PRÁCTICA 1: Análisis del impacto de la pandemia en la salud mental de usuarios/as de Redes Sociales (parte I)

Tal y como se indica en el título, en esta práctica trataremos de responder a una pregunta con un interés científico real: ¿Cómo ha afectado la pandemia a la salud mental de las personas? ¿Se puede extraer algún signo o pista mediante tratamiento automático de información sobre la evolución psicológica de las personas desde el inicio de la pandemia? Para ello, utilizaremos herramientas de Análisis de Textos, Recuperación de Información, Análisis de sentimiento y de extracción de dimensiones psicológicas que nos permitirán estimar si existe algún tipo de relación entre los eventos acaecidos en estos dos últimos años y lo que las personas publican en sus redes sociales. No todo el mundo tiene un perfil de redes sociales, pero estudios previos han demostrado que estos contenidos suelen aportar una buena pincelada de la sociedad en general.

Pasos a seguir:

1) como la gran mayoría de los proyectos de análisis de datos, el primer paso consiste en conseguir el acceso a los datos que se quieren tratar. En el caso de este proyecto, posteriormente se os proporcionará un dataset completo con el trabajar. Sin embargo, consideramos interesante que experimentéis por vuestra cuenta cómo obtener y procesar datos de una fuente como Twitter. Para ello, utilizaremos la librería Python Twint (https://github.com/MarcosFP97/twint), la cual permite obtener datos de forma masiva sin utilizar la API oficial de Twitter. Pasos para la instalación:

git clone https://github.com/MarcosFP97/twint.git
cd twint/
python setup.py install

El principal objetivo es obtener todos los tweets posibles (sin restricción de usuario) que hablen de un determinado tópico (p.ej. Afghanistan).

2) el conjunto de documentos que obtengáis en el paso anterior debe ser almacenado en disco en un formato adecuado. Esto permitirá la posterior reproducción de los análisis o procesamientos que se hagan sobre el corpus ya normalizado (u otros posteriores que se quiera hacer sobre los mismos datos). Para ello, definid un esquema CSV o JSON que permita almacenar toda la información disponible.

3) realizad un simple procesamiento del corpus anterior para vectorizar la colección y mostrar los términos con mayor ponderación tf/idf. Para ello:

a) instalad y familiarizaros con scikit-learn (http://scikit-learn.org/stable/) y, en particular, sus posibilidades para extraer características a partir de texto (sección 5.2.3 de la página http://scikit-learn.org/stable/modules/feature_extraction.html#feature-extraction) y el TfIdf Vectorizer:
http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

b) dado el corpus obtenido y considerando cada tweet como un documento individual, utilizad el TfIdf Vectorizer (filtrando stopwords y todas aquellas palabras que aparezcan en menos de 10 docs) para vectorizar la colección y seguidamente mostrar los 50 términos más “centrales” en la colección. Entendiendo como más centrales aquellos cuya suma acumulada de tf/idf sobre todos los documentos es mayor. Adicionalmente, mostrad también los 100 términos más repetidos en la colección (suma de su tf en los documentos).

4) Llegados a este punto, podemos proceder a descargarnos los datos que utilizaremos para nuestro análisis sobre la salud mental en la pandemia (https://nubeusc-my.sharepoint.com/:u:/g/personal/marcosfernandez_pichel_usc_es/EY7v-EkfJvJJkgFUJpDx1lkBymjTEEOcaQaK2cw1_k9JbA?e=VkfSZV), para esta práctica sólo nos interesa la US data. Se trata de una colección generalista de tweets localizados en EEUU y alrededores durante los años 2020 y 2021.

5) Una vez que tenemos la colección cargada, procederemos a hacer el análisis de sentimiento de los tweets utilizando la librería VADER (https://github.com/cjhutto/vaderSentiment). Esto nos dará una idea de la positividad, negatividad o neutralidad de los tweets publicados a lo largo de la pandemia.

6) Hacer algún tipo de representación gráfica de ese sentimiento, utilizando la granularidad temporal que mejor se ajuste para la extracción de conclusiones (pista: tened cuidado con qué representamos, entender bien el valor de "compound" en VADER).

Escribid en un par de líneas las conclusiones que se aprecian relacionando las variaciones en el sentimiento con eventos de la pandemia (inicio o fin de olas, aumento de restricciones, etc.).

7) (apartado opcional, 0.5 puntos) Utilizad el siguiente enlace (https://covid.cdc.gov/covid-data-tracker/#trends_dailycases) para descargar los datos diarios de infecciones por COVID-19 en EEUU.

Si añadimos la curva de contagios a la representación de sentimiento del apartado anterior, nos permitirá extraer mejores conclusiones.

Entregables:

1) Guión python (.py)
2) Python Notebook (.pynb)

Es fundamental que el Notebook sea autoexplicativo de todos los pasos (con celdas textuales acompañando a celdas con código y que contenga explícitamente los resultados -sin tener que ejecutar las celdas de nuevo-). Comprobad esto antes de enviar el Notebook. Cualquier proyecto de Analítica de Datos debe ser autodocumentado y sus experimentos fáciles de reproducir. Un aspecto clave en la evaluación de esta práctica reside en la calidad de las explicaciones y documentación que acompañéis al código dentro del Notebook.

- Valoración y Fecha de Entrega:
  - Esta práctica tiene una valoración de 4 puntos (sobre el total de 7 puntos de la parte práctica de la materia). 3.5 puntos se corresponden a la correcta realización de los apartados obligatorios -apartados de 1) a 6)- y el 0.5 se corresponde con la correcta realización del apartado optativo.

Fecha límite entrega: 12 de noviembre
Se permiten entregas retrasadas pero se reducirá la puntuación del siguiente modo:
- Cada día tarde reduce en un 10% de la máxima nota alcanzable (es decir, cada día tarde resta un 0.4 puntos de la nota que se os asigne al valorar la práctica)