Contexto Roadmap Actual

Para desarrollar un modelo que conecte imágenes y texto, es necesario representar las palabras como vectores. Esto requiere construir un **diccionario basado en la lista de medicamentos disponible**. En paralelo se trabaja con la parte de Optical Character Recognition.

Base de Datos

La base de datos fue generada gracias al **web scraping** implementado por Ale. A través de operaciones de cruce, cada medicamento en la base final contiene la siguiente estructura:

```
medicamento = {
        nombre: string,
        abreviaciones: list,
        dosis: string,
        marcas: list,
        principios_activos: list,
        alerta_LASA: boolean
}
```

Resultados del Web Scraping

El código generado por Ale entregó los siguientes resultados:

Listado Principios Activos.xlsx

- **Principio Activo**
- **Dosis**
- **Forma Farmacéutica**

Lista de Registros 2.csv

- Registro (único para entrar a Chile)
- Nombre
- Referencia de Trámite
- Equivalencia Terapéutica o Biosimilar

- Titular
- Estado del Registro
- Resolución "Inscríbase"
- Fecha "Inscríbase"
- Última Renovación
- Fecha Próxima Renovación
- Régimen
- Vía de Administración
- Condición de Venta
- Expende Tipo de Establecimiento
- Indicación
- Fecha Próxima Renovación
- Procesado

Ruido y Errores en Recetas

Agregar ruido al modelo, basado en distancias entre palabras, puede ser clave para simular **errores comunes** y mejorar la transparencia en alertas LASA.

Observaciones del Equipo QF

- Crear un diccionario con:
  - **Errores comunes de una palabra.**
  - **Abreviaciones comunes.**
- Verificar que las recetas controladas incluyan explícitamente:
  - Cantidad.
  - Dosis.
  - Forma de administración.
    Esto previene rechazos en farmacias y ayuda a usuarios finales, como adultos mayores, a verificar recetas en tiempo real.

Modelo Probabilístico

El modelo predice palabras con una probabilidad ajustable en base a:

- **Dosis.**

- **Vía de administración.**
- **Síntomas.**

Métodos Experimentales

1. **Filtros estrictos**: Solo incluir síntomas compatibles.
2. **Modificación de probabilidades**: Ajustar valores en caso de que los síntomas no coincidan exactamente, permitiendo flexibilidad en el modelo.

Consideraciones sobre Tokenización

Se discutieron tres enfoques:

1. **Modelo que prediga líneas completas**:
   - Ventaja: Capacidad de capturar contexto completo.
   - Desventaja: Difícil de entrenar y generalizar; requiere más capacidad computacional.
2. **Predicción de palabras individuales**:
   - Ventaja: Más sencillo.
   - Desafío: Manejar abreviaciones y errores comunes.
3. **Subtokenización (dividir palabras en partes)**:
   - Equipo QF considera que no es útil debido a la falta de estandarización y posibles confusiones semánticas (ej. "para" vs. "cetamol").

Beneficios de Implementar Word2Vec

- **Modelo generativo**: Usar los vectores de medicamentos, marcas y dosis para generar líneas sintéticas de recetas con modelos como **difusión generativa**.
- **Dataset sintético**:
   - Generar recetas completas utilizando OpenCV para combinar líneas sintéticas.
   - Verificar calidad y usar estas recetas para entrenar modelos con menos de 1000 datos reales, una técnica común en el estado del arte.

Tareas para la Próxima Semana

Ale:

- Incluir los principios activos en el pipeline.

Fabi:

- Implementar el modelo **Word2Vec** para medicamentos.
- Configurar **Tesseract** u otra herramienta de OCR para identificar cuadros de texto en recetas.

Dani:

- Estudiar los detalles presentes en las recetas, adoptando una visión holística del problema para identificar posibles problemáticas.

Notas Finales

Este esquema permitirá avanzar en el diseño del pipeline para representar medicamentos como vectores, generar datasets sintéticos y mejorar la capacidad de predicción del modelo.