**Diploma Universitario**

**en Ciencias Sociales Computacionales y Humanidades Digitales**

**Consigna Trabajo FInal**

**Opción 2 (Módulo 5)**

**Integrantes: Ignacio Bastías, Thelma Fratarelli y Beatriz Soria**

**Introducción**

El trabajo final del módulo 5 se basará en utilizar el corpus provisto (cuyas características se describen a continuación) y recorrer todo el flujo de trabajo visto en clase: preprocesamiento, generación de distribuciones de frecuencias de palabras y modelado de tópicos. Prestaremos especial atención en la corrección a la explicación y justificación de las decisiones tomadas, en tanto muestren manejo de los tópicos vistos en clase.

**Consignas**

A partir del corpus deberán:

* Cargar los datos
* Preprocesarlos (normalizar texto, eliminar stopwords)
* Generar una matriz token-por-fila para cada documento

A continuación deberán responder las siguientes consignas:

* ¿Cuáles son las palabras más utilizadas en cada uno de los medios? ¿Pueden verse diferencias? (Tener en cuenta las diferentes métricas trabajadas en el curso: tf, tf-idf, etc.) Generar las visualizaciones que considere más pertinentes para responder la pregunta
* ¿Cuáles son los tópicos principales en el corpus? ¿Pueden evidenciar diferencias en cada uno de los medios? Explicar qué método se utilizó para responder la pregunta, cuáles son los supuestos del mismo. Generar las visualizaciones más adecuadas para responder a las preguntas
* A continuación, seleccionar las noticias vinculadas a algún tópico relevante (por ejemplo, “Elecciones”) y construir un clasificador para predecir la orientación del diario[[1]](#footnote-1). Utilizar alguno de los modelos de clasificación vistos a lo largo de al Diplomatura (regresión logística, random forest, etc.). Utilizar como features el [“Spanish Billion Word Corpus and Embeddings”](https://crscardellino.ar/SBWCE/), analizado en clase (pueden descargar el embedding en formato .bin del link). ¿Qué resultados arroja el modelo? ¿Es posible mediante el texto de las noticias conocer la línea editorial del diario? Generar las visualizaciones y tablas correspondientes para una correcta evaluación del modelo.

**Entregables**

Se esperan dos entregables:

* Un documento en formato word, google doc (no pdf) en el que se desarrollan las respuestas a las consignas, se presentan los principales resultados (tablas, visualizaciones) y las interpretaciones.
* Un notebook/script en el que se realiza el procesamiento del texto y se generan los modelos y visualizaciones correspondientes.

La entrega se hará mediante el sistema Google Classroom

**Modalidad de trabajo**

El trabajo deberá hacerse en grupos de 2 personas. El trabajo deberá ser subido a Google Classroom *por solamente uno de los integrantes* del grupo. Deberá constar en el trabajo el nombre de ambos.

**Fecha de entrega**

La fecha última de entrega será el 17/3/2024 a las 23:59. No se concederán prórrogas.

#frente no la quité

**palabras\_vacias <- c("a", “aún”, "ante", "bajo", "con", "contra", "de", "desde", "durante", "en", "entre", "hacia", "hasta", "mediante", "para", "por",**

**"según", "sin", "sobre", "también", "más", "año", "años", "está", "están", "había", "después", "qué", “que”, "además", "así", “embed', 'leé', “cómo”, “como”, “sólo”, “través”, “tenía”, “sí”, “si”, “podría”,'cronica.com.ar', 'minutouno.com', 'comentar', 'jpg', 'loading', 'pristupluk', ' minutouno.com', 'guardar', ' páginai12', '01', 'l.l', 'loading', 'jpe', 'más', 'también', 'está', 'había', 'qué', 'así', 'están', 'además', 'día', '1', 'gusta', 'twitter', 'fuente', 'whatsapp', 'compartir', 'facebook', 'mail', 'él')**

1. Cabe destacar que la clasificación de orientación de la línea editorial es provisoria y puede ser modificada por ustedes sí les parece necesario. [↑](#footnote-ref-1)