¿Cuántas fuentes de datos existen?

Además de la fuente de conocimiento que constituyen los distintos miembros del proyecto (los cuáles se listan a continuación), gracias a su conocimiento, y a la posibilidad de transmitirlo para el avance del proyecto:

-BI Devs, con conocimientos en ETL, limpieza de datos, preprocesamiento de datos, modelamiento de datos, aprendizaje de máquina: supervisado, no supervisado, y profundo.

-BI Analyst, con conocimientos en estadística básica (medidas de tendencia central, de dispersión, distribuciones, pruebas de hipótesis, correlaciones), exploración visual de datos, evaluación y selección de modelos de datos.

-Data Scientist, con conocimientos avanzados en infraestructura y despliegue de servicios para consumo y almacenamiento de datos (basados en AWS), despliegue de APIs REST, despliegue de servicios de gestores de bases de datos como PostgreSQL y MongoDB. Recurrir a él para soporte técnico de infraestructura.

-Product Owner, y Scrum Master, conocedores de la metodología Scrum y familiarizados con las tareas técnicas a desarrollar en el proyecto; para definirlas y vigilar su ejecución. El product owner, naturalmente, también está al tanto de los requerimientos comerciales y de patrocinadores; y a partir de ello define las actividades del proyecto.

Tenemos también como fuentes de datos para el proyecto:

Un repositorio con archivos de texto con datos no estructurados obtenidos mediante webscrapping y una etapa inicial de procesamiento; la cual consistió en remover las etiquetas HTML y el contenido ajeno a la información divulgativa dispuesta para el usuario final del sitio web de donde fueron extraídos. Esta etapa inicial de procesamiento, sin embargo, no limpió completamente el resultado de caracteres especiales, y símbolos ajenos a las palabras del idioma español.

El contenido de los archivos del repositorio corresponde al texto de artículos de divulgación académica que fueron originalmente publicados por terceros en el portal Wikipedia. Los artículos de Wikipedia fueron seleccionados aleatoriamente; por tanto, no guardan relación.

Estos archivos fueron provistos por el cliente para que en el proyecto se pueda extraer un banco de palabras de acuerdo a los siguientes lineamientos:

1. No incluir palabras con más de una letra en mayúscula (acrónimos).

2. No incluir stop words.

3. No incluir palabras con menos de 4 caracteres.

2. No permitir letras con tildes. Reemplazar letras con tildes por sus homólogos sin ellas.

3. No permitir caracteres de formato especiales (ej: "\t", "\n", "\u200b"), símbolos (@, !, ?, etc...), y letras ajenas al lenguaje español. Reemplazarlos por espacios en blanco.

8. Finalmente todas las palabras obtenidas deben estar en minúscula.

El análisis del banco de palabras a construir con los archivos de la fuente de datos indicada, y su procesamiento, constituirá el conjunto de datos de referencia para los sistemas predictores de palabras (las palabras fuera de este banco no podrán ser predichas por los predictores a desarrollar), y será lo que fundamentará los lineamientos básicos de su lógica de predicción.

Como puede entenderse a partir de lo explicado anteriormente, los datos de la fuente descrita no se encuentran en un formato estructurado, deduplicado, ni conforme a los requerimientos del cliente. Por otro lado, el proceso de limpieza podría comprometer su completitud. Estas condiciones perjudican su calidad.

Adicionalmente, para futuras etapas del proyecto se contará con las siguientes fuentes de datos:

Base de datos tipo PostgreSQL alojada en Render; la cual contendrá los resultados de las pruebas de cada método predictor de palabras implementado, así como un directorio de datos con las palabras del banco de referencia asociadas a su id. Las tablas con los resultados de las pruebas deberán poblarse a medida que se realicen pruebas con un predictor, siguiendo los lineamientos definidos para almacenar datos de calidad en la misma.

Base de datos no SQL en MongoDB; la cual contendrá información sobre cada usuario de método predictor de palabras, y los intentos de predicción realizados por cada usuario; de acuerdo con una estructura definida previamente. Los documentos de esta base de datos se poblarán automáticamente a medida que un usuario consumas las API del sistema Wordle, siguiendo los lineamientos definidos por el cliente para almacenar datos de calidad en la misma. Una de las particularidades de esta base de datos, es que sus documentos registrarán las fechas en que un usuario consumió las Api del sistema Wordle; esto permitirá, más adelante, medir la agilidad con la que los usuarios logran predecir palabras.

Aunque estas bases de datos se actualizan de manera independiente, las tablas o documentos que almacenarán información sobre los intentos de predicción en cada una de ellas deberán compartir la hora en que se realizó cada intento, el id de la palabra que se intentaba predecir, y la palabra estimada en la predicción.