FACULTAD
DE INGENIERIA
Universidad de Buenos Aires

75.74 - Sistemas Distribuidos I

TP1 - Middleware y Coordinación de Procesos

Reddit Memes Analyzer

1° Cuatrimestre de 2022

Eliana Gamarra - 100016

Introducción

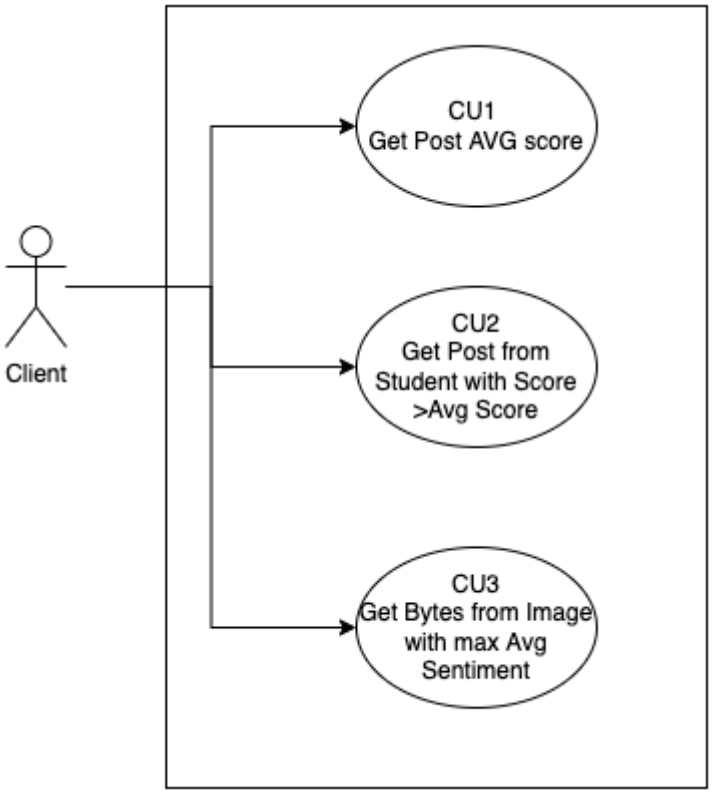
En el siguiente informe se presenta una solución a un sistema distribuido para el análisis de datos extraídos de Reddit.
Dicho sistema posee un middleware que permite la utilización de queues de RabbitMQ. Se lanzarán varios servicios en simultáneo que se coordinarán para escribir y leer los datos de las queues.
El cliente enviará en una conexión los datos de los posts y comments en diferentes queue y estos datos completaran el recorrido para obtener tres salidas de los datos:

- Promedio de score de todos los posts
- URLs de memes que gustan a estudiantes (con comments sobre university, college, student, teacher, professor y con score mayor al promedio)
- Descarga del meme con mejor sentiment promedio

A continuación se presentan una serie de diagramas para facilitar el entendimiento del sistema.

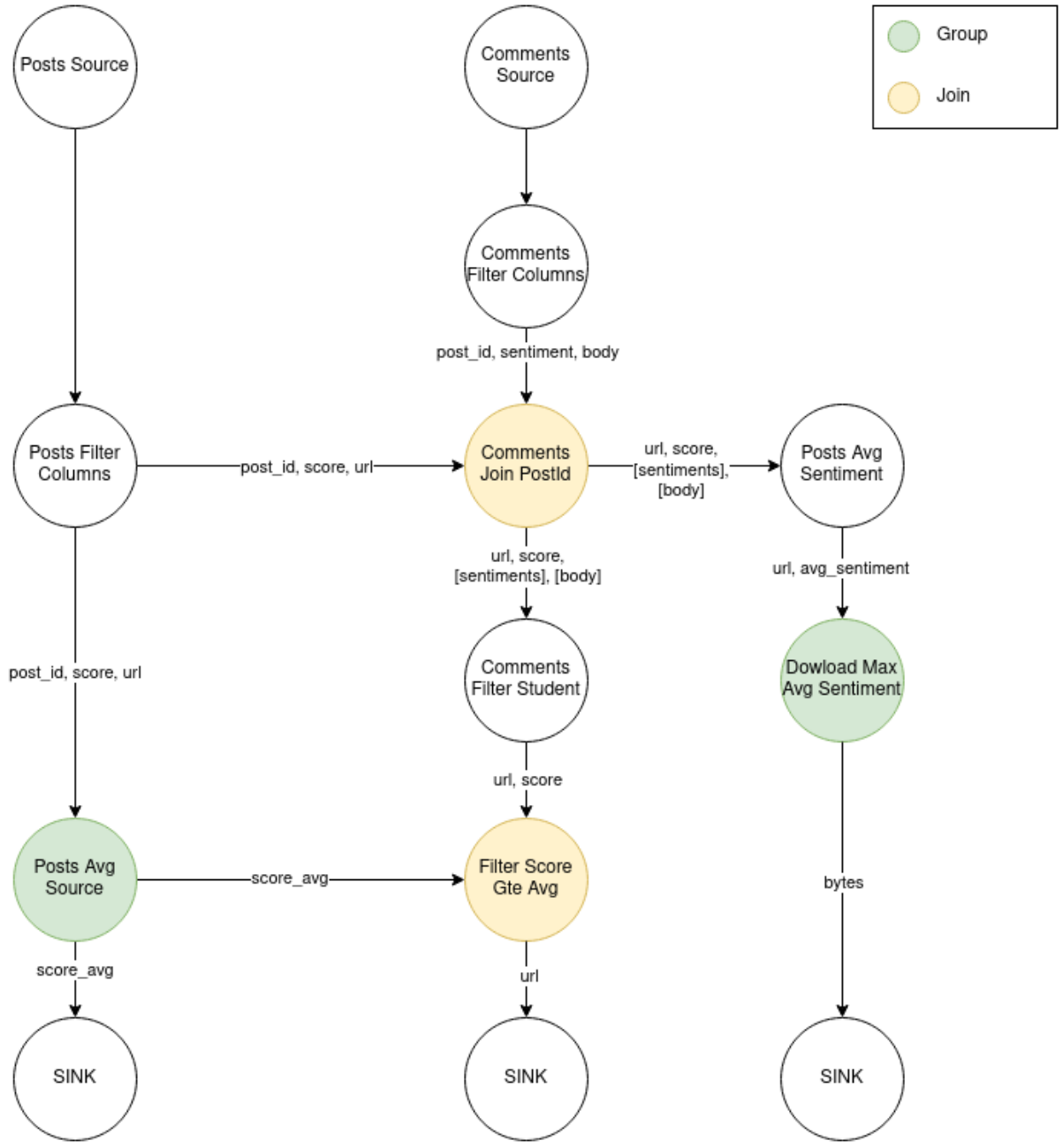
Escenario

Diagrama Casos de Uso



Vista lógica

DAG



Vista Física

Diagrama de Robustez

Diagrama de los distintos procesos involucrados y de qué manera se comunican por medio de queues de RabbitMQ.
Para establecer las diferentes conexiones con Rabbit se usó la librería Pika. Que crea una conexión y canal de comunicación, este será el Middleware
A la izquierda se tiene el punto de entrada del usuario a la app, donde el cliente comienza a enviar mediante chunks las lineas leidas del CSV de post y comments a sus respectivas queues.
Ya que el archivo tiene muchas cantidad de filas a procesar se eligió tener varios workers (la cantidad es configurable) para filtrar los datos innecesarios.
He de aclarar que los servicios que debían esperar la totalidad de los datos para poder enviar el resultado obtenido (como lo son: Join Comments With Post, Post Avg Score y Download Max Avg Sentiment) no fueron replicados.
La salida de cada procesamiento se vuelven a comunicar por medio de queues que recibirán los distintos procesos hasta que el cliente recibe los resultados obtenidos.

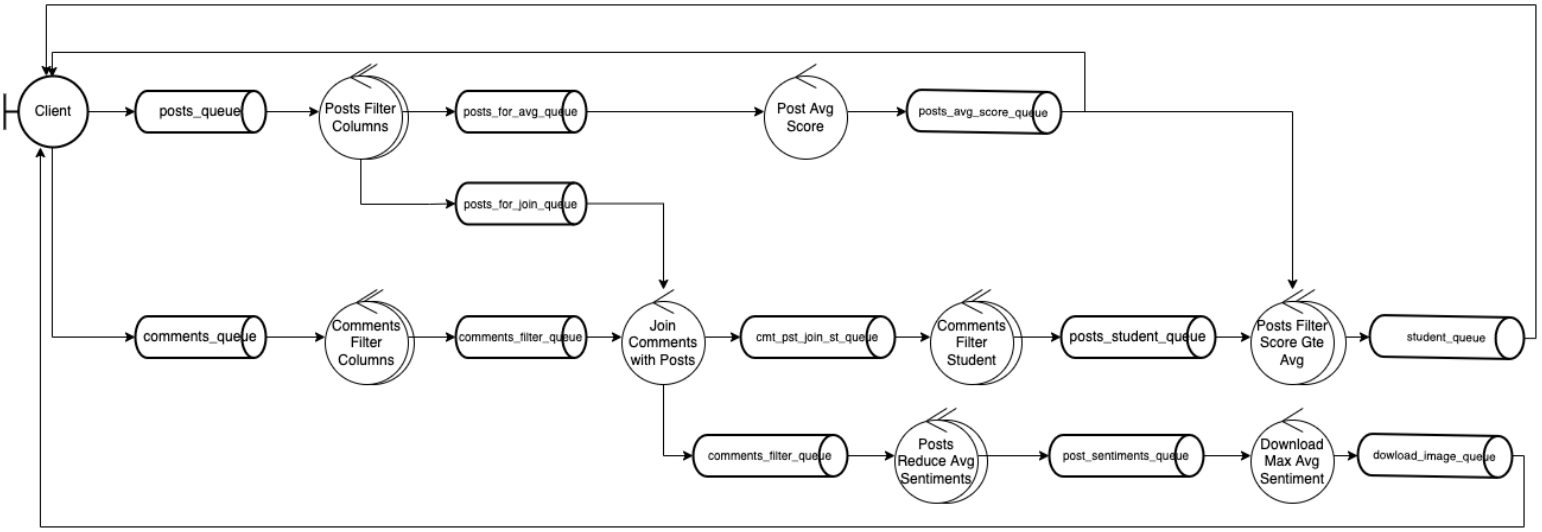
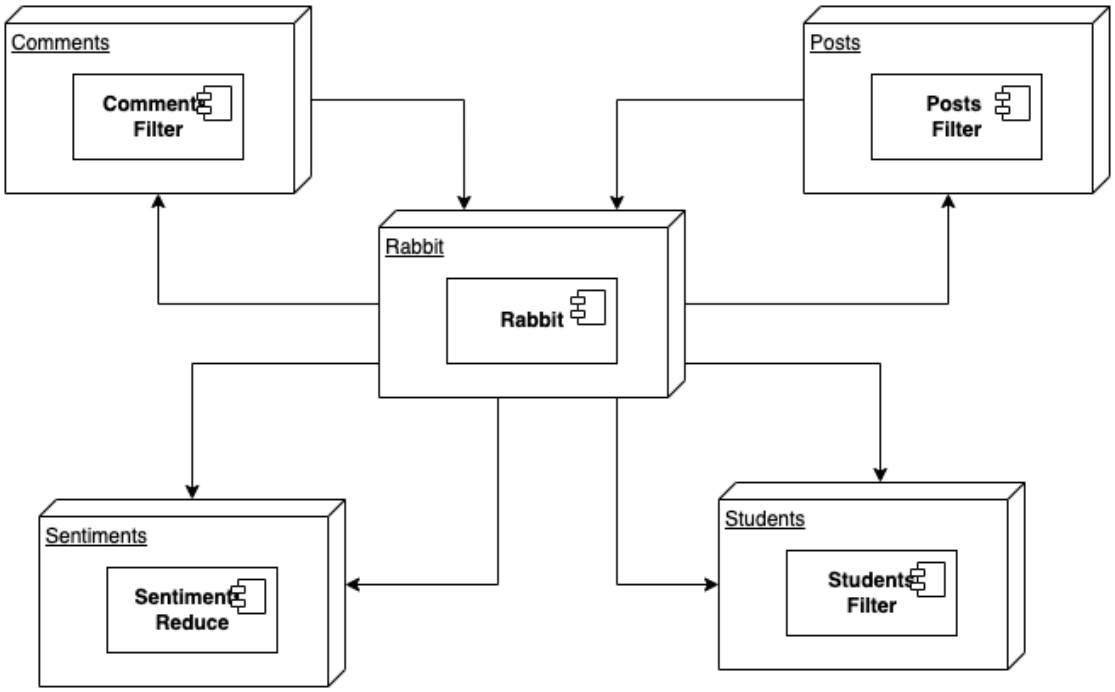


Diagrama de despliegue

Para el diagrama de despliegue se busco evitar que sea repetitivo al diagrama de robustez, por lo que se hizo más generico para representar y agrupar los distintos tipos de nodos.
Por un lado tenemos el nodo de Rabbit que es el middleware por el cual todos los nodos se comunican por eso es central a la arquitectura
Tambien tenemos un nodo de Posts, donde se encontraran Post Filter Columns y Post Avg Score, el nodo de Comments para Comments Filter Columns y el Join Post with Comments.

Luego los nodos Sentiment y Students que reciben la salida al terminar el join y realizan los filtros y reduce necesarios para generar las distintas salidas pedidas por enunciado



Vista de Procesos

Diagrama de Actividad

A continuación se muestra el diagrama de actividad para obtener los posts que cumplen gustarle a los estudiantes (con comments sobre university, college, student, teacher, professor) y con score mayor al promedio
No se tiene en cuenta en el diagrama el paso inicial de filtrar columnas, ya que no tiene significancia.

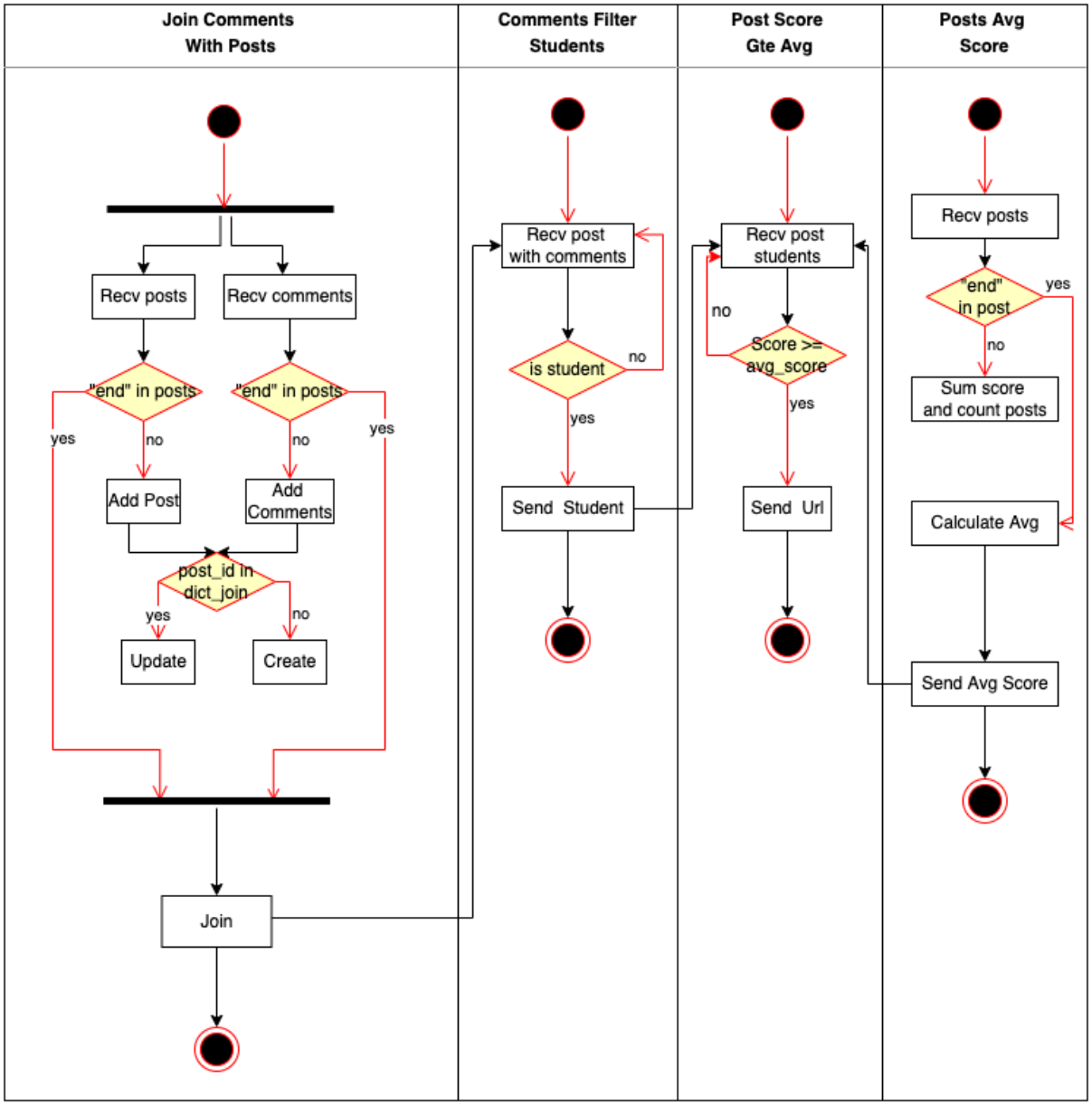
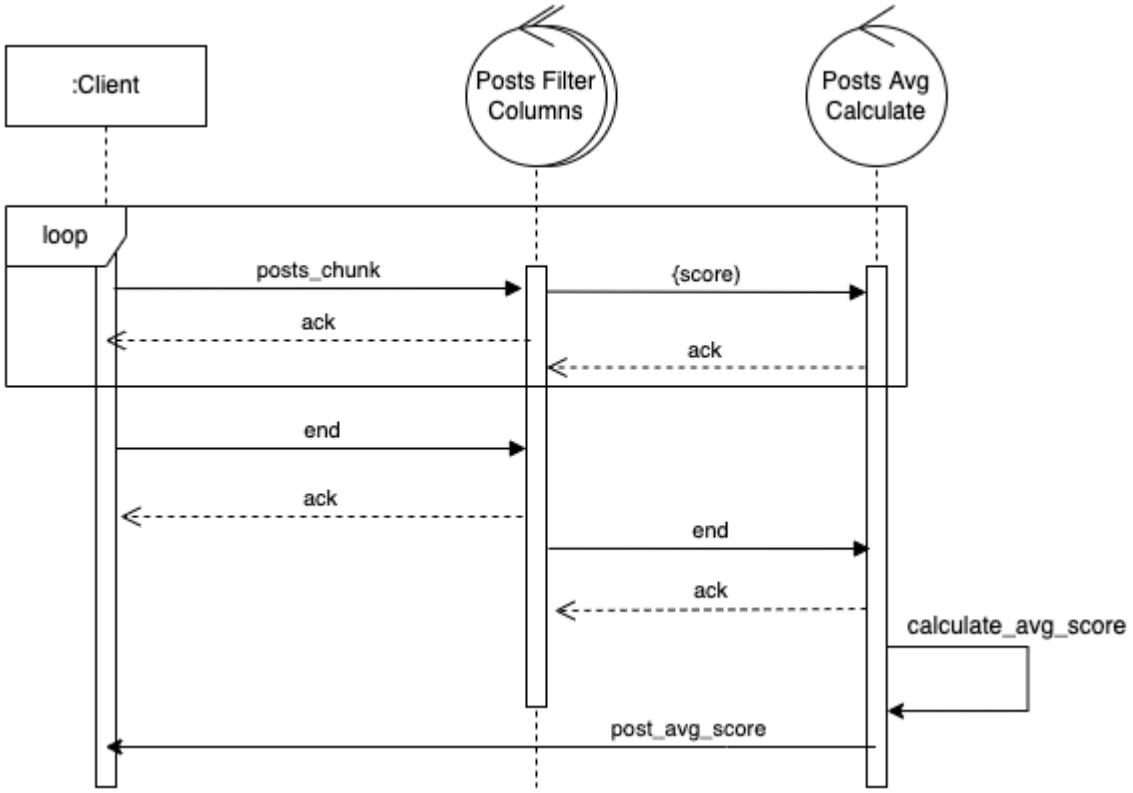


Diagrama de secuencia

Se muestra a continuación el diagrama de secuencia del recorrido hasta obtener el avg score de todos los posts.
El cliente comienza enviando todos los post del csv, que luego Post Filter Columns elimina los que fueron eliminados o no tienen url y se los que restan se lo envía al Post Avg Source. Este va recibiendo y sumando el score y contando la cantidad de post recibidos en variables internas. Cuando finalmente recibe el mensaje de finalización, calcula el avg_score y por último envía al cliente el resultado obtenido



Deuda Técnica

- CommentsFilter y PostFilter a pesar de tener varios workers, solo uno de ellos está consumiendo todos los mensajes
- Al enviar el último mensaje max_avg_sentiment se corta la conexión de rabbit y da error
- El cliente se comunica con el MOM y no debería
    - Agregar envio del csv por socket externo
- El cliente no está recibiendo los mensajes del sink