TP2
Retrieval Augmented Gen

Grupo 4:
- Nicolás Birsa
- Valentín Ye Li
- Salvador Castagnino

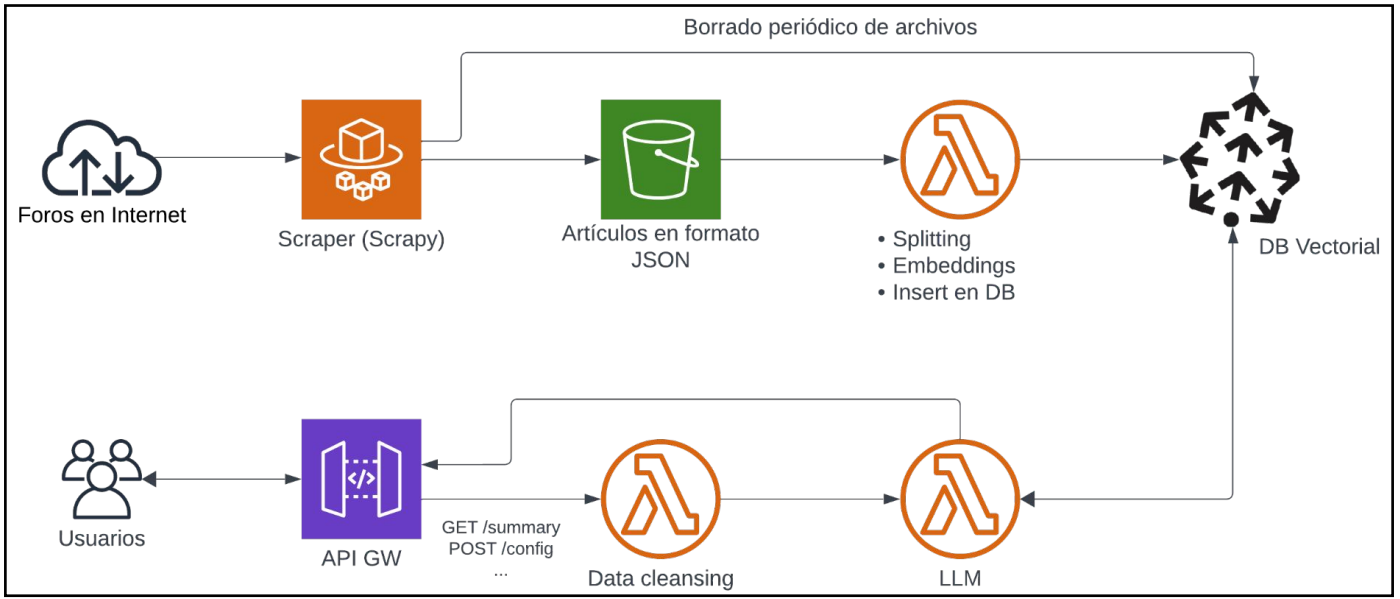
Problema

- Una persona puede estar interesada en seguir las novedades de un área, en particular lo que escriben ciertos **autores** sobre ciertos **temas**, pero no tener el tiempo para leer todo los artículos de manera comprensiva
- No es raro encontrar artículos que contengan **mucho ruido** y poco material de interés, como podemos extraer lo más valioso?

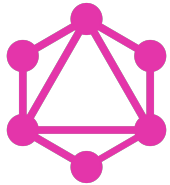
Solución

- El usuario elige un **conjunto de autores** y un **tema**, el sistema le provee con un listado de resúmenes
- Cada resumen corresponde a una publicación de uno de los autores seleccionados y explica qué es lo que dijo ese autor sobre ese tema en esta
- Solo se extraen las **publicaciones relevantes** con respecto al tema
- Utilizamos las publicaciones del **AI Alignment Forum**

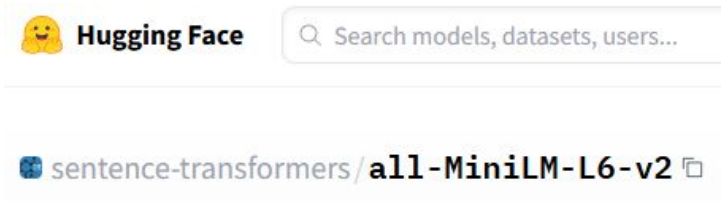
Arquitectura candidata



Scrapping y Carga



- Scrapeamos post de 30 días con Scrapy y Graphql
- Definimos un JSONLoader para matchear el esquema del scraping
- Usar **RecursiveCharacterTextSplitter** con un chunksize:1000 y overlap: 200 para mantener un contexto entre un mismo post
- Guardamos los chunks a pinecode con el embedding de HF “sentence-transformers/all-MiniLM-L6-v2”



Retrieval

- La **similitud** se calcula entre el embedding del tema elegido y el body de los chunks de las publicaciones, no nos interesa toda la publicacion sino solo los chunks que refieren al tema
- Los embedding se realizan con un **Sentence Transformer** (SBERT)
- Los chunks se **filtran por autor** para obtener solo de los autores seleccionados
- Se toma como relevante una publicación si la **suma de los scores** de sus chunks supera un **threshold** elegido por el usuario

Summarization

- Una vez filtrados los chunks se **agrupan por publicación** y por cada publicación se realiza 1 resumen
- Se realiza **prompt engineering** para tratar de obtener resúmenes mas acertados
- Los resúmenes se realizan con un LLM del **HuggingFaceEndpoint** utilizando **LangChain**
- Principal problema, el modelo que toma prompt genera resúmenes incoherentes

Summarization con Prompt...

```
4
5 Summary: strategies that do not rely on evaluating behavior on concrete inputs. In pa
  which we want to estimate the probability of a catastrophic tail event. Explain some
  contemporary AI systems. Discuss deceptive alignment as a particularly dangerous case
```

```
  functions f 0 , f 1 , . . . , f n
  can estimate the probability of ca
  several methods for estimating thi
10
11 We will
```

```
We will assume that C is also a neural network and expre
function which is 1 if and only if x is a catastrophic i
functions f 0 , f 1 , . . . , f n − 1 . Each f i represe
can estimate the probability of catastrophic behavior by
several methods for estimating this probability, each wi

We will assume that C is also a neural network and expre
function which is 1 if and only if x is a catastrophic i
functions f 0 , f 1 , . . . , f n − 1 . Each f i represe
can estimate the probability of catastrophic behavior by
several methods for estimating this probability, each wi
```

Pruebas y Resultados

Pruebas

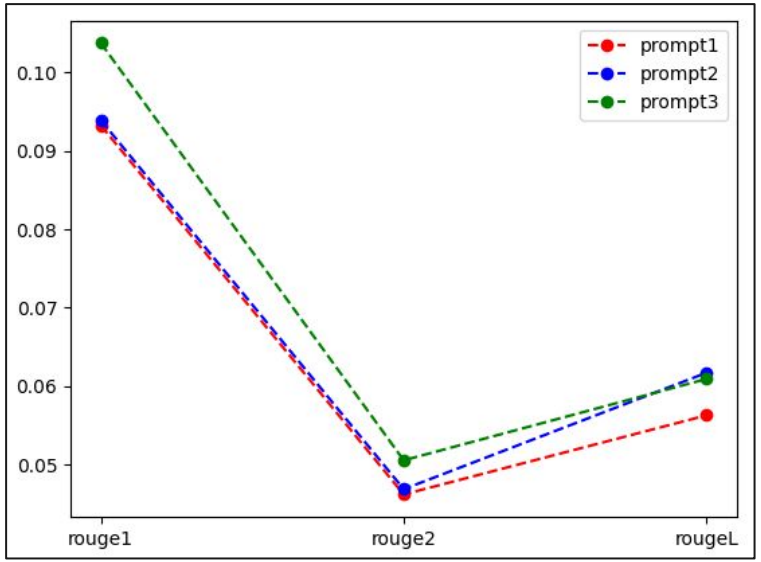
Se probaron 3 prompts distintos para generar los resúmenes:

1. **Make a summary of the following**
2. **Write a concise summary in third person of the following content**
3. **Write a cohesive short text in third person telling us what the author's thoughts regarding {topic} are in the following text**

Se evaluarán los resúmenes hechos en 7 documentos según dos métricas: una de similitud de contenido (**ROUGE**) y otra de legibilidad (**Flesh kincaid**).

Índices ROUGE - Con Prompt

Se enfocan en medir la similitud de N-gramas (ROUGE-1, ROUGE-2) y la subsecuencia de palabras más larga (ROUGE-L) entre el documento y sus resúmenes.



ROUGE-1 es útil para evaluar la cobertura general de las palabras clave.

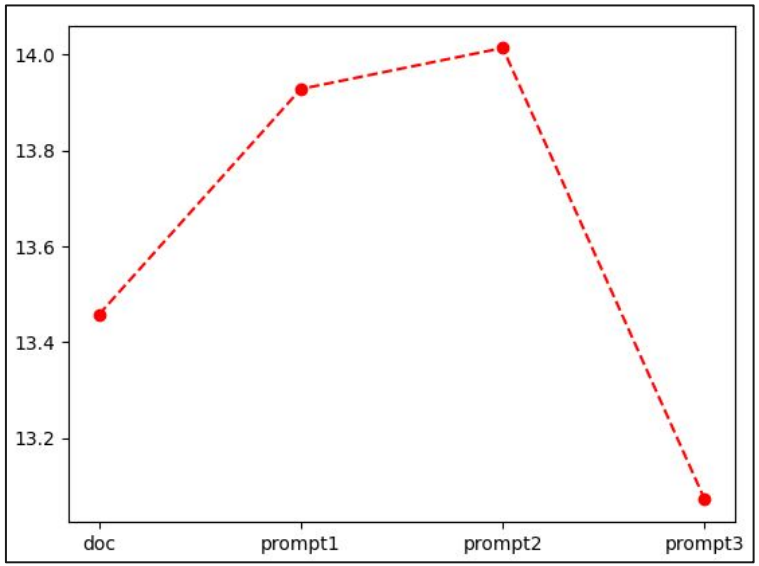
ROUGE-2 es más útil cuando el **orden y la cohesión** de las frases clave son importantes.

ROUGE-L es ideal para medir la **coherencia global** y **preservación de la estructura** del texto original.

En gral, >0,7 es muy bueno, 0,5-0,7 es aceptable, y <0,5 es malo/deficiente

Índice de legibilidad (Flesh kincaid) - Con Prompt

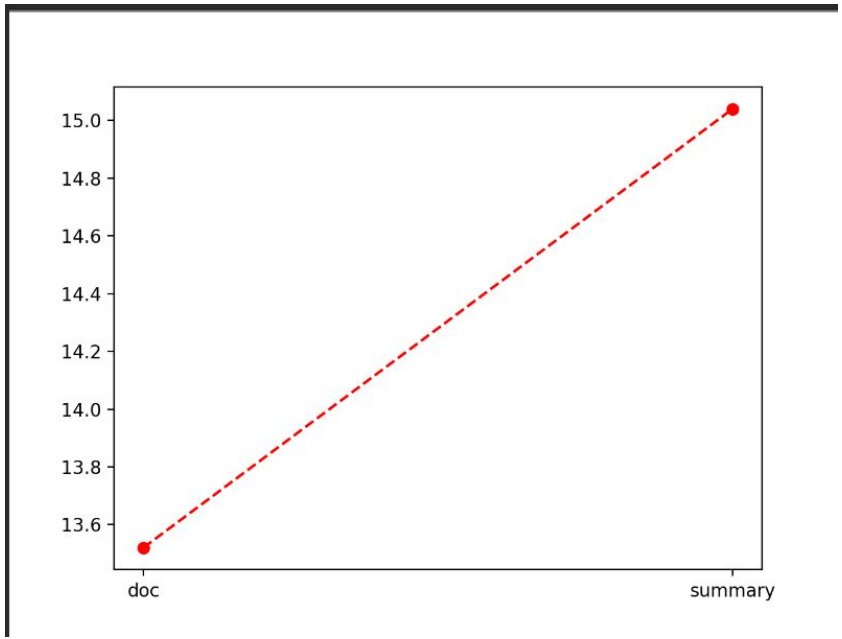
Es un valor que indica el grado de lectura del texto, donde a mayor valor, más fácil resulta leer el texto.



Score	School level (US)
100.00–90.00	5th grade
90.0–80.0	6th grade
80.0–70.0	7th grade
70.0–60.0	8th & 9th grade
60.0–50.0	10th to 12th grade
50.0–30.0	College
30.0–10.0	College graduate
10.0–0.0	Professional

Índice de legibilidad (Flesh kincaid) - Con Prompt

Los resúmenes utilizados tienen exclusivamente “high strictness”



Score	School level (US)
100.00–90.00	5th grade
90.0–80.0	6th grade
80.0–70.0	7th grade
70.0–60.0	8th & 9th grade
60.0–50.0	10th to 12th grade
50.0–30.0	College
30.0–10.0	College graduate
10.0–0.0	Professional

Conclusiones

- Es difícil llegar a conclusiones de las prompt cuando las respuestas del modelo son deficientes y semánticamente incoherentes sin importar la prompt o el contexto
- El ROUGE debería comprarse contra resúmenes generados por personas, por eso debe dar malos resultados
- La readability debe ser baja (graduate o profesional) ya que los post son altamente técnicos, observemos que los resúmenes tienden
- La readability sin prompt es mejor que con prompt, aun así por poco

Demo

Ejemplos - Demo

Ejemplo 1

- Authors: All
- Topic: superhuman ai
- Strictness: mid

Ejemplo 2

- Authors: All
- Topic: superhuman ai
- Strictness: high

Ejemplos - Demo

Ejemplo 3

- Authors: Sam Bowman, Dan H
- Topic: superhuman ai
- Strictness: mid