1. Exporta una conversación entre dos personas de Whatsapp, la forma de exportarla como txt es la siguiente :

   Primero seleccionas la opción "Más" sobre tu conversación , posteriormente seleccionas la opción de "Exportar chat" y finalmente eliges la opción "SIN ARCHIVOS" , de esta forma se exporta la conversación en .txt a algún correo.

Figure 1: PASO I

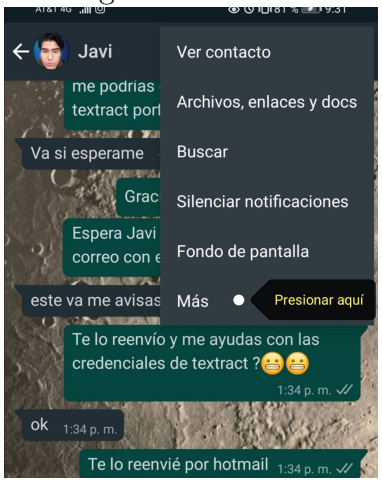
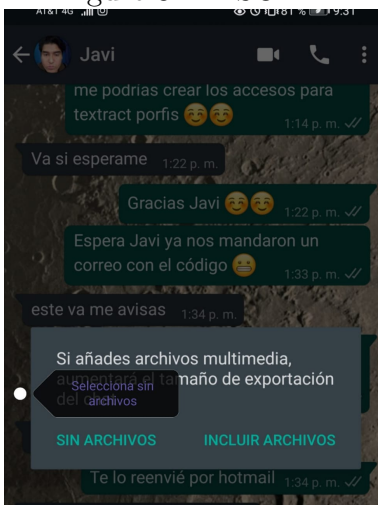


Figure 2: PASO II

Figure 3: PASO III



2. El documento con extensión .txt transformar en un DataFrame de la siguiente forma:

Figure 4: PASO III

	t_fecha	v_hora	t_texto	v_usuario
0	2019-01-12	9	Holi Jaime, te compartÃ una carpeta en drive...	Carla
1	2019-01-12	9	Hoy estarÃ© fuera d emi Casita pero si necesi...	Carla
2	2019-01-12	9	Holi Carlita, vientos, muchas gracias, eres l...	Jaime
3	2019-01-12	9	Bambi Bambi , graciaaas	Jaime
4	2019-01-12	16	Carlitaaa	Jaime
...	...	...	...	...
1128	2020-10-29	21	El de el creo si lo hizo chido	Jaime
1129	2020-10-29	21	Va va va ð□□¬	Carla
1130	2020-10-29	21	Se lo voa pedir â□º¡□â□º¡□	Carla
1131	2020-10-29	21	Si si si	Jaime
1132	2020-10-29	21	Mejor para no errarle xD	Jaime

1028 rows × 4 columns

- "t_fecha" es la fecha en la que el mensaje fue enviado , en un formato de año-mes-día
- "v_hora" es la hora en que el mensaje fue enviado , debe estar en un formato de 24 horas
- "t_texto" es el contenido de cada mensaje enviado
- "v_usuario" es el nombre de la persona que mando el mensaje

Aspectos importantes a considerar:

- La nomenclatura de las variables debe ser la misma.
- De acuerdo a la configuración de tu reloj, la hora puede venir en un formato de 12 o 24 horas. Si se tiene en un formato de 12 horas es necesario hacer uso de la información proporcionada en tu txt que indique si es "am" o "pm", por ejemplo :

Figure 5: Hora

	v_hora	t_texto
0	01/12/19 9:29 a.Â m.	Carla: Holi Jaime, te compartÃ una carpeta e...
1	01/12/19 9:31 a.Â m.	Carla: Hoy estarÃ© fuera d emi Casita pero si...
2	01/12/19 9:57 a.Â m.	Jaime: Holi Carlita, vientos, muchas gracias,...
3	01/12/19 9:58 a.Â m.	Jaime: Bambi Bambi , graciaaas
4	01/12/19 4:00 p.Â m.	Jaime: Carlitaaa
...	...	...
1128	29/10/20 9:33 p.Â m.	Jaime: El de el creo si lo hizo chido
1129	29/10/20 9:33 p.Â m.	Carla: Va va va ð□□¬
1130	29/10/20 9:33 p.Â m.	Carla: Se lo voa pedir â□º□ï¸□â□º□ï¸□
1131	29/10/20 9:33 p.Â m.	Jaime: Si si si
1132	29/10/20 9:33 p.Â m.	Jaime: Mejor para no errarle xD

1133 rows × 2 columns

En este caso el "txt" nos proporciona la información "a.A.m" o "p.A.m" con la cual podemos conocer que parte del día es, utilizamos esta información para añadirle 12 horas al valor original de la hora, de esta forma tendremos un formato de 24 horas.

3. Una vez teniendo la tabla en esa estructura procedemos a hacer lo siguiente:

   (a) Haz el conteo de missings por variable y crea un dataframe con el siguiente formato y guardarlo en una variable llamada "missings"

Figure 6: DataFrame Missings

	nombre_columna	total_missings
0	d_fecha	0
1	v_hora	0
2	t_texto	0
3	v_usuario	0

   En caso de tener "missings" eliminar los registros de tu DataFrame

   (b) Posteriormente haz el conteo de registros que tienen como valor "Multimedia omitido" en la variable "t_texto" y asignala a una variable llamada "total_multimedia", adicional a lo anterior elimina dichos registros. El número de registros mínimos para poder trabajar los siguientes pasos debería ser 200.

Figure 7: Multimedia Omitido

	d_fecha	v_hora	t_texto	v_usuario
17	2019-01-12	22	<Multimedia omitido>	Carla
23	2019-01-12	22	<Multimedia omitido>	Carla
51	2020-01-22	22	<Multimedia omitido>	Carla
112	2020-05-29	17	<Multimedia omitido>	Carla
115	2020-06-30	13	<Multimedia omitido>	Carla
...	...	...	...	...
919	2020-10-17	11	<Multimedia omitido>	Carla
949	2020-10-21	12	<Multimedia omitido>	Jaime
964	2020-10-21	16	<Multimedia omitido>	Jaime
967	2020-10-21	16	<Multimedia omitido>	Carla
1017	2020-10-27	19	<Multimedia omitido>	Carla

84 rows × 4 columns

   (c) Verifica la calidad de los datos

(d) Realiza la limpieza de la variable "v_usuario" , solo debes tener dos categorías en esta columna, el texto no debe contener caracteres especiales , en minúsculas y sin espacios

(e) La variable "d_fecha" conviertela a tipo "datetime"

(f) La variable "v_hora" conviertela a tipo entero

(g) Realiza la limpieza de la variable "t_texto"

(h) Elimina stop words

(i) Elimina hapaxes

(j) Realiza tokenización

(k) Realiza derivación (Stemmer) en español, se puede utilizar de la siguiente forma :
from nltk.stem import SnowballStemmer
spanish_stemmer = SnowballStemmer('spanish')
(spanish_stemmer.stem("texto_en_str"))

(l) Creación de variables(Se deben crear al menos 10 nuevas variables):
Utilice Count vectorizer o Tfidf Vectorizer para generar carácteristicas además de uno de los métodos vistos en clase para generar nuevas carácteristicas sobre el texto(longitud , análisis de sentimientos,reconocimiento de entidad).

Si se utiliza análisis de sentimientos se debería hacer lo siguiente:
Primero se debería traducir el texto a inglés, utilizando TextBlob
traducir = lambda x: TextBlob(x).translate(to="en")
Como paso siguiente se utilizará TextBlob(x).sentiment.polarity

NOMBRE DE LA TABLA FINAL : df_text

PARTE 2 – Ingeniería Continuas

Con el conjunto de datos "BOLSAA.MX.csv" realice las siguientes actividades:

(a) Etiquetado de las variables

(b) Calidad de datos, para verificar que no existan anomalías

(c) Verifique que los datos estén ordenados por fecha

(d) ¿Cuál es la diferencia promedio entre fechas?

(e) (TABLA 1) Generé las siguientes variables sin cambiar la frecuencia de los registros (con fechas originales, sin agrupaciones):

- Realice el desfase de la variable "Cierre" de tal forma que los valores por registro le corresponda el valor de "Cierre" de dos periodos posteriores
- Variable "diff_max_ min" que sea la diferencia de los valores High con los valores Low
- Variable "pct_cie_ape" cambio porcentual entre el valor de cierre y apertura
- Variable "ratio" que corresponda a los valores de cierre entre los valores de apertura
- Cree las siguientes variables para esta lista: ['Open', 'High', 'Low', 'Volume']
  -Columnas de diferencia de 1 periodo a 3 períodos para cada variable
  -Columnas de ventana de tiempo de 1 ,..., 3 para cada variable utilizando la media
  -Columnas de cambio porcentual de 1 periodo a 3 períodos para cada variable
  -Generar el cuadrado de cada variable

(f) (TABLA 2) Cambie la frecuencia de las fechas de tal forma que los registros sean cada 3 días y genere al menos cuatro nuevas variables por cada variable original. No olvide hacer el desfase de la variable objetivo.

NOMBRES DE LA TABLAS FINALES : df_ori, df_3_days