Project Brief: Visualising Streaming Data with Streamlit II

Introduction

In this project, you will create a real-time data visualization dashboard using Streamlit to analyze streaming data from Reddit to identify fraud in telecommunications. The project will involve connecting to Reddit's API, collecting real-time posts, processing the posts to extract useful information, and visualizing the data using Streamlit.

Problem Statement

Fraud in telecommunications is a significant problem that costs the industry billions of dollars annually. Fraudsters use various techniques to exploit telecom infrastructure weaknesses, including hacking into phone systems, stealing identities, and exploiting vulnerabilities in billing systems. The challenge for telecom companies is to detect and prevent fraud in real-time before it causes significant financial damage.

Your task is to develop a real-time data visualization dashboard that monitors Reddit for mentions of telecoms fraud and other related keywords, such as "telecoms scam", "phone fraud", "billing fraud", and "identity theft". You will extract useful information from the posts, such as the post text, user name, subreddit, and date/time, and use this information to analyze the data for patterns and trends related to telecom fraud.

Project Requirements

- Connect to Reddit's API and collect real-time posts related to telecom fraud and other related keywords.
- Process the posts to extract useful information, including the post text, user name, subreddit, and date/time.
- Analyze the data to identify patterns and trends related to telecom fraud and other related keywords.
- Use Streamlit to create an interactive data visualization dashboard that displays real-time information about telecom fraud and other related keywords.
- The dashboard should include at least one chart or graph that displays the data meaningfully, e.g., a bar chart showing the number of fraud mentions by subreddit or a line chart showing the frequency of fraud mentions over time.
- The dashboard should be easy to use and visually appealing, with clear and concise labels and instructions.

Project Steps

- Set up a Reddit Developer Account and create an app to access the Reddit API.
- Install the necessary Python packages, including praw, pandas, and Streamlit.
- Write a Python script to connect to Reddit API and collect real-time posts related to telecom fraud and other related keywords. The script should use the praw library to authenticate the API access and collect the posts using the Reddit Streaming API.
- Process the posts to extract useful information, including the post text, user name, subreddit, and date/time. Store the data in a Pandas DataFrame for further analysis.
- Analyze the data to identify patterns and trends related to telecom fraud and other related keywords. Use Pandas and NumPy libraries to perform statistical analysis on the data and visualize the results using Matplotlib or Seaborn libraries.
- Use Streamlit to create an interactive data visualization dashboard that displays real-time information about telecom fraud and other related keywords. The dashboard should include at least one chart or graph that displays the data meaningfully, e.g., a bar chart showing the number of fraud mentions by subreddit or a line chart showing the frequency of fraud mentions over time.
- Test the dashboard by running it locally and ensure it updates in real-time as new posts are collected.
- Deploy the dashboard to a cloud-based platform such as Heroku or AWS to make it publicly accessible.

Deliverables

- Python script to collect and process real-time posts from Reddit API.
- Interactive data visualization dashboard created using Streamlit.
- Deployment of the dashboard to a cloud-based platform.