlärning för Textgenerering

Maskininlärning spelar en central roll i detta projekt även om denna del av projekt misslyckas, men syftar att automatiskt generera artiklar baserade på brottsdata. Genom att använda transformer-baserade modeller som GPT eller T5 (Text-to-Text Transfer Transformer), kan man träna modellen på historiska brottsrapporter som är hämtat från Aftonbladet och sedan använda den för att generera textbaserad på aktuella händelser.

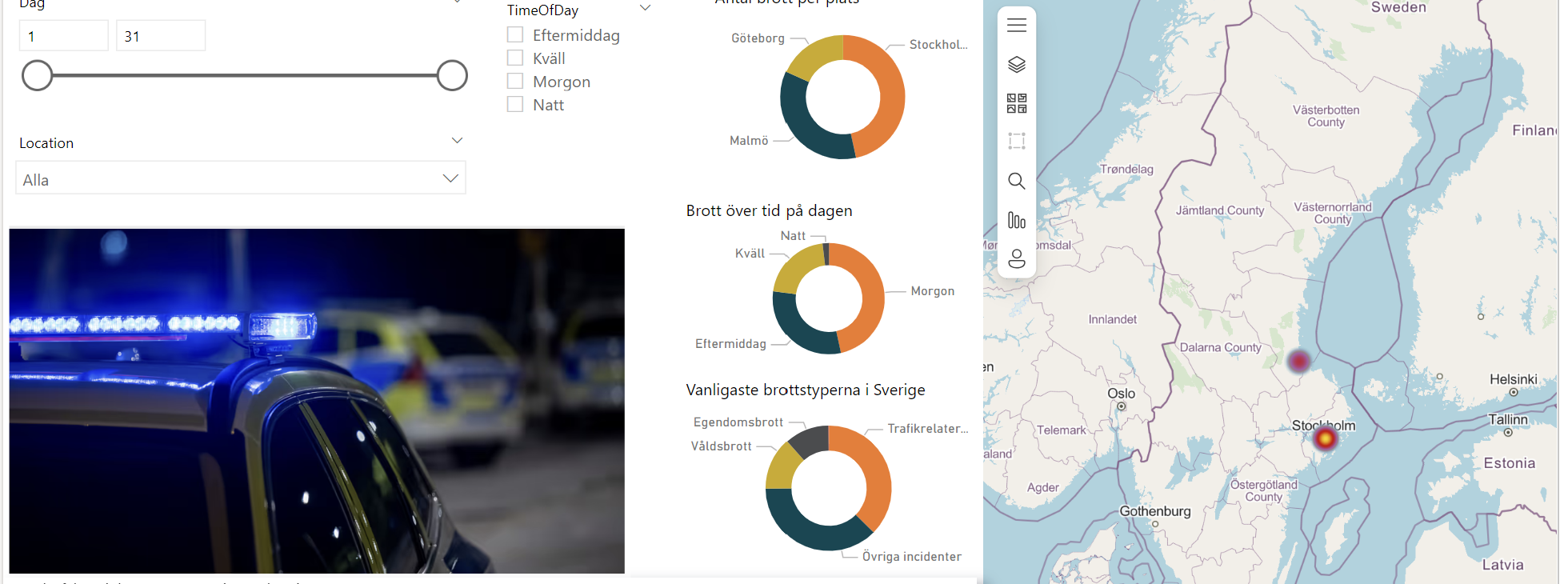


Transformermodeller är en avancerad typ av neurala nätverk som används inom språkteknologi. Modeller som GPT-2 eller GPT-3 kan ta en kort beskrivning av ett brott (t.ex. tid, plats, typ av brott) som input och generera en meningsfull nyhetstext som output. Detta gör det möjligt att automatisera skapandet av nyhetsartiklar, vilket kan vara värdefullt för att snabbt informera om nya händelser, Dock har det visat sig vara utmanande att uppnå bra resultat utan tillräcklig mängd träningsdata.

## Visualisering och Analys i Power BI

Efter att brottsdata samlats in och bearbetats används Power BI för att skapa interaktiva visualiseringar som ger insikter i brottstrender. Power BI är ett kraftfullt verktyg för att bygga instrumentpaneler som möjliggör analys av brottslighet över tid, i specifika geografiska områden eller för olika brottstyper.

Visualiseringarna kan inkludera stapeldiagram som visar de mest utsatta områdena, linjediagram för att följa förändringar över tid och map för att visuellt illustrera var brott är mest koncentrerade. Genom att koppla Power BI till SQL-databasen kan visualiseringarna enkelt uppdateras med aktuell data, vilket gör verktyget dynamiskt och användbart för brottsanalys och rapportering.



## Sammanfattning av Teori

Den teoretiska grunden för detta projekt täcker ett brett spektrum av tekniker, från automatiserad datainsamling och SQL-baserad datahantering, till maskininlärning för textgenerering och avancerade visualiseringar i Power BI. Genom att kombinera dessa tekniker skapas ett effektivt och kraftfullt system för att samla in, analysera och rapportera brottshändelser i realtid. Den automatiserade arbetsflödet förbättrar effektiviteten och ger beslutsfattare och intresserade parter bättre verktyg för att förstå och agera på brottsrelaterad information.

# Metod

För att genomföra detta projekt och uppfylla syftet med att automatisera datainsamling, analys, och visualisering, delades arbetet in i flera steg. Nedan beskrivs varje delmoment utförligt.

## **Datainsamling från Polisens API**

Det första steget i projektet var att automatiskt samla in brottsdata från polisens API. Detta gjordes genom att skapa ett Python-skript som använde requests-biblioteket för att hämta data i JSON-format från API. För att hantera tidsbegränsningar och kontinuerlig uppdatering av data, schemalades skriptet att köras regelbundet via Windows Task Scheduler. Data oformaterades och rensades för att vara lättare att bearbeta och lagra i en SQL-databas.

## **Lagring i SQL-databas**

Efter datainsamlingen lagrades informationen i en SQL-databas, vilket ger en struktur som möjliggör effektiv sökning och filtrering av data. En SQL-databas skapades och Python användes tillsammans med SQLAlchemy för att ansluta och lägga till informationen i en tabell. Varje daterad innehöll information såsom händelsetyp, plats, datum och tid, samt länk.

## Modellering för Textgenerering

Nästa steg i metodologin var att utveckla en maskininlärningsmodell som kunde generera artiklar baserade på brottstyper och händelser från polisens rapporter. För detta användes en förtränad T5-modell (Text-to-Text Transfer Transformer) som finjusterades med den insamlade brottsdata. Modellen tränades på att ta händelsetyp, plats och sammanfattning av händelsen som input och generera en sammanhängande artikel.

## 3.4 Analys och Visualisering i Power BI

Efter lagring och modellering genomfördes en analys av data genom SQL-frågor för att ta fram specifika trender och insikter om brottsligheten. Resultatet av dessa analyser importerades sedan till Power BI för att skapa interaktiva visualiseringar. Dessa visualiseringar användes för att få en tydligare bild av var brott var vanligast, vilken typ av brott som oftast förekom, samt andra mönster i data. Genom att använda Power BI kunde dynamiska dashboards skapas som underlättar för beslutsfattare att få insikter i realtid.

## 3.5 Modellutvärdering

Den generativa modellen utvärderades genom att jämföra dess utdata med existerande nyhetsartiklar för att avgöra om den kunde återspegla liknande språklig kvalitet och informativitet. Utvärderingen fokuserade på mått som språklig flyt, relevans och korrekthet i rapporteringen.

## Agilt

Vi hade dagliga möten för att diskutera projektets framsteg och eventuella ändringar som behövde göras. Detta reflekterar den agila principen om frekvent och effektiv kommunikation inom teamet, vilket hjälper oss att snabbt identifiera förbättringsområden och justera arbetet. Teammedlemmarna arbetade utifrån sina styrkor, vilket stämmer överens med agila principer om att ha självorganiserande team där varje individ bidrar utifrån sina unika färdigheter och erfarenheter. Vi visade flexibilitet genom att anpassa oss till skolans tidschema och omfördela uppgifter när någon inte kunde fullfölja en task. Detta säkerställde att projektet fortskred utan hinder, vilket reflekterar den agila förmågan att snabbt reagera på förändringar. Projektet delades upp i mindre, hanterbara delar som möjliggjorde individuellt arbete samtidigt som vi arbetade mot ett gemensamt mål. Detta speglar principen om att leverera fungerande delar av projektet löpande. Efter varje större del av projektet reflekterade vi över vårt arbetssätt och insåg utmaningar, som till exempel svårigheten att bygga en chatbot baserad på en kolumnist. Genom dessa reflektioner strävade vi efter att ständigt förbättra vårt arbetssätt och processer. Under projekten har vi ändrat våra mål några gånger beroende på svårigheter vi fick.

# Resultat och Diskussion

Resultatet av detta projekt visar potentialen i att automatisera hela processen från datainsamling till visualisering. Automatiseringen av insamling och lagring av brottsdata från polisens API fungerade väl och data kunde struktureras i en SQL-databas som möjliggjorde flexibel analys. Analysen i SQL, följt av visualisering i Power BI, gav tydliga insikter i brottstrender, såsom var brott oftast sker och vilka typer av brott som är mest vanliga.

## **4.1 Resultat av Datainsamling**

Datainsamlingen från polisens API var framgångsrik, och genom schemaläggning av Python-skriptet kunde data kontinuerligt uppdateras. Den insamlade informationen omfattade händelsetyp, plats, datum och tidpunkt, vilket var tillräckligt för att identifiera trender och skapa en modell för textgenerering.

## 4.2 Modellering för Textgenerering

Trots flera iterationer av träning och finjustering misslyckades modellen i att generera sammanhängande och meningsfulla artiklar för varje given händelse. Problem uppstod främst i modellens förmåga att korrekt tolka komplexa händelsetyper och skapa grammatiskt korrekta och kontextuellt relevanta texter. En möjlig orsak till detta kan vara begränsningar i datamängden och kvaliteten på de inmatade data, vilket kan ha påverkat modellens träningsprocess.

## 4.3 Visualisering och Analys

SQL-analysen och visualiseringarna i Power BI visade tydliga mönster för brott i olika områden. Med Power BI kunde man exempelvis se att vissa typer av brott var vanligare i specifika områden under olika tider på dygnet. Dessa insikter kan vara användbara för polisens resursplanering och preventionsstrategier.

# Slutsatser

I detta projekt lyckades vi automatisera processen att hämta och lagra brottsdata från polisens API till en SQL-databas, samt analysera och visualisera data i Power BI. För att besvara frågeställningarna kan vi dra följande slutsatser:

1. **Hur kan data från polisens API hämtas och lagras automatiskt i en SQL-databas?** Data från polisens API kan hämtas automatiskt genom att använda ett Python-skript som anropar API och lagrar informationen i en SQL-databas. Genom att schemalägga skriptet kan kontinuerlig uppdatering av databasen säkerställas.
2. **Hur kan en modell utvecklas för att generera automatiserade artiklar baserade på brottstyper och rapporter från polisen?** T5-baserad och gbt 2 är modell kan användas för att generera automatiserade artiklar, men utmaningar kvarstår när det gäller att hantera komplexa brottstyper och variera artiklarnas innehåll. Förbättringar kan göras genom längre träningstid, större och mer varierade träningsdata samt eventuell användning av en mer avancerad modell.
3. **Hur kan data från SQL-databasen användas i Power BI för att visualisera brottstrender och skapa insiktsfulla analyser?** Data från SQL-databasen importeras till Power BI sedan skapa många visualiseringar som ger insikter om brottstrender i olika regioner. Genom visualiseringarna blir det enklare att identifiera mönster och trender, vilket är värdefullt för beslutsfattande.

Projektet visade på vikten av datakvalitet och konsekvent datarensning, både för maskininlärningsmodeller och för datavisualisering. För framtida arbete skulle det vara fördelaktigt att fokusera på att samla in en större mängd varierade data samt utforska mer avancerade tekniker för textgenerering och analys.

# Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.

En av de största utmaningarna i detta projekt var att hantera den varierande datakvaliteten som hämtades från polisens API och hanterar skannade data sen att skapa en modell som kan generera text var svår

1. Vilket betyg du anser att du skall ha och varför.   
   jag har läggt mycket arbeta och önska vg
2. Något du vill lyfta fram till Antonio?

nej