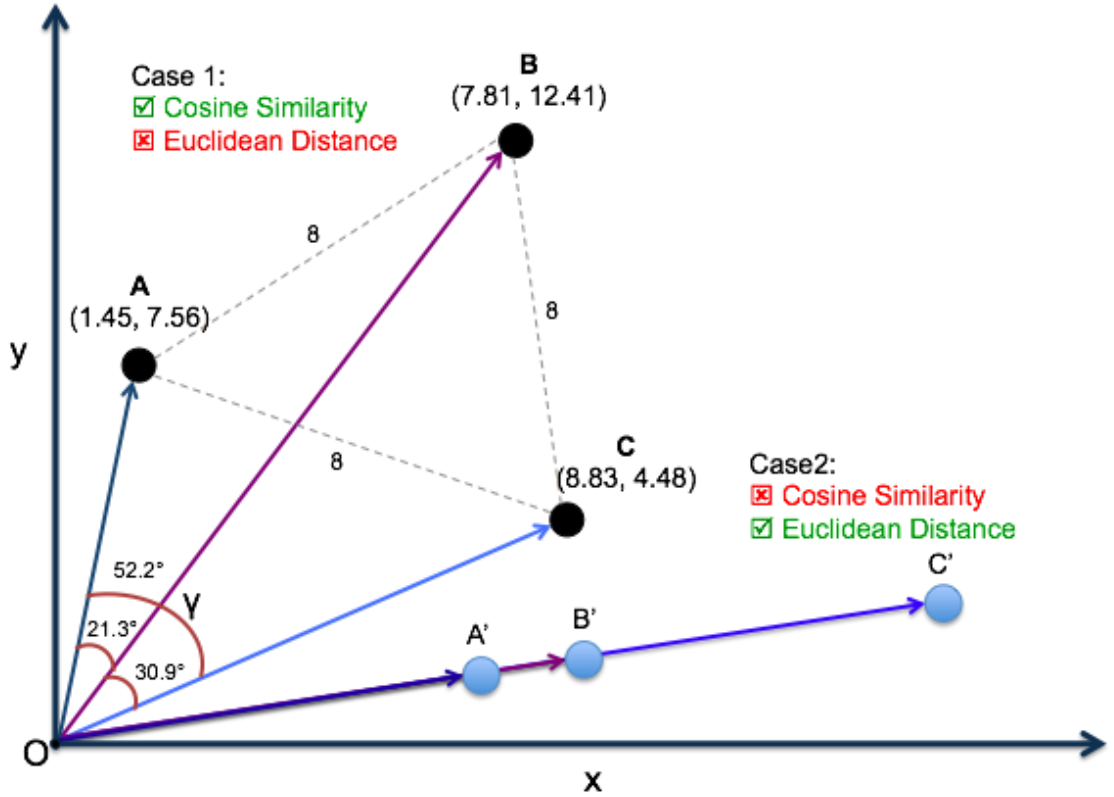
MSDS600 W7: Collecting social media data

REGIS UNIVERSITY

Review from W6

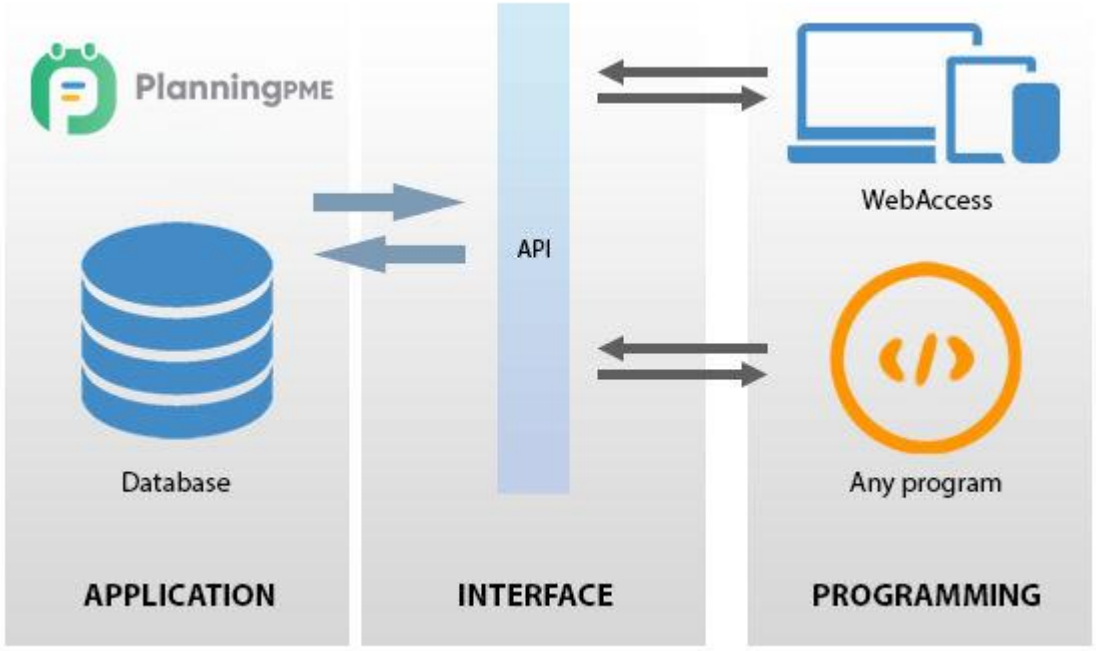
- Recommender systems:
  - Simple/default (e.g. top items)
  - Collaborative (data from all users) and content-based (data on items and one user's preferences and previous items)
- Cosine and Euclidean distance
- Big data strategies and Python packages
- Graph theory and Python packages



https://medium.com/@sasi24/cosine-similarity-vs-euclidean-distance-e5d9a9375fc8

This week’s topics

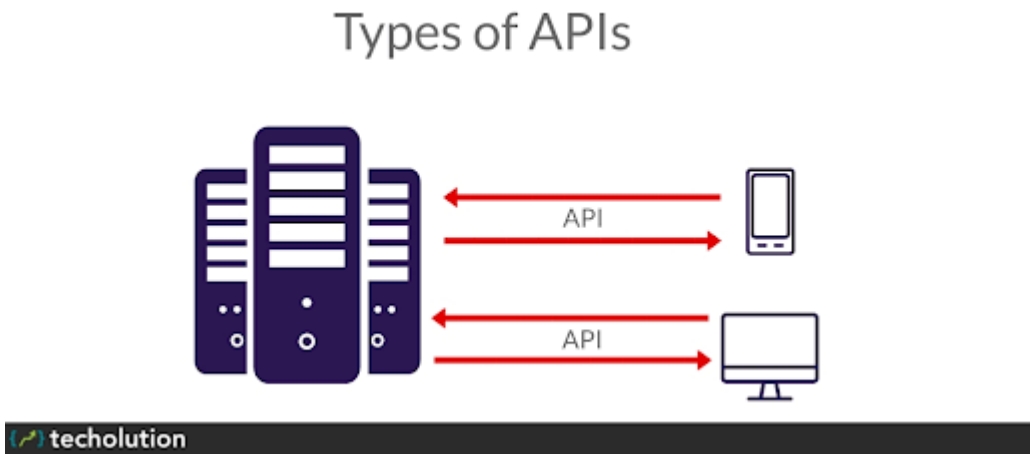
- Collecting social media data from Reddit with an API
- Storing data for later using SQL



https://www.planningpme.com/planningpme-api.htm

APIs (application programming interfaces)

- Allows us to interface with another software system
  - Allows two software systems to talk to one another
- E.g. web service (data) from Python or other programming languages
- Types of web APIs:
  - REST (common)
  - Websocket (newer, can stream data, common where fast streaming data is needed)
  - Other types
- There are limits: we can only request so much data at a time, and other limits



https://techolution.com/types-of-apis/

Using APIs in Python

- There are Python (and other programming language) API wrappers for many common websites and web services. For example:
  - Twitter (there are at least a few different packages)
  - Reddit (the one we will use)
  - Facebook
  - others
- There are also other Python wrappers and packages for APIs:
  - Cryptocurrency exchanges (this uses websockets)
- Some web scraper packages exist that bypass API limits (gets into grey ethical/legal areas that we discuss more in MSDS610)
  - https://pypi.org/project/facebook-scraper/
  - https://pypi.org/project/twitter-scraper/

It can be fun and interesting to analyze Twitter data. If you want to use the Twitter API for future classes (e.g. text analytics or the practicum), you can apply for an API key now. The Twitter API approval process can take weeks.

You can apply for access here:
https://developer.twitter.com/en/apply-for-access

Using APIs in Python

- Python API packages are usually wrappers, which means they take care of the tricky parts of using the API and allow us to use Python functions and classes to run it.
- A REST API call to collect data might look like:

```
https://www.reddit.com/r/php/search.json?q=oop
```

- We might also need to add authentication, and things can get complicated quickly (like this).
- Wrappers condense this down to a few lines of code, which we will learn how to use with Reddit.

```python
import praw

reddit = praw.Reddit(
    client_id="my client id",
    client_secret="my client secret",
    user_agent="my user agent",
)
```

```python
for submission in reddit.subreddit("learnpython").hot(limit=10):
    print(submission.title)
```

SDKs

- Closely related to APIs are SDKs (software development kits).
- These allow us to develop applications with a specific toolset.
  - e.g. Python SDKs for AWS, Google cloud, Facebook, and more.
- With this, we can automate processes from Python (such as starting up a cloud computing cluster, running some processing, and shutting it down).
- SDKs can allow us to get data like APIs sometimes.

https://industrialin.com/software-development-kits

SQL

https://www.tutorialrepublic.com/sql-tutorial/

SQL (structured query language) has been around since the 1970s and isn't going anywhere anytime soon.

The reason is SQL is a standard, structured language for querying databases.

NoSQL has appeared since the internet age began, but it has problems.

SQL databases are adapting to become more like NoSQL in some ways (e.g. scalability) and are used in big-time solutions (like AWS Redshift).

Knowing some basics with SQL will help a lot in data science. We will cover some in MSDS610 and you can learn more in MSDE631. It's also not a bad idea to do some extra SQL studying on your own, especially if you don't take MSDE631.

SQLite

There are many different SQL databases (MySQL, Oracle SQL, SQLite, etc).

They each follow the same base SQL language guidelines, although each type of database will have its own peculiarities and advantages/disadvantages.

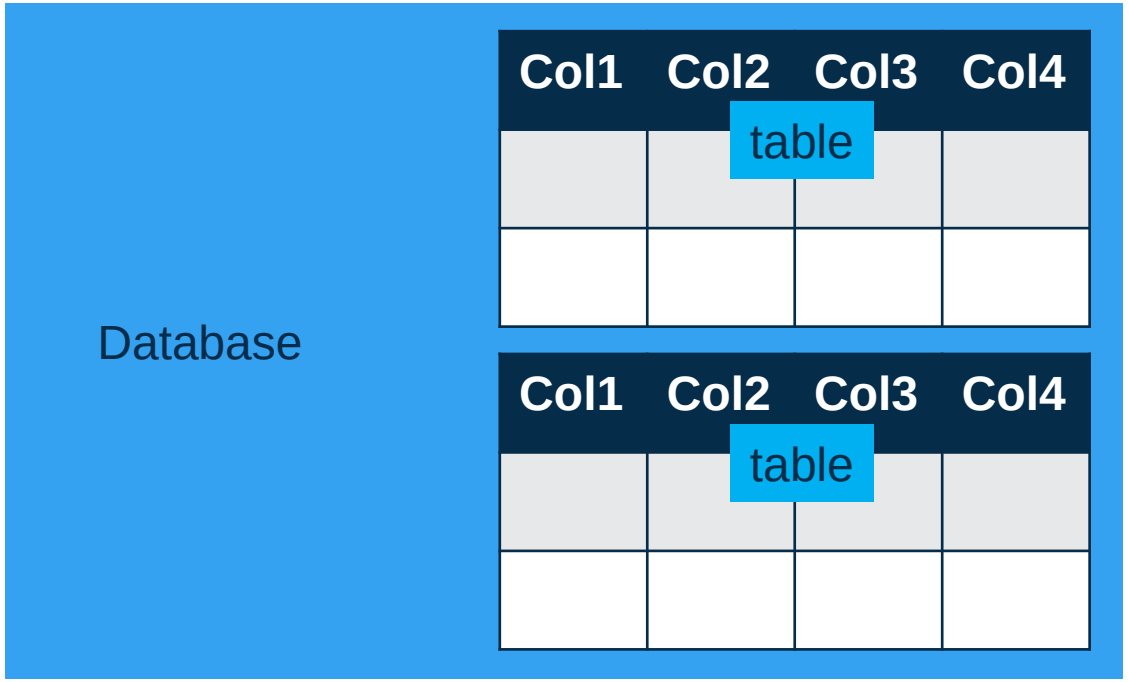
SQLite3 comes installed with Python, and we can use the built-in module from base Python.

SQLite simply stores our data to a file on our hard drive, so we don't have to bother with setting up a database system to run on our computer.

It's meant to be easy and quick to use, as well as lightweight.

SQL organization

- SQL databases are typically organized into databases, tables, and columns.
- A single database can have several tables, and each table can have several columns.
- There are different types of columns we can use (numeric, datetime, text, blob).
- With SQLite, we have a file that is our database, and tables within the file. Then each table has columns.



SQL queries

SQL commands and queries follow a pattern.

Typically there is a primary command (like SELECT) followed by options and subcommands. Some of the common commands are shown here.

We can also find the specific syntax for different SQL variants through their documentation, or sometimes online tutorials. For example, SQLite is shown here for the SELECT statement. The SQLite docs are a little confusing for the SELCET statement.

We use the SELECT statement often because that's how we retrieve data. It's important to have a basic understanding of it. The SQLite tutorial (linked above and on the right) can be helpful for getting started.

```
SELECT DISTINCT column_list
FROM table_list
  JOIN table ON join_condition
WHERE row_filter
ORDER BY column
LIMIT count OFFSET offset
GROUP BY column
HAVING group_filter;
```

https://www.sqlitetutorial.net/sqlite-select/

Our Social Media Analysis Plan

- Collect data from the Reddit API using the PRAW Python package.
- Collect data from at least one subreddit where we want to analyze the sentiment of the text.
- Store the data in a SQLite3 database file on our computer.
- Load the data and analyze it using sentiment analysis to understand what the sentiment in the subreddit looks like.

- Next week we will learn sentiment analysis and some other things about text analysis and natural language processing.

https://www.translatemedia.com/us/blog-usa/machine-translation-multilingual-sentiment-analysis-projects/