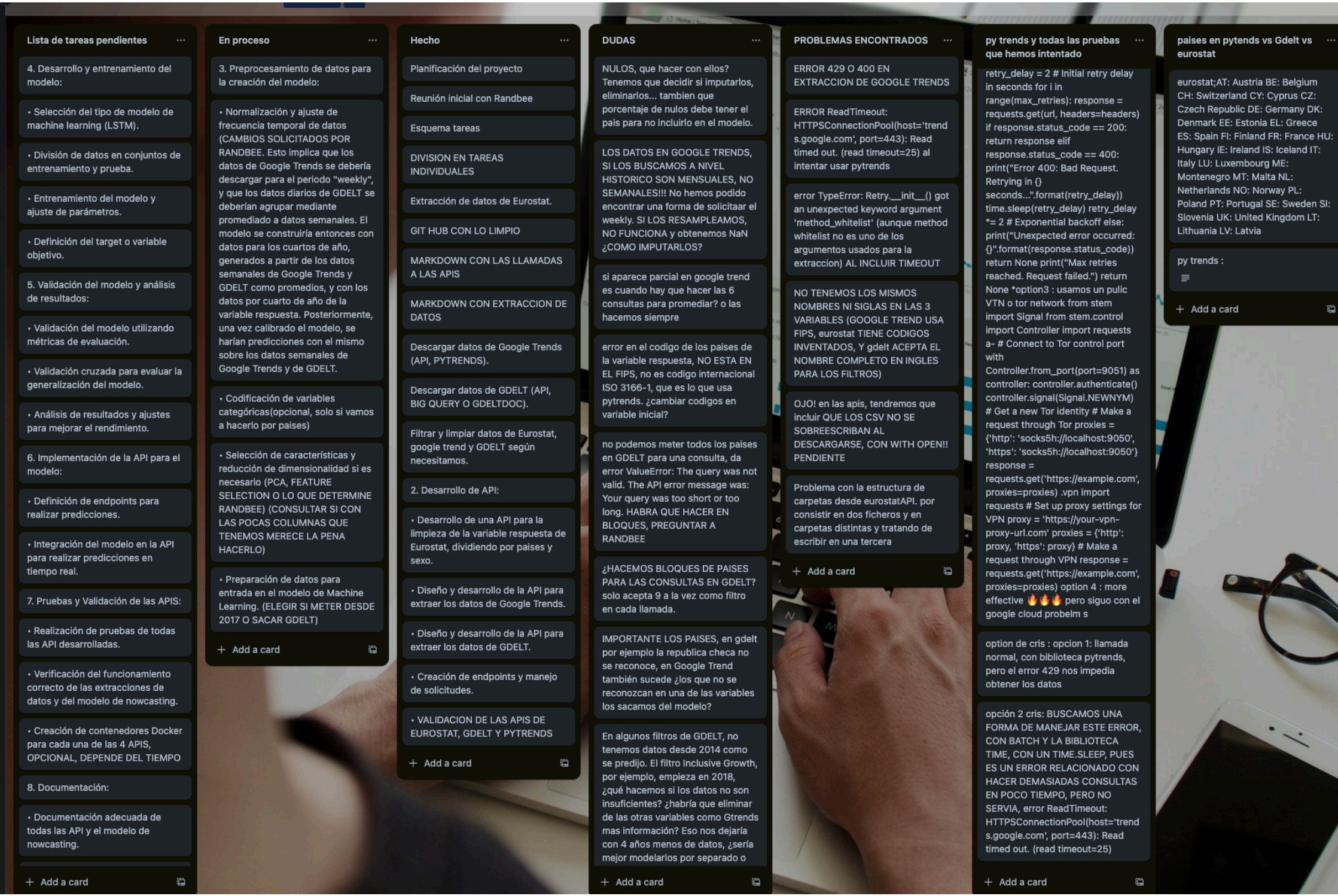

Lista de tareas pendientes

4. Desarrollo y entrenamiento del modelo:

• Selección del tipo de modelo de machine learning (LSTM).

• División de datos en conjuntos de entrenamiento y prueba.

• Entrenamiento del modelo y ajuste de parámetros.

• Definición del target o variable objetivo.

5. Validación del modelo y análisis de resultados:

• Validación del modelo utilizando métricas de evaluación.

• Validación cruzada para evaluar la generalización del modelo.

• Análisis de resultados y ajustes para mejorar el rendimiento.

6. Implementación de la API para el modelo:

• Definición de endpoints para realizar predicciones.

• Integración del modelo en la API para realizar predicciones en tiempo real.

7. Pruebas y Validación de las APIS:

• Realización de pruebas de todas las API desarrolladas.

• Verificación del funcionamiento correcto de las extracciones de datos y del modelo de nowcasting.

• Creación de contenedores Docker para cada una de las 4 APIS, OPCIONAL, DEPENDE DEL TIEMPO

8. Documentación:

• Documentación adecuada de todas las API y el modelo de nowcasting.

+ Add a card

En proceso

3. Preprocesamiento de datos para la creación del modelo:

• Normalización y ajuste de frecuencia temporal de datos (CAMBIOS SOLICITADOS POR RANDBEE. Esto implica que los datos de Google Trends se debería descargar para el periodo "weekly", y que los datos diarios de GDELT se deberían agrupar mediante promediado a datos semanales. El modelo se construiría entonces con datos para los cuartos de año, generados a partir de los datos semanales de Google Trends y GDELT como promedios, y con los datos por cuarto de año de la variable respuesta. Posteriormente, una vez calibrado el modelo, se harían predicciones con el mismo sobre los datos semanales de Google Trends y de GDELT.

• Codificación de variables categóricas(opcional, solo si vamos a hacerlo por paises)

• Selección de características y reducción de dimensionalidad si es necesario (PCA, FEATURE SELECTION O LO QUE DETERMINE RANDBEE) (CONSULTAR SI CON LAS POCAS COLUMNAS QUE TENEMOS MERECE LA PENA HACERLO)

• Preparación de datos para entrada en el modelo de Machine Learning. (ELEGIR SI METER DESDE 2017 O SACAR GDELT)

+ Add a card

Hecho

Planificación del proyecto

Reunión inicial con Randbee

Esquema tareas

DIVISION EN TAREAS INDIVIDUALES

Extracción de datos de Eurostat.

GIT HUB CON LO LIMPIO

MARKDOWN CON LAS LLAMADAS A LAS APIS

MARKDOWN CON EXTRACCION DE DATOS

Descargar datos de Google Trends (API, PYTRENDS).

Descargar datos de GDELT (API, BIG QUERY O GDELTDOC).

Filtrar y limpiar datos de Eurostat, google trend y GDELT según necesitamos.

2. Desarrollo de API:

• Desarrollo de una API para la limpieza de la variable respuesta de Eurostat, dividiendo por paises y sexo.

• Diseño y desarrollo de la API para extraer los datos de Google Trends.

• Diseño y desarrollo de la API para extraer los datos de GDELT.

• Creación de endpoints y manejo de solicitudes.

• VALIDACION DE LAS APIS DE EUROSTAT, GDELT Y PYTRENDS

+ Add a card

DUDAS

NULOS, que hacer con ellos? Tenemos que decidir si imputarlos, eliminarlos... tambien que porcentaje de nulos debe tener el pais para no incluirlo en el modelo.

LOS DATOS EN GOOGLE TRENDS, SI LOS BUSCAMOS A NIVEL HISTORICO SON MENSUALES, NO SEMANALES!!! No hemos podido encontrar una forma de solicitaar el weekly. SI LOS RESAMPLEAMOS, NO FUNCIONA y obtenemos NaN ¿COMO IMPUTARLOS?

si aparece parcial en google trend es cuando hay que hacer las 6 consultas para promediar? o las hacemos siempre

error en el codigo de los paises de la variable respuesta, NO ESTA EN EL FIPS, no es codigo internacional ISO 3166-1, que es lo que usa pytrends. ¿cambiar codigos en variable inicial?

no podemos meter todos los paises en GDELT para una consulta, da error ValueError: The query was not valid. The API error message was: Your query was too short or too long. HABRA QUE HACER EN BLOQUES, PREGUNTAR A RANDBEE

¿HACEMOS BLOQUES DE PAISES PARA LAS CONSULTAS EN GDELT? solo acepta 9 a la vez como filtro en cada llamada.

IMPORTANTE LOS PAISES, en gdelt por ejemplo la republica checa no se reconoce, en Google Trend también sucede ¿los que no se reconozcan en una de las variables los sacamos del modelo?

En algunos filtros de GDELT, no tenemos datos desde 2014 como se predijo. El filtro Inclusive Growth, por ejemplo, empieza en 2018, ¿qué hacemos si los datos no son insuficientes? ¿habría que eliminar de las otras variables como Gtrends mas información? Eso nos dejaría con 4 años menos de datos, ¿sería mejor modelarlos por separado o

+ Add a card

PROBLEMAS ENCONTRADOS

ERROR 429 O 400 EN EXTRACCION DE GOOGLE TRENDS

ERROR ReadTimeout: HTTPSConnectionPool(host='trends.google.com', port=443): Read timed out. (read timeout=25) al intentar usar pytrends

error TypeError: Retry.__init__() got an unexpected keyword argument 'method_whitelist' (aunque method whitelist no es uno de los argumentos usados para la extraccion) AL INCLUIR TIMEOUT

NO TENEMOS LOS MISMOS NOMBRES NI SIGLAS EN LAS 3 VARIABLES (GOOGLE TREND USA FIPS, eurostat TIENE CODIGOS INVENTADOS, Y gdelt ACEPTA EL NOMBRE COMPLETO EN INGLES PARA LOS FILTROS)

OJO! en las apis, tendremos que incluir QUE LOS CSV NO SE SOBREESCRIBAN AL DESCARGARSE, CON WITH OPEN!! PENDIENTE

Problema con la estructura de carpetas desde eurostatAPI. por consistir en dos ficheros y en carpetas distintas y tratando de escribir en una tercera

+ Add a card

py trends y todas las pruebas que hemos intentado

retry_delay = 2 # Initial retry delay in seconds for i in range(max_retries): response = requests.get(url, headers=headers) if response.status_code == 200: return response elif response.status_code == 400: print("Error 400: Bad Request. Retrying in {} seconds...".format(retry_delay)) time.sleep(retry_delay) retry_delay *= 2 # Exponential backoff else: print("Unexpected error occurred: {}".format(response.status_code)) return None print("Max retries reached. Request failed.") return None *option3 : usamos un pulic VTN o tor network from stem import Signal from stem.control import Controller import requests a- # Connect to Tor control port with Controller.from_port(port=9051) as controller: controller.authenticate() controller.signal(Signal.NEWNYM) # Get a new Tor identity # Make a request through Tor proxies = {'http': 'socks5h://localhost:9050', 'https': 'socks5h://localhost:9050'} response = requests.get('https://example.com', proxies=proxies) .vpn import requests # Set up proxy settings for VPN proxy = 'https://your-vpn-proxy-url.com' proxies = {'http': proxy, 'https': proxy} # Make a request through VPN response = requests.get('https://example.com', proxies=proxies) option 4 : more effective 🔥🔥🔥 pero siguo con el google cloud probelm s

option de cris : opcion 1: llamada normal, con biblioteca pytrends, pero el error 429 nos impedia obtener los datos

opción 2 cris: BUSCAMOS UNA FORMA DE MANEJAR ESTE ERROR, CON BATCH Y LA BIBLIOTECA TIME, CON UN TIME.SLEEP, PUES ES UN ERROR RELACIONADO CON HACER DEMASIADAS CONSULTAS EN POCO TIEMPO, PERO NO SERVIA, error ReadTimeout: HTTPSConnectionPool(host='trends.google.com', port=443): Read timed out. (read timeout=25)

+ Add a card

paises en pytends vs Gdelt vs eurostat

eurostat;AT: Austria BE: Belgium CH: Switzerland CY: Cyprus CZ: Czech Republic DE: Germany DK: Denmark EE: Estonia EL: Greece ES: Spain FI: Finland FR: France HU: Hungary IE: Ireland IS: Iceland IT: Italy LU: Luxembourg ME: Montenegro MT: Malta NL: Netherlands NO: Norway PL: Poland PT: Portugal SE: Sweden SI: Slovenia UK: United Kingdom LT: Lithuania LV: Latvia

py trends :

+ Add a card