LINEAMIENTOS PARA EL PROYECTO FINAL

Les dejo a continuación una serie de lineamientos para que se guíen al momento de realizar el Proyecto Final. Los mismos son de carácter orientativo para quienes no estén seguros de cómo aplicar los contenidos vistos hasta el momento, es decir, pueden quitar o incorporar técnicas más allá de las mencionadas de forma libre. Es importante que cuando busquen un dataset, consideren que contenga alguna columna con cadenas de texto para que puedan aplicar las técnicas de NLP vistas hasta el momento. Otro aspecto interesante es que los datos contengan una columna con valores que les permita crear un modelo de aprendizaje supervisado.

Si usas datos de tu trabajo te recomendamos que procures anonimizar los mismos para evitar problemas con la empresa para la cual prestás servicios. Quienes no tengan dataset recuerden que en el sitio https://www.kaggle.com/ pueden acceder a una gran cantidad de conjuntos de datos para desarrollar su proyecto. Les planteamos que estructuren, a fines meramente organizativos, el trabajo en dos etapas.

- ESTRUCTURA DEL TRABAJO.
    - Descripción del problema de negocio.
    - Objetivo general
    - Origen de los datos
    - Definición de las variables.
    - Librerías a utilizar.
    - Desarrollo*(ETAPA 1 Y 2)
    - Conclusiones (Ejemplo: Insight, observaciones, resultados obtenidos, rechazo o no de una hipótesis).
    - Perspectivas futuras del proyecto.

- Desarrollo - ETAPA 1: Trabajar con Procesamiento de Lenguaje Natural (Eligiendo algunas de las técnicas vistas en clase):

    - Lectura del documento
    - Quitar símbolos y signos de puntuación
    - Tokenizar
    - Convertir a minúsculas
    - Remover stopwords
    - Lematización o stemming
    - Crear una o más nubes de palabras
    - Crear N-gramas
    - Aplicar análisis de Sentimiento

- Desarrollo - ETAPA 2 - Vectorización y entrenamiento de modelo:
    - MACHINE LEARNING:
        - TF-IDF
        - Bag of Words (BOW)
    - DEEP LEARNING: Text to sequence de Keras.

Podés combinar estas técnicas con un modelo de aprendizaje Supervisado utilizando modelos vistos en cursos anteriores y/o practicando con Redes Neuronales Artificiales dentro de lo que sería Deep Learning.

    - Para facilitar esta etapa del trabajo, al momento de usar un modelo de Machine Learning, sugerimos que trabajes con algún modelo sencillo como puede ser un modelo clasificación como Regresión Logística.
    - Si contás con un dataset con una columna de carácter binomial podes usarla como label o variable a predecir del modelo.
    - Si trabajás con un dataset con una columna de puntaje y valores del 1 al 5 podés:
        - Convertir el puntaje de la columna “puntaje” a una variable binomial. Mantenés los valores de entre 5 y 4 como buenos(1), eliminas filas con valor 3 y convertís los valores 1 y 2 como malos (0).
        - Con esta nueva columna podés armar un modelo de clasificación utilizando un modelo visto en el curso de Data Science II.

- **CUESTIONES A TENER EN CUENTA**
    - Es fundamental que seas organizado con los pasos de tu proyecto, aprovechá la estructura sugerida.
    - Comentá el código para demostrar que entendés lo que se está haciendo en cada etapa.
    - Expresá insights, conclusiones, aspectos que valgan la pena señalar de los resultados obtenidos.
    - No se trata de crear modelos con un rendimiento excepcional ni de crear un conocimiento revolucionario para la industria, el proyecto final es una oportunidad para demostrar tus capacidades utilizando los contenidos vistos.
    - Es importante que te comprometas con el proyecto dado que el mismo puede terminar formando parte de tu portfolio como Científico de Datos.

Recordá que lo que buscamos es que aprendan a utilizar las herramientas vistas en clase, que sepan interpretar resultados y analicen los datos en su contexto. Cualquier duda aprovechen los canales habilitados.

¡Saludos!

Eze