Descripción Proyecto Final

Minería de Textos

Primavera 2024

1 Planteamiento del problema

Recientemente, el Signa_Lab del ITESO participó en el procesamiento y selección de preguntas para el debate presidencial de México. Esto debido a que el Instituto Nacional Electoral (INE) decidió que fueran los ciudadanos quienes plantearan las preguntas de acuerdo a sus intereses, preocupaciones y prioridades.

El INE también diseño un conjunto de requerimientos que el procesamiento debería cumplir para organizar, eliminar, depurar y seleccionar las preguntas que se les harían a los candidatos durante el debate.

2 Requerimientos del INE

2.1 Dataset con los registros de las preguntas

El dataset “crudo” entregado por el INE contiene los siguientes campos:

- ID del registro
- Entidad de Origen
- Edad
- Género
- Identificación con algún grupo en situación de discriminación
- Tema de la pregunta
- Texto de la pregunta
- Desea agregar otra pregunta
- Tema de la pregunta 2
- Texto de la pregunta 2
    - Se repiten los 3 campos anteriores hasta tener máximo 6 preguntas por cada ID
- Día y hora (timestamop) del envío de la pregunta

2.2 Condiciones para aceptación de una pregunta

1. Las preguntas deben apegarse a los temas definidos para el debate

2. La redacción de las preguntas no debe incluir: discurso de odio, inclinación partidista, ideológica o religiosa o cualquier manifestación de violencia o discriminación, o referirse a algún logro de gobierno y/o propaganda gubernamental
3. Las preguntas deben redactarse en forma general y abierta y no estar dirigidas a una candidatura específica.

3 Actividades para el proyecto

3.1 Limpieza del Dataset

La primera actividad consiste en la limpieza del dataset, de manera que éste cumpla con los requerimientos del "Tidy Data", es decir:

1. Cada renglón o fila del dataset corresponde a una observación o individuo
2. Cada columna del dataset corresponde a una y sólo una variable
3. No se mezcla información de naturaleza diferente

En el caso del dataset entregado por el INE, estos principios, sobre todo el número 1, no se respetan ya que en nuestro caso, una observación/individuo es una (y sólo una) pregunta.

El resultado de esta actividad es un dataset (almacenado en un archivo CSV) que contenga los siguientes campos:

- ID del registro
- Entidad de Origen
- Edad
- Género
- Identificación con algún grupo en situación de discriminación
- Tema de la pregunta
- Texto de la pregunta

Es decir, en los casos en los que una persona envió más de una pregunta, cada pregunta debe estar en su renglón con el resto de los datos de identificación.

3.2 Eliminación de preguntas que no cumplen los criterios

Con el dataset limpio, podemos comenzar a realizar un análisis utilizando algunas técnicas de NLP para eliminar aquellas preguntas que no cumplan con los criterios del INE establecidos en el punto 2.2.

El resultado de este paso serán:

- Un archivo que contenga el texto de las preguntas que no cumplen los criterios y una breve explicación de porqué no está cumpliéndolos (discurso de odio, violencia, malas palabras, etc.)

- Un archivo con las preguntas restantes (los campos serán los mismos que en el punto 3.1

3.3 Eliminación de preguntas repetidas

Ya con todas las preguntas que cumplen los requisitos del INE para ser tomadas en cuenta, vamos a hacer otra depuración para eliminar aquellas preguntas que están repetidas.

Es importante considerar que dos preguntas pueden ser iguales aún cuando estén formuladas con palabras diferentes (es decir, una puede ser paráfrasis de la otra).

El resultado de este paso serán dos archivos:

- Uno que contenga el texto de las preguntas repetidas, así como una métrica que mida su parecido. De estas preguntas, sólo se conservará 1 en el dataset de trabajo.
    - La pregunta que será conservada es aquella que fue enviada primero de acuerdo al timestamp
- Un archivo con las preguntas depuradas (eliminando las repeticiones)

3.4 EDA con las preguntas restantes

Ya con el dataset limpio y depurado, podemos comenzar a realizar análisis exploratorio para extraer información a partir de los registros y las preguntas.

Con el dataset depurado vamos a intentar responder a las siguientes preguntas, utiliza gráficos para ilustrar tus respuestas:

1 Distribución de las preguntas por:

1.1 Tema

1.2 Entidad

1.3 Género

1.4 Grupos vulnerables

1.5 Grupos de edad (13-18, 18-25, 25-40 y mayores)

2 Cruces entre las variables del punto anterior

2.1 Qué temas interesan más por género

2.2 Por entidad

2.3 Por grupo de edad

3 Separa las preguntas por tema y genera nubes de palabras y tablas de frecuencia para mostrar:

3.1 qué palabras, bi-gramas o tri-gramas son los que más se repiten por tema

3.2 Por entidad

3.3 Por género

4 Separa las preguntas por tema y realiza un análisis de sentimientos y de emociones.

4.1 Muestra los resultados en forma tabular o gráfica

3.5 Selección de preguntas según metodología del INE (opcional)

4 Entregables

- Documento con el reporte de hallazgos en el proceso:
  - Portada con el nombre de la materia, periodo y nombre del alumno
  - Introducción
  - Hallazgos en cada una de las actividades descritas en el punto 3
    - En su caso, gráficas generadas y explicación de las mismas
  - Piezas de código que se usaron para generar esos resultados
    - Explicación general de los modelos utilizados, los parámetros e hiperparámetros
    - Explicación general de las funciones definidas
  - Conclusiones y aprendizajes
  - Bibliografía
- Código (scripts en python o jupyter notebooks) usados en cada una de las actividades realizadas.
  - Señalar claramente las librerías utilizadas (de preferencia con número de versión incluido)
  - Documentar las funciones principales, así como los parámetros definidos para utilizar los modelos
- Archivos en formato CSV con los resultados de cada actividad señalada en el punto 3
  - Se debe conservar el dataset original y en cada etapa del proceso generar nuevos archivos según el caso.
  - Utilizar nombres descriptivos para los archivos

4.1 Formato de los entregables

Todos los entregables se deben adjuntar en un archivo comprimido (.zip) para subirse a la plataforma Canvas en el espacio correspondiente.