MCDI
MAESTRÍA EN CIENCIA DE DATOS E INFORMACIÓN

INFOTEC CENTRO DE INVESTIGACIÓN E INNOVACIÓN EN TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN

DIRECCIÓN ADJUNTA DE INNOVACIÓN Y CONOCIMIENTO
GERENCIA DE CAPITAL HUMANO
POSGRADOS

"ANÁLISIS EXPLORATORIO DE DATOS PARA LA DETECCIÓN DE PRODUCTOS TECNOLÓGICOS"

IMPLEMENTACIÓN DE UN PROYECTO LABORAL
Que para obtener el grado de MAESTRO EN CIENCIA DE DATOS E INFORMACIÓN

Presenta:

Alan Rubén García Pérez

Asesor:

Dr. Mario Graff Guerrero

Ciudad de México, diciembre, 2020.

CONACYT

INFOTEC POSGRADOS

INFOTEC
POSGRADOS

AUTORIZACIÓN DE IMPRESIÓN Y NO ADEUDO EN BIBLIOTECA

MAESTRÍA EN CIENCIA DE DATOS E INFORMACIÓN

Ciudad de México, 17 de mayo de 2021.
INFOTEC-DAIC-GCH-SE-0159/01.

La Gerencia de Capital Humano / Gerencia de Investigación hacen constar que el trabajo de titulación intitulado

ANÁLISIS EXPLORATORIO DE DATOS PARA LA DETECCIÓN DE PRODUCTOS TECNOLÓGICOS

Desarrollado por el alumno **Alan Rubén García Pérez** y bajo la asesoría del **Dr. Mario Graff Guerrero**; cumple con el formato de biblioteca. Por lo cual, se expide la presente autorización para impresión del proyecto terminal al que se ha hecho mención.

Asimismo se hace constar que no debe material de la biblioteca de INFOTEC.

Vo. Bo.

Lic. Susana Argelia Salomón Jalili
Coordinadora de Biblioteca

Anexar a la presente autorización al inicio de la versión impresa del trabajo referido que ampara la misma.

C.p.p Servicios Escolares

Agradecimientos

Mi agradecimiento a Dios y a la vida, por la oportunidad de alcanzar una meta más, por las personas que me guiaron y acompañaron durante este proceso, en especial por todo el apoyo y los conocimientos que me brindó mi asesor, el doctor Mario Graff, en cada una de las etapas de este proyecto para obtener los resultados esperados. De igual forma agradezco a INFOTEC, por brindarme los recursos y herramientas necesarios para desarrollar este trabajo.

A mi familia, por apoyarme en todo momento, por los valores y principios que me han inculcado y que me han llevado a ser la persona que soy.

Agradezco a Daniela Abigail, por estar presente en todo este proceso y ha sido persona clave para conseguir este sueño.

Y a ti, querido lector, que tomarás el tiempo de leer este trabajo.

Tabla de contenido

Índice de gráficos

Índice de cuadros

Siglas y abreviaturas

AMVO	Asociación Mexicana de Ventas Online.
CES	En inglés, Consumer Electronics Show.
INEGI	Instituto Nacional de Estadística y Geografía.
INFOTEC	Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación.
MSV	Máquinas de Soporte Vectorial.
SARS-CoV2	En inglés, Severe Acute Respiratory Syndrome Coronavirus 2.
TF-IDF	En inglés, Term Frequency Inverse Document Frequency.

Glosario

“A”

Algoritmo: En matemáticas, lógica y disciplinas relacionadas, es una serie finita de pasos no-ambiguos y ordenados para resolver un problema, como realizar un cómputo, procesar datos y llevar a cabo otras tareas o actividades.[2]

“B”

Bigramas: Es un grupo de dos letras, dos sílabas, o dos palabras; son utilizados comúnmente como base para el simple análisis estadístico de texto.[18]

BoxPlot: Es un método estandarizado para representar gráficamente una serie de datos numéricos a través de sus cuartiles.[20]

“C”

Ciencia de datos: Campo interdisciplinario que utiliza procesos, algoritmos y métodos científicos para extraer valor de los datos, se requiere una variedad de habilidades estadísticas, computacionales y conocimiento empresarial.[8]

Clustering: Es la tarea de agrupar objetos en grupos o conjuntos de manera que los miembros del mismo grupo tengan características similares.[3]

Corpus: Es un conjunto de textos o de datos relativamente grande destinado a la investigación de la lingüística general.[13]

Comercio electrónico: Consiste en la compra y venta de productos o servicios a través de internet, tales como redes sociales y otras páginas web.[4]

Cuartil: Son valores que dividen una muestra de datos en cuatro partes iguales y sirven para determinar de manera rápida la dispersión y tendencia central de dicho conjunto de datos.[21]

“I”

Internauta: Persona que utiliza los servicios de internet u otra red informática.[9]

“L”

Lematización: Proceso lingüístico que consiste en determinar el lema de una palabra dada su forma (en plural, en femenino, conjugada, etc) tal como la

encontraríamos en el diccionario, por ejemplo, el lema de “dije”, “dijo” y “dirá” es “decir”.[7]

“N”

Nube de palabras: Es una representación visual de las palabras que conforman un texto, en donde el tamaño es mayor para las palabras con un valor más alto.[10]

“O”

Outliers: Es una observación que es numéricamente distante del resto de los datos.[19]

“P”

Preprocesamiento: Es una etapa fundamental en el proceso de extracción de conocimiento, cuyo objetivo principal es obtener un conjunto de datos que sea de calidad y útil a través de la limpieza de datos, su integración, transformación y reducción para la siguiente fase de minería de datos.[1]

Python: Se trata de un lenguaje de programación multiparadigma, ya que soporta orientación a objetos, programación imperativa y, en menor medida, programación funcional.[17]

“R”

Recuperación de información: Es la ciencia de la búsqueda de información en documentos electrónicos y cualquier tipo de colección documental digital que describan su contenido, o también realizar la recuperación de textos, imágenes y sonido de manera pertinente y relevante.[11]

“S”

Stemming: Proceso que consiste en convertir una palabra en su raíz.[6]

Stopwords: Son palabras muy comunes y poco informativas tales como conjunciones (y, o, ni, que,etc.), preposiciones (a, en, para, por, etc.) que son filtradas en la mayoría de los casos para realizar análisis de texto.[5]

Streaming: Se refiere a cualquier contenido de medios, ya sea en vivo o grabado, que se puede disfrutar en computadoras y aparatos móviles a través de Internet y en tiempo real.[22]

“U”

Unigramas: Es una letra, una sílaba, o una palabra; son utilizadas comúnmente como base para el simple análisis estadístico de texto.[12]

“W”

Wearables: La tecnología vestible se trata de dispositivos electrónicos inteligentes incorporados a la vestimenta o usados corporalmente como implantes o accesorios que pueden actuar como extensión del cuerpo o mente del usuario.[23]

Introducción

Hoy en día un adecuado preprocesamiento[1] de datos es fundamental para alimentar sistemas de información, algoritmos[2] de aprendizaje computacional, sistemas expertos, visión artificial, reconocimiento de voz y análisis de texto. Por ejemplo, dando un tratamiento adecuado a los datos, se puede realizar una agrupación de documentos similares usando algoritmos de clustering[3] o incluso realizar análisis de sentimientos. Tales documentos pueden ser textos cortos como tuits, opiniones en blogs o comentarios de usuarios sobre productos en plataformas de comercio electrónico[4] (ecommerce), hasta colecciones de texto más extensas como noticias o libros digitales, estos documentos forman parte de los llamados datos no estructurados. También existen los llamados datos estructurados, cuya diferencia según Devin Pickell se explica a continuación: "Los datos estructurados están altamente organizados y formateados de tal manera que se pueden buscar fácilmente en bases de datos relacionales. Los datos no estructurados no tienen un formato u organización predefinidos, lo que hace que sea mucho más difícil de recopilar, procesar y analizar" [17]. Se puede pensar que la mayoría de los datos están bien organizados, sin embargo, se sabe que solo el 5% de la información es estructurada [16], el resto proviene de imágenes, audios, o documentos como los mencionados anteriormente [1].

El preprocesamiento consiste en usar métodos eficientes que homologuen o den forma óptima a los datos con el objetivo ser utilizados para su análisis, simplificar representaciones finales, enfocarse en palabras de carga semántica o remover complicaciones innecesarias dada una tarea, por ejemplo, en el caso de texto podemos transformar todas las letras a minúsculas, eliminar signos de

[1] Salvador García, et al. Big Data: Preprocesamiento y calidad de datos. [Septiembre 2020].

[2] Cormen et al. Introduction to algorithms [Mayo 2019].

[3] Garerth James et al. An introduction to statistical learning with applications in R. [Marzo 2020].

[4] AMVO. Recuperado de https://www.amvo.org.mx/glosario/e-commerce-comercio-electronico [Mayo 2020].

puntuación o símbolos raros (emojis, caracteres especiales), eliminar stopwords[5] y utilizar procesos de stemming[6] o lematización[7] con el fin de existan emparejamientos entre las características de un texto [1].

La ciencia de datos[8] se auxilia de estas y otras técnicas para generar información que ayude a una mejor toma de decisiones basadas en una gran cantidad y variedad de datos. Es por ello que la figura del científico de datos, quién es una persona que combina sus habilidades de programación, estadística, creatividad para contar historias y sobre todo la capacidad de extraer datos de cualquier fuente o medio, con el objetivo de transformarlos en información útil, se ha vuelto tan importante para las empresas y en general para cualquier sector que desee incurrir en lo digital [16].

Se tienen grandes referencias de empresas tecnológicas que han logrado aplicar y desarrollar metodologías que benefician a la mayoría de internautas[9] un claro ejemplo es Google y su potente motor de búsqueda; que gracias a las millones de consultas que se realizan diariamente por los usuarios, es capaz de realizar la corrección ortográfica sobre una búsqueda de manera inmediata. Otra empresa que sin duda ha sorprendido a muchos usuarios por sus desarrollos exponenciales es OpenAI, que a través de una gran cantidad de texto en internet ha desarrollado un algoritmo capaz de desempeñar varias tareas, tales como: traducción de texto, generación de preguntas y respuestas, comprensión y razonamiento de texto [13].

La finalidad de este trabajo es realizar un análisis exploratorio de datos capaz de identificar productos tecnológicos relevantes para comercios electrónicos, explorando tendencias generadas con tuits en español e inglés de un conjunto de palabras clave, se toma una muestra de los primeros 5 meses del año

[5] Medium. Recuperado de https://medium.com/qu4nt/reducir-el-número-de-palabras-de-un-texto-lematización-y-radicalización-stemming-con-python-965bfd0c69fa [Agosto 2020].

[6] Medium. [Agosto 2020].

[7] Medium. [Agosto 2020].

[8] The Economist. Data, data everywhere. Recuperado de https://www.economist.com/special-report/2010/02/27/data-data-everywhere [Marzo 2020].

[9] RAE. Recuperado de https://dle.rae.es/internauta [Junio 2020].

2020 para elegir términos más adecuados que permitirán continuar con el análisis de texto de noticias extraídas de Google dadas las fechas donde se obtuvieron los picos más altos en tendencias por semestre de los años 2018 al 2020.

El trabajo se estructura de la siguiente forma: Capítulo 1; Antecedentes y Objetivos, en este capítulo se describen estudios previos que dan pauta al desarrollo de este trabajo. Capítulo 2; Tendencias y Comparación de Palabras Clave, dentro de este capítulo se mostrarán las tendencias exploradas y la selección de términos que ayudarán a la búsqueda de noticias en Google. Capítulo 3; Nubes de Palabras Con Noticias de Google, en esta parte del trabajo se encontrarán las fechas donde se obtuvieron las frecuencias máximas de menciones de los términos del capítulo 2 y sus nubes de palabras[10] por semestre que ayudarán a detectar productos tecnológicos; por último, encontramos el apartado de Conclusiones a las que se llegaron con el desarrollo del trabajo.

[10] Educación 3.0. Recuperado de https://www.educaciontrespuntocero.com/recursos/crear-una-nube-tags-las-palabras-mas-usadas-texto [Agosto 2020].

Capítulo 1
Antecedentes y Objetivos

Capítulo 1. Antecedentes y Objetivos

Dentro del contenido de este capítulo se encuentra una breve explicación de investigaciones y aplicaciones previas que han dado pauta para el desarrollo de este proyecto, además se expone el planteamiento del problema que incluye los objetivos a los que se quiere llegar y las preguntas de investigación ligadas, justificación y la viabilidad que se tiene de la investigación y desarrollo del trabajo.

1.1 Antecedentes

En los últimos años, el número de usuarios activos en redes sociales se ha incrementado de manera exponencial, por ejemplo, en abril de 2016, Twitter registró 320 millones de usuarios y para abril de 2020 se tiene un registro de 386 millones [2]. Otra red social y quizá la más importante en la actualidad es Facebook, que pasó de 1,590 millones en abril 2016 a 2,498 millones de usuarios activos para 2020 [2], en tales plataformas las personas interactúan compartiendo contenido multimedia como videos e imágenes, acompañados en su mayoría de ideas en texto o con comentarios en las propias publicaciones.

Gracias a esta creciente interacción digital y avance en técnicas de minería de texto y recuperación de información[11], se han llevado a cabo investigaciones principalmente usando comentarios de Twitter, por ejemplo, el trabajo de Tellez et al. [1] cuyo objetivo fue identificar de un gran conjunto de combinaciones de modelado de texto, las transformaciones óptimas para un análisis de sentimientos con tuits en español, se utilizó lematización, derivación (stemming), eliminación de entidades, tokenizadores (modelos de n-gramas y q-gramas) y esquemas de pesado TF-IDF, estos últimos tuvieron mayor impacto en la precisión de un clasificador MSV (Máquinas de Soporte Vectorial), se probó exhaustivamente que el modelo de q-gramas fue superior frente a los n-gramas.

En el artículo Bjarke Felbo et al. [3] se usaron millones de tuits en inglés que contenían emojis, con el objetivo de detectar el sentimiento, emociones o sarcasmo, para el conjunto de entrenamiento se consideraron tuits que

[11] GlossariumBITri. Recuperado de https://sites.google.com/site/glosariobitrum/Home/recuperacion-de-informacion [Agosto 2020].

contuvieran al menos un token que no fuera puntuación, emoji o caracter especial. Encontraron que estos íconos son base fundamental para generar modelos pre entrenados capaces de aprender el contenido emocional.

En el trabajo expuesto por Sidorov et al. [7] se exploran cómo diferentes configuraciones de texto (tamaño de n-gramas, tamaño de corpus, número de clases de sentimiento, corpus balanceados vs no balanceados, etc.) afectan la métrica de precisión de algoritmos de clasificación (Naive Bayes, árboles de decisión y máquinas de soporte vectorial) utilizando tuits en español relacionados a temas de “celulares” y “elecciones presidenciales mexicanas”. Se obtuvo que la mejor configuración está conformada por: unigramas[12], un menor número de clases (positivo, negativo), usar al menos 3,000 tuits como conjunto de entrenamiento, corpus[13] con clases balanceadas genera ligeramente peores resultados y usando el clasificador de máquinas de soporte vectorial.

Además, en México se cuenta con aplicaciones e investigaciones por parte de instituciones gubernamentales, INEGI desarrolló una plataforma que determina el estado de ánimo de los tuiteros en México, se puede visualizar información desde el 1 de enero de 2016 por medio de series de tiempo y datos geográficos a nivel estado [14]. La aplicación desarrollada por Abel Coronado[14] [15] muestra como generar un pequeño lago de datos con el fin de realizar un tablero de control en el que se visualizan gráficos de noticias recolectadas de Google, en uno de sus análisis aplica el pesado TF-IDF para obtener las palabras más significativas de las noticias extraídas con el patrón de búsqueda “Covid México”.

Por otro lado, el comercio electrónico ha tenido un impresionante crecimiento en los últimos años, como se menciona en un video de Platzi[15] [8], alrededor de 6 millones de sitios web son ecommerce y las plataformas de redes

[12] Eric S. Tellez et al. A case study of Spanish text transformations for twitter sentiment analysis [Marzo 2020].

[13] Alicia San Mateo. A bigram corpus used as a grammar checker for Spanish native Speakers [Marzo 2020].

[14] Abel Coronado. Recuperado de https://abxda.wordpress.com/2020/05/06/analizando-el-big-data-de-las-noticias-con-tu-micro-data-lake-baterias-incluidas/ [Marzo 2020].

[15] Platzi. Recuperado de https://www.youtube.com/watch?v=kSwDHNjGIOY [Junio 2019].

sociales son utilizadas para promover venta de productos, las que destacan Facebook y Twitter.

Actualmente el mundo se encuentra en una de las crisis más grandes que ha afrontado la humanidad debido a una nueva enfermedad llamada SARS-CoV2 y que ha impactado en la economía mundial para los negocios tradicionales. Como se describe en el Reporte No. 2, Impacto COVID-19 en Venta Online México por parte de la AMVO[16] [11]. El impacto económico ya es palpable en la región debido a la vulnerabilidad cambiaria y una fuerte economía informal, lo cual pronostica una contracción económica en algunos países. Sin embargo, la pandemia podría verse como un catalizador para el impacto en los hábitos de compra en línea en los consumidores a largo plazo. Para ese mismo reporte se realizó un análisis de texto, en el Gráfico 1 se muestra la evolución de menciones con términos "ecommerce" + "covid", usando redes sociales como fuentes de datos (70% Twitter, 12% Blogs, 8% Noticias Online, 8% Facebook y 2% Otros), se muestra que, a consecuencia de estar en casa, se tuvo un aumento significativo en la actividad de usuarios en redes sociales.

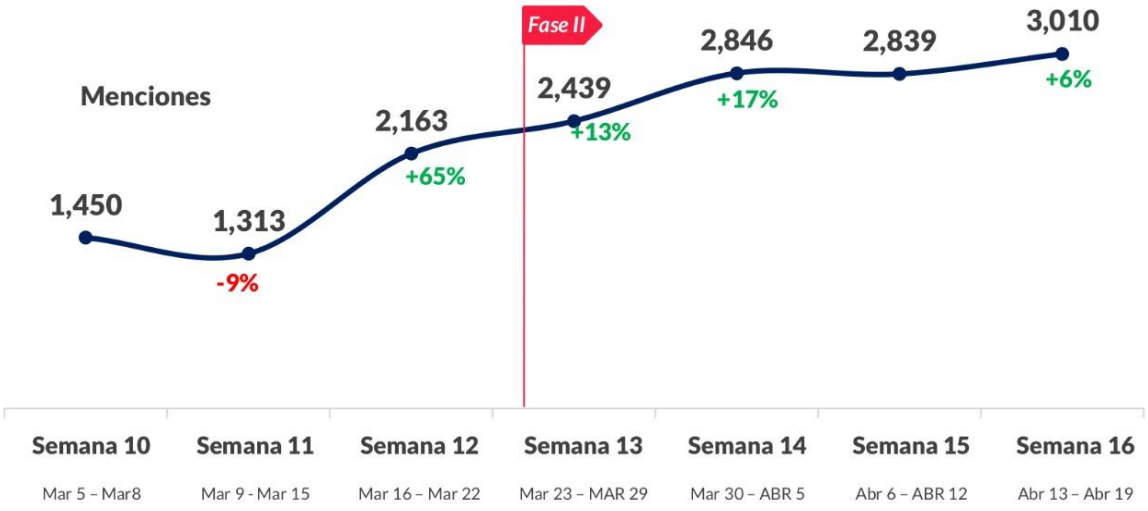


Gráfico 1. *Evolución de menciones de términos "ecommerce"+ "covid".*

Fuente: AMVO. Reporte No. 2, Impacto COVID-19 en Venta Online México.

[16] AMVO. Recuperado de https://www.amvo.org.mx/estudios/reporte-2-impacto-covid-19-en-venta-online-mexico [Mayo 2020].

Dentro del análisis de la AMVO [11], se determinó que existe un sentimiento neutral en la mayoría de comentarios, en promedio 65% de las menciones se concentran en apoyo a pequeños negocios y la transformación digital de las empresas. En promedio 17% fue sentimiento negativo, los comentarios describen retrasos por envíos o entregas, formas de pago e interacciones con el canal de ayuda.

Partiendo de los aportes que han realizado investigadores e instituciones, se quiere combinar técnicas para elaborar un análisis exploratorio de datos usando tuits y noticias de Google en inglés y español, para identificar productos tecnológicos mediante análisis de texto. A continuación, se plantea el objetivo que persigue esta investigación.

1.2 Objetivo General

Explorar tendencias de palabras clave usando tuits en español e inglés para identificar productos tecnológicos de gran relevancia mediante análisis de texto.

1.2.1 Objetivos Específicos

- Analizar tendencias de términos o palabras clave.
- Seleccionar palabras clave con las mayores frecuencias de menciones.
- Elegir la metodología de análisis de texto que permita identificar productos relevantes.

1.3 Preguntas de investigación

- ¿Qué términos o palabras clave tienen frecuencias de menciones más altas?
- ¿Qué metodología de análisis de texto es la más adecuada para identificar productos relevantes?
- ¿Qué productos tecnológicos se pueden identificar mediante la metodología elegida?

1.4 Justificación

Al existir un crecimiento exponencial en el uso de redes sociales desde el año 2016 hasta la actualidad donde la mayoría de las personas tienen que quedarse en casa para evitar el contagio por SARS-CoV2; lleva a que las interacciones diarias sean más frecuentes en redes sociales, por lo que existe una mayor cantidad de datos no estructurados como son los tuits. Se ha detectado la oportunidad de identificar productos tecnológicos relevantes mediante el análisis de tendencias usando palabras clave y sus menciones diarias en comentarios de Twitter, además de análisis de texto. Esto contribuirá con proponer una metodología que facilite la búsqueda de productos específicos.

1.5 Viabilidad

Dado que las principales fuentes de datos pueden ser consultadas de manera gratuita, así como las herramientas para manipulación de los datos, el desarrollo de este trabajo se puede llevar a cabo sin ninguna restricción al contar con los recursos, capacidad y conocimiento para su elaboración.

Expuestos los objetivos y la viabilidad para desarrollar este trabajo, se comenzará con el análisis de tendencias de palabras clave relacionadas a tecnología usando datos de los primeros meses del año 2020.

Capítulo 2
Tendencias y Comparación de Palabras Clave

Capítulo 2. Tendencias y Comparación de Palabras Clave

El desarrollo de este capítulo se centra en el análisis de tendencias y menciones de algunos términos relacionados a la tecnología contenidos en publicaciones de Twitter. Se hace uso de una librería de Python[17] que permite obtener información procesada de tuits desde diciembre de 2015 en forma de tuplas que contienen unigramas o bigramas[18] junto con su frecuencia diaria, documentada en el trabajo Mario Graff et al. [10]. En nuestro caso se usaron unigramas dado que se obtuvo la mayor cantidad posible de datos.

Se proponen sinónimos de dos o tres palabras en inglés y español a los que hemos denominado “Grupos”, por cada uno se seleccionará el término más relevante de acuerdo al análisis realizado, el objetivo es utilizarlo como insumo para la búsqueda de noticias en Google.

2.1 Tendencias

Para tener una mejor visualización de todos los términos y sus menciones diarias, se ha decidido que el periodo de extracción de datos fuera del 1 de enero al 31 de mayo de 2020 usando unigramas. Sabemos que en este intervalo de tiempo, la enfermedad SARS-CoV2 tuvo crecimiento en todo el mundo, como consecuencia las personas se vieron forzadas a permanecer en casa, por tanto, es posible que se observe un incremento en las menciones diarias como muestra el Gráfico 1. Los grupos y las palabras que se usaron para la extracción de datos en inglés y español se muestran en el Cuadro 1, en total son 10 palabras por idioma permitiendo explorar las diferentes configuraciones de la librería.

Idioma	**Grupo 1**		**Grupo 2**			**Grupo 3**			**Grupo 4**	
Español	televisión	tv	celular	teléfono	móvil	computadora	pc	laptop	inteligente	inteligencia
Inglés	television	tv	phone	telephone	mobile	computer	pc	laptop	smart	intelligence

Cuadro 1. *Palabras y sinónimos en inglés y español.*

Fuente: Elaboración Propia

[17] Python Land. Recuperado de https://python.land/python-tutorial/what-is-python [Octubre 2020].

[18] Alicia San Mateo [Marzo 2020].

En el Gráfico 2 se visualizan los resultados de las menciones diarias por palabra, cada recuadro presenta un término en español (línea azul) y sus valores en el eje derecho y su homólogo en inglés (línea naranja) cuyas frecuencias se muestran en el eje izquierdo.

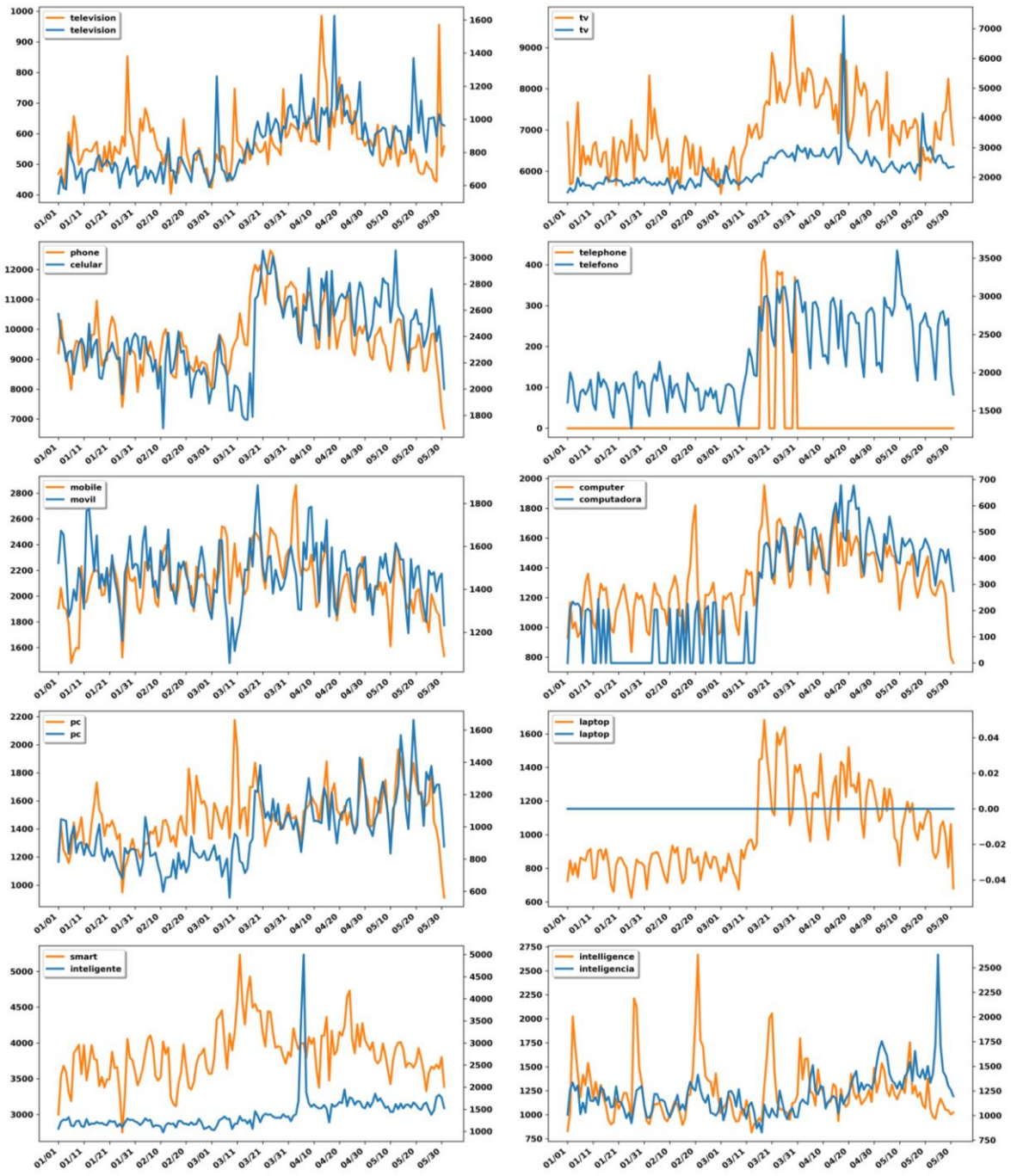


Gráfico 2. *Frecuencias de palabras en inglés y español, enero-mayo 2020.*

Fuente: Elaboración Propia

Es posible ver que la mayoría de las palabras presentan tendencia a la alza para ambos idiomas, con crecimiento en sus menciones diarias a partir de mitades del mes de marzo 2020, consecuencia de la interacción frecuente en redes sociales por cuarentena.

Si tomamos las menciones de todos los términos por idioma y comparamos el incremento de la segunda quincena del mes de marzo respecto a la primera, observamos un crecimiento de 32.3% para menciones en español y 19.59% en idioma inglés, sin embargo hay que notar que existen palabras en inglés o español con menciones iguales a 0, lo que provoca que la suma de valores para alguna quincena de marzo disminuyan. Nos referimos a los términos “telephone”, “computadora” y “laptop”, si removemos esas palabras y realizamos el mismo ejercicio, obtenemos que las menciones en español tuvieron crecimiento de 23.3% y en inglés un crecimiento de 16.78% de la segunda quincena respecto a la primera. Con este indicador podemos afirmar que el inicio de la pandemia influyó al incremento de menciones diarias, principalmente para los términos en español.

Lo siguiente a realizar es la comparación entre sinónimos con el objetivo de elegir el término más relevante de cada grupo del Cuadro 1.

2.2 Comparación de sinónimos

Para el desarrollo de esta sección se abordará un grupo por subsección, realizando una comparación de los términos que la componen. Para ello se observarán sus frecuencias en ambos idiomas, tendencias y elementos clave que proporcionen información necesaria para elegir el sinónimo más útil que permita continuar con una exploración más profunda. Se quiere obtener una única palabra por grupo, dado que la librería usada para la extracción de menciones solo permite unigramas y bigramas como entrada, se probó con ambos casos, sin embargo, usando unigramas se obtuvo el máximo de datos.

2.2.1 Comparación de términos del Grupo 1

El Grupo 1 considera los términos “televisión” y “tv”, de manera general ambos siguen una tendencia creciente sin importar el idioma. Al mirar las frecuencias del

Gráfico 2 para los términos en español, notamos que las menciones diarias relacionadas a “tv” se encuentran por arriba de 2,000, mientras que “televisión” apenas alcanzó un máximo de 1,600 después del inicio de la cuarentena. Se observa que las 2 palabras alcanzan su valor máximo entre el 10 y 20 de abril de 2020, lo que podría significar un evento o hecho importante para productos relacionados a estas palabras.

Tomando el promedio de menciones para términos en inglés, se encontró que “televisión” únicamente tuvo 567 contra las 6,968 menciones promedio obtenidas por “tv”, resultando una gran diferencia entre ambos términos. Dados estos resultados se ha decidido que “tv” sea el sinónimo seleccionado para este grupo.

2.2.2 Comparación de términos del Grupo 2

Comparando los términos “celular”, “teléfono” y “móvil”, inmediatamente se decidió descartar el término “teléfono”, pues “telephone” solo tuvo menciones en algunos días de marzo, lo que dificulta la comparación contra los otros sinónimos del Grupo 2. En cuanto a los términos restantes, es notable un mayor interés por parte de los usuarios al incluir “celular” o “phone” en sus tuits, pues las frecuencias diarias son superiores comparadas a las de “móvil” o “mobile”, incluso después del inicio de la cuarentena. De esta forma se elegirá a “celular” como la palabra con la que se desarrollarán análisis posteriores.

2.2.3 Comparación de términos del Grupo 3

Un caso similar al anterior se presenta para los términos “computadora”, “pc” y “laptop”, es decir, “laptop” para tuits en español no tuvo registro alguno de menciones durante el periodo seleccionado, por tanto, este término queda descartado para comparación.

Por otro lado, tanto “pc” como “computadora” comenzaron a subir el volumen de menciones posterior al 20 de marzo, sin embargo al observar las frecuencias de los términos en español previo a esta fecha, notamos que “computadora” no era relevante para los usuarios al registrar 0 menciones en

algunos días, esto no ocurrió para “pc” que pasó de 800 a 1,200 menciones promedio.

Para las menciones de términos en inglés, “computer” pasó de 1,200 a 1,400 menciones promedio, en cambio “pc” de 1,400 a 1,600, lo que hace que este último sea relevante para los usuarios. Con base en estos hechos, se decidió que la palabra “pc” sea la indicada para este grupo.

2.2.4 Comparación de términos del Grupo 4

El Grupo 4 contiene los términos “inteligente” e “inteligencia” y sus homólogos en inglés. Partiendo de las menciones en español, notamos outliers[19] en ambos casos, “inteligente” llegó cerca de las 5,000 menciones; 2,000 más que “inteligencia”. Si tomamos las menciones promedio de estos términos en ambos idiomas obtenemos que “inteligencia” tuvo 1,230, mientras que “inteligente” el valor es de 3,850 dando pauta a que se tenga preferencia por este último.

2.2.5 Selección de términos

Dadas las consideraciones y el análisis de los términos de cada grupo, se han obtenido las siguientes palabras:

1. tv
2. celular
3. pc
4. Inteligente

Con esta selección se realizará una nueva extracción de datos desde el año 2018, con el objetivo de comparar las menciones diarias agrupadas por semestres, lo que permitirá encontrar fechas relevantes y determinar si la cuarentena impactó significativamente en las frecuencias obtenidas.

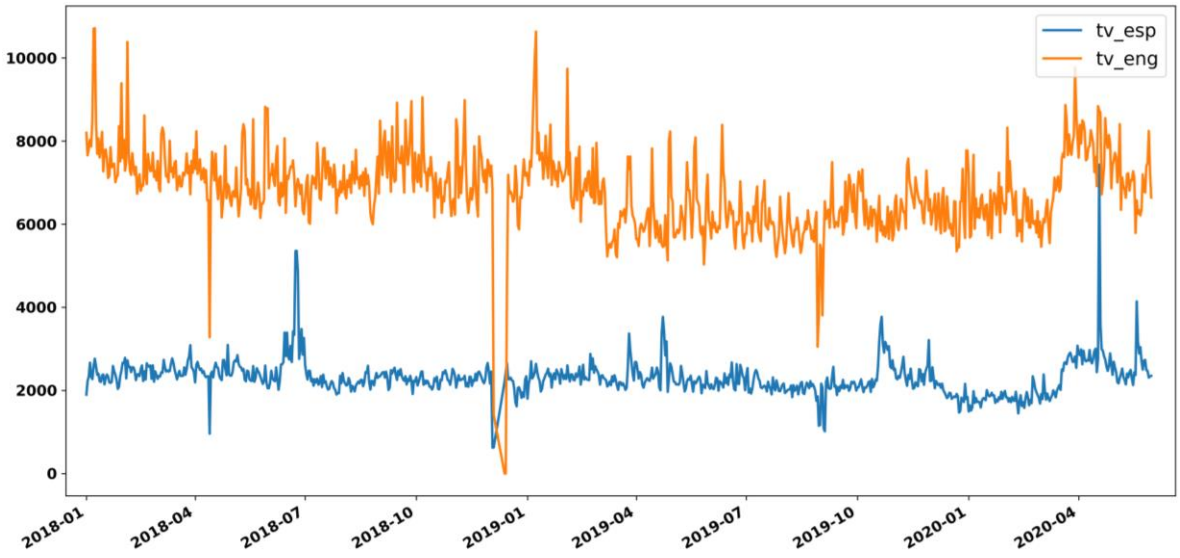
[19] Robert Johnson, et al. Estadística elemental: Lo esencial [Septiembre 2020].

2.3 Gráficos de Frecuencias y BoxPlots

En esta sección se encontrarán gráficos de las menciones diarias de las palabras seleccionadas previamente “tv”, “celular”, “pc” e “inteligente” contenidas en tuits de los idiomas inglés y español, el periodo que se abarca es del 1 de enero de 2018 al 31 de mayo de 2020, se incluyen “gráficos de cajas” (BoxPlots[20]) también conocidos como “gráficos de bigotes”, los cuales pretenden mostrar de manera descriptiva si existen diferencias significativas entre las menciones generadas durante el inicio de la pandemia y años anteriores.

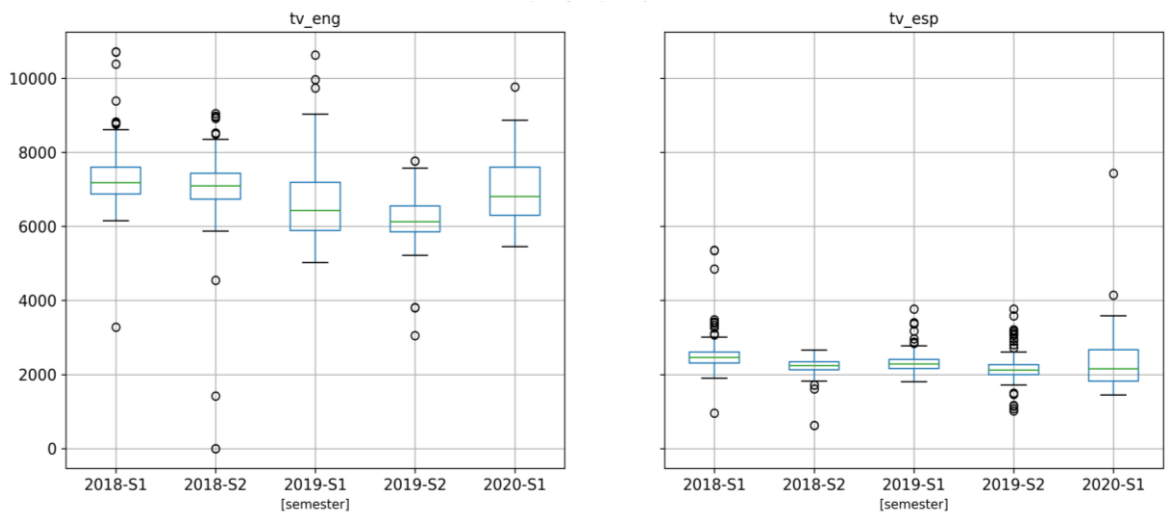
Cada término se analizará por separado en una subsección, mostrando sus frecuencias en inglés (línea naranja) y español (línea azul). Para el gráfico de cajas se decidió partir los años en semestres para obtener 5 grupos en los que sea posible visualizar de mejor forma diferencias que puedan existir entre semestres, con esta exploración se puede determinar si el inicio del coronavirus produjo un incremento en las menciones diarias de estos términos.

2.3.1 Análisis de “tv”



a) *Frecuencia de menciones diarias “tv”.*

[20] Garerth James et al. [Marzo 2020].



b) *BoxPlot "tv"*

Gráfico 3. *Frecuencias y diagrama de cajas para la palabra "tv"*
Fuente: Elaboración Propia

En ambos lenguajes es apreciable el crecimiento de menciones a mitades del mes de marzo 2020; visible en el Gráfico 3 (a), lo que pudo ocasionar que se tuvieran valores con mucha más dispersión tal y como se aprecia en los semestres *tv_eng 2020-S1* y *tv_esp 2020-S1* del Gráfico 3 (b). Estos últimos tuvieron frecuencias superiores comparados con el segundo semestre de 2019 de cada idioma y parece ser el semestre con los valores más bajos en menciones no contemplando valores atípicos (outliers).

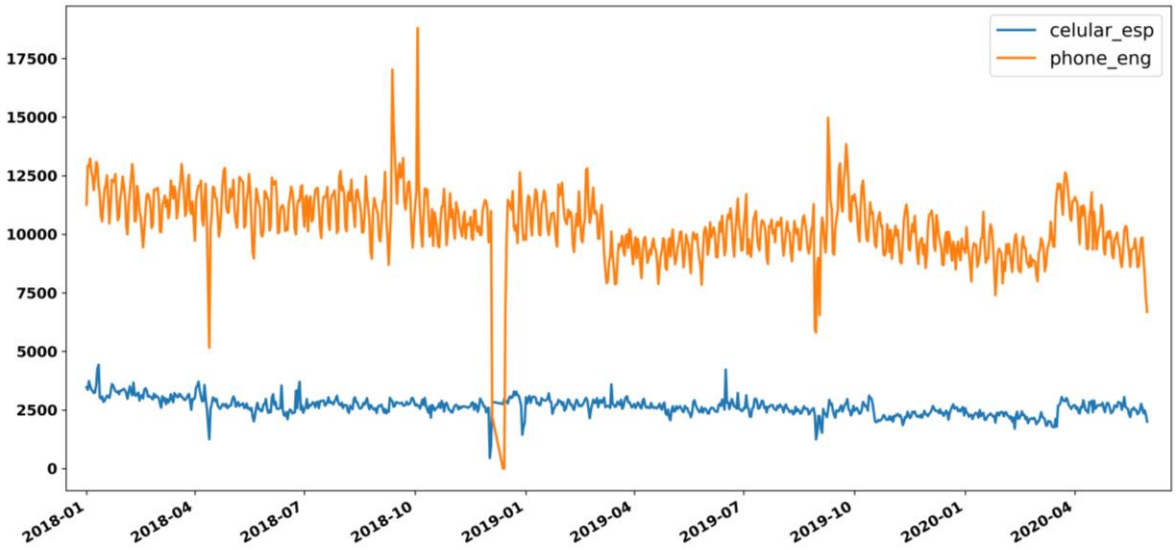
Vemos que el grupo *tv_esp 2018-S2* tiene una forma y dispersión similar a *tv_esp 2019-S2*, sin embargo, al compararlos mediante el cuartil[21] 3 (Q3) que representa el 75% de los datos, vemos que este último tiene un número de menciones menores a 2,265, mientras que el primero obtuvo menciones por debajo de 2,345, demostrando que el segundo semestre de 2019 fue el grupo con las menciones diarias más bajas para ambos idiomas.

Por último, en cada semestre del Gráfico 3 (b) existen valores muy grandes que sobresalen de las cajas, implicando algún suceso importante como la

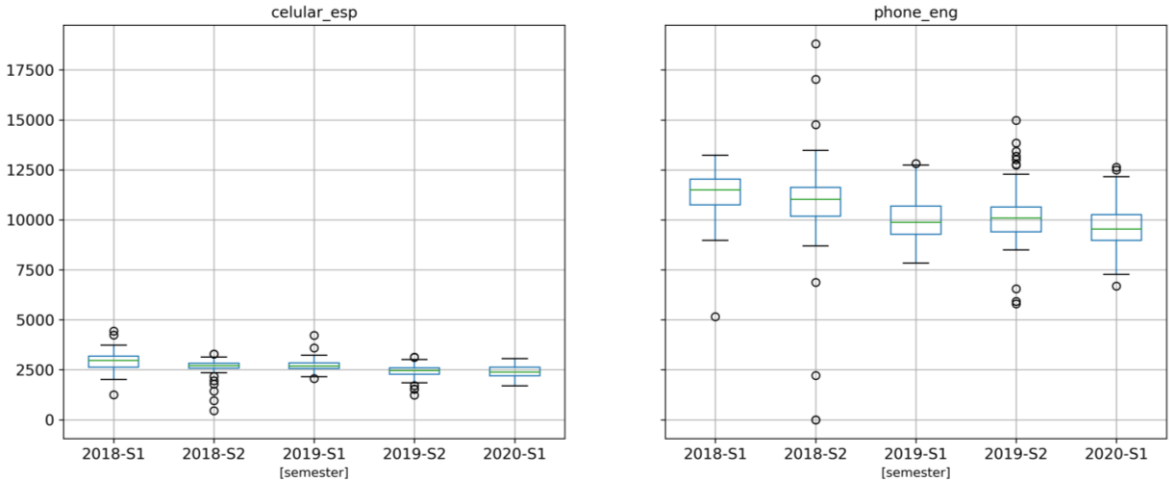
[21] Minitab 18. Recuperado de https://support.minitab.com/es-mx/minitab/18/help-and-how-to/graphs/how-to/boxplot/interpret-the-results/quartiles [Junio 2020].

realización de eventos, salida de una nueva marca o producto relacionado a televisiones; hay un caso que llama mucho la atención en la caja *tv_esp 2020-S1,* pues alcanzó el máximo de menciones a mitad de abril de 2020, alcanzando menciones reportadas de “tv” para tuits en inglés.

2.3.2 Análisis de “celular”



a) *Frecuencia de menciones diarias “celular”.*



b) *BoxPlot “celular”*

Gráfico 4. *Frecuencias y diagrama de cajas para la palabra “celular”*

Fuente: Elaboración Propia

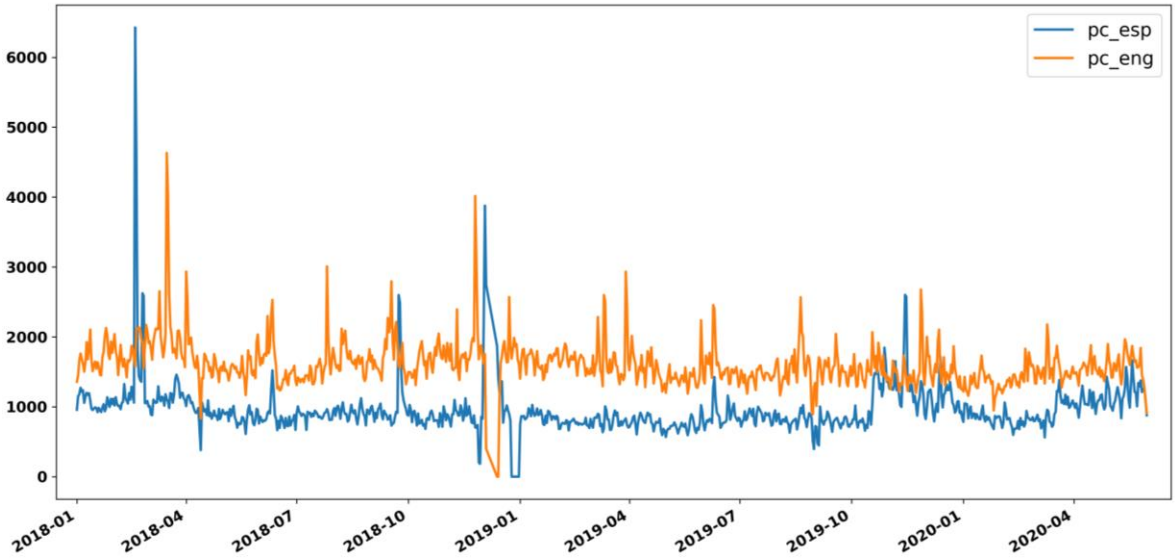
Comenzando con el análisis, nuevamente observamos el patrón de crecimiento para las menciones en marzo 2020 de acuerdo al Gráfico 4 (a), sin embargo en el caso de “phone” las frecuencias comienzan a descender rápidamente en días posteriores tal que las menciones diarias se asemejan a valores del año 2019. En cambio para el término en español “celular”, se observa una tendencia constante después del inicio de la cuarentena, con este crecimiento se han obtenido a frecuencias similares previas a octubre de 2019.

Una característica interesante del término en inglés son los picos que existen en el segundo semestre de los años 2018 y 2019, cuyas frecuencias son mayores al valor máximo registrado en 2020, se puede pensar que se trata de un patrón dado algún evento o lanzamiento de producto, pues estas frecuencias se encuentran entre septiembre y octubre. Para el año 2018 los picos se registraron el 12 de septiembre y 3 de octubre con menciones de 17,026 y 18,804 respectivamente, para 2019 las fechas correspondientes fueron el 9 y 24 de septiembre. En el capítulo 3 se hará el análisis de noticias registradas en estos días.

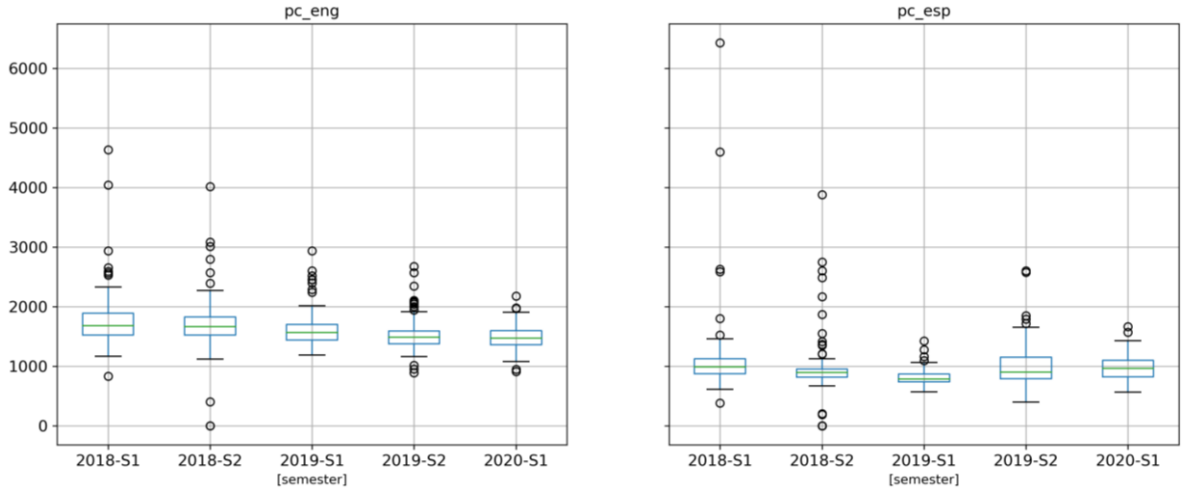
De acuerdo al Gráfico 4 (b) *celular_esp* no existió aumento significativo de menciones en el primer semestre de 2020 que evidencie algún efecto de la pandemia sobre este término, al compararlo con semestres de años anteriores las dispersiones se mantienen aparentemente constantes, destaca el semestre *2018-S1* al tener una mediana ligeramente mayor al resto de los grupos.

Para el caso del término en inglés, el Gráfico 4 (b) *phone_eng,* muestra que el inicio de la cuarentena no afectó las menciones significativamente respecto a semestres anteriores, incluso se observa una tendencia decreciente al pasar el tiempo.

2.3.3 Análisis de “pc”



a) *Frecuencia de menciones diarias “pc”.*



b) *BoxPlot “pc”*

Gráfico 5. *Frecuencias y diagrama de cajas para la palabra “pc”*

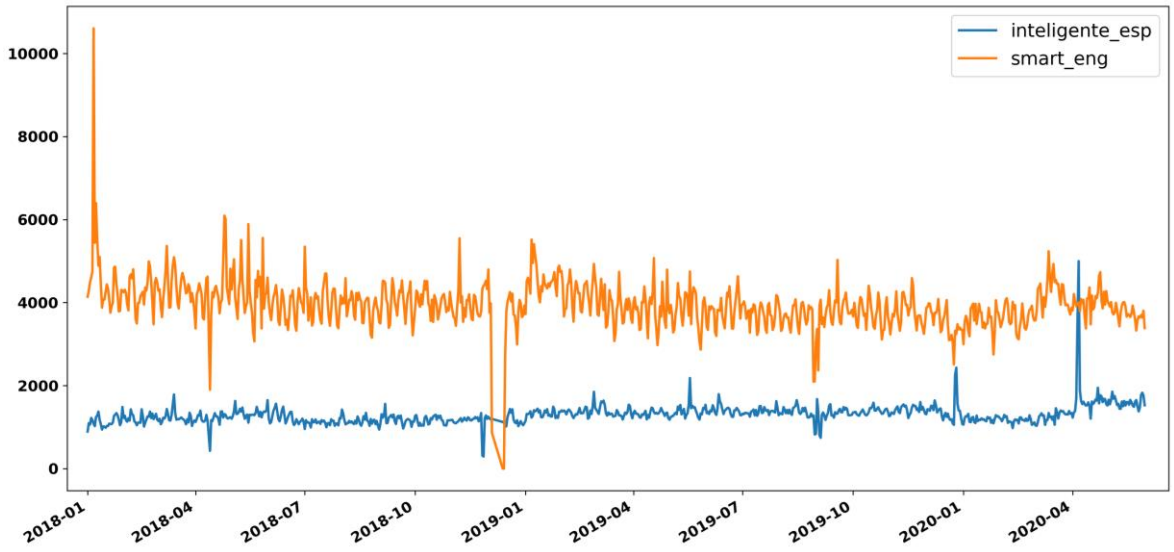
Fuente: Elaboración Propia

Al observar las frecuencias del Gráfico 5 (a) se tiene la particularidad que a diferencia de los términos anteriores, las menciones diarias para ambos idiomas no son muy distantes entre sí, lo que nos podría indicar que “pc” es similarmente usado por los usuarios tanto inglés como español. Observamos que en cuanto una

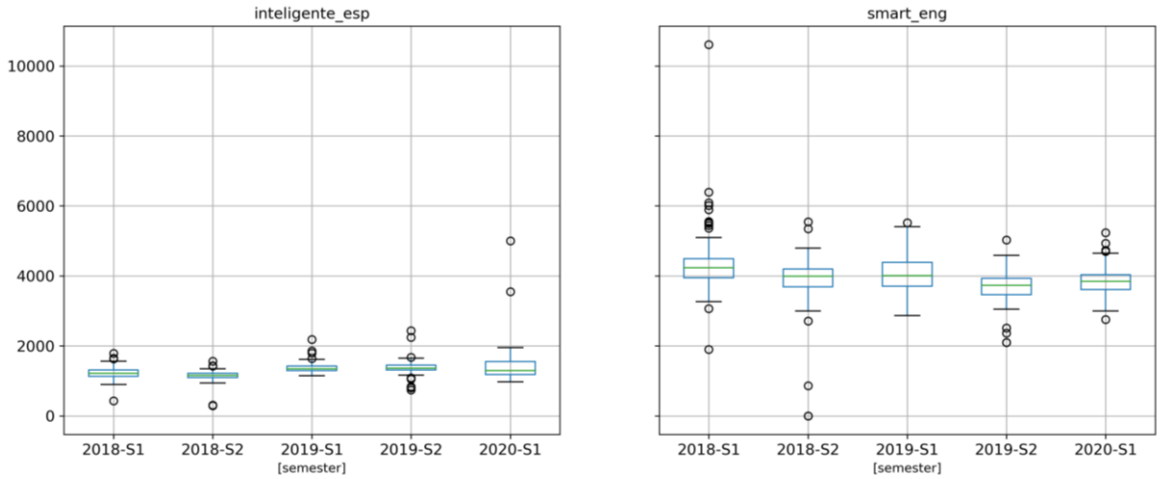
de las 2 líneas del Gráfico 5 (a) crece por encima del promedio, la otra tiende a crecer en días posteriores.

En cuanto a las frecuencias al inicio y durante del periodo de cuarentena es claro que no existe un crecimiento significativo para valores de la línea naranja, incluso vemos que los diagramas de cajas del Gráfico 4 (b) *pc_eng* tienen una tendencia decreciente al pasar de los semestres, esto claramente nos dice que la pandemia no tuvo efecto en este término para los tuits en inglés. Por el contrario *pc_esp*, mantuvo valores superiores y con tendencia creciente desde finales de marzo 2020 comparado con inicios de ese mismo año, sin embargo no fue suficiente para superar los valores de semestres anteriores, por ejemplo, al ver el valor del cuartil Q3, se tiene que es menor al del semestre *2019-S2* donde la pandemia por COVID-19 no era factor.

2.3.4 Análisis de "inteligente"



a) *Frecuencia de menciones diarias "inteligente".*



b) *BoxPlot "inteligente"*

Gráfico 6. *Frecuencias y diagrama de cajas para la palabra "inteligente"*

Fuente: Elaboración Propia

En este caso podemos ver el valor máximo en frecuencias a principios de 2018 para tuits con la palabra "smart", las menciones alcanzaron 10,607 mientras que "inteligente" generó 5,000 en abril de 2020, alcanzando las frecuencias de "smart".

Observando el Gráfico 6 (b) *inteligente_esp*, notamos una tendencia creciente al paso de los semestres, los valores de 2019 son ligeramente superiores a los del año 2018, el tema de la pandemia parece que sí tuvo impacto en frecuencias, pues el cuartil 3 del conjunto *2020-S1* es mayor que el de semestres anteriores, además se presentan outliers con los valores máximos.

Analizando las cajas de *smart_eng* es apreciable el siguiente patrón, existe un crecimiento en las frecuencias del primer semestre de cada año comparado con los segundos semestres, pero a pesar de la cuarentena el semestre *2020-S1* no tuvo menciones superiores comparado con los semestres *2018-S1 y 2019-S1,* el 75% de valores de *2020-S1* estaban por debajo de las 4,000, sin embargo para estos últimos apenas representaban el 25% y 50% respectivamente.

Capítulo 3

Nube de Palabras: Noticias de Google

Capítulo 3. Nube de Palabras: Noticias de Google

Con el uso de la herramienta googler [9], una librería de Python utilizada para consulta de búsquedas, noticias de diferentes sitios web y videos mediante la consola del sistema operativo sin necesidad de usar el motor de búsqueda de Google, se generarán nubes de palabras extrayendo noticias por cada semestre y en donde se haya captado el máximo de menciones con cada término del capítulo anterior en inglés y español. Como se ha descrito anteriormente, el objetivo es identificar productos tecnológicos haciendo un juicio del método que conviene utilizar para la visualización de las palabras más relevantes de acuerdo a nuestros propósitos, estos métodos se basan en Frecuencias o generando un pesado TF-IDF.

3.1 Identificación de valores máximos

Para cada palabra clave se identificó la fecha por semestre donde se haya obtenido el máximo de menciones dado que podría significar algún evento, noticia o lanzamiento de producto al mercado, además se amplió un rango de más menos 3 días con el propósito de captar el mayor número de noticias en torno a cada palabra, por ejemplo, la palabra “tv” contenida en tuits en inglés, el máximo de menciones fue el día 2018-01-08, por tanto el rango de extracción de noticias será del 2018-01-05 al 2018-01-11 tal como se observa en el Cuadro 2 (a). Las fechas siguen el formato AAAA-MM-DD (año, mes, día) admitido por la librería pandas de Python [12].

El Cuadro 2 muestra los rangos de fechas por semestre de las palabras clave en ambos idiomas, podemos notar que los máximos en menciones de las palabras contenidas en tuits en inglés se alcanzan en los primeros 3 meses de los años 2018 y 2019, principalmente en enero. En cambio las palabras contenidas en tuits en español no siguen el mismo patrón al tener sus máximos en meses diferentes. En el caso especial del periodo COVID-19, cuando la cuarentena iniciaba en marzo de 2020 para algunos países [20] todas las palabras en inglés tuvieron menciones máximas. Las palabras en español alcanzaron valores más

altos hasta los meses de abril y mayo, México iniciaba con la Fase 2 por contagio de la nueva enfermedad, alentando a la gente a quedarse en casa, ver Gráfico 1.

Para el segundo semestre de 2018 y 2019 la mayoría de las palabras en ambos idiomas alcanzaron máximos entre octubre y diciembre, puede resultar obvio, debido a las temporalidades donde las personas tienden a comprar más por eventos como Black Friday, Buen Fin, festividades navideñas y fin de año.

	Semestre				
Palabra	**2018-S1**	**2018-S2**	**2019-S1**	**2019-S2**	**2020-S1**
tv	2018-01-05 al 2018-01-11	2018-10-03 al 2018-10-09	2019-01-05 al 2019-01-11	2019-12-27 al 2020-01-02	2020-03-26 al 2020-04-01
phone	2018-01-01 al 2018-01-07	2018-09-30 al 2018-10-06	2019-02-18 al 2019-02-24	2019-09-06 al 2019-09-12	2020-03-21 al 2020-03-27
pc	2018-03-13 al 2018-03-19	2018-11-22 al 2018-11-28	2019-03-26 al 2019-04-01	2019-11-24 al 2019-11-30	2020-03-07 al 2020-03-13
smart	2018-01-03 al 2018-01-09	2018-11-04 al 2018-11-10	2019-01-03 al 2019-01-09	2019-09-15 al 2019-09-21	2020-03-09 al 2020-03-15

a) *Fechas de extracción palabras en inglés.*

	Semestre				
Palabra	**2018-S1**	**2018-S2**	**2019-S1**	**2019-S2**	**2020-S1**
tv	2018-06-21 al 2018-06-27	2018-11-26 al 2018-12-02	2019-04-20 al 2019-04-26	2019-10-18 al 2019-10-24	2020-04-15 al 2020-04-21
celular	2018-01-08 al 2018-01-14	2018-12-19 al 2018-12-25	2019-06-13 al 2019-06-19	2019-06-30 al 2019-07-06	2020-05-09 al 2020-05-15
pc	2018-02-15 al 2018-02-21	2018-11-30 al 2018-12-06	2019-06-07 al 2019-06-13	2019-11-11 al 2019-11-17	2020-05-16 al 2020-05-22
inteligente	2018-03-11 al 2018-03-17	2018-09-03 al 2018-09-09	2019-05-15 al 2019-05-21	2019-12-23 al 2019-12-29	2020-04-03 al 2020-04-09

b) *Fechas de extracción palabras en español.*

Cuadro 2. *Rango de fechas para extracción de noticias por semestre.*
Fuente: Elaboración Propia

Identificados los rangos de fechas se procede a elegir la mejor forma de obtener las nubes de palabras que muestren productos tecnológicos referentes a los términos anteriores, se analizarán gráficos usando frecuencia de aparición y el pesado TF-IDF.

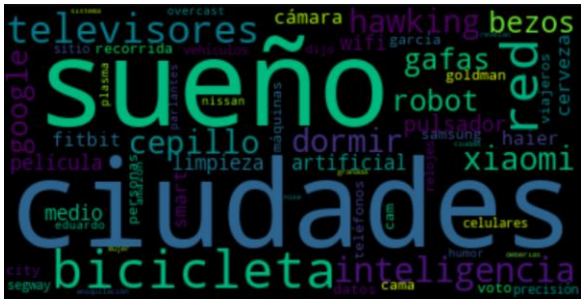
3.2 Elección de metodología para nubes de palabras

Para generar la nube de palabras, se extrajo el atributo “text” de todas las noticias, se aplicó un preprocesamiento básico que consistió en convertir el texto a minúsculas, se eliminaron signos de puntuación y se quitaron stopwords en inglés y español. Los números se mantienen dado que es posible encontrar versiones de productos o tipos de tecnología, por ejemplo, tecnología 4k. Se preservan los acentos, diéresis u otros signos diacríticos con el objetivo de no perder palabras

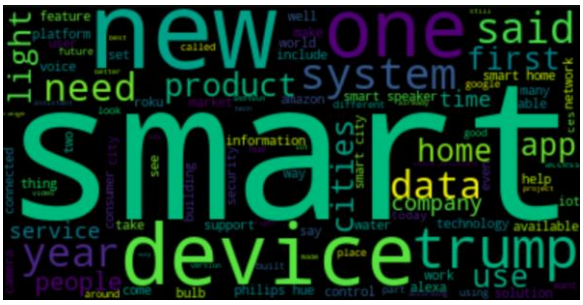
únicas que representen productos. Para más información de los pasos descritos, ver Anexo 1.

Con el objetivo de seleccionar la metodología adecuada se generaron nubes de palabras basadas en frecuencias y pesado TF-IDF. Se utilizaron las noticias del término “inteligente” en ambos idiomas y del primer semestre del 2018.

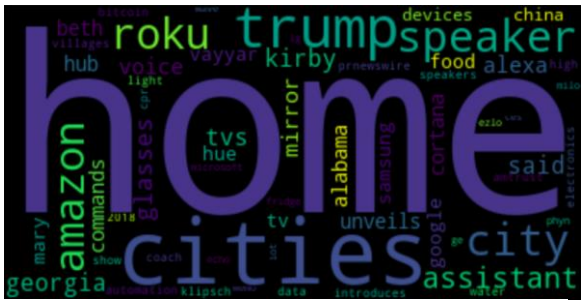
a) *Frecuencias en español*



b) *TF-IDF en español*



c) *Frecuencias en inglés*



d) *TF-IDF en inglés*

Gráfico 7. *Nubes de palabras de noticias del primer semestre 2018 usando frecuencia de aparición y pesado TF-IDF con el término “inteligente”.*

Fuente: *Elaboración Propia*

El Gráfico 7 (a) y (c) muestran que la palabra más sobresaliente es la misma con la que se realizaron las búsquedas de noticias, es decir “inteligente” y “smart”. Sobresalen aquellas con mayor repetición, por ejemplo, para los documentos en español se tiene: “persona”, “sistema”, “usuario”, etc. y en el caso del idioma inglés: “device”, “new”, “system”, “one”, etc. A pesar de ser las palabras más utilizadas dentro de las noticias recopiladas, no se muestran a primera vista

términos relacionados con productos tecnológicos, solo se alcanzan a visualizar las palabras “bulb”, ”app” y “home” en el Gráfico 7 (c).

Usando el pesado TF-IDF fue posible obtener palabras de productos que se visualizan fácilmente, ya que no solo se toma la frecuencia de cada término por documento, además se considera el número de documentos en donde el término aparece, es decir, una palabra tiene mayor peso mientras tenga una mayor frecuencia en un documento y menor aparición en el número de documentos, esto ayuda a distinguir palabras particulares. En el Gráfico 7 (b) se observan términos tales como:” televisores”, “gafas”, “cámara”, “bicicleta” etc., mientras que en inglés se distinguen “home”, “roku”, “speaker” y “glasses”,y marcas como “amazon”, “xiaomi” y “google”. Al usar frecuencias no fue posible visualizar estos términos rápidamente.

Por tanto, tomando en cuenta el análisis y el contenido de las visualizaciones, se ha decidido usar el pesado TF-IDF para continuar con las nubes de palabras del resto de términos por semestre.

3.3 Nubes de palabras

A continuación, se mostrarán las nubes de palabras aplicando el pesado TF-IDF a las noticias dado el resultado de la subsección anterior, se hará por término de búsqueda para ambos idiomas y por semestre de acuerdo al siguiente orden: “tv”, “celular”, “pc”, “smart”.

3.3.1 Nube de palabras “tv”

Gráfico 8. *Nubes de palabras semestrales palabra “tv”.*

Fuente: Elaboración Propia

Las nubes de palabras en español mayormente están relacionadas con eventos deportivos como fútbol y Fórmula 1. Podemos encontrar palabras de equipos de México; “tigres”, “guadalajara”, “cruz azul” y de España; “barcelona”, incluso se visualiza el término “mundial” en la nube *tv_esp_2018-S2*, año en el que se llevó a cabo el mundial Rusia 2018. También se encuentran nombres de empresas de televisión, actores y personajes famosos.

La nube *tv_esp_2020-S1* muestra “clases” como una palabra muy relevante en este periodo debido a la nueva forma de educación desde casa a través de televisión, por el tema de la contingencia en México, se refuerza esta idea pues también se muestran las palabras “Gatell” y “López” refiriéndose al subsecretario de salud Hugo López-Gatell. Los términos hallados en estas nubes, son de poca utilidad al no obtener términos de productos tecnológicos, solo se encuentra “apple” del cual se hablará más adelante.

En el caso de nubes de palabras en inglés, sí se aprecian términos que ayudan al objetivo del trabajo, por ejemplo, empresas de televisiones, como “lg” y “samsung”, son relevantes desde el primer semestre de 2018 hasta el segundo semestre de 2019, sin embargo en la nube *tv_eng_2020-S1* no figuran entre las principales, dando lugar a nuevas formas de entretenimiento como “streaming”[22] y “netflix”. También se visualizan “panasonic”, “philips” y “sony” con menos relevancia que las primeras 2, pues solo se muestran desde tv_*eng_2018-S1* hasta *tv_eng_2019-S1* y su tamaño se reduce al paso de los semestres. Vemos que a partir del segundo semestre de 2018 se muestra un tipo de tecnología aplicado a televisores, es el famoso “4k” y más recientemente “8k”.

Por último, para ambos idiomas desde el semestre *2019-S1* “apple” tuvo una relevancia tal que es la única en común hasta *2020-S1*, esta palabra hace referencia al producto Apple TV, un receptor digital diseñado para reproducir contenido multimedia con iTunes en una televisión de alta definición.

[22] AVG. Recuperado de https://www.avg.com/es/signal/what-is-streaming [Agosto 2020].

3.3.2 Nube de palabras “celular”

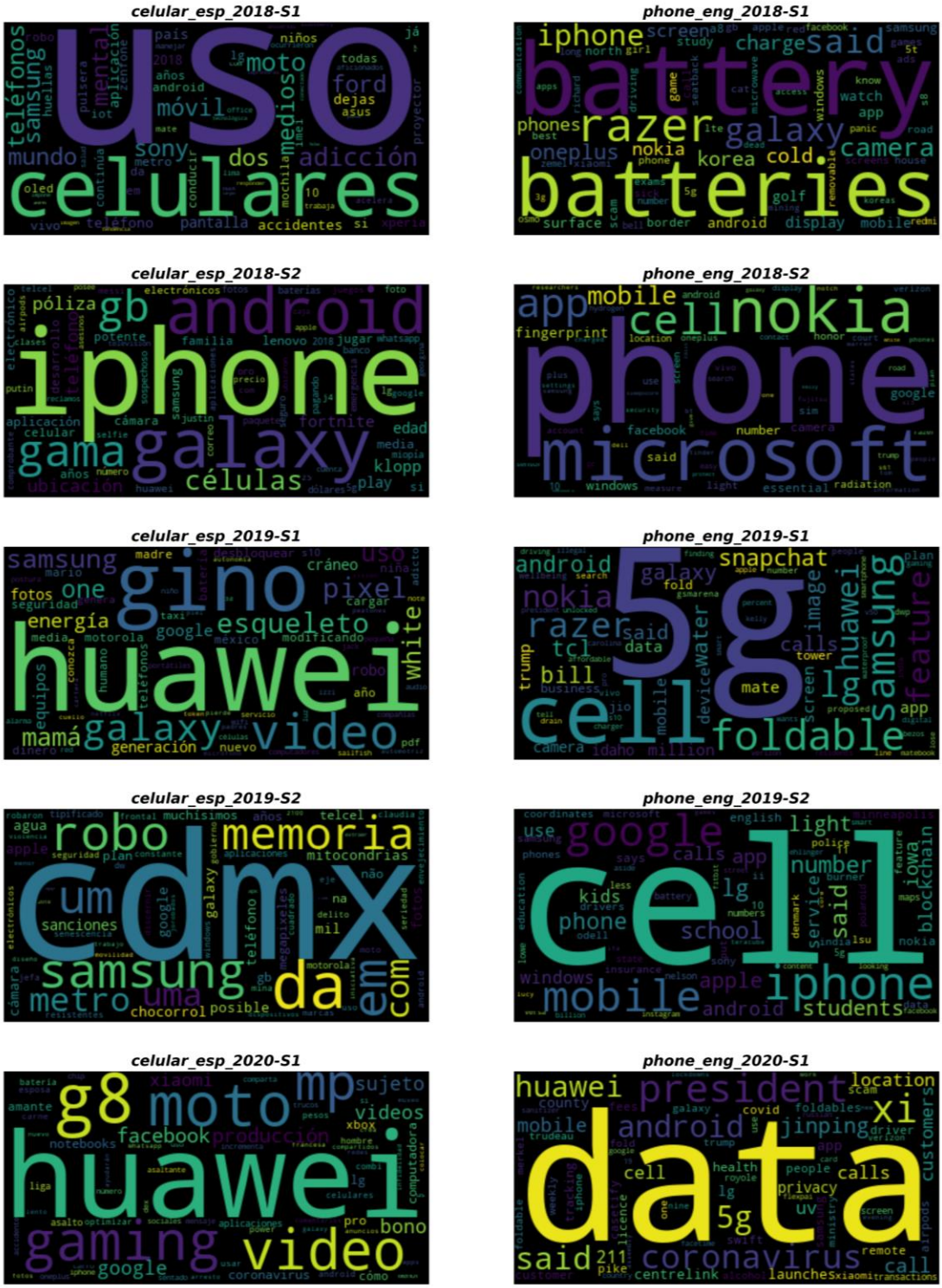


Gráfico 9. *Nubes de palabras semestrales palabra “celular”.*

Fuente: Elaboración Propia

El Gráfico 9 presenta gran cantidad de marcas y modelos de celular que podrían apoyarnos con nuestro objetivo. Comenzando con las nubes en español, se presentan dos gráficos cuyas palabras más relevantes son atípicas a lo que se busca obtener, nos referimos a los semestres *2018-S1* y *2019-S2;* en la primera destaca la palabra “uso”, profundizando en el contenido y encabezado de las noticias nos encontramos con temas relacionados a accidentes vehiculares causados por el uso del celular mientras se maneja. Además existen noticias describiendo el uso correcto del celular, para que niños y jóvenes no caigan en una “adicción”, palabra que también observamos en esa nube de palabras.

La nube *celular_esp_2019-S2* presenta términos como: “cdmx”, “metro” y “robo”, se detectaron 11 noticias relacionadas con hechos delictivos y presentación de nuevas normas aplicables por robo de celulares en la Ciudad de México a bordo del transporte colectivo Metro. Esto nos muestra la importancia del análisis de fuentes no estructuradas, como son las noticias, nos permitió detectar rápidamente hechos particulares que humanamente tomarían tiempo analizar.

Continuando con el resto de nubes en español, observamos que “huawei” aparece como una palabra con gran relevancia, los encabezados de las noticias muestran lanzamientos sobre nuevos celulares para esta marca y también noticias haciendo una comparación con otras marcas de celular como “galaxy”, “iphone”, “moto” y “sony” apreciables en las nubes de palabras.

Analizando las nubes de palabras en inglés, desde el primer semestre de 2019 se tenía en mente el tema de la red “5g” y de celulares que se doblan, esto por la palabra “foldable”, también se observan “nokia” y “samsung”. La nube *phone_eng_2018-S1* muestra “battery” y “batteries” como las palabras con mayor peso, el contenido de las noticias describe consejos para recargar tu celular, se propone mantener una carga por encima del 50% y no dejar sin recargar por periodos largos de tiempo ya que eventualmente no será capaz de mantener la carga completa, tal y como ha ocurrido con laptops. También encontramos nombres de empresas figurando en diferentes semestres, como ejemplo “razer”, compañía que fabrica hardware para videojuegos y celulares de gran capacidad, específicamente para atender la demanda de jugadores. Otra empresa es

“microsoft”, una de las más grandes de software en el mundo, a pesar de haber lanzado su propio celular integrando Windows como sistema operativo, no resultó tener éxito en el mercado.

En el gráfico *phone_eng_2020-S1* existe el término “data” como el más relevante del semestre, de igual forma es apreciable “coronavirus”, estos términos refieren a noticias donde el gobierno de algunos países ha propuesto seguir la dispersión del virus entre la población usando y analizando datos geográficos mediante el GPS de los celulares con el objetivo de tomar decisiones para controlar esta enfermedad.

Recordando el segundo pico más alto de la Gráfica 4 (a) que se registró el 12 de septiembre del año 2018 para el término “phone” se realizó la extracción de noticias con el objetivo de identificar algún hecho que haya provocado menciones tan altas para ese término, el proceso y el gráfico se exponen en el Anexo 1. Las palabras más relevantes que se obtuvieron en el gráfico fueron “iphone” y “xs”, esto claramente hace referencia a la gama de celulares iPhone Xs que fueron lanzados por primera vez al mercado el 12 de septiembre de 2018, causando sensación entre los usuarios.

3.3.3 Nube de palabras “pc”

Gráfico 10. *Nubes de palabras semestrales palabra “pc”.*

Fuente: Elaboración Propia

En este caso no es difícil identificar términos ya conocidos por los usuarios dado su gran renombre en la industria, las palabras relacionadas a “pc” son principalmente de videojuegos, empresas de software, plataformas digitales y de streaming tales como: “steam”, “stadia” y “twitch”, además encontramos consolas muy famosas mencionadas en el semestre *2020-S1* como los términos más relevantes, posiblemente por el tema de contingencia donde el consumo de videojuegos aumentó un 36% [22], también en este periodo se dieron lanzamientos de nuevas consolas como son: “xbox” en la nube de palabras en español y “playstation” para la nube de palabras en inglés, podría pensarse que para hablantes del español es más relevante la consola Xbox, mientras que para personas de habla inglesa sería PlayStation. Se pueden observar algunos juegos para estas consolas y pcs, por ejemplo, en el caso del idioma español se tienen “halo” y “death stranding”, mientras que en el idioma inglés es más relevante “fortnite”. Un caso curioso es la palabra “huawei” que aparece como una de mayor peso para el rango de fechas del primer semestre de 2019 para noticias en español, el contenido describe que esta empresa ha lanzado una tecnología para que el celular sea usado como una pc a través de su PC Mode. Parece que la mayoría de los usuarios usan el sistema operativo “windows” como el principal para sus pcs dado que es una palabra que se mantiene en todos los semestres para ambos idiomas, sin embargo, al pasar de los años las noticias respecto a este sistema han sido cada vez menos relevantes.

En cuanto hardware existe relevancia en procesadores tales como: “razen” e “intel”, unidad de almacenamiento “ssd” y unidades gráficas, principalmente “rtx” indispensables para tener los mejores gráficos en videojuegos.

3.3.4 Nube de palabras “inteligente”

Gráfico 11. *Nubes de palabras semestrales palabra “inteligente”.*

Fuente: Elaboración Propia

Para los temas relacionados al término “inteligente” encontramos objetos e inmuebles, se tienen palabras como “bicicleta”, que solo aparece en noticias de español para el primer semestre de 2018 dada su innovación por estaciones donde se pueden recoger bicicletas, otras noticias mencionan que mediante una aplicación podrás encontrar la más cercana para usarla. El término “espejo” relevante para noticias en inglés durante el semestre 2019-S1 y presentado en el CES de ese año, ha sido una innovación que permite conectarse con asistentes inteligentes y que dentro de sus múltiples funcionalidades cuenta con sensores que monitorean la salud cardiovascular mediante cambios en la apariencia física del usuario.

Otro objeto útil para convertir tu baño convencional en uno inteligente es el referido al término “inodoro” también fue presentado en el CES 2019 y es controlado por asistentes inteligentes, permite analizar orina y heces para detectar enfermedades de los usuarios. Se ha descrito el uso de asistentes inteligentes para controlar otros objetos y en el Gráfico 11 se muestran algunos como “alexa” y el asistente de “google”.

También podemos detectar wearables[23], entre ellos “pulsera” o “reloj” para monitoreo de la actividad física, uso rápido de aplicaciones de mensajería y “gafas” usadas para la visualización de realidad virtual y aumentada.

Un término que aparece en la nube *inteligente_esp_2019-S2* y poco común es “invernadero”, al identificar las noticias relacionadas vemos que se trata de un invernadero doméstico con el que se busca cultivar hortalizas y verduras de manera sustentable y durante todo el año.

Para el caso de los inmuebles, la palabra más relevante es “ciudad” que se muestra en la mayoría de las nubes de palabras por semestre y en ambos idiomas, también encontramos “casas”, “edificios”.

---

[23] Universidad Internacional de Valencia. Recuperado de https://www.universidadviu.com/es/actualidad/nuestros-expertos/que-es-wearable-y-que-tipos-de-dispositivos-existen [Agosto 2020].

3.4 Productos tecnológicos

Una vez analizadas las nubes de palabras y de haber mostrado los términos más relevantes por semestre, el Cuadro 3 resume los productos tecnológicos detectados por cada palabra clave con la que se realizó la extracción de noticias, en algunos casos se han determinado categorías, por ejemplo, los objetos del término “pc”, se dividieron en consolas, juegos y hardware, el orden en el que aparecen los productos representa su relevancia en las nubes de palabras.

Palabra Clave	Categoría	Producto
Tv	Televisores	• LG con tecnología 4k. • Samsung con tecnología 4k. • Panasonic • Philips • Sony
	Reproductores multimedia	• Apple TV
Celular		• Huawei • Iphone • Galaxy • Moto • Sony • Samsung
Pc	Consolas	• Xbox • Playstation
	Juegos	• Halo • Fornite • Death stranding

	Hardware	• Procesador Razen • Procesador Intel • Tarjetas gráficas RTX • SSD
Inteligente	Asistentes	• Alexa • Google
	Wearables	• Reloj • Pulsera • Gafas

Cuadro 3. *Productos tecnológicos por término de búsqueda.*

Fuente: Elaboración Propia

Conclusiones

Conclusiones

Gracias a la librería text models, obtener datos de menciones diarias en tuits de un término en específico ha sido sencillo, solo fue requerido ingresar unigramas en inglés y español, es decir, nuestras palabras clave. Esta herramienta ahorró el desarrollo de extracción de tuits en crudo, el preprocesamiento y la generación de una estructura de datos adecuada para el análisis. Se logró visualizar paralelamente el comportamiento de las menciones a través del tiempo de un término relacionado a tecnología en ambos idiomas, por tanto fue sencillo detectar eventualidades que detonan el uso de esa palabra en cierto rango de fechas, vimos que la mayoría de palabras en inglés tienen un número de menciones superior al de palabras en español, mostrando que usuarios de Twitter tienen un mayor interés por expresar opiniones o referencias acerca de la tecnología en inglés.

Mediante los gráficos de menciones en el periodo COVID, se pudo seleccionar por grupo de palabras la mejor opción que permitiera un análisis más profundo tal y como se realizó con las nubes de palabras. Por medio de la librería googler se extrajeron noticias entorno a los días donde se tuvieron valores máximos en menciones de una palabra. Al realizar el preprocesamiento se descubrió que un pesado TF-IDF es la metodología que permite mostrar los términos más relevantes de productos tecnológicos, ya que usando Frecuencias, dificultaba la visualización de los términos de interés.

Por último, describir que esta metodología no solo puede ser usada para encontrar palabras específicas, también para visualizar términos que podrían darnos pistas relacionadas a eventos, lanzamientos de productos o como en este caso, que se descubrieron hechos delictivos referentes a robo de celulares en el Metro de la CDMX, herramienta útil que podría ayudar a detectar hechos no denunciados por la población. En el caso de los términos tecnológicos, esta metodología sería útil para los ecommerce, pues podrían basar sus decisiones sobre qué productos ofertar de acuerdo con la tendencia e interés de los usuarios.

Bibliografía

[1] Eric S. Tellez, Sabino M,J., Mario Graff, Daniela M., Oscar S. Sjordia, Elio A. Villaseñor. (2017. 15 Septiembre). A case study of Spanish text transformations for twitter sentiment analysis. Expert Systems with Applications, Volumen 81, Páginas 457-471.

[2] Statista. (2020, 24 Noviembre). Global social networks ranked by number of users 2020. Recuperado Agosto 2020, de https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users

[3] Bjarke Felbo, Alan Mislove, Ander Sogaard, Iyad Rahwan, Sune Lehmann. (2017, 7-11 Spetiembre). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. Proceedings of the 2017 Conference on Empirical Methods in Natural Languaje Processing, páginas 1615-1625 Copenhagen, Denmark.

[4] Mario Graff, Sabino M. J. (2019). Clasificación de Texto. Recuperado Septiembre 2019, de http://ingeotec.mx/%7Emgraffg/clasificacion_texto.html

[5] Rose. D. (2016). Data Science: Create Teams That Ask the Right Questions and Deliver Real Value. Capítulo 12, páginas 106-114. Primera edición, Atlanta: Apress.

[6] Mathangi Sri, Digital Vidya, Lifecycle of a Data Science Project. (2018, 2 Septiembre). [Archivo de video]. Recuperado Mayo 2019, de https://www.youtube.com/watch?v=s4YY02dRm4Q

[7] Sidorov, G., Miranda Jiménez, S., Viveros J., F., Gelbukh, A., Castro Sanchez, N., Velásquez F., & Gordon, J. (2012, Octubre). Empirical study of machine learning based approach for opinion mining in tweets. In Mexican international conference on Artificial intelligence. Páginas 1-14. Springer, Berlin, Heidelberg.

[8] Freddy Vega, Platzi. El estado del eCommerce en el mundo en el 2019. (2019, 8 Marzo). [Archivo de vídeo]. Recuperado Mayo 2019, de https://www.youtube.com/watch?v=kSwDHNjGIOY

[9] Henri Hakkinen, Arun Prakash J., Zhiming W., Johnathan J., SZ Lin, Jerko S. (2020, 27 Julio). “googler V4.2”. Recuperado Agosto 2020, de https://github.com/jarun/googler

[10] Mario Graff, Daniela M., Sabino Mirando-Jiménez, Eric S. Tellez. (2020, 3 Septiembre). A Python Library for Exploratory Dara Analysis and Knowledge Discovery on Twitter Data. Recuperado Septiembre 2020, de https://arxiv.org/abs/2009.01826

[11] Asociación Mexicana de Ventas Online (AMVO). Reporte No. 2. Impacto COVID-19 en Venta Online México, (2020, 23 Abril). Versión Pública. Recuperado Junio 2020, de https://www.amvo.org.mx/estudios/reporte-2-impacto-covid-19-en-venta-online-mexico/#0

[12] Pandas documentation - pandas 1.1.4 documentation. (2020). Recuperado Julio de 2020, de https://pandas.pydata.org/pandas-docs/stable/index.html

[13] Tom B. B., Benjamin M., Nick R., Melanie S., Jared K., Prafulla D., Arvind N., Pranav S., Girish S., et al. (2020, 22 Julio). Languaje Models are Few-Shot Learners. OpenAI.

[14] Instituto Nacional de Estadística y Geografía (México). Estado de ánimo de los tuiteros en México. Recuperado Agosto 2019, de https://www.inegi.org.mx/app/animotiutero/#app/multiline

[15] Abel Coronado. Analizando el Big Data de las Noticias con tu Micro Data Lake (Baterías Incluidas). (2020, 6 Mayo). Recuperado Mayo 2020, de https://abxda.wordpress.com/2020/05/06/analizando-el-big-data-de-las-noticias-con-tu-micro-data-lake-baterias-incluidas/

[16] The Economist. (2010, 26 Febrero). Data, data everywhere. Recuperado Marzo 2020, de https://www.economist.com/special-report/2010/02/27/data-data-everywhere

[17] Devin Pickell, (2018, 16 Noviembre). Structured vs Unstructured Data What´s the Difference? Learning Hub. Recuperado Junio 2019, de https://learn.g2.com/structured-vs-unstructured-data

[18] Salvador García, Sergio Ramírez-Gallego, Julián Luengo, Francisco Herrera. (2016, Julio-Octubre). Big Data: Preprocesamiento y calidad de datos.

Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada (España), novática #237

[19] Lino Alberto Urdaneta Fernández. (2019, 24 Mayo). Reducir el número de palabras de un texto: lematización y radicalización (stemming) con Python. Medium. Recuperado Agosto 2020, de https://medium.com/qu4nt/reducir-el-número-de-palabras-de-un-texto-lematización-y-radicalización-stemming-con-python-965bfd0c69fa

[20] Ximena Castro Sardi, Diego Cagüeñas Rozo, Diana Patricia Quintero Mosquera, Juan José Fernández Dusso y Rafael Silva Vega. (2020). Ensayos sobre la pandemia, Cali, Universidad Icesi.

[21] Fueller, C. (2020, 17 Marzo). Así te contamos el avance del coronavirus a nivel mundial este lunes. Recuperado Septiembre 2020, de https://www.laprensa.com.ni/2020/03/16/internacionales/2651461-este-es-el-minuto-a-minuto-sobre-los-avances-del-coronavirus-a-nivel-mundial

[22] Cueto, H. (2020, 1 Mayo). El consumo de videojuegos se dispara por la pandemia de coronavirus. Recuperado Septiembre 2020, de https://businessinsider.mx/la-pandemia-de-coronavirus-dispara-el-consumo-de-videojuegos-pero-tambien-trae-consecuencias-para-la-industria

[23] Thomas M. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein. (2009). Introduction to algorithms. MIT, Tercera Edición.

[24] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2013, 24 Junio). An introduction to statistical learning with applications in R. Springer Science and Business Media.

[25] Robert Johnson, Patricia Kuby.(2008). Estadística elemental: Lo esencial. Cengage Learning, Décima Edición.

ANEXOS

ANEXO 1

Este anexo muestra parte del proceso para la generación de nubes de palabras expuestas en el capítulo 3, desde la extracción de noticias pasando por el preprocesamiento, hasta la generación del gráfico. Nos enfocaremos principalmente en la fecha del 12 de septiembre de 2018, día en que se obtuvo menciones con frecuencias muy altas de la palabra “celular” en inglés y que se visualiza en el Gráfico 4 (a). Recordar que el código para desarrollar esta aplicación se trata del lenguaje de programación Python.

Extracción de noticias

Comenzaremos con la extracción de noticias, por lo que haremos uso de un archivo con extensión “sh” y que contendrá el siguiente comando:

python googler_new.py -n 100 --news --json --lang "$8" --from "$3"/"$4"/"$2" --to "$5"/"$6"/"$2" "$1" > "$1"_"$8"_"$7".json

Cada valor de la forma “$n”, con n de 1 a 8; representa una variable que haremos llegar su valor usando la librería subprocess de Python que iniciará con nuestra extracción, primero hay que importarla de la siguiente forma:

import subprocess

Ahora estableceremos los argumentos que necesitamos, están ordenados de acuerdo a las variables del archivo sh al que hemos decido nombrar “news_extractor.sh”, por ejemplo, *search_topic* representa a “$1”, *execution_year* es “$2”, etc. El siguiente pedazo de código hará el trabajo de traernos las noticias que le solicitamos.

Establecer parámetros
search_topic="phone"
execution_year="2018"

start_execution_month="9"
start_execution_day="12"
end_execution_month="9"
end_execution_day="12"
semester="S2"
lang="en"
Comenzar con extracción
subprocess.run(['sh','news_extractor.sh',search_topic,execution_year,\start_execution_month,start_execution_day,end_execution_month,\end_execution_day,semester,lang])

Preprocesamiento

Una vez extraídas las noticias en el archivo “phone_en_S2.json” haremos la lectura y el preprocesamiento de texto, para ello se realizarán funciones que facilitarán el flujo del proceso. Lo primero que haremos será cargar las stopwords tanto en inglés como en español de la siguiente forma:

from nltk.corpus import stopwords
sw_eng=stopwords.words('english')
sw_esp=stopwords.words('spanish')

Los objetos sw_esp y sw_eng son listas que contienen las palabras más comunes de cada idioma y serán utilizadas para removerlas de las noticias dado que no son términos que nos interese visualizar. La siguiente función tiene como objetivo homologar las noticias, convierte cada una de las palabras a minúsculas, se sustituye por un espacio en blanco todo lo que comienza con “https” o “@”, es decir los links o menciones que lleguen a aparecer, además de todo aquello que no sea dígito, letra o espacio en blanco, por ejemplo caracteres como “$”,”%”,”&”, etc.

def remove_punctuation(text):
    import re

```
    text=text.lower()
    text=re.sub(r'https\S+|@\S+|[^\w\s]|_',' ',text)
    text=re.sub(r'\s+',' ',text).strip()
    return text
```

La siguiente función únicamente se encarga de eliminar las stopwords presentes en los textos de las noticias usando las listas que se cargaron previamente.

```
def remove_sw(text):
    texto=text.copy()
    sw=sw_eng+sw_esp
    for w in text:
        if w in sw:
            texto.remove(w)
    return texto
```

Por último, se realizó una función que leerá el archivo de noticias y hará uso de las funciones anteriores para completar el preprocesamiento que consiste en homologar texto y eliminar stop words. Para ello será necesario que el archivo de noticias se mantenga en el mismo directorio del archivo de Python. Analizando el siguiente código se debe mencionar que el objeto news_data almacena todo el contenido del archivo que se leyó, posteriormente en el ciclo for se recorre cada una de las noticias y se evalúa si el atributo text contiene una longitud menor a 2, entonces se tomará el título concatenado con el abstract como la noticia, en caso contrario será el atributo text. En la última parte del código se aplica el preprocesamiento ya descrito anteriormente.

```
def text_generator(filename):
    import json
    x=filename+".json"
    with open(x) as f:
        news_data = json.load(f)
    news_text=[]
```

```
    for n in news_data:
        if len(n["text"])<2:
            n=n["title"]+" "+n["abstract"]
        else:
            n=n["text"]
        news_text.append( " ".join(news for news in remove_sw( remove_punctuation( n ).split() )))
    return news_text
```

Pesado TFIDF

Ahora convertiremos nuestro texto en vectores usando el pesado TFIDF con TfidfVectorizer parte de la librería sklearn, se carga de la siguiente forma:

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

Se procederá a realizar la transformación mediante el siguiente pedazo de código, en X se tendrán los valores del pesado de cada término por cada noticia, además también se recogen las posiciones y palabras de todo el corpus en pos_tokens y tokens respectivamente.

```
corpus = text_generator("phone_en_S2")
vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(corpus)
X = X.toarray()
pos_tokens = np.array(list(vectorizer.vocabulary_.values()))
tokens = np.array(list(vectorizer.vocabulary_.keys()))
```

Lo siguiente es tomar los términos más relevantes por cada noticia de acuerdo al pesado TFIDF, se ordenará de mayor a menor pesado iterando sobre X, se seleccionarán únicamente los 10 términos con mayor pesado, los valores de TFIDF almacenarán en tfidf_values, mientras que los términos en tfidf_tokens. Usaremos un DataFrame para agrupar por término y se hará la suma de sus pesados, se ordenará de mayor a menor, el proceso se lleva a cabo

en el objeto most_relevant, los 70 términos más relevantes serán agregados a un diccionario llamado tfidf_terms dado que es la estructura de datos útil para nuestra nube de palabras. Todo lo descrito se presenta en el siguiente código.

```
import numpy as np
import pandas as pd

tfidf_values = np.array([])
tfidf_tokens = np.array([])
    for i in range(X.shape[0] ):
        x = np.argsort(-1*X[i,:],)[:10] #Se ordenan pesados por noticia.
        tfidf = X[i,x] # tfidf values por noticia
        t_ = tokens[[np.where(pos_tokens==y)[0][0] for y in x]] # tfidf tokens
        tfidf_values = np.append(tfidf_values,tfidf )
        tfidf_tokens = np.append(tfidf_tokens,t_)

most_relevant=pd.DataFrame({"tfidf_tokens":tfidf_tokens,"tfidf_values":tfidf_value})

most_relevant=most_relevant.groupby("tfidf_tokens").agg("sum").sort_values(by="t
fidf_values",ascending=False)
most_relevant = most_relevant.head(70).reset_index()
most_relevant["Freq"] = np.floor(most_relevant["tfidf_values"]*10)
most_relevant = most_relevant.iloc[:,[0,2] ]
most_relevant["Freq"] = most_relevant["Freq"].astype(int)

tfidf_terms = {}
    for k,v in most_relevant.iterrows():
        tfidf_terms[v[0]] = v[1]
```

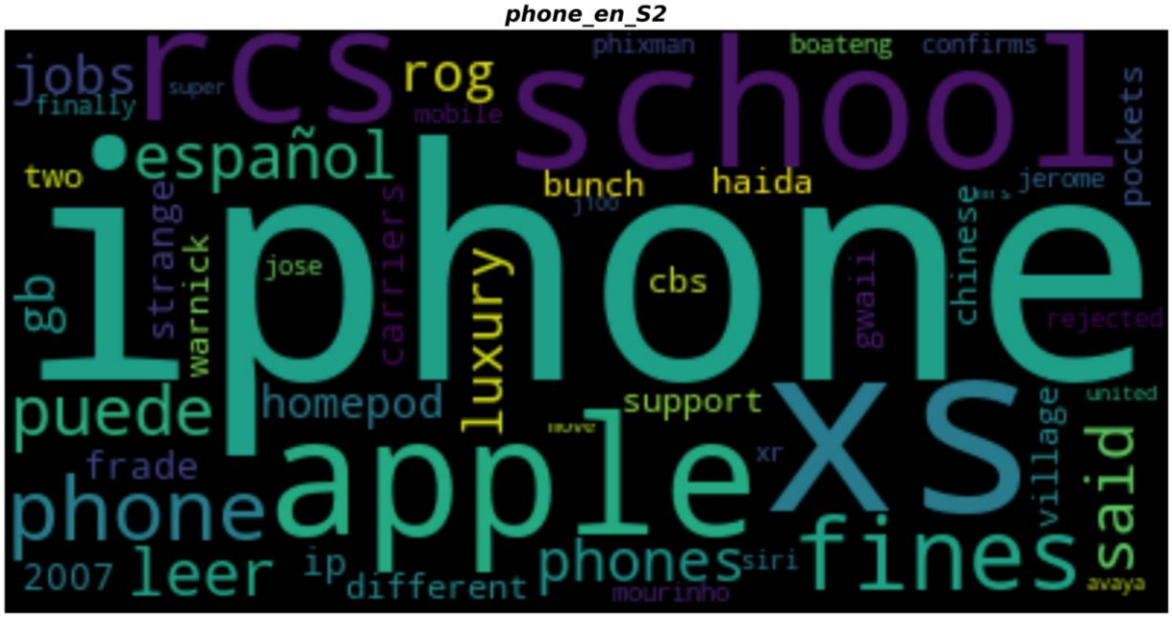
Nube de palabras

Lo último a realizar es la nube de palabras, para ello haremos uso de las siguientes funciones:

from wordcloud import WordCloud
import matplotlib.pyplot as plt

El código de abajo nos generará la nube con los términos del diccionario tfidf_terms, es un parámetro que recibe la función WordCloud por medio de su método generate_from_frequencies, ejecutando el código, obtenemos el resultado esperado.

wordcloud=WordCloud(max_font_size=200,max_words=50).generate_from_frequencies(tfidf_terms)
figsize(15,12)
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.title(filename)



Índice de términos