Whatsapp Chat Analysis

WhatsApp es la aplicación de mensajería instantánea para smartphones más popular del mundo con más de 2000 millones de usuarios activos. Permite conversar utilizando además del texto tradicional emoticonos, multimedia, urls… por lo que se generan datos de forma masiva en cada conversación de los usuarios de esta aplicación.

Para este proyecto, se ha decidió hacer un análisis de datos de un chat de grupo y se ha optado por el chat de grupo de la clase del máster, pensando que podría extraerse información valiosa e interesante de la misma. Se trata de un grupo con 52 participantes y creado el 7 de septiembre de 2020.

Lo primero ha sido exportar el chat desde el móvil en forma de txt, y se ha hecho uso de la aplicación *Jupyter Notebook* para poder trabajar con ella. Será necesario importar las librerías necesarias que se indican en el propio notebook.

Una vez se lee el fichero, se extraen los datos y se generan las variables y *dataframes* oportunos, se puede empezar con el análisis del contenido del chat. Ha sido necesario cargar un diccionario en castellano, puesto que es el idioma en el que se comunican los miembros de este chat.

Members activity

Lo primero ha sido identificar a los miembros del grupo que más participación han tenido durante este tiempo, al igual los que menos. Para ello, se han agrupado por miembro y ordenados de forma descendente en función de la suma. Se han seleccionado los primeros diez y los últimos diez, y representándolo gráficamente se obtiene lo siguiente:

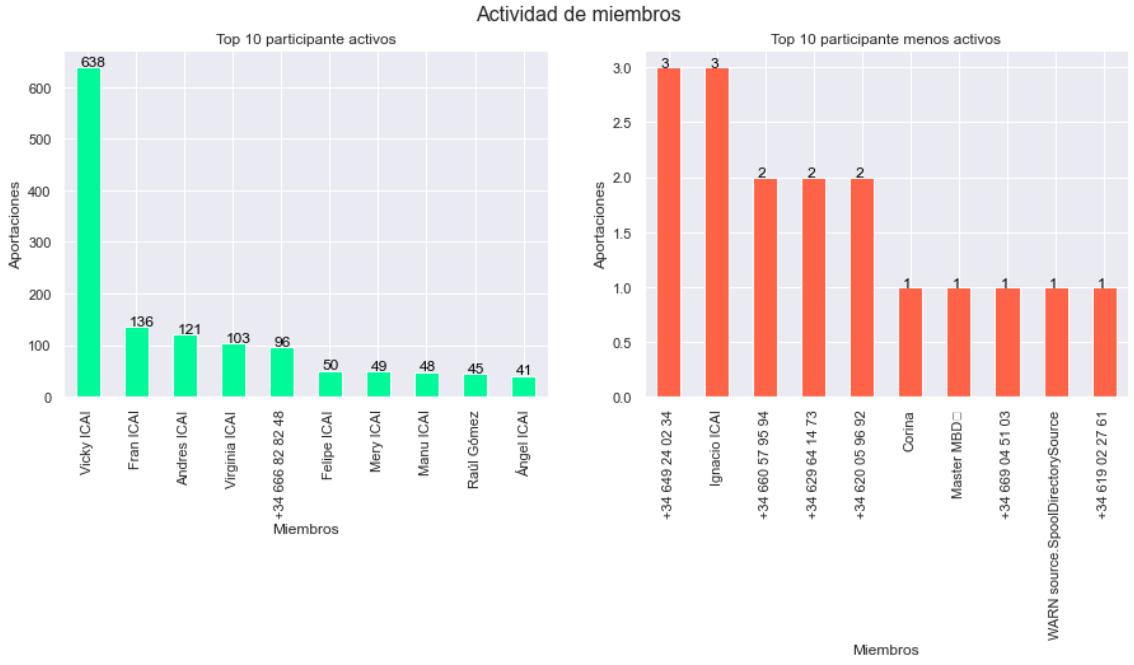

Ilustración 1 Actividad por miembros del grupo.

En el *barchart* de la izquierda de color verde se pueden observar los 10 miembros que han sido más activos en el chat, y en el gráfico de la derecha, los miembros menos activos, en color rojo.

Heatmap

Si por otro lado, se quiere analizar cuáles han sido los momentos álgidos de la conversación se tendrán en cuenta el día y la hora en el que se han enviado los mensajes. Para empezar, se representa un *heatmap* en función del día de la semana y de la hora del día en el que se ha hecho el envío:

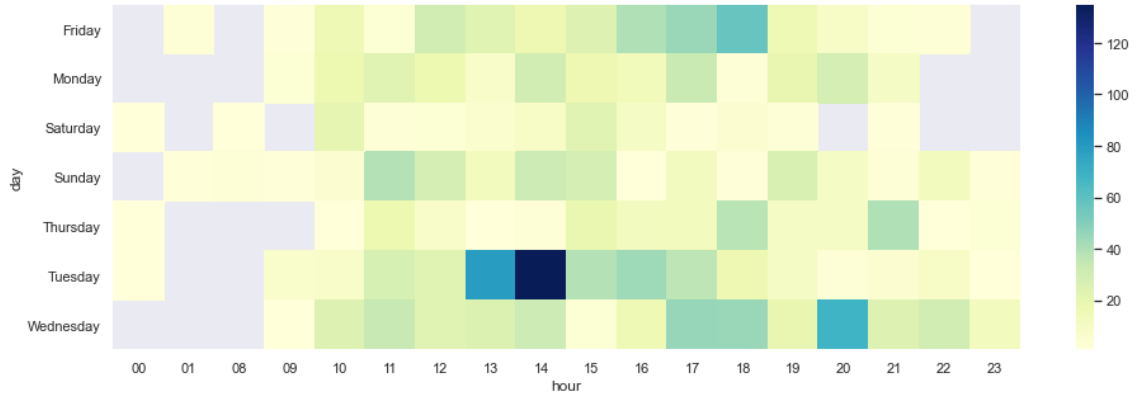

Ilustración 2 Heatmap.

En los mapas de calor se puede percibir la variación en base al color. Se identifica un valor atípico los martes a las 14 horas. Se puede interpretar que el numero más alto de mensajes se genera en esa franja. La diferencia entre esa hora con el resto del día parece considerable, pero para poder verlo de forma más clara se ha decidido generar otro gráfico donde se tengan en cuenta las horas del día y la cantidad de chats.

Active hours

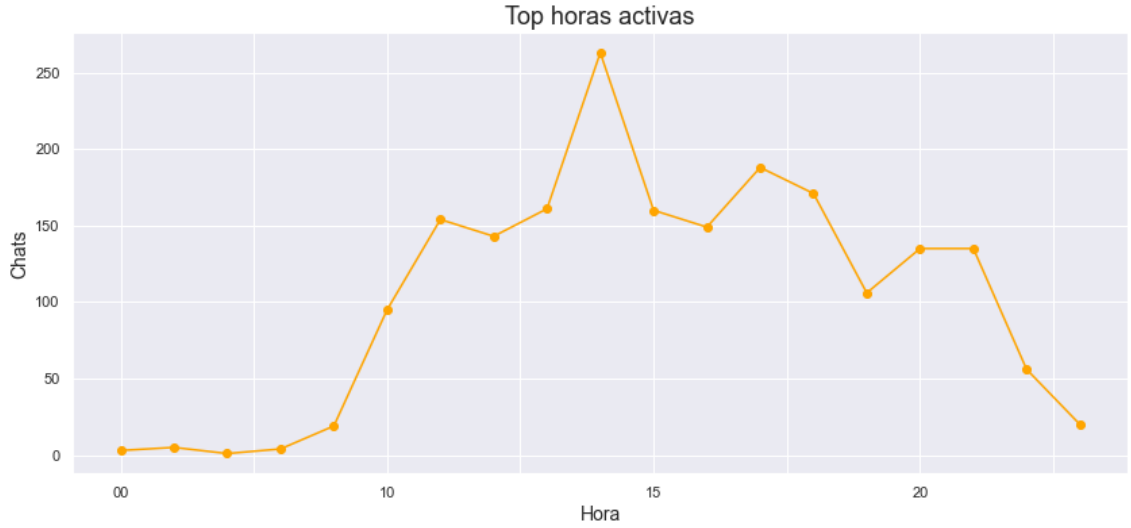

Ilustración 3 Diagrama lineal de horas más activas

Efectivamente, se confirma que el pico se da a las 14 horas. Además de eso, se pueden sacar conclusiones como que durante la noche no hay apenas actividad y a partir de las 9 de la mañana empiezan a mandarse mensajes por el grupo. A las 15:00 ocurre una bajada importante, que coincide con el comienzo de las clases por lo que parece lógica dicha bajada. Por otro lado, a las 18:00 vuelve a haber un ligero incremento, a pesar de que seguimos en horario de clase y a partir de las 21:00 vuelve a decrecer con el fin de las clases.

Active weekdays

En el mapa de calor mostrado anteriormente, el día de la semana con más actividad parecía ser el martes, sin embargo, parece que respecto al resto de días no hay demasiada diferencia en función de la cantidad de mensajes. Para analizarlo detenidamente se ha generado un radar que refleje la actividad por día de la semana:

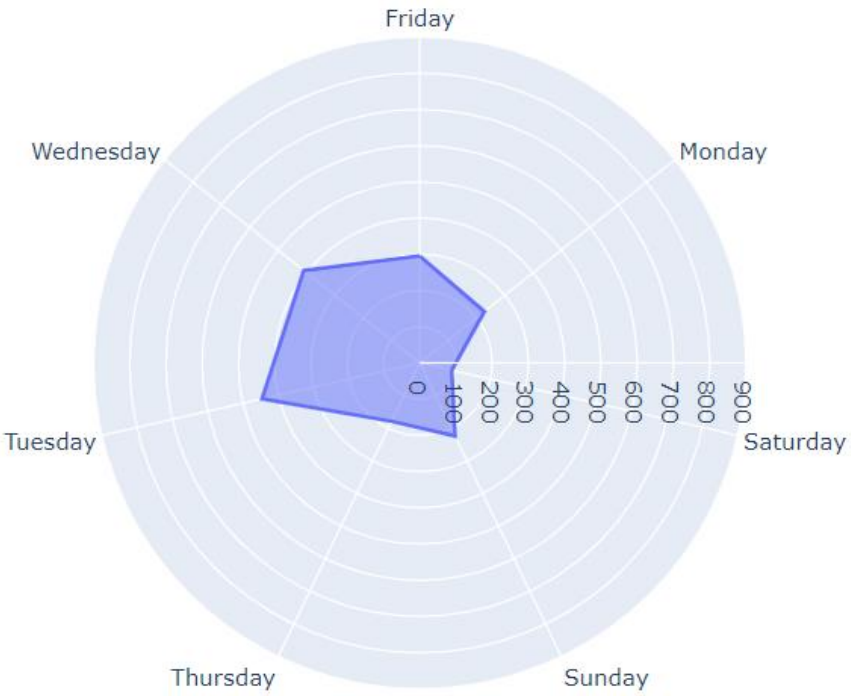

Ilustración 4 Radar de días de la semana.

Se vuelve a corroborar lo que decía el heatmap. En cambio, con esta representación, dicha diferencia respecto al resto de días de la semana no parece tan excesiva, por ejemplo respecto a los miércoles. Se puede ver que los sábados son los días con menor actividad en el chat de grupo, lo cual parece razonable teniendo en cuenta que es el día de la semana que la mayoría de las personas utiliza para desconectar y tomarse un descanso del trabajo o estudios. El domingo vuelve a haber actividad, probablemente relacionado a los trabajos, entregas o exámenes que se tengan la semana posterior.

Type of attachments sent

Como se ha mencionado, Whatsapp permite adjuntar archivos a las conversaciones para enriquecerlas.

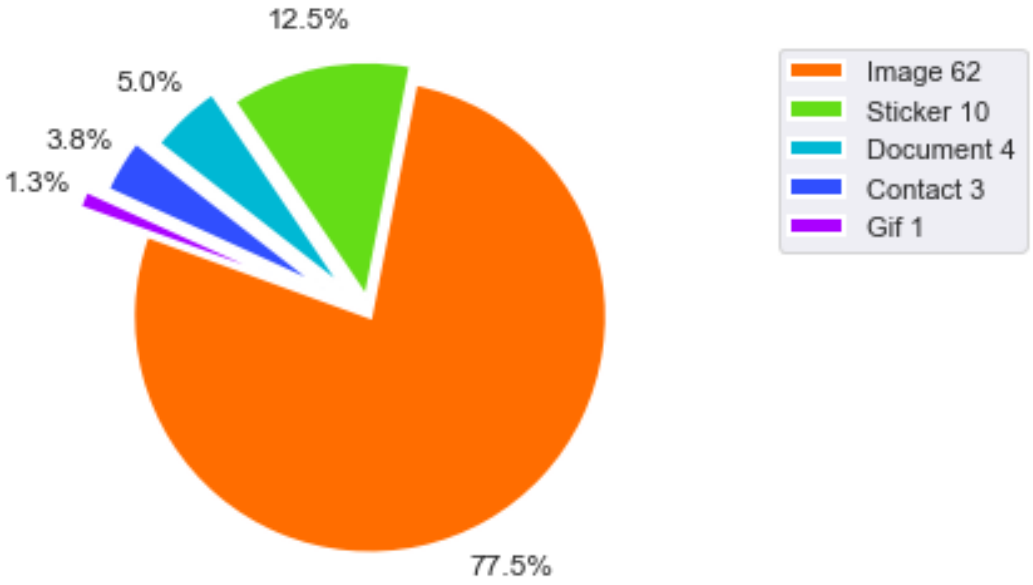

Ilustración 5 Diagrama de tarda de tipos de archivos adjuntados.

En este *pie chart*, vemos que el tipo de archivo que más se ha adjuntado en el chat ha sido la imagen, concretamente se han adjuntado 62 imágenes. También se conoce que se han enviado 10 *stickers*, 4 documentos, 3 contactos y por último, un único GIF.

Top Websites

Otro tipo de información que se puede extraer desde la conversación son las url que se han enviado. Se pueden identificar mediante una expresión regular que detecte la estructura que siguen los dominios web. Una vez se han identificado las urls y cuantas veces se han nombrado se puede graficar de la siguiente manera:

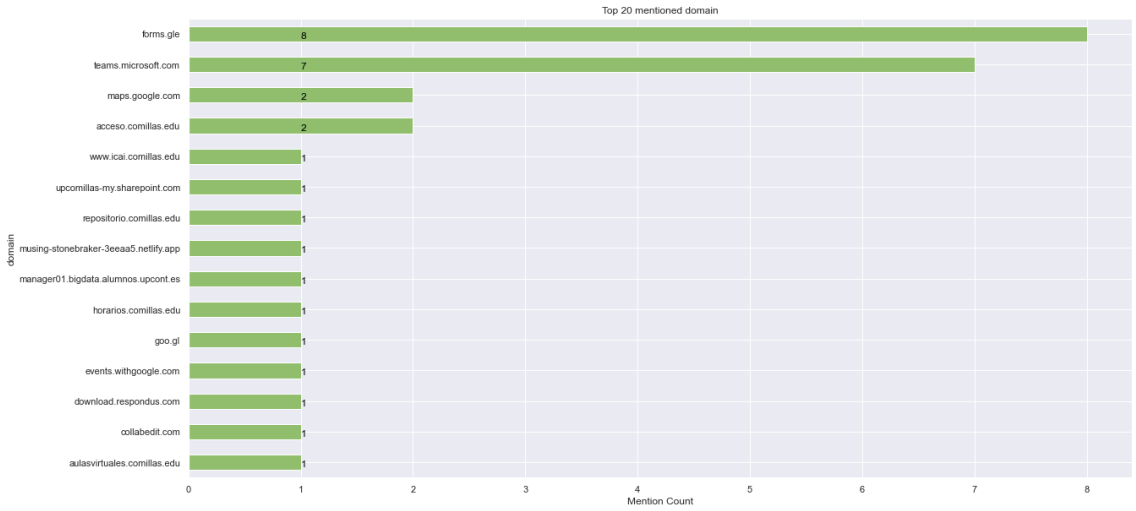

Ilustración 6 Diagrama de barras de dominios web.

Se pueden ver los 20 dominios más mencionados. Se puede observar que únicamente los primeros 4 han sido enviados más de una vez, con el primero de ellos enviado 8 veces. Como era de esperar, al tratarse del grupo de clase, todas las páginas web están relacionadas con la universidad o alguna asignatura.

Wordcloud

Se ha querido ir más allá y analizar las palabras que se hayan utilizado en las conversaciones. Lo primero será conocer cuáles son las palabras más repetidas para ver si nos dice algo más. Para ello se ha generado un *wordcloud* o nube de palabras. En esta representación el tamaño es mayor para las palabras que aparecen con más frecuencia.

Antes de ello, será necesario filtrar las *stopwords* o palabras vacías que no aportan valor e información relevante. Una vez hecho esto, se representa el wordcloud:

Ilustración 7 Nube de palabras más frecuentes.

En él, se concluye, que las palabras más utilizadas por los miembros del grupo son examen, clase o grupo. Una vez más, acorde a lo esperado, y en relación a la temática de universidad. Para poder ver de forma más concisa las palabras más utilizadas se h a querido hacer un recuento del número de repeticiones de cada palabra y ver cuales son realmente las alabaras más populares y si coinciden con lo que se ha visto en esta nube de palabras.

Top 5 words

Una vez se genera el diccionario con cada palabra y el número de veces que ha aparecido en el chat se ordena de forma descendente, y se seleccionan los primero cinco para después representarlos de la siguiente forma:

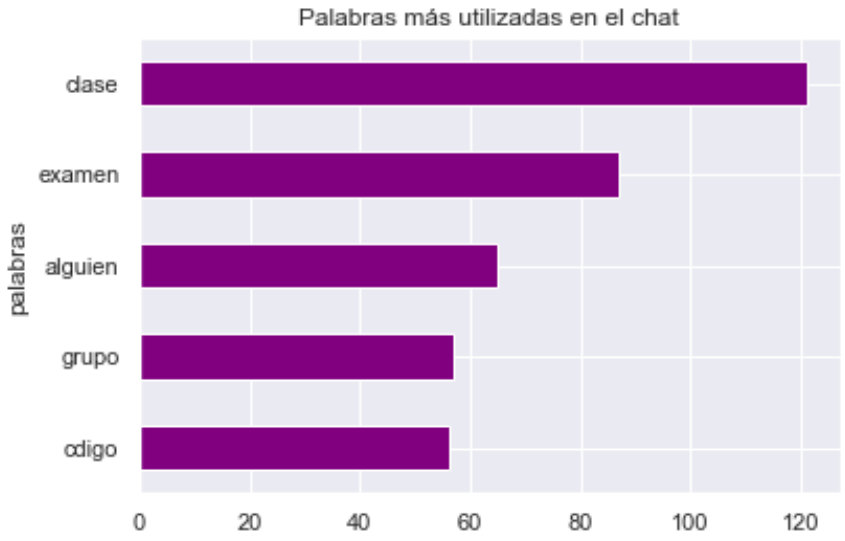

Ilustración 8 Diagrama de barras de palabras más repetidas.

Como ya se ha podido intuir en el *wordcloud*, la palabra más frecuente es clase, seguida por examen. Se encuentran también entre las palabras más repetidas alguien, grupo y por último ‘cdigo’ que probablemente haga referencia a la palabra del castellano código. Todas las palabras relacionadas de cierta forma con el contexto que se está tratando.

Swear words

Para terminar con el análisis, se ha querido estudiar si ha habido uso de tacos. Para ello, se ha añadido un fichero externo con las palabrotas más comunes del castellano y se ha hecho un barrido de todas las palabras del chat para ver si había coincidencias. Se ha visto que se han encontrado las siguientes:

	Palabras	resultado
28	coger	True
30	cojones	True
48	huevos	True
52	joder	True
93	puta	True

Solamente se han detectado 5 palabrotas en todo el chat analizado, por lo que aparentemente parece ser una conversación por lo general educada y respetuosa. Además, una de ellas, coger, no se considera un taco o palabrota en España, por lo que podría decirse que han sido 4 solamente.