UNIVERSIDAD DE MÁLAGA

E.T.S. INGENIERÍA
INFORMÁTICA
UNIVERSIDAD DE MÁLAGA

INGENIERÍA INFORMÁTICA

CLASIFICACIÓN Y RESPUESTA AUTOMÁTICA DE CORREOS ELECTRÓNICOS MEDIANTE HERRAMIENTAS DE ANÁLISIS DE DATOS

CLASSIFICATION AND AUTOMATIC REPLY OF E-MAILS BY MEANS OF DATA ANALYSIS

Realizado por
JOSE MANUEL RUIZ LUQUE

Tutorizado por
NICOLÁS MADRID LABRADOR
ÁNGEL MORA BONILLA

Departamento
MATEMÁTICA APLICADA
UNIVERSIDAD DE MÁLAGA

MÁLAGA, AGOSTO 2020

ÍNDICE

Resumen 5

Summary 5

1. INTRODUCCIÓN 7

1.1. Metodología de trabajo 12

2. LEY ORGÁNICA DE PROTECCIÓN DE DATOS 15

3. MINERÍA DE TEXTO 19

4. ANÁLISIS DE CONCEPTOS FORMALES 21

5. ANÁLISIS ESTADÍSTICO DE TESTEOS 27

6. CLASIFICADOR DE CORREOS ELECTRÓNICOS 31

6.1. Descripción del Problema 31

6.2. Implementación 31

7. CONCLUSIONES Y LÍNEAS FUTURAS 43

8. BIBLIOGRAFÍA 45

ANEXO 47

Estadística.R 47

Clasificador.R 53

CargaContextoFormalIngles.R 59

ClasificarIdioma.R 61

CargaContextoFormal.R 65

ScriptClasificadorIngles.R 67

PesoCancelacion.R 71

PesoNoCancelacion.R 73

Resumen

Con este estudio nuestro objetivo es reducir el tiempo de trabajo de los empleados de una empresa, haciendo que se centren en la respuesta a correos más individualizados mientras que los correos generales se respondan automáticamente. Por ello, hemos desarrollado un clasificador que es capaz de filtrar las palabras y dar una respuesta estándar si encuentra ciertas palabras clave usando la teoría de conceptos formales y la minería de datos como base. De esta forma, permitimos que una empresa un servicio más eficiente a sus clientes, los cuales recibirán una respuesta más rápida y en caso de ser un correo más personal, ya será tratado por un trabajador, que solo prestará atención a este tipo de correos. Así, mejoraremos la imagen de empresa al utilizar las nuevas tecnologías y los sistemas de información en los servicios que se ofrecen al cliente.

Summary

By the end of this research, we will get a reduction of employees' worktime, by making them to focus on an individualized reply to all the emails received and get an automatic reply for all the cancellations emails. To achieve this, we built a classifier that is able to filter words and give a classification if it finds some keywords using formal concept analysis and data mining. This tool would help to a company to provide to its customers a customized service, giving them either a faster answer or a more detailed answer in case of his email is related to another topic different to cancellations. With this solution, a company will get a better reputation using the latest technologies and information systems to offer better services.

Palabras clave: clasificador, análisis conceptos formal, minería datos, idioma.

Keywords: classifier, formal concept analysis, data mining, language

1. Introducción

Durante los últimos años los sistemas de información constituyen uno de los principales ámbitos de estudio en el área de organización de empresas, ya que el entorno que rodea a las organizaciones es cada vez más complejo y dinámico, y deben ir adaptándose a todos los cambios de una manera rápida y eficiente. Por ello, la información se ha convertido en un elemento clave en la gestión de empresas, debido a que la información es poder y esto es vital para hacer frente a la creciente globalización e internacionalización de las empresas, a la feroz competencia en los mercados, al avance continuo en el desarrollo de las Tecnologías de Información y Comunicación (TIC), a la incertidumbre que genera el entorno cambiante en el que vivimos y a la reducción de los ciclos de vida de los productos. Y no solo es un elemento clave para la gestión, si no que la información es un factor productivo más a tratar por las empresas como lo son el trabajo y el capital, siendo así un recurso con el que poder asegurar la supervivencia de la compañía y el crecimiento de esta.

Con relación a la importancia de la información, es fundamental saber adaptarse a los nuevos tiempos y a la constante evolución de la informática junto con las diferentes herramientas que ofrece. Por esta razón, utilizar las nuevas tecnologías en nuestras organizaciones es vital, y el uso de herramientas como podrían ser el Internet de las Cosas, el manejo de grandes volúmenes de datos (Big Data) o la Inteligencia Artificial, entre otras, adquieren gran relevancia.

A la vez que la tecnología realiza avances en el proceso de información, también se avanza en los medios de comunicación, que cada vez cuentan con mejoras en este sentido. Las vías de comunicación tradicionales del cliente con la empresa eran las cartas escritas y las llamadas telefónicas. Pero con el paso del

tiempo esto ha ido cambiando, y con la llegada del correo electrónico, estos medios de comunicación han quedado en un segundo plano.

Como consecuencia, el correo electrónico es el medio más utilizado para la atención al cliente de las empresas, ya que para los clientes es fundamental recibir una respuesta personalizada e individualizada en cualquier momento. Esto es valorado positivamente por los clientes y favorece la mejora de la calidad del servicio de atención del cliente de la empresa y su imagen de marca, lo cual genera un beneficio intangible de gran valor para la empresa. Pero esta vía de comunicación con los clientes puede suponer un problema para las empresas, ya que la gran mayoría recibe una cantidad superior a su capacidad de respuesta. Esto es debido a que una amplia parte de los clientes no suelen buscar la información disponible sobre algunos aspectos que se encuentran contemplados y expuestos, por ejemplo, en las secciones FAQ's de la página web de la empresa o en otros apartados de esta, sino que esperan recibir una respuesta más directa y esto supone un gran volumen de correos con preguntas básicas y que no se refieren a un problema o duda particular y para las cuales ya existe una respuesta clara.

Aparte de esta razón, y teniendo en cuenta el momento actual que atravesamos con la situación de emergencia sanitaria, para compañías de servicios aéreos, hoteleros, reservas de coche, etc. que ha supuesto la cancelación la gran mayoría de sus reservas. Por ello, la cantidad de correos electrónicos con solicitudes de cancelación ha desbordado las capacidades de la mayoría de las empresas. Por consiguiente, ser capaces de dar respuesta a todos ellos de manera manual se ha hecho inviable. Pero aplicando lo que la informática y sus herramientas nos ofrece podríamos conseguir automatizar el proceso, para responder de manera automática a una serie de preguntas frecuentes en los diferentes correos electrónicos recibidos, para lo que es esencial el uso de un sistema de información.

Las funciones de un sistema de información constan de distintos elementos clave para su funcionamiento. El primer elemento clave que necesita un sistema de información es un conjunto de datos de entrada, del que, con distintos procesos, seremos capaces de extraer información. Este procesado consta de dos partes. La primera de ellas es la del almacenamiento de los datos en una base de datos. El segundo proceso es el más importante, que es la etapa de procesamiento de los datos, en el que, con una serie de operaciones sobre el conjunto de entrada, obtendremos como resultado una información final. La información que extraemos del sistema podremos usarla para toma de decisiones, pero a su vez, nos servirá de ayuda para retroalimentar al mismo.

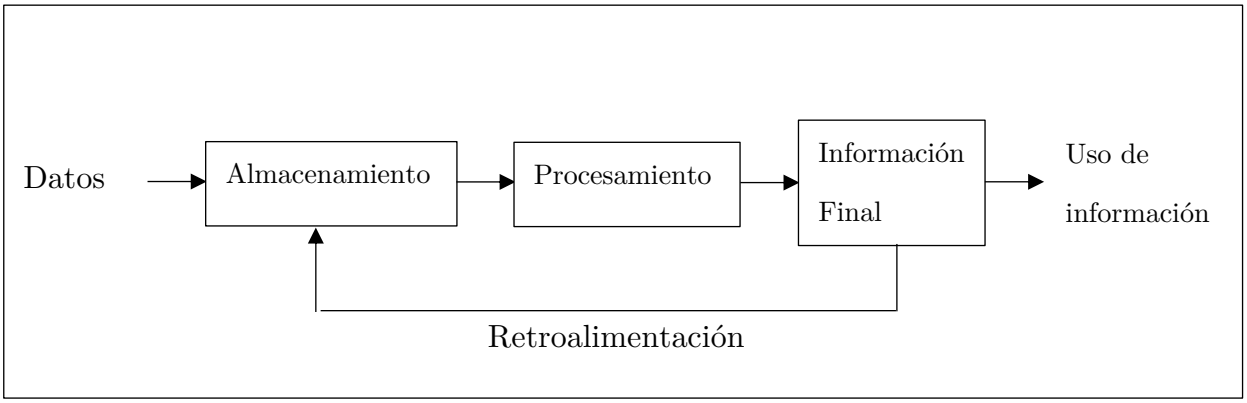


Figura 1. Sistemas de Información de la Organización empresarial

Usando el concepto de sistema de información podremos dar solución al problema de la respuesta automática de correos electrónicos. Como ya se ha mencionado anteriormente, este problema es de gran interés cuando se espera un gran volumen de correos electrónicos recibidos. En nuestro caso, nuestro sistema será capaz de diferenciar si un correo electrónico trata sobre una cancelación o sobre otro tema indeterminado. Para ello, necesitaremos dos clasificadores para cuando se reciba un correo electrónico. El primero de ellos nos dirá el idioma en el que está escrito; consideraremos dos idiomas, castellano e inglés. El otro clasificador nos dirá el tema del que trata el correo, dando como resultado si se

trata de un correo electrónico de cancelación o de otro tema. Este clasificador, que resulta simple a primera vista, nos da como resultado un idioma y un tema, lo que nos hace posible redactar una respuesta común a los correos recibidos que cumplan con ese idioma y ese tema.

Con la obtención de este clasificador, resolveremos el problema de no poder responder a cada uno de los correos electrónicos manualmente. Además, su uso puede evitar inconvenientes tales como la pérdida de tiempo de efectividad de los trabajadores que responden correos manualmente, efectos negativos en cuanto a la reputación de la empresa ya que existe el riesgo de que algunos clientes no reciban respuesta en un tiempo óptimo o prudencial, etc.

Sin embargo, su uso nos puede generar grandes beneficios, como la posibilidad de responder a todos los clientes de manera rápida, ya que los trabajadores que responden manualmente sólo tendrían que responder a los correos que no tratan sobre cancelaciones o que no estén redactados ni en inglés ni en castellano. Incluso en un futuro, esto se podría reducir, ya que añadir temas al clasificador e idiomas puede ser relativamente fácil. No solo obtendríamos este gran beneficio, sino que también aumenta la velocidad de respuesta en las gestiones, y esos trabajadores que antes respondían manualmente podrían realizar las gestiones post respuesta de estos emails. Como serían, en el caso de nuestro clasificador, anular las reservas, pagar el importe, etc. en un tiempo más reducido, lo que se reflejaría en una mejora de la reputación de la empresa y de la calidad de sus servicios.

La idea principal en base a la cual vamos a desarrollar nuestra herramienta es el procesamiento de los correos electrónicos, para el que recurriremos a diferentes técnicas pertenecientes a la minería de texto y al análisis de conceptos formales. Con la primera de ellas eliminaremos artículos, preposiciones, entre otros tipos de palabra, y conseguiremos, de este modo, filtrar las palabras, y dejar

solo aquellas con semántica, y la siguiente nos proporciona las palabras clave de cada tema a clasificar.

Además, de estas técnicas teóricas de la ciencia de la computación, para el desarrollo de nuestra herramienta vamos a usar correos electrónicos reales de la empresa Fetajo, S.A., cuya actividad empresarial se basa en el alquiler de vehículos en la zona de la Costa del Sol. Debido a que vamos a usar datos de clientes reales, es de vital importancia respetar la legislación vigente en cuanto a protección de datos. Por tanto, debemos cumplir en todo momento con lo que se estipula en la Ley Orgánica de Protección de Datos, ya que existe información sensible que puede ser utilizada en nuestro estudio y debe ser tratada de un modo correcto.

Algunos de los artículos más importantes que tratan el análisis de contextos formales son:

- M. Bogatyrev (2016) en *Conceptual Modeling with Formal Concept Analysis on Natural Language Texts*
- J. Poelmans, P. Elzinga, S. Viaene, G. Dedene (2010) en *Formal Concept Analysis in Knowledge Discovery: A Survey*
- B. Liu, L. Zhang (2012) *A Survey of Opinion Mining and Sentiment Analysis. In Mining Text Data*
- S.O. Kuznetsov (2004) *Machine Learning and Formal Concept Analysis*
- A. Onishchenko, O. Prokasheva, S. Gurov (2013) en *Classification methods based on formal concept analysis.*

En cuanto a Minería de Datos se refiere, esta técnica es estudiada en diferentes materias del grado referidas a la inteligencia artificial, lo cual hace que este término sea un concepto básico en el aprendizaje de la informática. Profundizaremos más adelante en cada una de estas herramientas y en las

diferentes operaciones que nos permiten realizar y que son útiles para nuestro desarrollo objetivo.

Como resultado final, esperamos obtener un proceso de respuesta automatizada sobre cancelaciones de servicios escritos en inglés o castellano, con el que después de clasificarlo, enviaremos una respuesta al cliente solicitando que nos envíen la información necesaria para que esta cancelación se lleve a cabo con éxito.

1.1. Metodología de trabajo

El estudio se llevará a cabo en 6 fases en las que trataremos de estimar un tiempo para cada una de ellas, de forma que la planificación y organización de este se simplifique. Este proyecto comienza con una fase de aproximación al tema con una lectura sobre retículos de conceptos y cómo usarlos para clasificar los distintos correos electrónicos, en la que estimaremos unas 66 horas que se basarán en la búsqueda y lectura de artículos y libros científicos.

Tras esta primera fase, pasaremos al preprocesamiento de los correos proporcionados por la empresa. Para comenzar nos dispondremos a anonimizar los correos para cumplir la ley de protección de datos. A continuación, procederemos con una clasificación manual tanto si son correos referidos a cancelaciones de reservas o no, y si están redactados en inglés o en castellano. Para esta fase estimamos que tiene una duración de unas 20 horas.

Una vez obtenemos este conjunto de correos electrónicos clasificado y anonimizado, el siguiente paso es la extracción de palabras clave, en el que nos ayudaremos del software *'RStudio'* y la librería *'tm'* de Minería de Texto. Para obtener las palabras clave de todos estos correos y la generación de un contexto para el retículo de concepto, que se estima el uso de unas 30 horas.

Con el contexto obtenido anteriormente podemos pasar a la generación del retículo de concepto, la reducción de palabras clave y el clasificador, con un tiempo estimado de unas 90 horas, que nos dirá si el correo esta redactado en inglés o castellano y si trata sobre una cancelación o no, ya que pasaremos por un proceso de aprendizaje con el software y ambos paquetes utilizados.

Una vez tenemos el correo recibido clasificado, queremos enviar una respuesta automática con los datos necesarios para realizar la cancelación. Estos campos suelen ser el número de reserva, número de cuenta (IBAN), número BIC/SWIFT y el nombre del cliente, al que está la reserva. La estimación de la redacción de la respuesta es de unas 2 horas, que, con las horas estimadas anteriormente, y las 88 horas estimadas para la redacción de esta memoria, suman un total de 296 horas estimadas.

2. Ley Orgánica de Protección de Datos

Con el avance de la tecnología, y a que los sistemas de información cada vez más recogen más información de los usuarios, es necesario una ley que defienda la privacidad y derecho del tratamiento de los datos de cada usuario. De esta necesidad nace la Ley Orgánica de Protección de Datos, cuyo nombre completo es Ley Orgánica de Protección de Datos y Garantía de Derechos Digitales (LOPDGDD). Esta ley entró en vigor el 6 de diciembre de 2018, sustituyendo a la antigua Ley Orgánica 15/1999 de Protección de Datos de Carácter Personal. Su objetico es adaptar la legislación española a la normativa europea, regida por el Reglamento General de Protección de Datos (RGPD), vigente desde el 25 de mayo de 2018.

La ley establece los requisitos y obligaciones de las empresas sobre cómo tratar la información personal, junto con los derechos de los consumidores. Como hemos comentado ya, su finalidad es la de proteger la intimidad, privacidad e integridad del individuo, cumpliendo con el artículo 18.4 de la Constitución Española. Asimismo, regula las obligaciones del individuo en todo proceso de transferencia de datos para garantizar la seguridad del intercambio. Ente los datos protegidos, se consideran datos personales aquellos que permitan identificar a una persona, como podrían ser nombre, correo electrónico, religión, salud personal, etc.

Otra de sus principales finalidades es establecer un marco legislativo para la protección de datos personales en Internet, incorporando puntos para tener en cuenta como el derecho al olvido o la portabilidad, además de cambios en la obtención del consentimiento para recoger y usar la información personal. Entre las principales modificaciones de la LOPDGDD respecto a su anterior versión de 1999 está la modificación de los requisitos para la obtención de la información,

guardarla o compartirla, y el establecimiento de cambios en el tratamiento de datos de usuarios en Internet.

La esencia de la ley es la de adaptar el ordenamiento español al RGPD, por ello, el modelo español ha tenido que añadir novedades importantes, como nuevas obligaciones sobre tratamiento de datos personales en procedimientos transfronterizos. Algunas de estas novedades son la posible rectificación o supresión de los datos de personas fallecidas, el dar consentimiento de los menores de 14 años a sus padres o tutores legales, el tratamiento de datos por obligación legal, interés público o ejercicio de poderes públicos, necesidad de más requisitos para poder tratar datos de ideología, afiliación sindical, religión como el cumplimiento de obligaciones y el ejercicio de derechos específicos del responsable del tratamiento o del interesado, proteger intereses vitales del interesado o de otra persona física, en el supuesto de que el interesado no esté capacitado física o jurídicamente para dar su consentimiento, entre los requisitos necesarios.

No sólo trata estas categorías, sino que, además, cuando su finalidad sea profesional, los datos de contactos de empresarios tienen como base jurídica el interés legitimo para poder tratar estos datos, la video vigilancia para garantizar la seguridad de las personas y bienes, así como de sus instalaciones. Además, es licito el tratamiento en las Administraciones Públicas de datos con fines de archivo basado en el interés público o las infracciones penales, llevadas a cabo por órganos competentes, abogados o procuradores. Asimismo, se añade una responsabilidad activa, un registro de actividades del tratamiento de datos, el bloqueo de los datos, la delegación de protección de datos y una normativa electoral. Entre las novedades también se encuentran los nuevos derechos digitales aprobados por el congreso como los derechos de neutralidad de Internet, el acceso universal al mismo, derecho a la seguridad y educación digital, protección de los menores en este medio, etc.

Al trabajar con datos sensibles, como en nuestro caso son los correos electrónicos reales que usamos en nuestra herramienta, es necesario tener en cuenta en todo momento la Ley Orgánica de Protección de Datos, ya que todo este tipo de información esta sujeta a dicha legislación. Por ello, para efectuar un análisis de datos correcto sobre los correos electrónicos e imprescindible anonimizar toda la información y eliminar aquellos datos sensibles que permiten identificar a la persona que remite el correo electrónico, para que no puedan ser usados de ninguna otra forma ni por ningún otro usuario.

Ejemplo de correo electrónico para explicar el proceso de anonimizar.

Thank you for the reply. I'm sure this is all a nightmare and we are so **** to have to cancel due to the 2 ****** quarantine. We were really looking ****** to visiting and using ****** again.**

I would prefer a refund as I won't be **** out until next summer. Rest assured I ****** be rebooking with ****** next summer yet again.**

The numbers you require are.

BIC ********

IBAN **********

****

Para cada campo en el que aparecen asteriscos hemos borrado la información referente al cliente, evitando así que el mismo pueda ser identificando.

3. Minería de Texto

Cada vez más el tratamiento de la información, teniendo en cuenta el punto anterior, da mayor facilidad a la toma de decisiones en las organizaciones. Por ello, ser capaces de entender lo que la gente expresa sobre un producto o saber de qué trata un texto es esencial en el mundo de los sistemas de información. Para conseguir esto, se hace uso de herramientas de análisis de datos, como pueden ser el análisis de sentimientos, la minería de datos, etc. En nuestro caso haremos uso de la herramienta de minería de texto, cuya herramienta es una rama específica de la minería de datos referida al proceso de analizar la información que aparece en un texto, mediante la identificación de patrones o correlaciones entre los términos.

Esta ofrece la posibilidad de realizar operaciones sobre cierto texto con el fin de extraer de él información, como palabras que pueden expresar sentimientos favorables o negativos por un producto o en nuestro caso, la obtención de palabras de un texto. En este caso, usaremos esta técnica para el cribado de palabras con el que poder formar un contexto al que aplicar el análisis de conceptos formales, con la que obtendremos las palabras clave.

Esta técnica requiere la consecución de diferentes pasos para completar el proceso deseado, cuyo primer escalón es la recolección de datos, que proceden de diferentes fuentes como pueden páginas web, libros, correos electrónicos, entre otras. Tras la obtención de estos datos, realizamos un preprocesamiento de estos, para agilizar su uso más adelante. En este paso se incluye la supresión de partes del texto que no son necesarias, como es el caso de la información sensible, por ejemplo, las direcciones de correo electrónico, los datos bancarios, etc. Esto es crucial ya que dichos datos apenas aportan información relevante o simplemente son palabras como conjunciones, determinantes, etc.

Otras etapas que cumplimentar son el enriquecimiento como añadir etiquetas a las palabras como pueden sustantivo, adjetivo, verbo, etc. Esto favorecerá la eficiencia de extracción de palabras importantes. Cuanto mejor sea nuestro conjunto de palabras extraídas, mejores decisiones seremos capaces de tomar con un mejor resultado. La ultima de las etapas sería la transformación de los datos, con el que seremos capaces de extrapolar información de los datos, como puede ser obtener el conjunto de palabras que mas se repiten en un conjunto de textos, sacar patrones en repetición de palabras en textos que hacen referencia a un tipo u otro. Esto se consigue con técnicas como la de obtener un vector de aparición de palabras, con el que más tarde podemos pasar a su cribado para extraer del mismo palabras clave, con las que somos capaces de diferenciar si un texto pertenece a una categoría u otra, etc. en la última de las etapas, la de extracción de características, como sería la clasificación de un texto si es de una temática u otra. Esta técnica se puede usar a la hora de extraer información de grandes cantidades de texto con la posibilidad de analizar los sentimientos por un determinado producto en caso de opiniones de producto o propias publicaciones en redes sociales, para crear un modelo predictivo o clasificador de textos con el que obtener palabras clave sobre textos, elaboración de resúmenes o simplemente resaltar información relevante de un texto.

Para esta tarea, hacemos uso de la librería tm – "*Text Mining*" que ofrece el entorno de desarrollo rStudio. Este paquete ofrece una serie de operaciones para llevar a cabo este proceso, el cual detallaremos más adelante, pero a modo de ejemplo, estas operaciones son la conversión de textos a minúsculas, extracción de palabras comunes en una serie de textos, etc.

4. Análisis de Conceptos Formales

Para llevar a cabo nuestro clasificador de correos electrónicos nos basaremos en la teoría matemática del Análisis de Conceptos Formales, que, además, es un método para el análisis de datos en cuanto a sus relaciones y estructura. Con su uso, el objetivo es que los datos se organicen, sin dejar de responder a la exigencia de rigor de un modelo matemático, se adapten mejor a la forma en que se organiza el pensamiento humano en relación con los conceptos y a su orden. Su base se asienta en la teoría de retículos y en la teoría matemática del orden. Por ello, al inicio del trabajo, vamos a adquirir conocimiento en el ámbito de los retículos de conceptos, los cuales nos ayudarán a construir nuestro clasificador.

El análisis de conceptos formal estudia las relaciones existentes en conjunto de datos y forma una estructura para los mismos. Definimos el análisis de conceptos formales a través de un conjunto M de atributos, un conjunto G de objetos u observaciones y una relación binaria $I \subseteq G$ x M tal que, con un objeto $g \in G$ y un atributo $m \in M$, obtenemos que la tupla $(g, m) \in I$ si y solo si el objeto g tiene el atributo m. El triple $K:= (G, M, I)$ recibe el nombre de contexto formal. Los operadores de derivación, teniendo que se definen para un subconjunto de objetos $A \subseteq G$ y un subconjunto de atributos $B \subseteq M$, son definidos por:

- $A' := \{m \in M \mid gIm \text{ para todo } g \in A\}$
- $B' := \{g \in G \mid gIm \text{ para todo } m \in B\}$

Podemos definir un concepto formal del contexto K que sea un par (A,B) que satisface las siguientes condiciones:

- $A \subseteq G$
- $B \subseteq M$
- $A' = B$

- $B' = A$

El conjunto A recibe el nombre de extensión y B el de intensión del concepto (A, B). Los conceptos formales pueden ordenarse a través de la siguiente relación de orden $(A_1, B_1) \geq (A_2, B_2) \Leftrightarrow A_1 \supseteq A_2)$, formando así. una red completa, denominada la red conceptual del contexto K. Otra de las definiciones a las que llegamos con esta investigación es al uso de la Hipótesis JSM la que se define a continuación. Además de los atributos del conjunto M, consideramos un atributo objetivo w $\notin$ M, junto con la partición del conjunto G de todos los objetos en tres subconjuntos, definidos como:

- G_+, que hace referencia al subconjunto de objetos de G que tienen la propiedad w (ejemplos positivos).
- G_-, que hace referencia al subconjunto de objetos de G que no tienen la propiedad w (ejemplos negativos).
- G_τ, que hace referencia al subconjunto de objetos de G, los cuales se desconoce si tienen o no la propiedad w.

A partir de estos tres subconjuntos del conjunto G, podemos definir 3 subcontextos del contexto K distintos:

- $K_+ := (G_+, M, I_+)$
- $K_- := (G_-, M, I_-)$
- $K_\tau := (G_\tau, M, I_\tau)$

donde, para $\varepsilon \in \{+, -, \tau\}$ tenemos $K_\varepsilon = I \cap (G_\varepsilon \times M)$ y los correspondientes operadores de derivación denotados por $(\cdot)^+$, $(\cdot)^-$, $(\cdot)^\tau$, respectivamente.

Una vez tenemos esto, la intensión del contexto formal es el conjunto de atributos compartidos por algunos de los objetos observados. Con el fin de formar una hipótesis sobre causas estructurales del atributo objetivo w, estamos interesados en obtener un conjunto de atributos que sean comunes a algunos ejemplos positivos, pero no a ejemplos negativos. De esta forma, una hipótesis

positiva h para w (llamada "hipótesis prohibitiva de contraejemplo" en el método JSM) como intensión de K_+, tal que $\mathrm{h}^+ \neq \emptyset$ y $h \subseteq g^+ \coloneqq \{m \in M \,|(g,m) \in I_-$ para ningún ejemplo negativo de un objeto $g \in G_-$

La intensión de K_+que esté contenido en la intensión de un ejemplo negativo se le llamara un **(+)-generalización falsificada**. De forma similar son definidas las hipótesis negativas. Estas hipótesis pueden usarse para clasificar estos ejemplos indefinidos. Si la intensión $g^t \coloneqq \{m \in M \,|(g,m) \in I_\tau\}$ de un objeto g $\in G_\tau$ contiene una hipótesis positiva, pero no negativa, entonces g^tes clasificada positivamente. Las clasificaciones negativas se definen de forma similar a estas. Si g^t contiene hipótesis de ambos tipos, o si g^tno contiene hipótesis, entonces la clasificación es contradictoria o indeterminada, respectivamente. En ese caso, podemos aplicar técnicas estándar de probabilidad conocidas en machine learning o data mining (voto mayoritario, aproximación bayesiana, etc.).

Del articulo [4], sabemos que de los estudios [6,7] realizados en el mismo articulo, se puede restringir a la hipótesis mínima (inclusión w.r.t. $\subseteq$), tanto positiva como negativa, ya que la intensión de un objeto contiene una hipótesis positiva si y solo si contiene una hipótesis mínima positiva. Del mismo artículo [4] de Kuznetsov (2004), extraemos el ejemplo 1.

G/M	Color	Forma	Firmeza	Suave	Objetivo
Manzana	amarillo	redondo	no	sí	+
Pomelo	amarillo	redondo	no	no	+
Kiwi	verde	ovalado	no	no	+
Ciruela	azul	ovalado	no	sí	+
Cubo Rubik	verde	cubico	sí	sí	−
Huevo	blanco	ovalado	sí	sí	−
Pelota de Tenis	blanco	redondo	no	no	−

Tabla 1. Tabla de ejemplo resuelto.

Este conjunto de datos o contexto multivaluado puede ser reducido a un contexto de la forma presentada anteriormente escalando, p.ej. como sigue:

G/M	b	a	v	z	f	¬f	s	¬s	r	o	¬o	Target
Manzana		×				×	×		×	×		+
Pomelo		×				×		×	×	×		+
Kiwi			×			×		×		×		+
Ciruela				×		×	×			×		+
Cubo Rubik			×		×		×				×	−
Huevo	×				×		×			×		−
Pelota de Tenis	×					×		×	×			−

Tabla 2. Tabla contexto reducido por escalado.

Hacemos uso de abreviaciones en esta tabla como 'b' para blanco, 'a' para amarillo, 'v' para verde, 'z' para azul, 's' para suavidad, 'f' para firmeza, 'r' para redondo, 'o' para ovalado y '¬ m' para $m \in \{b, a, v, z, f, s, r, o\}$. Una vez tenemos el conjunto de objetos G, el conjunto de atributos M y el contexto I que establece la relación ente objetos y atributos, estamos listos para generar nuestro pequeño contexto formal, el cuál da como resultado el siguiente retículo de conceptos de la figura 2.

Con el retículo de conceptos obtenido, hemos estudiado las distintas relaciones existentes en el conjunto de datos, junto con su estructura. Los objetos con base a sus características o atributos se organizan en grupos que coinciden en cuanto a esas características. Estos grupos se vuelven a subdividir, con base en otras características, de lo que resulta una estructura jerárquica que podemos

ilustrar por medio de un diagrama de orden como el de la figura 1. Cada uno de los grupos de objetos determinados por sus características comunes se define como una extensión del concepto, y el conjunto correspondiente de todas las características comunes como la intensión de este. Ambas partes en un conjunto, es decir, respectivamente, cada extensión con su correspondiente intensión, conforman un concepto formal. El tipo de orden de los conceptos formales se manifiesta como una estructura ordenada en forma de malla con ramificaciones. Se puede demostrar que estos órdenes poseen características especiales y bien estudiadas.

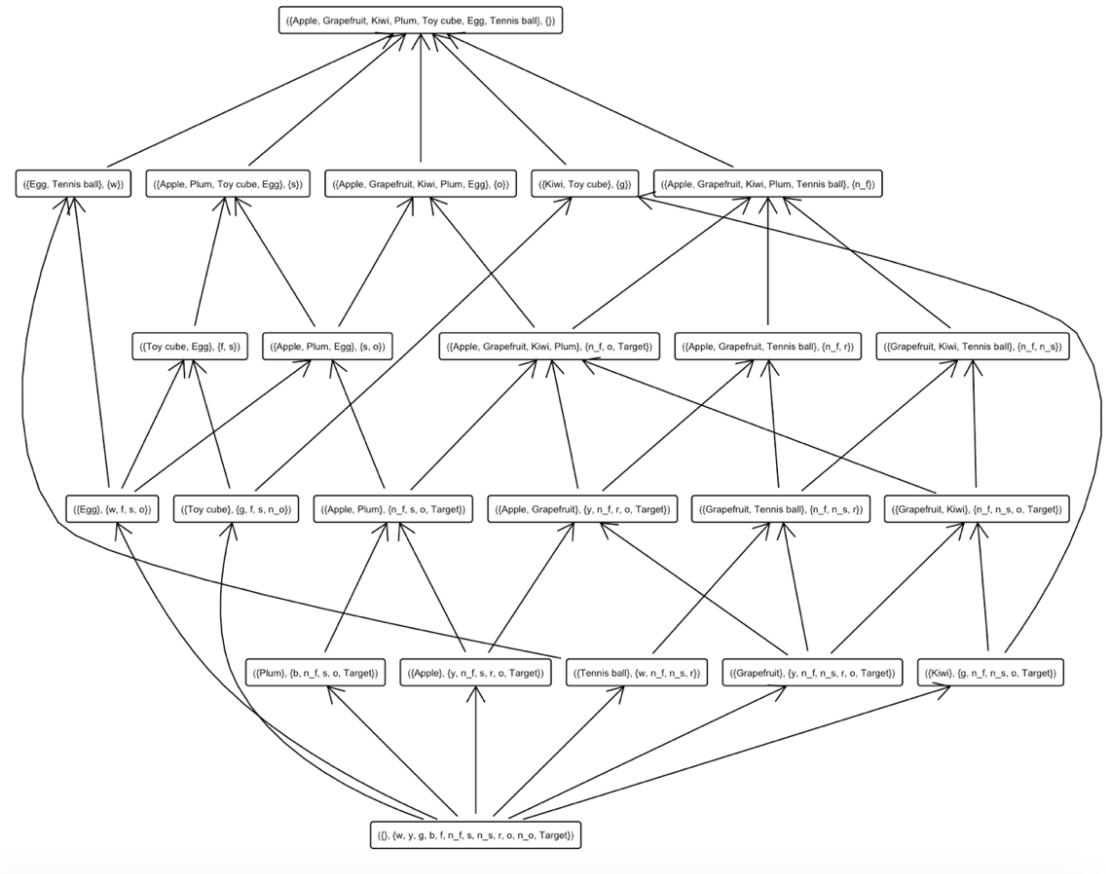


Figura 2. Red de conceptos relacionada con el ejemplo 1

5. Análisis Estadístico de Testeos

Conocer o evaluar nuestro sistema clasificador, es esencial para determinar la eficacia de este, por eso haremos uso de la estadística, en concreto de la probabilidad. Se define la incertidumbre como la falta de certeza o seguridad al completo de asegurar un hecho. La causa de la existencia de esta incertidumbre es la variabilidad asociada a los resultados. Por ello, necesitamos basarnos en esta teoría para determinar cuanto de probable es que un correo recibido se clasifique como es debido.

Para analizar la eficiencia del clasificador aplicaremos la teoría de Pruebas Diagnósticas, con la que obtendremos las probabilidades de que los correos sean de un idioma u otro, o si es de un tipo u otro. Necesitamos la definición de diferentes elementos necesarios para ver como de efectivo y eficiente es el clasificador. [10]

Definiciones

Experiencia Aleatoria. Se denomina experiencia aleatoria (E) a una prueba o fenómeno que presenta variación en sus resultados, y que antes de su realización no es posible predecir con seguridad cuál será el resultado particular.

Espacio Muestral. Un espacio muestral se define como el conjunto o lista de todos los resultados posibles que pueden ser observados al realizar una experiencia aleatoria definida previamente.

Suceso. Un suceso es el subconjunto de resultados del espacio muestral S. A su vez, definen espacio muestral como el conjunto o lista de todos los resultados posibles que pueden ser observados al realizar una experiencia aleatoria definida previamente.

Probabilidad. La probabilidad es una función que mide la expectativa de que ocurra un suceso asignando un valor entre cero y uno.

Con estas definiciones podemos pasar a determinar el valor de las probabilidades que corresponden a un suceso A cualquiera, pudiendo realizarlo de dos posibles formas:

- "a priori" por la regla de Laplace, que será la que usemos, en la que la probabilidad de que se de el suceso A es igual al número de casos favorables partido el número de casos posibles.

$$P(A) = (nº\ casos\ favorables)/(nº\ casos\ posibles)$$

- "a posteriori" por la ley de los grandes números, que es una definición empírica de probabilidad: en una serie larga de tiradas o repeticiones de una experiencia, la frecuencia relativa (fr) observada de un suceso se aproxima a su probabilidad:

$$fr(A) = (nº\ de\ veces\ que\ ocurre\ A)/(nº\ de\ repeticiones\ de\ la\ experiencia)$$

Prevalencia. Es una medida de los casos existentes de una condición en una población. Lo más usual es calcular la prevalencia puntual que se define como la probabilidad de que un elemento de la población sea un caso en un momento de tiempo dado:

$$Prevalencia = (nº\ de\ casos\ observados\ en\ el\ tiempo\ t) /(tamaño\ de\ la\ población\ en\ el\ tiempo\ t) = P(E)$$

Prueba Diagnóstica. Cualquier tecnología que pueda servirnos para detectar un síntoma o patrón que se relacione con el caso de interés.

El objetivo final del diagnóstico es clasificar un elemento de la población en uno de los posibles casos. Sin embargo, las pruebas no son perfectas, dando lugar a las definiciones de falso positivo (caso diagnosticado como positivo, siendo en realidad negativo) o falso negativo (caso diagnosticado como negativo, siendo este positivo).

La fiabilidad de una prueba, probabilidad de que funcione bien, se cuantifica por las proporciones de los resultados acertados:

- SENSIBILIDAD: verdaderos positivos en el conjunto de pruebas positivas.
- ESPECIFIDAD: verdaderos negativos en el conjunto de pruebas negativas.

Cuanto mayor sea la sensibilidad de nuestro sistema, mayor será la eficiencia del clasificador, al igual que ocurre con la especificidad.

El Teorema de la probabilidad total nos permite calcular la probabilidad de un suceso a partir de probabilidades condicionadas, cuya fórmula es:

$$P(B) = \sum_{i=0}^{n} A_i \cdot P(B/A\_i\,)$$

Aplicando el teorema de la probabilidad total a nuestro caso quedaría la siguiente expresión:

$$P(Clasifique\ Bien) = P((Clasifique\ positivo)/(Sea\ positivo)) \cdot P(Positivo) + P((Clasifique\ negativo)/(Sea\ negativo)) \cdot P(Negativo)$$

Es decir, la probabilidad de que la herramienta clasifique bien es igual a la suma de la posibilidad de que clasifique bien un positivo multiplicada por la probabilidad de que sea positivo (sensibilidad), junto con la probabilidad de que clasifique bien un negativo multiplicada por la probabilidad de que sea negativo (especifidad). Estos sucesos forman un sistema completo, ya que la suma de las posibilidades y sus complementos suman el 100%, es decir, la unión de los sucesos

forma el espacio muestral, requisito necesario para aplicar el teorema. Una vez tenemos como calcular la eficiencia del clasificador, podemos pasar a su desarrollo.

6. Clasificador de Correos Electrónicos

6.1. Descripción del Problema

Como hablamos en la introducción, es un problema habitual que una empresa reciba una avalancha de correos electrónicos y que esta no sea capaz de dar respuesta a todos en un tiempo prudencial ya que en la mayoría de las empresas se realiza una respuesta individualizada para cada uno de ellos a mano. Esto puede dar lugar a que queden correos en el olvido y sin contestar, generando malestar entre los consumidores.

Casi siempre el contenido de los correos electrónicos suele preguntar dudas que pueden resolverse en un apartado en la propia web de la empresa como preguntas frecuentes, pero los clientes prefieren escribir un correo electrónico, pues les darán la información que quieren sin apenas tener que buscarla. Por ello, disponer de una herramienta que clasifique y responda a los correos automáticamente es de gran importancia para la gestión de empresas.

Por problemas de tiempo, nuestro trabajo se centrará en obtener un sistema que al recibir un correo electrónico genere una respuesta automática para correos que tratan sobre la cancelación de un servicio. Contamos con un conjunto de emails, el cual recibiremos sin procesar, y a partir de él, generaremos un sistema que nos permita dar respuesta a nuestros clientes.

6.2. Implementación

Para el desarrollo de nuestro sistema haremos uso del entorno de desarrollo rStudio, el cual ofrece un entorno con 4 vistas esenciales que son en las que trabajaremos. Cuenta con una primera ventana y principal donde desarrollaremos la mayoría del estudio, la ventana de generación de scripts.

Para nuestro estudio dividiremos el trabajo en distintos ficheros, como son:

- Un script principal donde se desarrolla el análisis y clasificación del email recibido.
- Un segundo script el cual generará nuestra estructura de Contexto Formal
- Y los dos últimos los cuales generaran una proporción de cuantas palabras de cada tipo aparecen, según el idioma, por eso 2 ficheros. Uno en para correos electrónicos escritos en inglés y otro para los escritos en castellano

Junto con la herramienta, haremos uso de los paquetes de r como son "fcaR", el cual nos ayudará a generar el contexto, y, "tm", la librería que nos ofrecerá métodos para analizar cada uno de los correos recibidos tanto para aprender como para clasificar.

Comenzando por la librería tm, nos ofrece métodos como la posibilidad de pasar todo el texto a minúsculas, eliminar puntuaciones, patrones como tabulaciones, saltos de línea, etc. Junto con la posibilidad de eliminar palabra sin significado como a, algo, ante, bastante, conmigo, contigo, etc. También, ofrece métodos para extraer la frecuencia de las palabras, el cual usaremos para determinar los porcentajes de aparición.

```
> c(stopwords("english"))
  [1] "i"          "me"         "my"         "myself"
  [5] "we"         "our"        "ours"       "ourselves"
  [9] "you"        "your"       "yours"      "yourself"
 [13] "yourselves" "he"         "him"        "his"
 [17] "himself"    "she"        "her"        "hers"
 [21] "herself"    "it"         "its"        "itself"
 [25] "they"       "them"       "their"      "theirs"
 [29] "themselves" "what"       "which"      "who"
 [33] "whom"       "this"       "that"       "these"
 [37] "those"      "am"         "is"         "are"
 [41] "was"        "were"       "be"         "been"
 [45] "being"      "have"       "has"        "had"
 [49] "having"     "do"         "does"       "did"
 [53] "doing"      "would"      "should"     "could"
 [57] "ought"      "i'm"        "you're"     "he's"
 [61] "she's"      "it's"       "we're"      "they're"
 [65] "i've"       "you've"     "we've"      "they've"
 [69] "i'd"        "you'd"      "he'd"       "she'd"
 [73] "we'd"       "they'd"     "i'll"       "you'll"
```

Imagen 1. Palabras sin significado en inglés.

Hablando sobre la librería "fcaR", esta librería nos aporta los métodos necesarios para la generación de un retículo de conceptos formal con la que obtener nuestras palabras clave a partir de un contexto como el de la **Tabla 2** y encontrar las implicaciones entre objetos y atributos y así obtener nuestra red de conceptos. A la hora de aplicar en análisis de conceptos formales a la clasificación la extensión del conjunto de objetos que queremos clasificar (es decir los que tienen la característica objetivo) en nuestro caso, los correos electrónicos, obtendremos los atributos comunes a estos objetos, es decir, obtendremos un conjunto de características de las que podemos decir que si un objeto las posee, entonces tiene también el atributo objetivo.

Antes de comenzar a desarrollar la herramienta, debíamos realizar un preprocesamiento previo del conjunto de entrenamiento que recibimos de los correos electrónicos, eliminando toda clase de información sensible de clientes como nombres, apellidos, números de teléfono, números de reserva, fechas de obtención del permiso de conducir, fechas de nacimiento, números de vuelo, países o ciudades de origen, etc. Además de palabras que no aportaban conocimiento al sistema. En esta tarea observamos que dos de los emails estaban redactados en francés, idioma que queda fuera del estudio, por lo tanto, esos emails no nos servían para poder aportarle conocimiento a la herramienta, por ello quedan eliminados. Además, separamos el conjunto de correos electrónicos en dos subconjuntos, para usar uno como conjunto de pruebas para testear la eficiencia de la herramienta de 300 correos, y el segundo de ellos como conjunto de entrenamiento con el resto.

Tras esta primera vuelta aplicamos un primer intento de minería de texto para obtener un conjunto de palabras clave, pero este no nos ayudaría a conseguir clasificar los correos electrónicos que recibamos en un futuro. Al ser conscientes de este problema, volvimos a limpiar cada uno de los correos de palabras que no

aportan conocimiento al sistema, como palabras de saludo o despedida, verbos, adjetivos, sustantivos, etc. Con esta segunda limpieza, conseguimos un conjunto de correos de prueba apto con el que poder empezar a trabajar en las siguientes fases.

Como mencionamos anteriormente, realizamos un primer intento en el que extraíamos un conjunto de palabras clave bastante pobre, ya que palabras como "confirm", que debería estar asociada como palabra clave de correos tipo "no-cancelación", era palabra clave asociada a correos de cancelaciones. En los primeros intentos por obtener un conjunto de palabras clave óptimo para darle conocimiento a nuestro sistema, no conseguimos clasificar ni el idioma del correo ni el tipo de este. Por lo que continuamos trabajando en este conjunto, obteniendo el siguiente conjunto de palabras clave en con su respectivo número de apariciones en todos ellos.

	percent_cancel	percent_no_cancel	percent_word_appearance
address	12.50	87.50	4.12
advanc	53.33	46.67	2.58
advis	62.50	37.50	2.75
afternoon	21.05	78.95	3.26
ago	20.00	80.00	2.58
airport	15.62	84.38	5.50
alreadi	35.71	64.29	2.41
amount	48.00	52.00	4.30
appreci	25.81	74.19	5.33
arrang	28.57	71.43	2.41
arriv	5.88	94.12	2.92
ask	46.67	53.33	2.58
attach	21.05	78.95	3.26
beca	44.44	55.56	4.64
better	6.67	93.33	2.58
bic	15.38	84.62	4.47
book	55.74	44.26	40.38
busi	12.50	87.50	2.75
call	21.05	78.95	3.26
cancel	76.81	23.19	35.57
chang	23.53	76.47	5.84
check	6.25	93.75	2.75
collect	30.77	69.23	2.23
come	21.43	78.57	2.41
confirm	31.33	68.67	14.26
contact	33.33	66.67	3.09

Tabla 3. Porcentajes de aparición de cada palabra según el tipo de correo electrónico y porcentaje de aparición total.

Con este último conjunto de palabras que más se repiten en los correos electrónicos, pasamos a generar un contexto I con el que poder generar nuestro Retículo de Conceptos Formal y clasificar nuestros correos en si son cancelaciones o son correos de otro tipo, además de clasificar el idioma de estos entre inglés y castellano. Para ello comenzamos con la detección del idioma, obteniendo una serie de palabras clave para cada idioma, con el que compararemos el porcentaje de palabras clave en inglés y el de palabras clave en castellano, y a partir de ese porcentaje clasificaremos el correo en un idioma.

Antes de sacar las palabras clave se realizan una serie de operaciones a los correos electrónicos escritos en inglés que recibimos en el conjunto de entrenamiento, entre las cuáles se encuentran operaciones como la conversión del texto a minúsculas, la eliminación de signos de puntuación y elementos escritos en código ASCII como es el caso de "\u2028", etc. Tras ello, cogemos las palabras que más se repiten y las añadimos a nuestro conjunto de palabras más repetidas en inglés; y a continuación repetiríamos el proceso con el conjunto de emails en castellano.

Para probar la eficacia del clasificador del idioma, cogimos un conjunto de emails de prueba, en el que para cada correo le aplicábamos las mismas operaciones anteriormente comentadas y a partir de ahí, con las palabras clave del propio correo electrónico, comprobamos cuantas palabras aparecían en cada conjunto dando como valores:

- *probabilidad_castellano*=$(nº\ palabras\ clave\ que\ aparecen\ en\ el\ correo)/(nº\ total\ de\ palabras\ clave\ del\ conjunto\ castellano)$, *y*
- *probabilidad_ingles*=$(nº\ palabras\ clave\ que\ aparecen\ en\ el\ correo)/(nº\ total\ de\ palabras\ clave\ del\ conjunto\ inglés)$.

Con la comparación de ambos valores obtenemos el idioma, en el que obtenemos las siguientes estadísticas del clasificador del idioma:

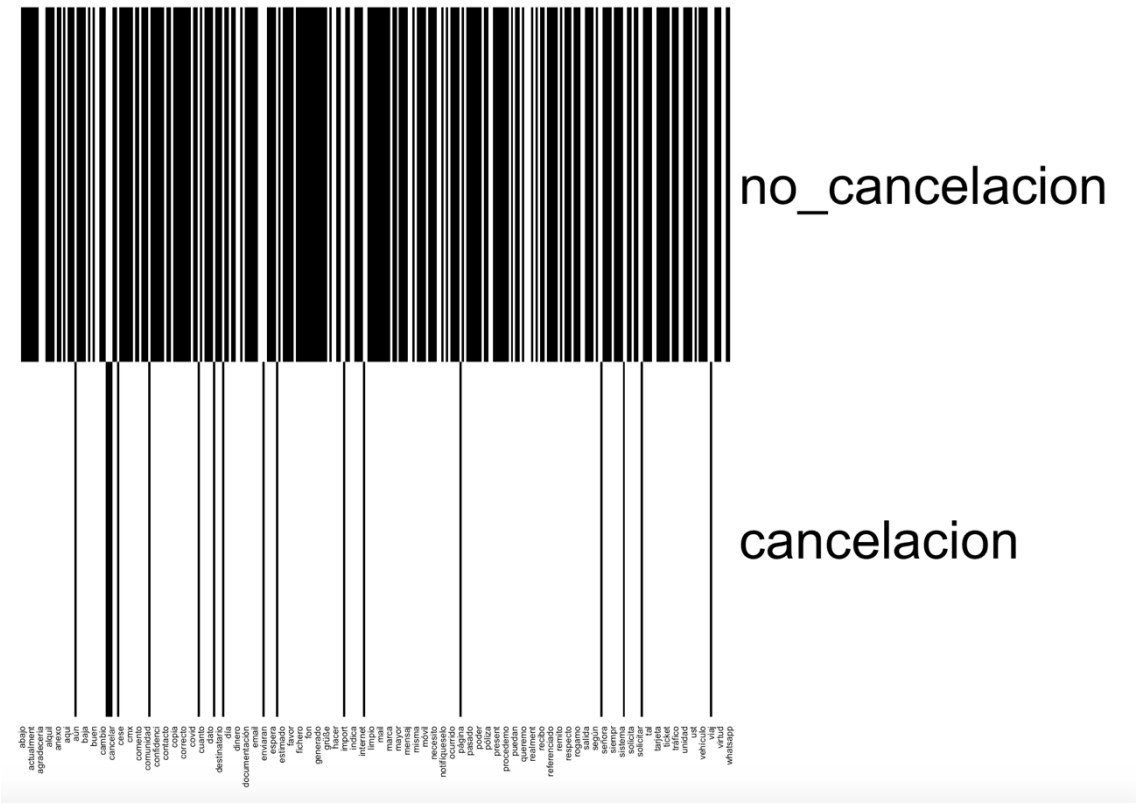
- % de correos clasificados como “castellano” y “bien clasificados”: 100%
- % de correos clasificados como “castellano” y “mal clasificados”: 0%
- % de correos clasificados como “ingles” y “bien clasificados”: 99,21%
- % de correos clasificados como “ingles” y “mal clasificados”: 0,79%
- % de correos clasificados como “indeterminado”: 22%

Este último conjunto de correos que no conseguimos clasificar en ningún idioma pasará a ser correos como no cancelación, para que pasen a tener una respuesta manual. Como vemos la posibilidad de estar bien clasificado a través del teorema de la probabilidad total, prácticamente es de valor 1, lo que muestra como la clasificación del idioma es prácticamente perfecta.

Con el conjunto de palabras clave y el de correos, formamos una matriz en la que si la casilla m_ij contiene un 1, decimos que el correo i contiene la palabra j, y si contiene un 0 significa lo contrario.

	address	advanc	advis	afternoon	ago	airport	alreadi
cancelacion	0	0	0	0	0	0	0
no_cancelacion	1	0	0	1	1	1	0

Tabla 5. Matriz de contexto



Gráfica 1. Conjunto de palabras clave asociadas a cada tipo. (Contexto Formal)

Continuaremos con la generación de nuestro contexto formal y red de conceptos. Para ello haremos uso de la matriz de contexto que hemos generado, que relaciona nuestro conjunto de objetos G de los correos electrónicos con el conjunto de atributos M de palabras clave. Con este contexto formal sacaremos qué palabras clave tienen en común los correos que son de tipo "cancelación", y también el idioma en el que están escritos estos (castellano o ingles). Por lo que nuestro conjunto de atributos objetivo estará compuesto por las variables Ingles y Cancelación, las cuales no forman parte de conjunto de atributos del concepto formal.

Pasamos entonces a la definición formal del contexto formal, siendo G el conjunto de correos electrónicos, M el conjunto de atributos e I una relación binaria G x M. Con estas consideraciones y la ayuda de la librería "*fcaR*" de "*rStudio*", podemos buscar los conceptos que definen este contexto, junto con sus

implicaciones, lo que nos da como resultado una red de conceptos muy grande. Tanto es así, que su número de conceptos es de 2000 y el numero de implicaciones entre estos es de 3000.

Con estos conceptos e implicaciones, cogeremos el correo recibido y junto con la librería mencionada en el apartado 2.3, extraeremos de él sus palabras, y nos quedaremos con aquellas que estén en nuestro conjunto de atributos o palabras clave. Si este correo contiene las palabras que obtenemos como comunes a los correos del tipo "cancelación" podremos clasificar este como un correo de este tipo, además de asignarle el idioma en el que está redactado, necesario para decidir en qué idioma responder al correo.

Con la clasificación final obtenemos las siguientes estadísticas recopiladas en la siguiente tabla:

- % de correos clasificados como "cancelación" y "bien clasificados": 66,67%
- % de correos clasificados como "cancelación" y "mal clasificados": 33,33%
- % de correos clasificados como "no_cancelación" y "bien clasificados": 85,14%
- % de correos clasificados como "no_cancelación" y "mal clasificados": 14,86%
- % de correos clasificados como "indeterminado": 0%

Al aplicar el teorema de probabilidad total ya comentado anteriormente, obtenemos que la probilidad de estar bien clasificado es de un 81,4%, siendo este un porcentaje de clasificación bastante alto.

$$P(Clasifique\ Bien) = P\left(\frac{Clasifique\ positivo}{Sea\ positivo}\right) \cdot P(Positivo) + P\left(\frac{Clasifique\ negativo}{Sea\ negativo}\right) \cdot P(Negativo) = 0.6667 \cdot 0.2024 + 0.8514 \cdot 0.7976 = 0.1349 + 0.6791 = 0.814$$

Es interesante comparar estos resultados con los resultados que obtendríamos usando sólo el conjunto de emails utilizados para el entrenamiento, que son los siguientes. Si tenemos en cuenta que, si se usa el mismo conjunto de datos para el aprendizaje y el testeo, los resultados pueden mejorarse considerablemente. A continuación, mostramos los resultados obtenidos al considerar todo el conjunto de emails como conjunto de aprendizaje y testeo:

- % de correos clasificados como "cancelación" y "bien clasificados": 88,89%
- % de correos clasificados como "cancelación" y "mal clasificados": 11,11%
- % de correos clasificados como "no_cancelación" y "bien clasificados": 93,33%
- % de correos clasificados como "no_cancelación" y "mal clasificados": 6,67%
- % de correos clasificados como "indeterminado": 0%

$$P(Clasifique\ Bien) = P\left(\frac{Clasifique\ positivo}{Sea\ positivo}\right) \cdot P(Positivo) + P\left(\frac{Clasifique\ negativo}{Sea\ negativo}\right) \cdot P(Negativo) = 0.8889 \cdot 0.2024 + 0.9333 \cdot 0.7976 = 0.1799 + 0.7444 = 0.9243$$

Como podemos observar, los datos mejoran considerablemente, ya que el porcentaje de emails bien clasificados sube al 92,43%. Sin embargo, es importante tener en cuenta que estos datos no son realistas, ya que el sistema de clasificación puede haber tomado características propias del conjunto de aprendizaje que no son generales y que por lo tanto no son detectadas al realizar el testeo de un sistema de clasificación con el mismo conjunto de aprendizaje, lo que da lugar a resultados engañosos.

Una vez tenemos el tipo de correo y su idioma, finalmente pasamos a redactar una respuesta común que se repite cada vez que se recibe un correo de

cancelación. Finalmente, una vez tenemos el correo clasificado, tenemos que responder a lo que el cliente reclama, una cancelación.

En el caso de recibir una cancelación, la empresa siempre pide una información concreta para llevar a cabo la cancelación de la reserva. En el correo le pediremos al cliente que nos envíe esta información, redactando uno en castellano y otro en inglés. En el caso de la empresa colaboradora, la información que debemos requerir en ambos correos es la siguiente:

- Numero de reserva
- Nombre titular de la reserva
- Número de cuenta para realizar el reembolso de su reserva, con el IBAN
- Número BIC/SWIFT

Los correos electrónicos que enviaremos estarán redactados en castellano e inglés y serán los siguientes:

- Correo electrónico en castellano

Estimado cliente,

Nuestro sistema ha detectado que desea realizar una cancelación de su reserva. Si no fuera así, responda directamente a este correo y uno de nuestros agentes se pondrá en contacto con usted.
Para llevar a cabo la cancelación de su reserva necesitamos cierta información:

- *Número de reserva*
- *Nombre completo del titular de la reserva*
- *Número de cuenta junto con su IBAN y su código BIC/SWIFT.*

Esperamos verle usando nuestro servicio en un futuro.
Un saludo,
Fetajo.

- Correo electrónico en inglés

Dear Costumer

The system has detected that you are interested on canceling one reservation. If this is not the case, please reply this email and our staff will contact you with your request.

In order to proceed with the cancelation procedure and provide you a total refund, we need some information about your reservation:

- reference number of reservation;
- the full name of the person who made the reservation;
- the IBAN code and BIC/SWIFT code of the bank account where we have to send the money back.

We are looking forward to hearing from you.

Best regards
Fetajo

Una vez tenemos estos correos podemos pasar a implementar nuestro sistema en la herramienta de servicio de mensajería, que por problemas de tiempo y el retraso a la hora de recibir los correos electrónicos, no hemos podido añadirlo

7. Conclusiones y Líneas Futuras

Tras el desarrollo de nuestro estudio podemos concluir que hemos creado una herramienta eficaz y eficiente que supone una mejora cualitativa en beneficio de la empresa que lo implemente dentro de su organización. Como hemos podido comprobar, la clasificación de correos electrónicos con el clasificador desarrollado es eficiente. Este clasificador puede ser útil para cualquier empresa con el objetivo de ahorrar tiempo en sus trabajadores, el cual podría ser usado para responder a los correos que necesitan de una respuesta manual.

De esta forma, se mejora la imagen de la empresa puesto que los clientes se sentirán valorados al recibir una respuesta rápida en el caso de las respuestas automatizadas, así como igualmente rápida en el caso de respuestas más individualizadas que han debido ser redactadas a mano por algún trabajador. En este último caso, el trabajador habrá buscado la mejor respuesta y solución a la pregunta o problema que el cliente haya planteado, por lo que de este modo el cliente se sentirá satisfecho con el trato recibido y seguirá confiando en la empresa en futuras ocasiones. Los clientes que reciben la respuesta automática también se sentirán informados de una forma adecuada y rápida, de manera más eficiente que si un trabajador tuviera que redactar uno por uno todos los correos de respuesta a preguntas frecuentes y reiterativas.

Además, nuestros trabajadores se sentirán más realizados en el sentido de que tendrán que abordar temas y problemas de mayor envergadura que la simple redacción de correos estándar. Debido a ello, deberán buscar soluciones de aspectos que realmente necesiten un trabajo interno dentro de la empresa para resolver las inquietudes de nuestros clientes y ofrecerles la mejor solución a sus problemas. Los trabajadores intentarán dar una respuesta rápida y solventar los problemas de los clientes sin tener que discernir si el correo que han recibido es

de importancia o es un correo que se puede responder de manera automática, ya que este filtro estará pasado con nuestro clasificador, y directamente llegan a nuestros trabajadores aquellos correos electrónicos que si necesitan una atención individualizada y personalizada. Esto incrementa la motivación de los trabajadores ya que sentirán que su trabajo es de una mayor responsabilidad y relevancia, y se involucrarán en mayor medida en los objetivos de la empresa.

Como sabemos la relevancia de este tipo de trabajos, hace posible que empresas de cualquier sector puedan ofrecer un servicio posventa o de servicio al cliente que hace que se mejoren las relaciones con el cliente, un elemento crucial para cualquier negocio hoy día.

Por eso, avanzar con la mejora en el medio de comunicación de los correos electrónicos y de la relación con los clientes podría mejorar este sistema añadiendo más tipos de correos electrónicos, para que el sistema les genere una respuesta automática. Como por ejemplo en el caso tratado, conseguir un clasificador de correos electrónicos que aparte de clasificar si trata sobre cancelaciones, que también identifique sobre precios de los vehículos, preguntas sobre retrasos en la hora de recogida o de entrega de los vehículos, etc.

También se podría continuar con la investigación sobre como implementar la herramienta en un sistema de mensajería como Outlook o Gmail, que, por temas de gestión y tiempo, no se han podido llevar a cabo en este trabajo.

8. Bibliografía

[1] M. Bogatyrev. *Conceptual Modeling with Formal Concept Analysis on Natural Language Texts*. CIn Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016). 2016

[2] J. Poelmans, P. Elzinga, S. Viaene, G. Dedene. *Formal Concept Analysis in Knowledge Discovery: A Survey*. Lecture Notes in Computer Science, vol 6208. 139-153. 2010

[3] B. Liu, L. Zhang. *A Survey of Opinion Mining and Sentiment Analysis. In Mining Text Data*. Springer, Boston, MA 415-463. 2012

[4] S.O. Kuznetsov. *Machine Learning and Formal Concept Analysis*. In Lecture Notes in Computer Science, vol 2961. 287-312. 2004

[5] A. Onishchenko, O. Prokasheva, S. Gurov. *Classification methods based on formal concept analysis*. In Proceedings of the Workshop Formal Concept Analysis Meets Information Retrieval (FCAIR 2013), páginas 95–104, 2013.

[6] M. Botta, A. Giordana, L. Saitta, and M.Sebag, Relational Learning as Search in a Critical Refion, *Journal of Machine Learning Research,* 2003, **4**, 431-463

[7] A.P. Budunova, V.V. Poroikov, V.G. Blinova, and V.K. Finn, The JSM-method of hypothesis generation: Application for the analysis of the relation "Structure – hepatoprotective detoxifying activity", Nauchno-Tekhnicheskaya Informatsiya, no. 7, pp.12-15, 1993 (En ruso)

[8] Protección de Datos: https://protecciondatos-lopd.com/empresas/nueva-ley-proteccion-datos-2018/

[9] Text mining: https://www.cambridgeassessment.org.uk/Images/466185-text-mining-an-introduction-to-theory-and-some-applications-.pdf

[10] Estadistica: C. Gómez González y J.F. Pérez Castán *Capítulo 8: Pruebas diagnósticas. Concordancia,* CURSO DE INTRODUCCIÓN A LA INVESTIGACIÓN CLÍNICA, SEMERGEN. 2007;33(10):509-19

[11] https://es.wikipedia.org/wiki/An%C3%A1lisis_formal_de_conceptos

[12] Tm library package for rStudio: https://cran.r-project.org/web/packages/tm/index.html

[13] Fca library package for rStudio: https://cran.r-project.org/web/packages/fcaR/

Anexo

Estadística.R

```
# Cargamos la ruta de los ficheros
path = dirname(rstudioapi::getSourceEditorContext()$path)

# Cargamos los emails recibidos y los dividimos por idioma
emails <- read.csv(file = gsub(" ", "", paste(path, "/emailsTFG.csv")), head = TRUE, sep=";")

# numero de elementos sobre el que realizaremos pruebas para comprobar la
# eficiencia del clasificador
primeros_x = 300

emails_prueba = head(emails,primeros_x)

library(dplyr)
emails_entrenamiento = anti_join(emails,emails_prueba)

# Cargamos los contextos formales para cada idioma, que nos ayudaran a
# determinar el idioma
conjunto_entrenamiento = subset(emails_entrenamiento,  emails_entrenamiento$ingles == 1)

source(gsub(" ", "", paste(path, "/CargaContextoFormalIngles.R")))

conjunto_entrenamiento = subset(emails_entrenamiento,  emails_entrenamiento$ingles == 0)

source(gsub(" ", "", paste(path, "/CargaContextoFormalCastellano.R")))


# Inicializacion de variables para obtener las estadisticas
estadisticas_idioma = c("Castellano, Bien"=0,"Castellano, Mal"=0,"Ingles, Bien"=0,"Ingles, Mal"=0,"Indeterminado"=0)

estadisticas = c("Cancelacion, Bien"=0,"Cancelacion, Mal"=0,"No Cancelacion, Bien"=0,"No Cancelacion,
Mal"=0,"Indeterminado"=0)

# Procesamiento de emails para su clasificacion
for (i in seq(from = 1, to = nrow(emails_prueba))){
  # Inicializacion de variables
  idioma=""
  clasificacion_original <- emails_prueba[i,3]
  idioma_original <- emails_prueba[i,2]
  email_clasificar <- emails_prueba[i,1]

  # COMIENZO CLASIFICADOR IDIOMA

  # Script que clasifica el idioma
  source(gsub(" ", "", paste(path, "/clasificarIdioma.R")))

  # Segun los porcentajes, clasificamos los correos en un idioma
  if(porcentaje_castellano > porcentaje_ingles) {
    idioma = "castellano"
  }else if(porcentaje_castellano < porcentaje_ingles) {
    idioma = "ingles"
  } else {
    idioma = "undetermined"
  }

  # Si el idioma coincide con el idioma original, incrementamos su valor
```

```
if(idioma == "castellano" & idioma_original == 0) {
  estadisticas_idioma[1] = estadisticas_idioma[1] + 1
} else if (idioma == "castellano" & idioma_original == 1) {
  estadisticas_idioma[2] = estadisticas_idioma[2] + 1
} else if(idioma == "ingles" & idioma_original == 1) {
  estadisticas_idioma[3] = estadisticas_idioma[3] + 1
} else if (idioma == "ingles" & idioma_original == 0) {
  estadisticas_idioma[4] = estadisticas_idioma[4] + 1
} else {
  estadisticas_idioma[5] = estadisticas_idioma[5] + 1
}


# COMIENZO CLASIFICADOR CANCELACIÓN
#Inicializamos variables
clasificacion=""
cancelKeywords=0
noCancelKeywords=0
porcentaje_cancelacion=0
porcentaje_no_cancelacion=0
peso_cancelacion=0
peso_no_cancelacion=0

# Segun el idioma previamente clasificado, cargamos un conjunto u otro con el
# que generaremos el contexto formal
if(idioma == "ingles") {
  conjunto_entrenamiento <- subset(emails_entrenamiento, emails_entrenamiento$ingles==1)
} else if(idioma == "castellano") {
  conjunto_entrenamiento <- subset(emails_entrenamiento, emails_entrenamiento$ingles==0)
}

# Script generador de contexto formal segun el idioma del correo
source(gsub(" ", "", paste(path, "/CargaContextoFormal.R")))

if(idioma == "ingles") {

  # Script calculador de las varibles necesarias para clasificar el correo
  source(gsub(" ", "", paste(path, "/ScriptClasificadorIngles.R")))

  # Procedemos a clasificar el correo, en el que primero, comparamos sus
  # porcentajes, y dependiendo si es menor el de cancelacion que el de
  # no_cancelacion, diremos que el correo es de tipo no_cancelacion.
  # En caso de que sea mayor cancelacion que no_cancelacion, observamos sus
  # pesos, que dependiendo de si uno es mayor que otro hemos detectado
  # distintos patrones para su clasificacion. en caso de que sus pesos y sus
  # porcentajes son iguales, su clasificacion sera no_cancelacion

  if (porcentaje_cancelacion < porcentaje_no_cancelacion) {

    clasificacion="no_cancelacion"

  } else if (porcentaje_cancelacion > porcentaje_no_cancelacion) {

    # Calculamos los pesos de los tipos disponibles
    source(gsub(" ", "", paste(path, "/PesoCancelacion.R")))
    source(gsub(" ", "", paste(path, "/PesoNoCancelacion.R")))

    if(peso_cancelacion > peso_no_cancelacion) {
      if(abs(peso_cancelacion-peso_no_cancelacion) < 2){
        if(cancelKeywords > noCancelKeywords) {
          clasificacion="cancelacion"
```

```
      }else {
        clasificacion="no_cancelacion"
      }
    } else {
      clasificacion="cancelacion"
    }
  } else if(peso_cancelacion < peso_no_cancelacion) {
    if(cancelKeywords < noCancelKeywords) {
      if(abs(cancelKeywords + peso_no_cancelacion - noCancelKeywords - peso_cancelacion) > 2) {
        if(abs(peso_cancelacion-peso_no_cancelacion) < 5 && abs(peso_cancelacion-peso_no_cancelacion) > 0.5) {
          clasificacion = "cancelacion"
        } else {
          clasificacion = "no_cancelacion"
        }
      } else {
        if(cancelKeywords * 2 <= noCancelKeywords) {
          if(abs(peso_cancelacion-peso_no_cancelacion) < 2) {
            clasificacion = "cancelacion"
          } else {
            clasificacion = "no_cancelacion"
          }
        } else {
          clasificacion = "cancelacion"
        }
      }
    } else if(cancelKeywords > noCancelKeywords) {
      if(cancelKeywords * peso_cancelacion > noCancelKeywords*peso_no_cancelacion) {
        if(abs(peso_cancelacion-peso_no_cancelacion) < 1) {
          clasificacion = "no_cancelacion"
        } else {
          clasificacion = "cancelacion"
        }
      } else {
        if(abs(peso_cancelacion-peso_no_cancelacion) < 1) {
          clasificacion = "no_cancelacion"
        } else {
          clasificacion = "cancelacion"
        }
      }
    } else {
      if(abs(peso_cancelacion-peso_no_cancelacion) < 2 && abs(peso_cancelacion-peso_no_cancelacion) > 1) {
        clasificacion = "no_cancelacion"
      } else {
        if(peso_cancelacion * 2 < peso_no_cancelacion) {
          clasificacion="no_cancelacion"
        } else {
          clasificacion = "cancelacion"
        }
      }
    }

  } else {
    clasificacion = "no_cancelacion"
  }

} else {
  clasificacion="no_cancelacion"
}

# Si el correo esta escrito en castellano se aplican los mismos metodos que
# para los esritos en ingles
```

```
} else if(idioma == "castellano") {

 source(gsub(" ", "", paste(path, "/ScriptClasificadorCastellano.R")))

  if (porcentaje_cancelacion < porcentaje_no_cancelacion) {

    clasificacion="no_cancelacion"

  } else if (porcentaje_cancelacion > porcentaje_no_cancelacion) {

    source(gsub(" ", "", paste(path, "/PesoCancelacion.R")))
    source(gsub(" ", "", paste(path, "/PesoNoCancelacion.R")))

    if(peso_cancelacion > peso_no_cancelacion) {
      if(abs(peso_cancelacion-peso_no_cancelacion) < 2){
        if(cancelKeywords > noCancelKeywords) {
          clasificacion="cancelacion"
        }else {
          clasificacion="no_cancelacion"
        }
      } else {
        clasificacion="cancelacion"
      }
    } else if(peso_cancelacion < peso_no_cancelacion) {
      if(cancelKeywords < noCancelKeywords) {
        if(abs(cancelKeywords + peso_no_cancelacion - noCancelKeywords - peso_cancelacion) > 2) {
          if(abs(peso_cancelacion-peso_no_cancelacion) < 5 && abs(peso_cancelacion-peso_no_cancelacion) > 1.6) {
            clasificacion = "cancelacion"
          } else {
            clasificacion = "no_cancelacion"
          }
        } else {
          if(cancelKeywords * 2 <= noCancelKeywords) {
            if(abs(peso_cancelacion-peso_no_cancelacion) < 2) {
              clasificacion = "cancelacion"
            } else {
              clasificacion = "no_cancelacion"
            }
          } else {
            clasificacion = "cancelacion"
          }
        }
      } else if(cancelKeywords > noCancelKeywords) {
        if(cancelKeywords * peso_cancelacion > noCancelKeywords*peso_no_cancelacion) {
          if(abs(peso_cancelacion-peso_no_cancelacion) < 1) {
            clasificacion = "no_cancelacion"
          } else {
            clasificacion = "cancelacion"
          }
        } else {
          if(abs(peso_cancelacion-peso_no_cancelacion) < 1) {
            clasificacion = "no_cancelacion"
          } else {
            clasificacion = "cancelacion"
          }
        }
      } else {
        if(abs(peso_cancelacion-peso_no_cancelacion) < 2 && abs(peso_cancelacion-peso_no_cancelacion) > 1) {
          clasificacion = "no_cancelacion"
        } else {
          if(peso_cancelacion * 2 < peso_no_cancelacion) {
```

```
            clasificacion="no_cancelacion"
          } else {
            clasificacion = "cancelacion"
          }
        }
      }
    } else {
      clasificacion = "no_cancelacion"
    }
  } else {
    clasificacion="no_cancelacion"
  }
 } else {
  clasificacion = "no_cancelacion"
 }

 # Una vez tenemos la clasificacion, al igual que con el idioma, incrementamos
 # las variables si la clasificacion es igual a la original o no

 if(clasificacion == "cancelacion" & clasificacion_original == 1) {
  estadisticas[1] = estadisticas[1] + 1
 } else if (clasificacion == "cancelacion" & clasificacion_original == 0) {
  estadisticas[2] = estadisticas[2] + 1
 } else if(clasificacion == "no_cancelacion" & clasificacion_original == 0) {
  estadisticas[3] = estadisticas[3] + 1
 } else if (clasificacion == "no_cancelacion" & clasificacion_original == 1) {
  estadisticas[4] = estadisticas[4] + 1
 } else {
  estadisticas[5] = estadisticas[5] + 1
 }

}

# Una vez procesamos el conjunto de prueba para obtener la especifidad y sensibilidad
# del clasificador, procesamos estos datos para obtener la informacion requerida

resultados_finales_cancelacion = c("Cancelacion, Bien"=0,"Cancelacion, Mal"=0,"No Cancelacion, Bien"=0,"No
Cancelacion, Mal"=0,"Indeterminado"=0)

resultados_finales_cancelacion[1] = estadisticas[1]/(estadisticas[1]+estadisticas[2])
resultados_finales_cancelacion[2] = estadisticas[2]/(estadisticas[1]+estadisticas[2])
resultados_finales_cancelacion[3] = estadisticas[3]/(estadisticas[3]+estadisticas[4])
resultados_finales_cancelacion[4] = estadisticas[4]/(estadisticas[3]+estadisticas[4])
resultados_finales_cancelacion[5] = estadisticas[5]


resultados_finales_idioma =  c("Castellano, Bien"=0,"Castellano, Mal"=0,"Ingles, Bien"=0,"Ingles,
Mal"=0,"Indeterminado"=0)

resultados_finales_idioma[1] = estadisticas_idioma[1]/(estadisticas_idioma[1]+estadisticas_idioma[2])
resultados_finales_idioma[2] = estadisticas_idioma[2]/(estadisticas_idioma[1]+estadisticas_idioma[2])
resultados_finales_idioma[3] = estadisticas_idioma[3]/(estadisticas_idioma[3]+estadisticas_idioma[4])
resultados_finales_idioma[4] = estadisticas_idioma[4]/(estadisticas_idioma[3]+estadisticas_idioma[4])
resultados_finales_idioma[5] = estadisticas_idioma[5]

resultados_finales_idioma
resultados_finales_cancelacion
```

Clasificador.R

```
email_clasificar <- "
> Subject: Cancellation - ref *********
>


>
> Unfortunately we are no longer ****** to travel to Spain as the UK Government
> has changed its advise and have said that we should not travel to Spain.
> This means that we need to cancel our ****** hire booking. ****** ****** you
> cancel booking ******* - ****** we ****** a refund?
>
> ************
>"

# Cargamos la ruta de los ficheros
path = dirname(rstudioapi::getSourceEditorContext()$path)

# Cargamos los emails recibidos con palabras basicas como hola, good morning... y los dividimos por idioma
emails_idioma <- read.csv(file = gsub(" ", "", paste(path, "/emailsTFGidioma.csv")), head = TRUE, sep=";")

# Cargamos los emails recibidos y los dividimos por idioma
emails <- read.csv(file = gsub(" ", "", paste(path, "/emailsTFG.csv")), head = TRUE, sep=";")

# Cargamos los contextos formales para cada idioma, que nos ayudaran a
# determinar el idioma
conjunto_entrenamiento = subset(emails_idioma,  emails$ingles == 1)

source(gsub(" ", "", paste(path, "/CargaContextoFormalIngles.R")))

conjunto_entrenamiento = subset(emails_idioma,  emails$ingles == 0)

source(gsub(" ", "", paste(path, "/CargaContextoFormalCastellano.R")))

# COMIENZO CLASIFICADOR IDIOMA

idioma = ""

# Script que clasifica el idioma
source(gsub(" ", "", paste(path, "/clasificarIdioma.R")))
```

```
# Segun los porcentajes, clasificamos los correos en un idioma
if(porcentaje_castellano > porcentaje_ingles) {
  idioma = "castellano"
}else if(porcentaje_castellano < porcentaje_ingles) {
  idioma = "ingles"
} else {
  idioma = "undetermined"
}

# Segun el idioma previamente clasificado, cargamos un conjunto u otro con el
# que generaremos el contexto formal
if(idioma == "ingles") {
  conjunto_entrenamiento <- subset(emails, emails$ingles==1)
} else if(idioma == "castellano") {
  conjunto_entrenamiento <- subset(emails, emails$ingles==0)
}

# Script generador de contexto formal segun el idioma del correo
source(gsub(" ", "", paste(path, "/CargaContextoFormal.R")))

if(idioma == "ingles") {

  # Script calculador de las varibles necesarias para clasificar el correo
  source(gsub(" ", "", paste(path, "/ScriptClasificadorIngles.R")))

  # Procedemos a clasificar el correo, en el que primero, comparamos sus
  # porcentajes, y dependiendo si es menor el de cancelacion que el de
  # no_cancelacion, diremos que el correo es de tipo no_cancelacion.
  # En caso de que sea mayor cancelacion que no_cancelacion, observamos sus
  # pesos, que dependiendo de si uno es mayor que otro hemos detectado
  # distintos patrones para su clasificacion. en caso de que sus pesos y sus
  # porcentajes son iguales, su clasificacion sera no_cancelacion

  if (porcentaje_cancelacion < porcentaje_no_cancelacion) {

    clasificacion="no_cancelacion"

  } else if (porcentaje_cancelacion > porcentaje_no_cancelacion) {

    # Calculamos los pesos de los tipos disponibles
    source(gsub(" ", "", paste(path, "/PesoCancelacion.R")))
```

```
source(gsub(" ", "", paste(path, "/PesoNoCancelacion.R")))


if(peso_cancelacion > peso_no_cancelacion) {
  if(abs(peso_cancelacion-peso_no_cancelacion) < 2){
    if(cancelKeywords > noCancelKeywords) {
      clasificacion="cancelacion"
    }else {
      clasificacion="no_cancelacion"
    }
  } else {
    clasificacion="cancelacion"
  }
} else if(peso_cancelacion < peso_no_cancelacion) {
  if(cancelKeywords < noCancelKeywords) {
    if(abs(cancelKeywords + peso_no_cancelacion - noCancelKeywords - peso_cancelacion) > 2) {
      if(abs(peso_cancelacion-peso_no_cancelacion) < 5 && abs(peso_cancelacion-peso_no_cancelacion) > 0.5) {
        clasificacion = "cancelacion"
      } else {
        clasificacion = "no_cancelacion"
      }
    } else {
      if(cancelKeywords * 2 <= noCancelKeywords) {
        if(abs(peso_cancelacion-peso_no_cancelacion) < 2) {
          clasificacion = "cancelacion"
        } else {
          clasificacion = "no_cancelacion"
        }
      } else {
        clasificacion = "cancelacion"
      }
    }
  } else if(cancelKeywords > noCancelKeywords) {
    if(cancelKeywords * peso_cancelacion > noCancelKeywords*peso_no_cancelacion) {
      if(abs(peso_cancelacion-peso_no_cancelacion) < 1) {
        clasificacion = "no_cancelacion"
      } else {
        clasificacion = "cancelacion"
      }
    } else {
      if(abs(peso_cancelacion-peso_no_cancelacion) < 1) {
        clasificacion = "no_cancelacion"
```

```
        } else {
          clasificacion = "cancelacion"
        }
      }
    } else {
      if(abs(peso_cancelacion-peso_no_cancelacion) < 2 && abs(peso_cancelacion-peso_no_cancelacion) > 1) {
        clasificacion = "no_cancelacion"
      } else {
        if(peso_cancelacion * 2 < peso_no_cancelacion) {
          clasificacion="no_cancelacion"
        } else {
          clasificacion = "cancelacion"
        }
      }
    }

  } else {
    clasificacion = "no_cancelacion"
  }

} else {
  clasificacion="no_cancelacion"
}

# Si el correo esta escrito en castellano se aplican los mismos metodos que
# para los esritos en ingles
} else if(idioma == "castellano") {

source(gsub(" ", "", paste(path, "/ScriptClasificadorCastellano.R")))

if (porcentaje_cancelacion < porcentaje_no_cancelacion) {

  clasificacion="no_cancelacion"

} else if (porcentaje_cancelacion > porcentaje_no_cancelacion) {

  source(gsub(" ", "", paste(path, "/PesoCancelacion.R")))
  source(gsub(" ", "", paste(path, "/PesoNoCancelacion.R")))

  if(peso_cancelacion > peso_no_cancelacion) {
    if(abs(peso_cancelacion-peso_no_cancelacion) < 2){
```

```
    if(cancelKeywords > noCancelKeywords) {
      clasificacion="cancelacion"
    }else {
      clasificacion="no_cancelacion"
    }
  } else {
    clasificacion="cancelacion"
  }
} else if(peso_cancelacion < peso_no_cancelacion) {
  if(cancelKeywords < noCancelKeywords) {
    if(abs(cancelKeywords + peso_no_cancelacion - noCancelKeywords - peso_cancelacion) > 2) {
      if(abs(peso_cancelacion-peso_no_cancelacion) < 5 && abs(peso_cancelacion-peso_no_cancelacion) > 1.6) {
        clasificacion = "cancelacion"
      } else {
        clasificacion = "no_cancelacion"
      }
    } else {
      if(cancelKeywords * 2 <= noCancelKeywords) {
        if(abs(peso_cancelacion-peso_no_cancelacion) < 2) {
          clasificacion = "cancelacion"
        } else {
          clasificacion = "no_cancelacion"
        }
      } else {
        clasificacion = "cancelacion"
      }
    }
  } else if(cancelKeywords > noCancelKeywords) {
    if(cancelKeywords * peso_cancelacion > noCancelKeywords*peso_no_cancelacion) {
      if(abs(peso_cancelacion-peso_no_cancelacion) < 1) {
        clasificacion = "no_cancelacion"
      } else {
        clasificacion = "cancelacion"
      }
    } else {
      if(abs(peso_cancelacion-peso_no_cancelacion) < 1) {
        clasificacion = "no_cancelacion"
      } else {
        clasificacion = "cancelacion"
      }
    }
```

```
      } else {
        if(abs(peso_cancelacion-peso_no_cancelacion) < 2 && abs(peso_cancelacion-peso_no_cancelacion) > 1) {
          clasificacion = "no_cancelacion"
        } else {
          if(peso_cancelacion * 2 < peso_no_cancelacion) {
            clasificacion="no_cancelacion"
          } else {
            clasificacion = "cancelacion"
          }
        }
      }
    } else {
      clasificacion = "no_cancelacion"
    }
  } else {
    clasificacion="no_cancelacion"
  }
} else {
  clasificacion = "no_cancelacion"
}

cat("El corrreo ha sido clasificado como ",clasificacion, " y está escrito en ", idioma)
```

CargaContextoFormalIngles.R

La carga del contexto formal en castellano se realiza de forma similar.

```
# Cargamos paquete text mining
library(tm)

#construimos una nueva variable corpus llamada corpus
corpus = VCorpus(VectorSource(conjunto_entrenamiento$text))

#convertimos el texto a minusculas
corpus = tm_map(corpus, content_transformer(tolower))
corpus <- tm_map(corpus, content_transformer(gsub), pattern = "[\n]+", replacement = " ")
corpus <- tm_map(corpus, content_transformer(gsub), pattern = "[[\u2028] ]+", replacement = " ")
corpus <- tm_map(corpus, content_transformer(gsub), pattern = "[[:punct:] ]+", replacement = " ")

#eliminamos todas las puntuaciones(.,!?;:'"...) del corpus
corpus = tm_map(corpus, removePunctuation)

#eliminamos todos los stopwords ingleses (articulos y demas)
#del corpus
corpus = tm_map(corpus, removeWords, stopwords("en"))

corpus <- tm_map(corpus, content_transformer(gsub), pattern = "[[:space:]]+", replacement = " ")

#extraemos las palabras del corpus
corpus = tm_map(corpus, stemDocument)



#Ahora estamos listos para extraer la frecuencia de palabras para
#usarlas en nuestro problema de prediccion
#Crearemos un doc donde la filas corresponderan a los emails y
# las colmnas a las palabras
dtm = DocumentTermMatrix(corpus)

#eliminamos los terminos que no aparecen tanto (sparse terms)
#nos quedamos con los terminos que aparecen al menos en el 5%
# de los documentos
spdtm = removeSparseTerms(dtm, 0.98)

#convertimos spdtm a un dataframe
emailsDataFrame = as.data.frame(as.matrix(spdtm))
```

```
v0 = colSums(subset(emailsDataFrame))
v3 = colSums(subset(emailsDataFrame, conjunto_entrenamiento$Cancelacion == 1))
v4 = colSums(subset(emailsDataFrame, conjunto_entrenamiento$Cancelacion == 0))

tabla =
cbind(percent_cancel=round(v3/v0*100,2),percent_no_cancel=round(v4/v0*100,2),percent_word_appearance=round(
v0/nrow(conjunto_entrenamiento)*100,2))
I = rbind(cancelacion=round(v3/v0*100,2),no_cancelacion=round(v4/v0*100,2))



#creamos contexto asignando 1s a aquellos valores cuyo porcentaje supera el 60%
for (i in seq(from = 1, to = nrow(I)))
{
  for(j in seq(from = 1, to = ncol(I)))
    if(I[i,j] > 75){
      I[i,j]=1
    } else {
      I[i,j]=0
    }
}

# Cargamos libreria fcaR
library(fcaR)

# Construimos contexto formal
fc3 <- FormalContext$new(I = I)

#Extraemos las implicaciones y conceptos del contexto formal
fc3$find_implications()
```

ClasificarIdioma.R

```
# Cargamos paquete text mining
library(tm)

if(email_clasificar == "") {
  porcentaje_castellano = 0
  porcentaje_ingles = 0
}else {
  #construimos una nueva variable corpus llamada corpus
  corpus1 = VCorpus(VectorSource(email_clasificar))

  #convertimos el texto a minusculas
  corpus1 = tm_map(corpus1, content_transformer(tolower))
  corpus1 <- tm_map(corpus1, content_transformer(gsub), pattern = "[\n]+", replacement = " ")
  corpus1 <- tm_map(corpus1, content_transformer(gsub), pattern = "[[\u2028] ]+", replacement = " ")
  corpus1 <- tm_map(corpus1, content_transformer(gsub), pattern = "[[:punct:] ]+", replacement = " ")

  #eliminamos todas las puntuaciones(.,!?;:'"...) del corpus
  corpus1 = tm_map(corpus1, removePunctuation)

  #eliminamos todos los stopwords ingleses (articulos y demas)
  #del corpus
  corpus1 = tm_map(corpus1, removeWords, stopwords("en"))

  corpus1 <- tm_map(corpus1, content_transformer(gsub), pattern = "[[:space:]]+", replacement = " ")

  #extraemos las palabras del corpus
  corpus1 = tm_map(corpus1, stemDocument)



  #extraemos la frecuencia de palabras del email
  dtm1 = DocumentTermMatrix(corpus1)

  #convertimos dtm1 a un dataframe
  emailDataFrame = as.data.frame(as.matrix(dtm1))



  if(ncol(emailDataFrame) < 1) {
    porcentaje_castellano = 0
```

```
  porcentaje_ingles = 0
} else {
  for (i in seq(from = 1, to = ncol(emailDataFrame))) {
    if(emailDataFrame[1,i] > 1) emailDataFrame[1,i] = 1
  }

  # Obtenemos las palabras que nos indican si es cancelacion en ingles
  concepto_ingles = fc3$concepts[4][[1]]

  atributos_ingles = fc3$intent(concepto_ingles$get_extent())

  atributos_ingles

  # Obtenemos las palabras que nos indican si es cancelacion en castellano
  concepto_castellano = fc2$concepts[4][[1]]

  atributos_castellano = fc2$intent(concepto_castellano$get_extent())

  atributos_castellano

  # Procesamos el correo recibido obteniendo diferentes variables necesarias para el estudio
  v_ing = as_vector(atributos_ingles)

  v_ing = v_ing[v_ing == 1]

  if(length(emailDataFrame) > 1) {
    inglesKeywords = length(emailDataFrame[ , which(names(emailDataFrame) %in% names(v_ing))])
  } else {
    inglesKeywords = length(emailDataFrame)
  }

  v_cas = as_vector(atributos_castellano)

  v_cas = v_cas[v_cas == 1]

  if(length(emailDataFrame) > 1) {
    casteKeywords = length(emailDataFrame[ , which(names(emailDataFrame) %in% names(v_cas))])
  } else {
    casteKeywords = length(emailDataFrame)
  }
```

```
    porcentaje_castellano = round(casteKeywords / length(v_cas),2)

    porcentaje_ingles = round(inglesKeywords / length(v_ing),2)
  }
}
```

CargaContextoFormal.R

```
# Cargamos paquete text mining
library(tm)

#construimos una nueva variable corpus llamada corpus
corpus = VCorpus(VectorSource(conjunto_entrenamiento$text))

#convertimos el texto a minusculas
corpus = tm_map(corpus, content_transformer(tolower))
corpus <- tm_map(corpus, content_transformer(gsub), pattern = "[\n]+", replacement = " ")
corpus <- tm_map(corpus, content_transformer(gsub), pattern = "[[\u2028] ]+", replacement = " ")
corpus <- tm_map(corpus, content_transformer(gsub), pattern = "[[:punct:] ]+", replacement = " ")

#eliminamos todas las puntuaciones(.,!?;:'"...) del corpus
corpus = tm_map(corpus, removePunctuation)

#eliminamos todos los stopwords ingleses (articulos y demas)
#del corpus
if(idioma=="ingles") {
  corpus = tm_map(corpus, removeWords, stopwords("en"))
} else {
  corpus = tm_map(corpus, removeWords, stopwords("es"))
}

corpus <- tm_map(corpus, content_transformer(gsub), pattern = "[[:space:]]+", replacement = " ")

#extraemos las palabras del corpus
corpus = tm_map(corpus, stemDocument)



#Ahora estamos listos para extraer la frecuencia de palabras para
#usarlas en nuestro problema de prediccion
#Crearemos un doc donde la filas corresponderan a los emails y
# las colmnas a las palabras
dtm = DocumentTermMatrix(corpus)

#eliminamos los terminos que no aparecen tanto (sparse terms)
#nos quedamos con los terminos que aparecen al menos en el 5%
# de los documentos
spdtm = removeSparseTerms(dtm, 0.98)
```

```
#convertimos spdtm a un dataframe
emailsDataFrame = as.data.frame(as.matrix(spdtm))


v0 = colSums(subset(emailsDataFrame))
v3 = colSums(subset(emailsDataFrame, conjunto_entrenamiento$Cancelacion == 1))
v4 = colSums(subset(emailsDataFrame, conjunto_entrenamiento$Cancelacion == 0))

tabla =
cbind(percent_cancel=round(v3/v0*100,2),percent_no_cancel=round(v4/v0*100,2),percent_word_appearance=round(
v0/nrow(conjunto_entrenamiento)*100,2))
I = rbind(cancelacion=round(v3/v0*100,2),no_cancelacion=round(v4/v0*100,2))


#creamos contexto asignando 1s a aquellos valores cuyo porcentaje supera el 60%
for (i in seq(from = 1, to = nrow(I)))
{
  for(j in seq(from = 1, to = ncol(I)))
    if(I[i,j] > 75){
      I[i,j]=1
    } else {
      I[i,j]=0
    }
}

# Cargamos libreria fcaR
library(fcaR)

# Construimos contexto formal
fc <- FormalContext$new(I = I)

#Extraemos las implicaciones y conceptos del contexto formal
fc$find_implications()
```

ScriptClasificadorIngles.R

La carga del contexto formal en castellano se realiza de forma similar.

```
if(email_clasificar == "") {
  porcentaje_cancelacion = 0
  porcentaje_no_cancelacion = 0

  #cat("undetermined") ;
}else {
  #construimos una nueva variable corpus llamada corpus
  corpus1 = VCorpus(VectorSource(email_clasificar))

  #convertimos el texto a minusculas
  corpus1 = tm_map(corpus1, content_transformer(tolower))
  corpus1 <- tm_map(corpus1, content_transformer(gsub), pattern = "[\n]+", replacement = " ")
  corpus1 <- tm_map(corpus1, content_transformer(gsub), pattern = "[[\u2028] ]+", replacement = " ")
  corpus1 <- tm_map(corpus1, content_transformer(gsub), pattern = "[[:punct:] ]+", replacement = " ")

  #eliminamos todas las puntuaciones(.,!?;:'"...) del corpus
  corpus1 = tm_map(corpus1, removePunctuation)

  #eliminamos todos los stopwords ingleses (articulos y demas)
  #del corpus
  corpus1 = tm_map(corpus1, removeWords, stopwords("en"))

  corpus1 <- tm_map(corpus1, content_transformer(gsub), pattern = "[[:space:]]+", replacement = " ")

  #extraemos las palabras del corpus
  corpus1 = tm_map(corpus1, stemDocument)


  #extraemos la frecuencia de palabras del email
  dtm1 = DocumentTermMatrix(corpus1)

  #convertimos dtm1 a un dataframe
  emailDataFrame = as.data.frame(as.matrix(dtm1))



  #filtramos las columnas que no son palabras clave de nuestro sistema
  palabras_clave <- c(fc$attributes)
```

```
if(ncol(emailDataFrame) < 1) {
  porcentaje_cancelacion = 0
  porcentaje_no_cancelacion = 0
  #cat("undetermined")
} else {
  for (i in seq(from = 1, to = ncol(emailDataFrame))) {
    if(emailDataFrame[1,i] > 1) emailDataFrame[1,i] = 1
  }

  concepto_cancelacion = fc$concepts[2][[1]]

  atributos_cancelacion = fc$intent(concepto_cancelacion$get_extent())

  concepto_no_cancelacion = fc$concepts[3][[1]]

  atributos_no_cancelacion = fc$intent(concepto_no_cancelacion$get_extent())

  v_canc = as_vector(atributos_cancelacion)

  v_canc = v_canc[v_canc == 1]




  if(length(emailDataFrame) > 1) {
    cancelKeywords = length(emailDataFrame[ , which(names(emailDataFrame) %in% names(v_canc))])
  } else {
    cancelKeywords = length(emailDataFrame)
  }


  v_no_canc = as_vector(atributos_no_cancelacion)

  v_no_canc = v_no_canc[v_no_canc == 1]

  if(length(emailDataFrame) > 1) {
    noCancelKeywords = length(emailDataFrame[ , which(names(emailDataFrame) %in% names(v_no_canc))])
  } else {
    noCancelKeywords = length(emailDataFrame)
  }
```

```
    palabrasPesoCancelacion=names(emailDataFrame[which(names(emailDataFrame) %in% names(v_canc))])
    palabrasPesoNoCancelacion=names(emailDataFrame[which(names(emailDataFrame) %in% names(v_no_canc))])


    porcentaje_cancelacion = round(cancelKeywords / length(v_canc),2)

    porcentaje_no_cancelacion = round(noCancelKeywords / length(v_no_canc),2)

    #cat(c("division cancelacion: ", porcentaje_cancelacion)) #resultado de que aparezcan x de nuestras keywords
    #asociadas a cancelacion entre el numero total de keywords asociadas a cancelacion

    #cat(c("division no cancelacion: ", porcentaje_no_cancelacion)) #resultado de que aparezcan x de nuestras keywords
    #asociadas a no cancelacion entre el numero total de keywords asociadas a no cancelacion
  }

}
```

PesoCancelacion.R

```
if(length(palabrasPesoCancelacion) == 0) {
  #Si no tenemos palabras de cancelacion su valor es 0
  peso_cancelacion = 0
} else {
  suma_peso_porcentaje = 0

  # calculamos pesos de las palabras del correo asociadas a cancelacion
  for (i in seq(from = 1, to = length(palabrasPesoCancelacion))){
    palabra = palabrasPesoCancelacion[i]

    suma_peso_porcentaje = suma_peso_porcentaje + tabla[palabra,1]
  }

  peso1 = suma_peso_porcentaje/length(palabrasPesoCancelacion)

  suma_peso_porcentaje=0
  # calculamos pesos de las palabras del correo asociadas a no_cancelacion
  if(length(palabrasPesoNoCancelacion) != 0) {
    for (i in seq(from = 1, to = length(palabrasPesoNoCancelacion))){
      palabra = palabrasPesoNoCancelacion[i]

      suma_peso_porcentaje = suma_peso_porcentaje + tabla[palabra,1]
    }
    peso2 = suma_peso_porcentaje/length(palabrasPesoNoCancelacion)
  } else {
    peso2 = 0
  }

  #sumamos ambos pesos y lo dividimos por el numero total de palabras que tenemos
  peso_cancelacion = (peso1+peso2)/(length(palabrasPesoNoCancelacion)+length(palabrasPesoCancelacion))

}
```

PesoNoCancelacion.R

```
if(length(palabrasPesoNoCancelacion) == 0) {
  #Si no tenemos palabras de no_cancelacion su valor es 0
  peso_no_cancelacion = 0
} else {
  suma_peso_porcentaje = 0
  # calculamos pesos de las palabras del correo asociadas a no_cancelacion
  for (i in seq(from = 1, to = length(palabrasPesoNoCancelacion))){
    palabra = palabrasPesoNoCancelacion[i]

    suma_peso_porcentaje = suma_peso_porcentaje + tabla[palabra,2]
  }

  peso1 = suma_peso_porcentaje/length(palabrasPesoNoCancelacion)

  suma_peso_porcentaje=0
  # calculamos pesos de las palabras del correo asociadas a cancelacion
  if(length(palabrasPesoCancelacion) != 0) {
    for (i in seq(from = 1, to = length(palabrasPesoCancelacion))){
      palabra = palabrasPesoCancelacion[i]

      suma_peso_porcentaje = suma_peso_porcentaje + tabla[palabra,2]
    }
    peso2 = suma_peso_porcentaje/length(palabrasPesoCancelacion)
  } else {
    peso2 = 0
  }
  #sumamos ambos pesos y lo dividimos por el numero total de palabras que tenemos
  peso_no_cancelacion = (peso1+peso2)/(length(palabrasPesoNoCancelacion)+length(palabrasPesoCancelacion))

}
```

UNIVERSIDAD DE MÁLAGA | uma.es

E.T.S de Ingeniería Informática
Bulevar Louis Pasteur, 35
Campus de Teatinos
29071 Málaga

E.T.S. DE INGENIERÍA INFORMÁTICA