/r/technology Subreddit Analysis

Digital Futures Capstone Project
Anthony Squire

Agenda

01
Project Proposal
Project Requirements
Expected Outputs

02
Project Brief
Objective
Aims

03
Project Plan
Timeline
Project Steps
Data Flow Diagram

04
Extraction Script
Extraction Flow Diagram

05
Demonstrations
Script Demonstration
Streamlit App Demo

06
Summary
Reflections
Q&A

01

Project Proposal

Project Requirements

1. Choose a website, API or RSS feed that updates frequently
2. Create a script or process that obtains the data.
3. Put the data into a table in an SQL environment or similar
4. Create a visualisation app that demonstrates the data

Expected Outputs

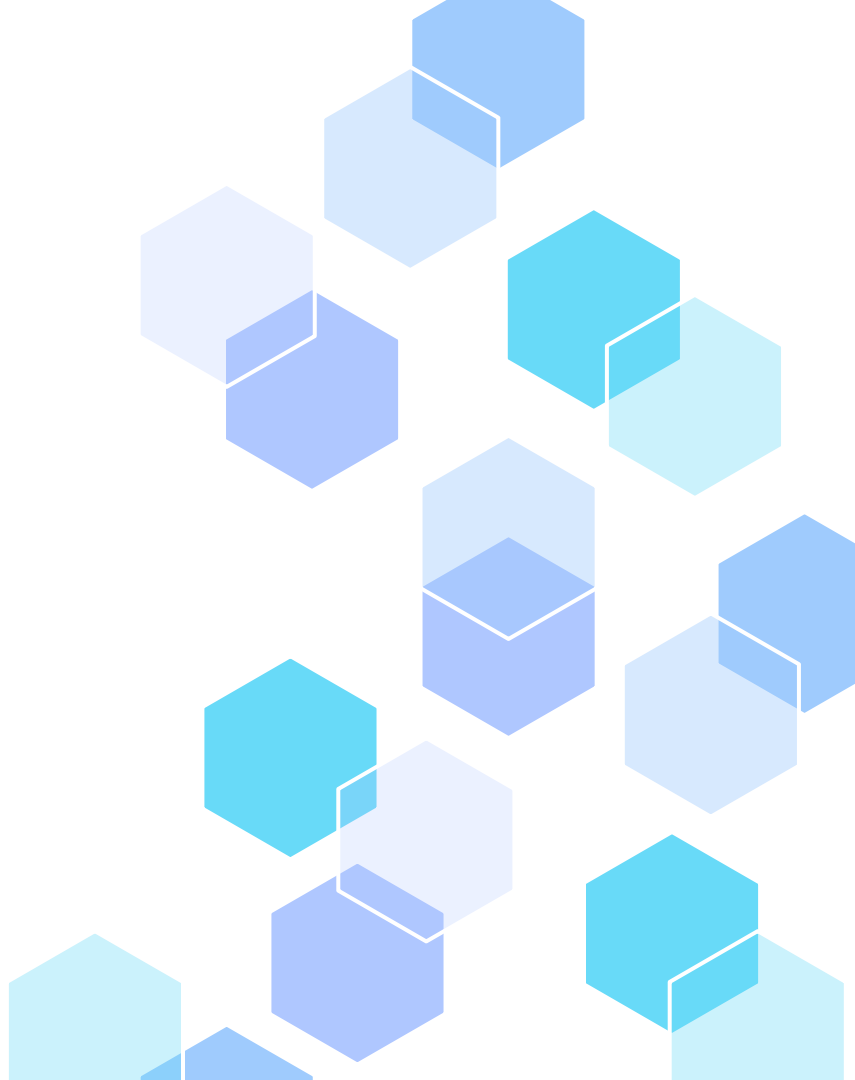1. Project plan

2. Data flow diagram

3. Demo of the script and visualisations

4. Presentation of process and product

02
Project Brief

Objectives & Aims

Objectives

Build a pipeline that regularly collects data from the /r/technology subreddit and stores it in a postgreSQL database.

Conduct analysis on the data and generate visualisations.

Produce a Streamlit app to showcase the data and visualisations.
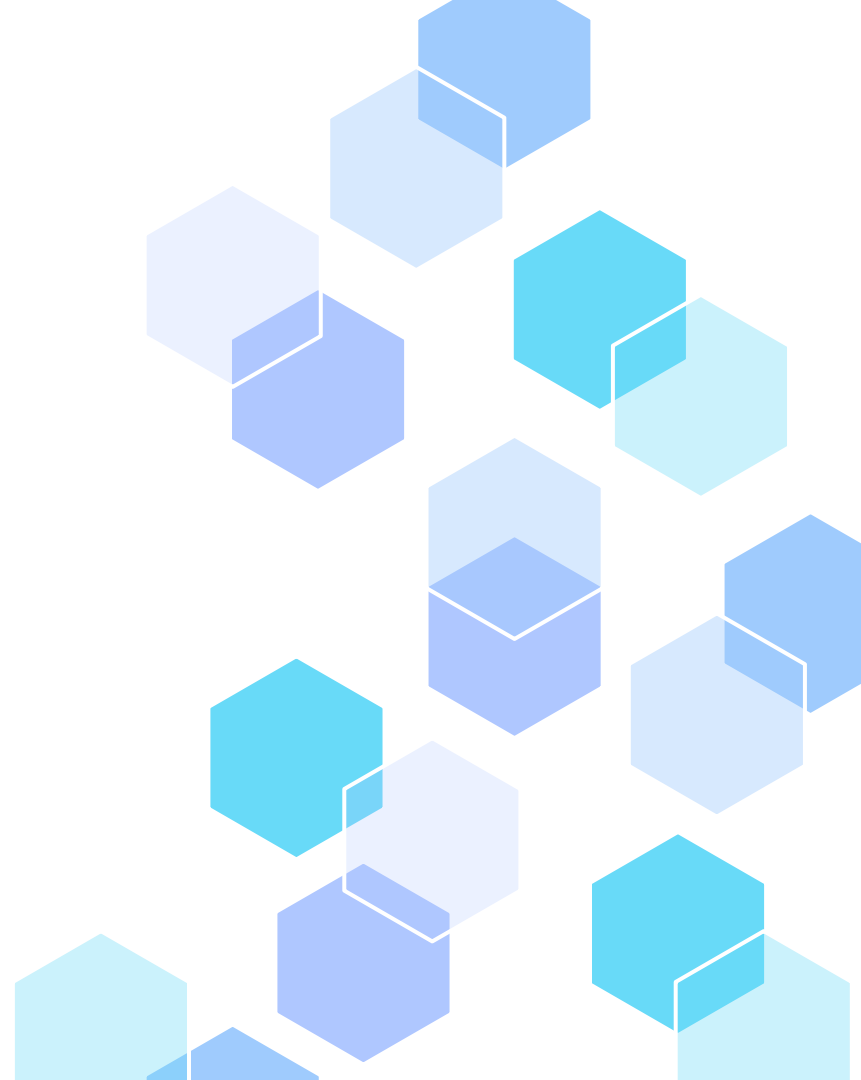
Aims

Uncover trends and insights from the subreddit through analysing the hot posts and comments.

Identify recurring themes, trending topics, and their tone using word clouds and sentiment analysis.

Utilise knowledge gained from the academy while improving and showcasing my skills.

03
Project Plan

Project Timeline

Project Step	Mon	Tue	Wed	Thu	Fri
January Week 4					
Pick an API	■				
Project Plan	■	■			
Explore API	■	■			
Produce Extraction Script		■			
Data Analysis		■	■	■	
Develop App			■	■	
Produce Presentation				■	■
Deliver Presentation & Demo					■

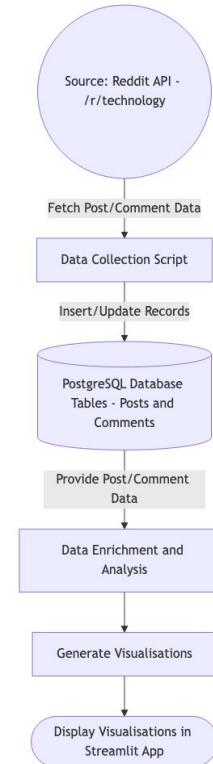# Project Plan Overview

Utilise the reddit API

- Use the python reddit API wrangler (praw)
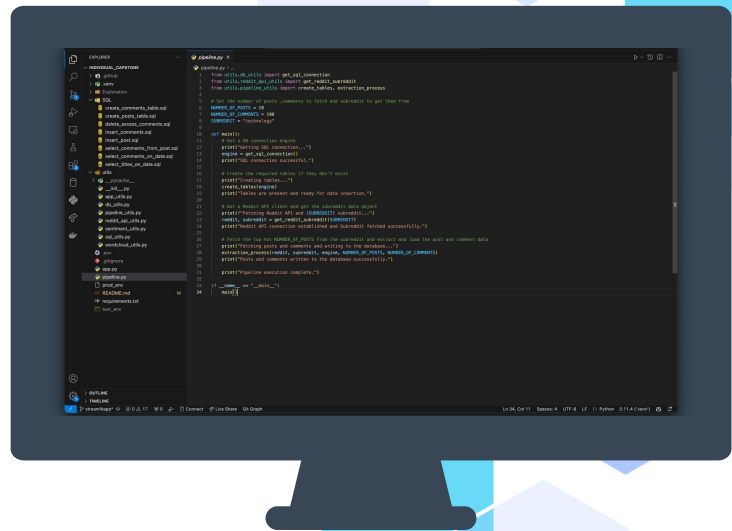- Query the /r/technology subreddit

Produce an Extraction Script

- Create a posts and comments table with appropriate fields
- Get the top 10 hot posts data
- Get the top 100 top level comments data on each post
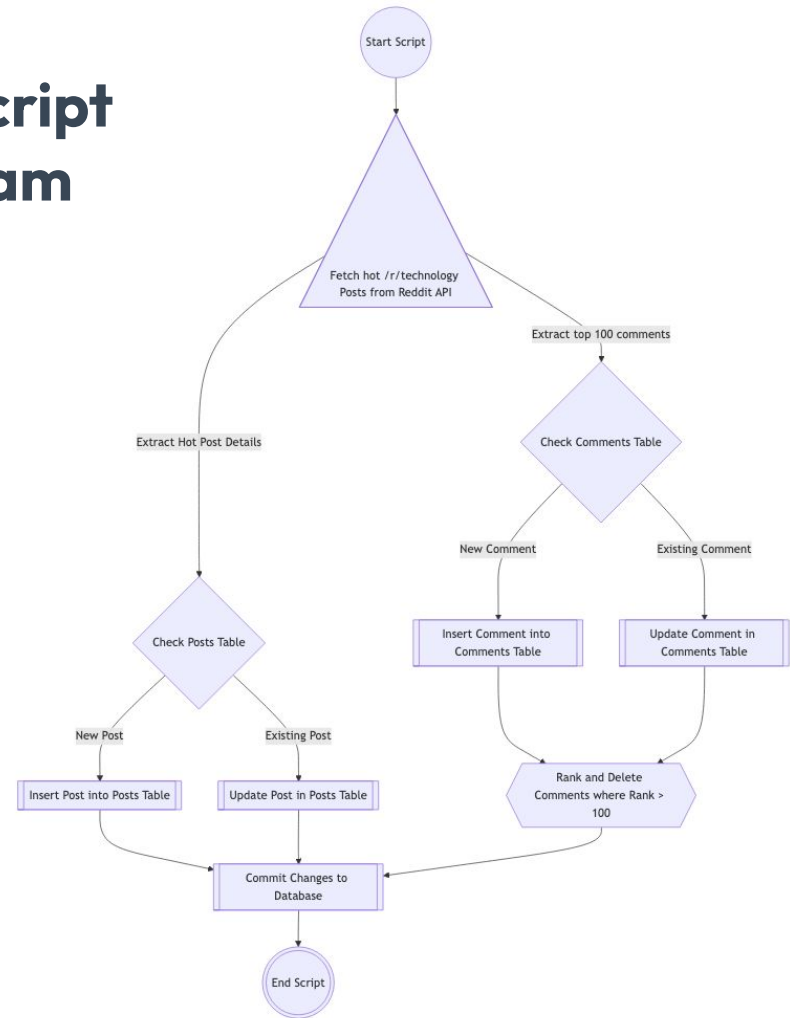- Write the data to the database

Create and Deploy App

- Generate word clouds of titles and comments.
- Perform sentiment analysis on titles and comments
- Identify trending subjects/words each day

Source: Reddit API - /r/technology

Fetch Post/Comment Data

Data Collection Script

Insert/Update Records

PostgreSQL Database Tables - Posts and Comments

Provide Post/Comment Data

Data Enrichment and Analysis

Generate Visualisations

Display Visualisations in Streamlit App

04
Extraction
Script

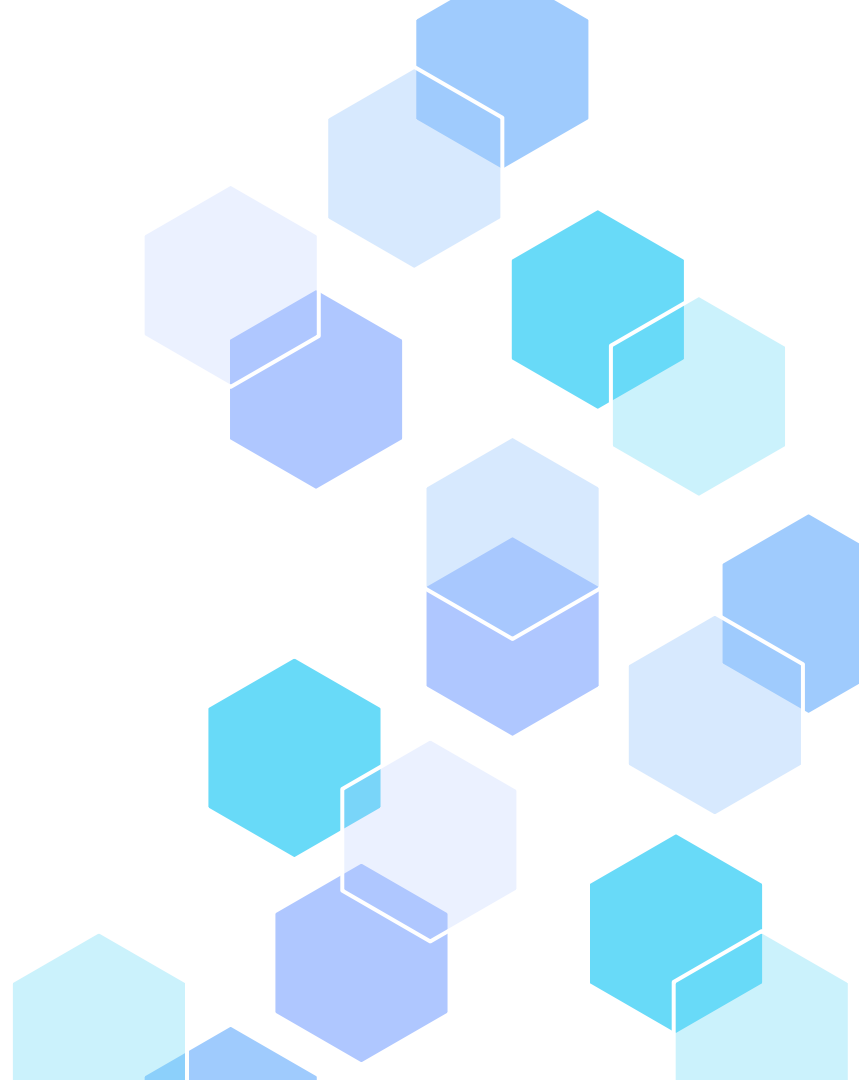# Extraction Script Flow Diagram

05
Demonstrations

06

Summary

Project Summary

Project requirements met

- API chosen
- Script made and executed
- Data in pagila database
- Visualisations app made

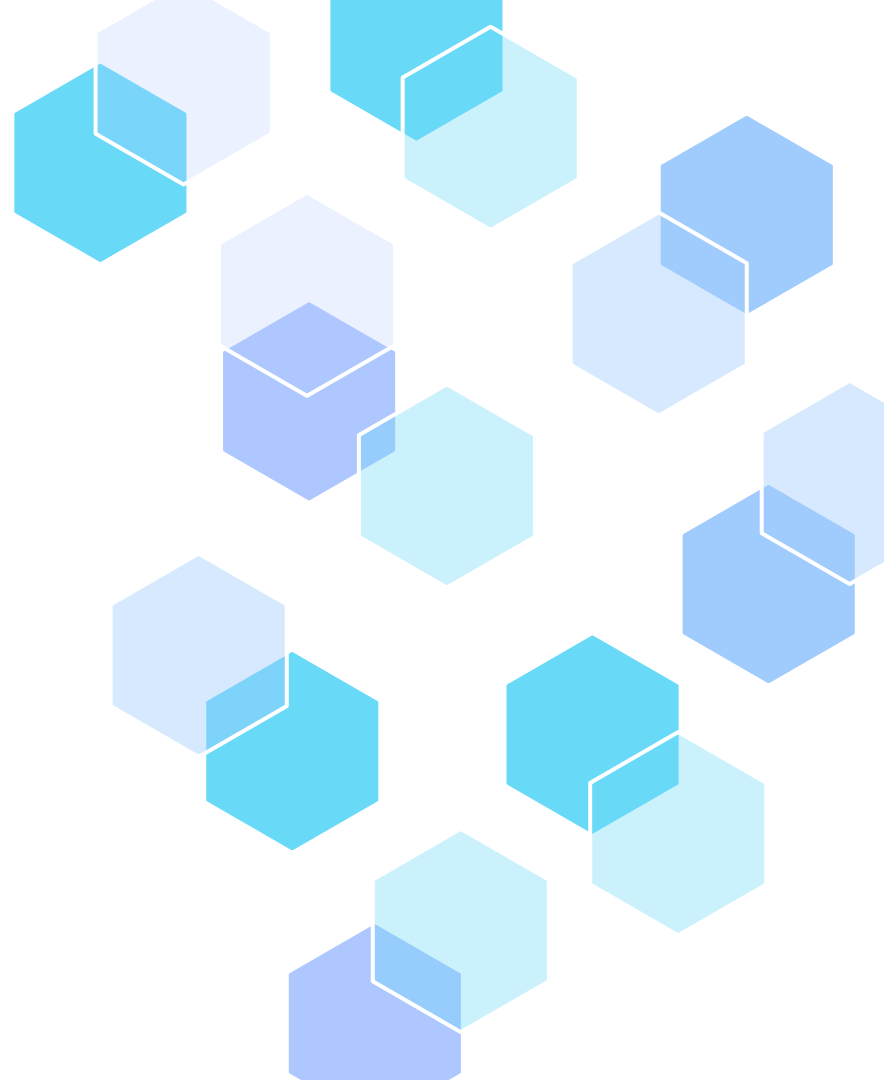## Expected outputs produced

- Plan
- Data flow diagram
- Demo of script and app
- Finished with this presentation!

Possible Improvements

- Testing!
- Script efficiency – seems slow to run
- Refactor parts of the streamlit app and functions
- Better sentiment modelling
- New features:
  - Weekly trends/comparison
  - More comment levels
  - Other subreddits

Personal Reflections

- Enjoyed the project
  - Particularly producing the pipeline script
- Need to improve testing knowledge and skills
- Time management is very important
- More agile methodology practice/knowledge, particularly in the planning stage would be beneficial
- Challenges
  - Decision making – best methods etc.
  - Lone working

Thanks!

Do you have any questions?

If you would like to try the app please visit:

reddit–technology–analysis.streamlit.app