Universidad
Carlos III de Madrid

Teoría de la Señal y Comunicaciones

PROYECTO FIN DE CARRERA

Desarrollo de un sistema para el análisis del posicionamiento sentimental de marcas en Internet

Autor: García Pérez, Fernando

Director: Bousoño Calzón, Carlos

Leganés, Julio de 2013

Título: DESARROLLO DE UN SISTEMA PARA EL ANÁLISIS DEL POSICIONAMIENTO SENTIMENTAL DE MARCAS EN INTERNET

Autor: Fernando García Pérez

Director: Carlos Bousoño Calzón

EL TRIBUNAL

Presidente: _____

Vocal: _____

Secretario: _____

Realizado el acto de defensa y lectura del Proyecto Fin de Carrera el día 24 de Julio de 2013 en Leganés, en la Escuela Politécnica Superior de la Universidad Carlos III de Madrid, acuerda otorgarle la CALIFICACIÓN de

Fdo: Presidente Fdo: Secretario Fdo: Vocal

Agradecimientos

Primero, quisiera dar las gracias a mi tutor, Don Carlos Bousoño Calzón, por haberme dirigido este Proyecto Fin de Carrera. Carlos, muchas gracias por su atención, ayuda y paciencia.

Segundo, agradecer a mis padres todos los sacrificios realizados para que mi hermano y yo hayamos podido tener la educación que hemos recibido. A mi hermano, Javier, que sigue mis pasos, gracias por ser así. Sois los mejores.

Gracias a todos los compañeros que he conocido a lo largo de la carrera. Han sido muchos los momentos vividos con vosotros en clase, en las interminables prácticas en las aulas informáticas y en las temidas épocas de exámenes. Ha sido un orgullo haber compartido todos estos momentos junto a vosotros, siempre os recordaré.

Por último, quisiera agradecer todo su apoyo a esa persona que es muy especial en mi vida. Lidia, muchas gracias por tu paciencia, por creer en mí y estar siempre a mi lado. Este verano deseo recuperar todo el tiempo que no hemos podido pasar juntos.

Resumen

La World Wide Web es, hoy día, un enorme escaparate en el que los consumidores observan, consultan y compran productos. El rápido acceso a la información que proporcionan los buscadores web ha motivado que los consumidores busquen una segunda opinión en la Web, mediante la cual, reducir el riesgo percibido en la compra.

Este Proyecto Fin de Carrera tiene como objetivo el diseño y la implementación de un sistema que permita analizar el contenido de artículos online, concretamente, artículos de coches publicados en sitios especializados en el mundo del motor y la actualidad del automóvil, e identificar las relaciones de similitud existentes entre marcas y sentimientos.

Palabras clave:

Buscadores web, recuperación de la información, procesamiento lenguaje natural, latent semantic analysis y análisis de sentimientos.

Abstract

Nowadays, the World Wide Web is a huge storefront in which consumers observe, see and buy products. The fast access to information provided by search engines has prompted consumers to seek a second opinion on the Web in order to reduce risk purchase.

The goal of this project is the design and implementation of a system to allow analyze reviews, specifically, car reviews from automobile magazines sites, and recognize relationships between car brands and feelings.

Keywords:

Search engines, information retrieval, natural language processing, latent semantic analysis and sentiment analysis.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción y objetivos

Índice Capítulo 1

1.1. Introducción

Hoy en día, la WWW *(World Wide Web)* es un enorme escaparate en el que los consumidores observan, consultan y compran productos.

El auge del boca a boca digital y la consulta de artículos online han generado un gran impacto en las decisiones de compra de los consumidores. Las recomendaciones personales ya no son una garantía de compra y los consumidores buscan una segunda opinión en la Web, mediante la cual, reducir el riesgo percibido en la compra. Mike Hollywood, director de nuevos medios de la firma de investigación *Cone* manifestó que «los consumidores actuales quieren estar seguros antes de 'rascar sus bolsillos', y que solo las recomendaciones personales no son suficientes para garantizar una compra». El rápido acceso a la información que proporcionan los buscadores web (*Google*, *Yahoo!*, *Bing*, etcétera.) ha motivado que la gran mayoría de los usuarios utilice Internet para encontrar información, donde los artículos online están inmediatamente disponibles en las páginas de resultados de estos buscadores.

Los mensajes que las compañías emiten en los medios de comunicación tradicionales (televisión, radio, etcétera.) ya no son suficientes para controlar la imagen que se quiere mostrar al mundo sobre una determinada marca o producto. Ahora cualquier usuario, a través de Internet, puede expresar sus opiniones y/o experiencias con una marca/producto, teniendo influencia en la decisión de compra de los usuarios que buscan información. Un estudio realizado en 2011 por *PowerReviews* revela en este sentido que el 90% de los usuarios consumidores reconoció que este tipo de contenidos generan un gran impacto en sus decisiones de compra, y de los cuales el 60% afirma que ejerce mayor influencia respecto a otras fuentes. Como consecuencia, Internet se ha convertido en una nueva ventana donde observar la imagen que tienen los usuarios (detractores y promotores) de una marca o producto.

Dado el fuerte incremento de usuarios experimento en Internet, las compañías han enfocado sus esfuerzos en los últimos años en el marketing online, siendo ya Internet el segundo medio que acoge más publicidad dentro del grupo de medios convencionales. El último estudio *InfoAdex* de la Inversión Publicitaria en España 2013 revela que Internet es la segunda opción preferida por las marcas para invertir en publicidad, superando por primera vez a los diarios y posicionándose justo por debajo del primer medio, la televisión.

En el actual desarrollo del marketing online, el papel de los sentimientos cada vez es más importante para las marcas de una empresa. El posicionamiento relativo de unas marcas frente a otras en el espacio sentimental de los potenciales clientes es un aspecto decisivo a la hora de vender los productos asociados a esa marca. Cuando parece que las rebajas en los precios del artículo y las ofertas agresivas devienen en la única estrategia de venta para marcas y empresas, la *Social Media Strategist & Brander*

Gaby Castellanos insiste en que la fidelización de la comunidad y la interacción con la misma son factores claves para generar clientes. En una reciente entrevista concedida a la agencia *Webpositer*, esta líder de opinión en los Social Media asevera que «si eres capaz de enamorar a tu consumidor, podrás hacer que te elija ante diez productos de la competencia en un lineal del supermercado. Si no lo enamoras, elegirá el más barato o más conveniente».

Internet puede ser visto como herramienta de investigación de mercado para las compañías, con el objetivo principal de conocer cómo están posicionados sus productos en la mente del consumidor, y más importante aún, conocer cómo lo están los de la competencia. Lejos de los cuestionarios colocados en páginas y entrevistas enviadas por correo, la comparativa de posicionamiento sentimental de unas marcas frente a otras puede ser realizada directamente mediante el análisis de los sentimientos expresados en el texto de las páginas de la Web 1.0 y las opiniones de los usuarios (en blogs, redes sociales, foros de discusión, etcétera) de la Web 2.0.

La promesa de una mejor y más rápida toma de decisiones y una ventaja competitiva fruto de la capacidad de analizar y actuar sobre la información en el momento oportuno ha impulsado la demanda de soluciones de análisis de negocio. Un estudio de la *International Data Corporation (IDC)* refleja que en el 2012 el mercado mundial de software de análisis de negocio creció un 8.7% interanual con unos ingresos que alcanzaron los 34.5 billones de dólares, superando al mercado de software general que creció un 3.6% interanual ese mismo año. Dan Vesset, vicepresidente de programas de *IDC*, asegura que «existe una creciente evidencia cuantificable de que la toma de decisiones impulsada por datos permitida por soluciones de análisis de negocio proporciona una diferencia competitiva», además de que «esto, junto con un amplio interés en las grandes colecciones de datos, ha llevado la tecnología a la cima de muchas agendas ejecutivas y su introducción en el mercado convencional».

1.2. Objetivos

El objetivo de este Proyecto Fin de Carrera es el diseño e implementación de un sistema que de apoyo a las compañías en la toma de decisiones de marketing. Este sistema permitirá rastrear la Web, en concreto sitios web especializados en el mundo del motor y la actualidad del automóvil, aplicando técnicas del campo de la búsqueda y recuperación de la información (en inglés *Information Search and Retrieval (IRS)*) para completar la operación de rastreo y técnicas para el procesamiento del lenguaje natural para analizar el contenido de los artículos recolectados.

Se decidió que el sistema se enfocase en encontrar las relaciones de similitud existentes entre las marcas automovilísticas y los sentimientos recogidos en estos artículos, pero es fácilmente configurable a otros mercados (ropa, alimentación, higiene, etcétera.) u otros ámbitos (destinos turísticos, políticos, etcétera.). El sistema debería permitir conocer a la compañía: 1) cómo están posicionados sus marcas y/o productos en la mente de los usuarios y especialistas del sector, 2) cómo lo están los de la competencia y 3) cómo estas marcas y/o productos se diferencian de los de la competencia. Conocer esta información permitirá a la compañía tomar ventaja competitiva en la captación de nuevos clientes con respecto a sus rivales, identificando nichos de mercado que no están lo suficientemente atendidos por sus rivales o por la propia compañía.

La motivación del proyecto no es puramente económica, sino que también es académica, dado que se analizarán las dificultades encontradas en la implementación de un sistema de esta índole, proporcionando conocimiento futuro para próximos proyectos con objetivos relacionados.

Los principales retos afrontados en el desarrollo de este proyecto se enumeran a continuación:

- La construcción de un corpus de documentos adecuado para conseguir los objetivos propuestos fruto de la operación de rastreo.
- La extracción de la información relevante contenida en los textos de los artículos recolectados, obteniendo una representación de los mismos que facilite su procesado.
- La presentación de los resultados devueltos por el procesado, de modo que ésta sea entendible y permita extraer conclusiones.

1.3. Estructura de la memoria

El presente documento se ha estructurado en los 7 capítulos que se describen a continuación:

Capítulo 1. Introducción: En este punto se introduce al lector sobre el propósito del proyecto y se enumeran los objetivos establecidos.

Capítulo 2. Estado del arte. En este punto se describe lo que existe sobre el propósito que se aborda, así como tecnologías y aplicaciones similares a las empleadas en el presente proyecto.

Capítulo 3. Diseño y arquitectura del sistema: En este punto se explica la arquitectura desarrollada, detallando cada una de las fases en las que se compone el diseño del sistema.

Capítulo 4. Experimento y resultados: En este punto se detalla el planteamiento y metodología seguida en la consecución del experimento, las fases en las que se ha dividido el experimento en función de su tamaño y los resultados obtenidos en su realización.

Capítulo 5. Conclusiones y líneas futuras: En este punto se describen las impresiones alcanzadas tras la realización del proyecto y se presentan posibilidades de mejora y de ampliación del trabajo aquí completado en las que seguir trabajando.

Capítulo 6. Anexo I: Herramientas desarrolladas ad hoc: En este punto se detallan todos los aspectos relacionados con la implementación de cada uno de los módulos que conforman el sistema desarrollado, así como las herramientas de programación y lenguajes utilizados.

Capítulo 7. Anexo II: Manual de usuario: En este punto se detalla, por cada módulo, la interfaz de usuario diseñada, los parámetros de entrada y salida requeridos y los pasos a seguir para su ejecución.

Capítulo 2

Estado del Arte

Índice Capítulo 2

2.1. Introducción

La WWW *(World Wide Web)* puede ser vista como un enorme repositorio de información no estructurada, que desencadena la necesidad de herramientas eficientes para gestionar, recuperar y filtrar la información de la Web (R.Baeza-Yates, and B. Ribeiro-Neto, 1999). La aproximación utilizada más habitualmente para recuperar y localizar información en la Web la constituyen los buscadores basados en técnicas de rastreo (Manuel Álvarez Díaz, 2007). Estos rastreadores son programas software que, de manera automatizada, recorren la Web recolectando rápida y eficientemente tantas páginas útiles como les sea posible.

Puesto que la información contenida en la Web no está estructurada, se necesita de tecnologías que puedan arrojar información que facilite la interpretación del texto. La extracción de información es un campo del procesamiento de lenguaje natural (NLP) que permite extraer automáticamente conocimiento estructurado a partir de información existente en texto no estructurado en lenguaje natural. El primer paso es el reconocimiento de entidades con nombre (NER), también denominado etiquetado semántico, que permite detectar y clasificar los elementos del texto en categorías predefinidas.

La evolución de la Web hacia la conocida Web 2.0 facilita la máxima interacción entre los usuarios, permitiendo la generación de contenidos donde expresar su opinión. Todos los días, en redes sociales, blogs y foros de discusión los usuarios postean una enorme cantidad de mensajes informales donde expresan su opinión sobre los productos y marcas que consumen. Conocer la polaridad de estas opiniones sobre una marca o producto determinado resulta de principal interés para las compañías. El análisis de sentimientos (en inglés *sentiment analysis*) es el campo que se ocupa de la extracción de opiniones positivas o negativas de un texto no estructurado (Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A.,2010)

El análisis de sentimientos habitualmente sucede en dos o tres fases. Primero, la entrada de texto es dividida en frases y cada frase es analizada para ver si contiene algún sentimiento, clasificando la frase como subjetiva u objetiva. Segundo, las frases subjetivas son analizadas para detectar su polaridad sentimental (positiva o negativa) y, finalmente, se identifica el objeto hacia el cual la opinión es dirigida (Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A.,2010).

Daedalus y *CyberEmotions* son dos referentes a nivel empresarial y académico en el procesamiento de lenguaje natural cuyas tecnologías en este campo han servido de referente en la realización del presente proyecto. A continuación, se describe qué son y a qué se dedican, así como las tecnologías más significativas en el procesamiento del lenguaje natural.

2.2. Daedalus

Daedalus es una empresa constituida en 1998 cuya actividad se centra en torno a las tecnologías de búsqueda, las tecnologías del lenguaje y la gestión avanzada del conocimiento (inteligencia de negocio). *Daedalus* simplifica los escenarios más complejos de procesamiento y análisis de la información mediante sus herramientas de procesamiento lingüístico, semántica e inteligencia de negocio.

Entre sus productos destaca *STILUS*[1], familia completa de productos de tecnología lingüística para realizar cualquier tipo de procesamiento lingüístico que se desee sobre documentos de una gran variedad de idiomas.

2.2.1. Base léxica

Daedalus dispone de una base léxica de gran calidad y cobertura para el castellano. Compuesto por diccionarios, la base léxica además de palabras individuales incorpora más de 27.000 expresiones multipalabra, como por ejemplo "con respecto a", "Juan Carlos I", etcétera. En total, la base léxica da cuenta de más de 130.000 lemas distintos de palabras en castellano que en total dan lugar a más de 6 millones de palabras en castellano (por ejemplo: *"pequeñ+ito"*, *"compándo+se+lo"*, etcétera.). A partir de ellas, el tratamiento de la afijación con prefijos nominales o verbales permite reconocer un número superior a los 15 millones (por ejemplo: *"sobre+actuar"*, *"super+pequeño"*, etcétera.). Por último, esta base léxica es ampliable con diccionarios temáticos a medida, disponiendo STILUS de diccionarios de economía, astronomía, música, etcétera., así como de palabras comunes del español en diferentes zonas de la comunidad hispanohablante.

Daedalus también cuenta con distintas versiones de esta base léxica para otros idiomas como el inglés, francés e italiano.

2.2.2. Recuperación de información

Daedalus integra tecnología de filtrado de documentos que hace posible el reconocimiento de numerosos formatos electrónicos distintos (HTML, PDF, texto, XML, etcétera.), además de interpretar contenidos multimedia como vídeos o ficheros

[1] http://www.daedalus.es/productos/stilus/

de audio como *DALI* (*Digital Audio Library Indexing*) o *CBIR* (búsqueda de imágenes basadas en contenido).

Daedalus también se centra en aumentar la precisión de los resultados de un proceso de búsqueda explotando el etiquetado definido por la web semántica, ampliando el conjunto de palabras empleado para representar un documento, o la explotación de las tecnologías que puedan arrojar información que facilite la interpretación del texto como la extracción de información que permitan representar el texto a través de campos de datos.

2.2.3. Extracción de información

El campo de las tecnologías de la lengua o procesamiento del lenguaje natural es una de las áreas de trabajo más relevantes para *Daedalus*. Entre las líneas en las que este campo puede dividirse destacan la corrección de texto y la extracción de información. La extracción de información incluye algoritmos, métodos y procesos centrados en la identificación de información dentro de un texto. La posibilidad de reconocer automáticamente la aparición de un nombre propio en un texto es una de las aplicaciones más útiles de la extracción de información. *Daedalus* aparte de reconocer estas entidades permite realizar su categorización, distinguiendo cuándo se habla de una persona, una organización o de un lugar.

Daedalus afronta la dificultad de que estas entidades puedan aparecer en diferentes formas, por ejemplo: *Banco Santander Central Hispano* puede también aparecer como *Banco Santander*, *BSCH*, etcétera., o los problemas de la ambigüedad para su clasificación, por ejemplo la entidad *Sevilla* podría hacer referencia a la ciudad o al equipo de fútbol.

El etiquetado semántico del texto se realiza a partir de diccionarios de entidades con nombres. Primero se realiza la segmentación del texto en unidades o *"tokens"* (palabras o entidades multipalabra) y posteriormente se marcan como entidades candidatas aquellas unidades que aparezcan en alguno de los diccionarios de entidades del sistema. Si para una forma se tiene más de una entidad candidata, se realiza una desambiguación basada en heurísticos. Esta desambiguación puede centrarse en la frecuencia de aparición de la entidad en el texto o la presencia de unidades lingüísticas que preceden a la entidad (como preposiciones o artículos).

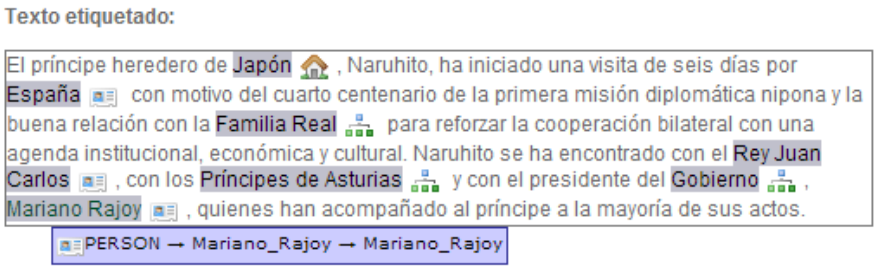


Figura 1. Ejemplo de extracción de entidades con nombre empleando Daedalus

Daedalus además del reconocimiento de entidades con nombre también ofrece identificación de estructuras en un texto como números de teléfono o correos electrónicos, identificación de palabras clave o la elaboración de resúmenes.

2.2.4. Análisis morfosintáctico y sintáctico

Daedalus utiliza el análisis morfosintáctico y sintáctico en la identificación de datos en textos no estructurados.

Dada una palabra o grupo de palabras, se obtienen todos sus análisis morfosintácticos:

- La o las categorías morfosintácticas que puede adoptar la palabra, codificada según el etiquetario de *STILUS*.
- El o los lemas que corresponden a cada categoría gramatical.
- La información semántica de la palabra.
- La forma canónica de la palabra (mayúsculas/minúsculas).

Para no devolver todos los análisis posibles de una palabra, se aplica un proceso de desambiguación para filtrar los análisis inválidos en el contexto donde aparece dicha palabra. Por ejemplo, *"casa"* tiene los siguientes tres análisis: 1) nombre femenino *"casa"*, 2) verbo *"casar"* en 3ª persona del presente o 3) el verbo en imperativo singular. Si tenemos el grupo de palabras *"la casa roja"*, teniendo en cuenta el contexto lingüístico, el análisis como verbo no tendría ningún sentido. Esta desambiguación está basada en reglas.

Además del análisis morfosintáctico, también se dispone de funcionalidad para realizar un análisis sintáctico superficial del texto (por cada una de las frases que componen el texto), detectando sintagmas nominales, verbales, preposicionales o adverbiales, como su posible función dentro de cada frase.

2.3. CyberEmotions

CyberEmotions es un dominio de investigación que estudia el papel de las emociones colectivas en la creación, la formación y la disolución de comunidades electrónicas.

El dominio se centra en el análisis de mensajes online usando metodologías como el análisis de sentimientos para facilitar el acceso a las señales emocionales en muestras grandes. Uno de sus programas, *SentiStength*, permite extraer la fuerza sentimental positiva y negativa de textos electrónicos cortos e informales escritos en inglés. El resultado se muestra en una escala de 5 puntos, de 1 (no positivo) a 5 (extremadamente positivo) para definir el rango de fuerza sentimental positiva identificada en el texto, y de -1 (no negativo) a -5 (extremadamente negativo) para la negativa. El algoritmo es aplicado a comentarios en *MySpace* y es capaz de detectar emociones positivas con un 60% de exactitud y negativas con un 72,8%.

La detección de sentimientos online presenta la dificultad de que en el ciberespacio se ignoran las reglas de gramática y ortografía (Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A.,2010). *SentiStrength* emplea un diccionario de sentimientos con medidas de fuerzas asociadas *(the sentiment word strength list)* y explota una serie de grafías no convencionales reconocidas y otros métodos textuales comunes de expresar sentimientos en el ciberespacio.

La colección está formada por 289 términos positivos y 465 negativos clasificados por su fuerza positiva o negativa con un valor entre 2 y 5 puntos. Los términos son palabras estándar inglesas y otras no estándar que son comunes en MySpace como *"luv"*, *"lol"* o *"muah"*, entre otras. La fuerza sentimental es específica del contexto en el cual la palabra tiende a ser usada. *SentiStrength* emplea un algoritmo de aprendizaje automático para optimizar la fuerza sentimental de las palabras de la lista, en el que se incrementa o decrementa la fuerza en 1 punto dependiendo de cómo se incremente la exactitud de las clasificaciones.

La fuerza de la emoción de una secuencia de palabras puede verse alterada por las palabras que las preceden en el texto. *SentiStrength* recoge una lista de palabras amplificadores que aumentan o reducen la emoción de una secuencia de palabras posteriores, ya sea esta positiva o negativa. Cada palabra de la lista incrementa la fuerza de la emoción 1 o 2 puntos (por ejemplo, las palabras *"very"* o *"extremely"*) o la reduce en 1 punto (por ejemplo, la palabra *"some"*). El algoritmo también cuenta con una lista de palabras negadoras que invierten la polaridad de la emoción de palabras posteriores, incluyendo cualquier palabra amplificadora precedente (por ejemplo, *"very happy"* tiene una fuerza positiva de 4 puntos, pero *"not very happy"* tendrá una fuerza negativa de 4 puntos).

SentiStrength emplea un algoritmo de correción ortográfica que identifica las grafías estándar de palabras que han sido mal escritas por la inclusión de letras repetidas. El algoritmo también tiene en cuenta que el uso de letras repetidas es común para expresar emoción o energía en los comentarios de *MySpace*, por lo que antes de ser corregidas ortográficamente son usadas para aumentar en 1 punto la emoción de las palabras, siempre y cuando sean al menos dos letras adiccionales (una sola letra repetida comunmente aparece por un error tipográfico).

El empleo de emoticonos es común en las redes sociales, por tanto la lista de fuerzas de sentimientos se suplementa con una lista de emoticonos con fuerzas asociadas (positivas o negativas de 2 puntos). Además cualquier frase con marca de exclamación es asignada con un mínimo de fuerza positiva.

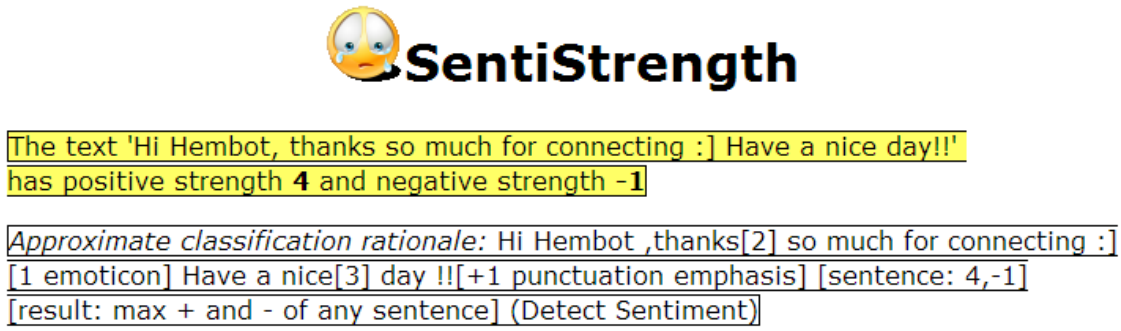


Figura 2. Ejemplo de detección de emociones empleando SentiStrength

Por último, cabe destacar que *SentiStrength* no cuenta con una desambiguación semántica para palabras ambiguas debido a los problemas causados por la gramática no estándar que se emplea en los comentarios y que requieren un esfuerzo altamente computacional. Un ejemplo válido serían las frases *"you rock!!!"* y *"do you listen to rock music?"*. Mientras que la primera frase cuenta con una emoción fuertemente positiva la segunda frase es neutra en emoción.

2.4. Motivación

El sistema se ha diseñado teniendo en cuenta la limitación de recursos al estar encuadrado en el esfuerzo propio de un Proyecto Fin de Carrera. Esta limitación ha servido para cubrir el interés en comprobar si una aproximación con menos base lingüística, y por lo tanto más ágil y susceptible de uso en otras lenguas o contextos semióticos, podría dar resultados razonables.

1) Base léxica

La base léxica de nuestro sistema está diseñada para identificar en el texto únicamente marcas y sentimientos, siendo las entidades de interés categorizadas en diccionarios, por tanto quedará lejos de las dimensiones de la base léxica de la que dispone Daedalus.

2) Recuperación de información

El único formato electrónico que reconoce el sistema es el HTML, por tanto sólo trabajará con documentos en este formato.

3) Extracción de información

El etiquetado semántico que realizará nuestro sistema se asemeja al realizado por *Daedalus* y parte de los diccionarios que conforman el léxico, identificando las marcas y sentimientos de interés contenidos en el texto y siendo la segmentación del texto realizada en palabras.

4) Análisis morfosintáctico y sintáctico

La aproximación con menos base lingüística imposibilita la realización de análisis morfosintácticos y sintácticos que posibilitan la desambiguación semántica en la extracción de información que solucionen el problema de la ambigüedad en la clasificación de entidades del texto.

5) Correción ortográfica

El sistema carece de un algoritmo de correción ortográfica como el utilizado por *CyberEmotions* que solucione las faltas gramaticales y ortográficas que cometen los usuarios al publicar sus opiniones en Internet.

Capítulo 3

Diseño y arquitectura del sistema

Índice Capítulo 3

3.1. Introducción

La arquitectura desarrollada está basada en un sistema en cascada que, tras pasar por una fase de procesamiento lingüístico sobre una serie de artículos online, aplica el algoritmo *Latent Semantic Analysis* para identificar las relaciones existentes entre las marcas y sentimientos recogidos en estos artículos.

El experimento se inicia con una colección de documentos. Esta colección puede ser construida manualmente por un experto o mediante un robot que recorra de manera automatizada la Web recolectando páginas. Se decidió que el sistema contemplase esta última posibilidad, implementando lo que se conoce en el campo de recuperación de la información como rastreador o araña web.

Los documentos se deben transformar de su formato inicial a una representación que pueda ser usada por el algoritmo empleado por el sistema. La representación de los documentos elegida es la del modelo de espacio vectorial, donde se asigna un peso a las palabras que los conforman, siguiendo el modelo *"bags-of-words"* ampliamente utilizado en el campo de análisis de textos.

No todos los elementos que aparecen en los documentos serán útiles ni todos los documentos relevantes para el propósito del experimento (sobre todo si la recolección de documentos es realizada por el rastreador web). Por esta razón, durante la fase de procesado lingüístico, son empleados diccionarios de términos (léxico a utilizar) y umbrales de decisión para realizar esta selección.

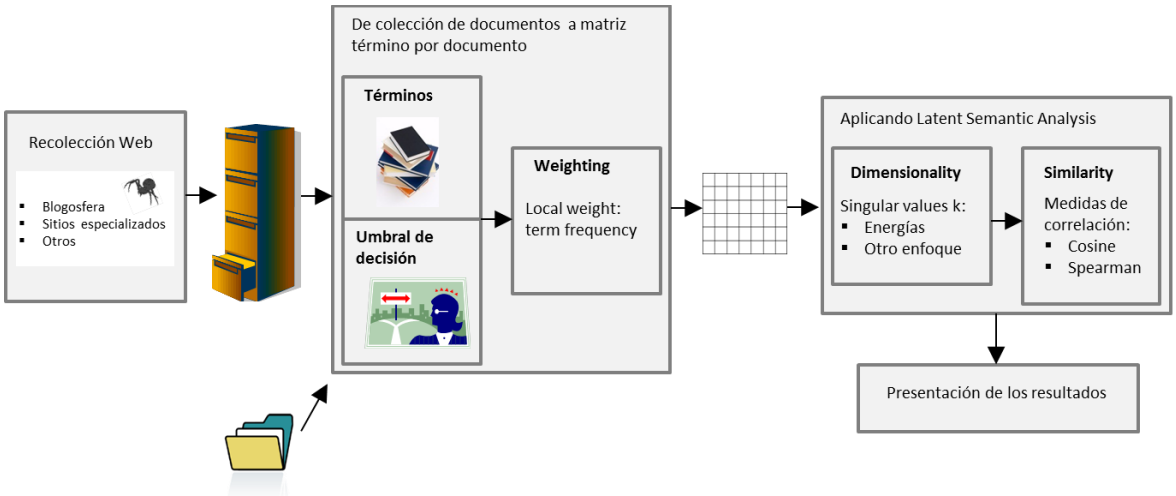


Figura 3. Arquitectura del sistema desarrollado

3.2. Fases del diseño

Las fases imprescindibles en las que puede dividirse el diseño del sistema desarrollado son las siguientes:

- **Recolección de documentos:** La adquisición de documentos puede ser manual, realizada por un experto, o automatizada, mediante una araña web que recolecte páginas de Internet.

- **Procesado lingüístico:** Obtención de una representación adecuada de los documentos para poder ser empleada por el algoritmo *Latent Semantic Analysis*. Se emplea el modelo de espacio vectorial para la representación.

- **Procesado semántico:** Uso del algoritmo *Latent Semantic Analysis* para inferir relaciones semánticas entre los términos del léxico presentes en los documentos y capturar la similitud en significado.

- **Presentación de los resultados:** Representación en un espacio de 2 dimensiones para visualizar los grupos semánticos en los que se agrupan los sentimientos y las relaciones de similitud identificadas entre estos grupos y las marcas de estudio.

3.3. Recolección de documentos

La obtención de la colección de documentos o corpus sobre el que poder realizar el procesado se presenta como punto de partida.

Para facilitar el trabajo se ha implementado en el sistema un rastreador que de manera automatizada recorra la Web recolectando rápida y eficientemente tantas páginas útiles como le sea posible. En el campo de recuperación de la información a estos rastreadores web (en inglés *crawlers*) también se les conoce con el nombre de arañas web o robots.

Un documento es la vista lógica de una página web recolectada cualquiera, estando formada la colección por tantos documentos como páginas ha recolectado el rastrador web. La estructura de acceso que permite una localización rápida de los documentos se denomina índice y para su construcción se ha empleado la tecnología que facilita Apache Lucene.

Apache Lucene utiliza una estructura de índice invertido. La idea general de un índice invertido sería la de una lista de términos donde cada término es asociado con una lista de punteros hacia los documentos donde este aparece. El rastreador web es quien crea el índice de documentos y quien lo alimenta con las páginas que va recolectando durante la operación de rastreo.

La operación comienza con una o más URLs que constituyen el conjunto semilla. Se escoge una dirección URL de este conjunto y se obtiene su página web. La página obtenida es analizada para extraer tanto el texto como los enlaces de la página, donde cada uno de ellos apunta a otra dirección, con los que proseguir con la operación.

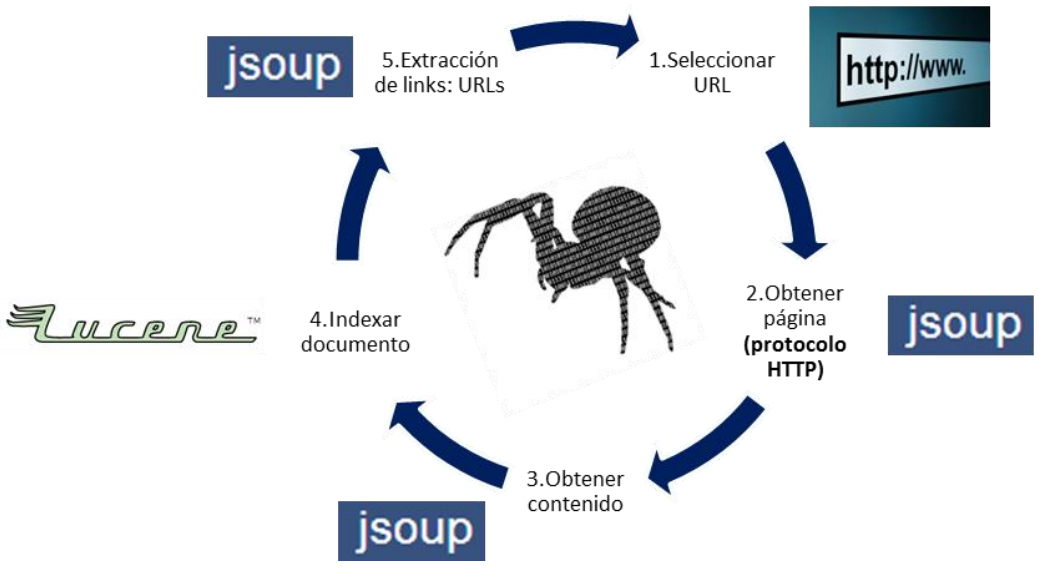


Figura 4. Operación de rastreo

La obtención de la página como la extracción del contenido es realizada con JSOUP. Esta librería permite obtener el documento de hipertexto de la página (no confundir con su vista lógica) a partir de la URL y trabajar con las etiquetas HTML que conforman el documento.

La obtención de la página es realizada mediante el protocolo HTTP. La comunicación HTTP entre el rastreador web y el servidor sería la siguiente:

1. El rastreador web solicita una página (solicitud).
2. El servidor recibe la comunicación y comprueba si existe el fichero solicitado.
3. El servidor envía el fichero (respuesta).
4. El rastreador obtiene y analiza el fichero HTML.

Tanto en la solicitud como en la respuesta, se envía en las cabeceras del HTTP información para el control y especificación de las comunicaciones. En la solicitud se incluye información de control para decir en qué condiciones el rastreador acepta el fichero. Se ha decidido que el rastreador comunique las siguientes condiciones:

Campos cabeceras	Definición	Condiciones solicitadas
Accept	Especifica los tipos de información (*media*) que son aceptables por respuesta.	*text/html*
Accept-Charset	Indica qué conjuntos de caracteres son aceptables por respuesta.	*ISO-8859-1,utf-8*
Accept-Encoding	Indica qué formatos de comprensión son admitidos por respuesta. El servidor que recibe esta cabecera podrá enviar los ficheros solicitados en formato comprimido, reduciéndose el tiempo de obtención de la página. Los ficheros serán enviados comprimidos o no dependiendo de cómo esté configurado el servidor.	*gzip, deflate*

Tabla 1. Condiciones indicadas en cabecera solicitud HTTP

Dado que el presente proyecto pretende trabajar con artículos online sobre un determinado tema o *topic* (artículos de coches), se ha intentado asemejar el rastreador aquí construido con un rastreador dirigido (en inglés *focused crawler*) con el fin de mejorar su precisión. A diferencia de un rastreador de propósito general que recolecta la mayor cantidad de páginas posibles dentro de un cierto intervalo de tiempo, un rastreador dirigido busca selectivamente páginas que son relevantes para un conjunto de temáticas específicas.

El rastreador dirigido ideal recupera el conjunto de páginas relevantes de un dominio atravesando el menor número de documentos irrelevantes. El conjunto semilla que dé comienzo a la operación de rastreo debe estar formado de páginas relacionadas con la temática determinada. Como se quiere recolectar páginas que contengan artículos de coches, un buen conjunto semilla estaría formado por páginas de sitios web especializados en el mundo del motor y la actualidad del automóvil.

El rastreador juzga la relevancia y calidad de la página obtenida en base al tema específico de búsqueda. Durante el proceso de recolección, un clasificador binario de hipertexto clasifica la página como relevante o no relevante en función de unas palabras clave asociadas. Dado que el tema de búsqueda es 'artículos de coches', se decidió asociar a esta temática las palabras clave *coche*, *vehículo* y *motor*. Si ninguna de estas palabras clave es identificada en el texto extraído del documento de hipertexto, la página es clasificada como no relevante y, por tanto, no es indexada. Una posterior evaluación de la precisión demostrada por el rastreador determinará si la elección de las palabras clave fue la más correcta.

Si el rastreador sólo utilizase el texto extraído de las páginas obtenidas para determinar el recorrido a realizar correría el riesgo de perder páginas relevantes. Que una página sea clasificada por el clasificador como no relevante no significa que no pueda tener enlaces a otras páginas que sí están relacionadas con la temática de búsqueda. Por ejemplo, en el rastreo se puede llegar a páginas que presentan poco texto pero que están formadas por un gran conjunto de enlaces relacionados con la temática de búsqueda.

El rastreador tendrá previamente en cuenta otras características de la página para decidir si debe visitar sus hijos. Los metadatos integrados en el documento de hipertexto son un buen punto de partida para realizar este juicio ya que ofrecen información variada sobre la página. Sin embargo, dado que no existe ninguna regla acerca de la información que se puede o no se puede incluir en los *metatags*, el rastreador podría en ocasiones no tener suficiente información para tomar la decisión de descartar o no los enlaces de la página. En estos casos, el rastreador omitirá este control y procederá a extraer el texto y las URLs contenidas en la página.

El idioma del documento, identificable mediante el atributo *lang* de la etiqueta *<html>*, es la primera información que determina si las URLs contenidas en la página deben ser tenidas en consideración. Si el idioma de la página es diferente al español (o al seleccionado por el usuario del sistema en la interfaz) muy posiblemente los enlaces conduzcan a páginas con ese mismo idioma, por tanto tampoco serán relevantes para el clasificador.

```
<!DOCTYPE html>
<html lang="en">
<head>
<title>Swapping Songs</title>
</head>
<body>
<h1>Swapping Songs</h1>
<p>Tonight I swapped some of the songs I wrote with some
friends, who
gave me some of the songs they wrote. I love sharing my
music.</p>
</body>
</html>
```

Figura 5. Ejemplo de identificación de idioma del documento HTML (Fuente: http://www.w3.org/TR/html5/semantics.html#the-html-element)

Otra información a analizar son las palabras clave que resumen de forma significativa el contenido de la página, identificables mediante el atributo *content* de las etiquetas <*meta*> con atributo *name "keywords"*. El rastreador solo tiene en cuenta las palabras clave de la página de inicio de la web de un sitio ya que en esta circunstancia ofrecen un resumen más global del contenido ofrecido en el sitio. Por ejemplo, en la página de inicio de la web del sitio *Hobby Consolas*[2] se identifican como palabras clave las palabras *videojuegos*, *revista de videojuegos* o *videoconsolas*, por tanto, al ser improbable encontrar en este sitio páginas con artículos de coches, las URLs contenidas en esta página son ignoradas por el rastreador. El atributo *"description"* ofrece un resumen más amplio del contenido de la página que el atributo *"keywords"* y sería igual de válido para realizar esta comprobación.

```
<head>
  <title>World Wide Web Consortium (W3C)</title>
  <meta http-equiv="Content-Type" content="text/html;
charset=utf-8" />
  <meta http-equiv="Content-Style-Type" content="text/css"
/>
  <meta name="copyright" content="© W3C" />
  <meta name="author" lang="en" content="" />
  <meta name="robots" content="Index,Follow" />
  <meta name="description"
        content="The World Wide Web Consortium (W3C) is an
international community
        where Member organizations, a full-time staff,
        and the public work together to develop Web
standards." />
  <meta name="keyword" content="W3C, HTML, CSS, SVG, Web
standards" />
</head>
```

Figura 6. Ejemplo de identificación de etiquetas <meta> del documento HTML (Fuente: http://www.w3.org/wiki/HTML/Elements/meta)

[2] http://www.hobbyconsolas.com/

Las páginas marcadas como relevantes por el clasificador binario son añadidas al índice. El documento a indexar es creado con Apache Lucene y en él se han definido los siguientes campos a rellenar:

- URL: URL de la página.
- Padre: URL de la página padre.
- Profundidad: Profundidad de búsqueda en la cual la página ha sido indexada.
- Título: Contenido de la etiqueta <*title*>.
- Texto: Concatenación de los contenidos de las etiquetas <*p*> {CAMPO INDEXADO}.

El índice de documentos se construye con los términos del campo *Texto*, siendo este campo el único indexado. Los campos *URL*, *Padre*, *Profundidad* y *Título*, serán metadatos asociados al documento, es decir, campos almacenados pero no indexados. Apache Lucene segmenta el texto del campo a indexar en unidades (términos), suprime los signos de puntuación y realiza la conversión a letras minúsculas. Este pre-procesamiento de texto es importante en el presente proyecto en el momento de realizar el procesamiento lingüístico de los documentos del índice.

Doc_Id	2419
URL	http://www.autobild.es/pruebas/lamborghini-aventador-prueba-176227
Padre	http://www.autobild.es/sportscars
Profundidad	2
Título	Lamborghini Aventador: el mejor de todos - Autobild.es
Texto	Rss Feed Facebook Encuéntranos en Google+ Twitter Pinterest Tuenti Desenvaina el cronómetro. En el tiempo en el que lees estas dos frases, el Lamborghini Aventador LP700-4 ya vuela a 200 km/h (aquí tienes el vídeo del Aventador en el 0 a 100 km/h). Exactamente han sido 8,9 segundos, durante los cuales tus ojos han ido de la primera palabra a la última y tu sonrisa delata que sabes de qué estoy hablando. Quien pase estos 8,9 segundo a bordo de este Lambo, sentirá muchas cosas. Notará que está en el séptimo cielo, pero a la vez destilará tensión a causa de la emoción. Porque la manera en que acelera el nuevo Lamborghini, cómo ataca tu sistema nervioso, cómo resopla, restalla, cómo aúlla de éxtasis desde la cámara de combustión de seis litros y medio... Esto es drama, comedia y thriller. Y lo es todo al mismo tiempo. Un teatro de aceleración para todos los sentidos, con la coreografía a cargo de los expertos en movimiento de Sant 'Agata Bolognese, que muestran su locura por la gasolina en todo su esplendor, para deleite del mundo entero. Dejando a un lado, claro, que una entrada para este espectáculo cuesta 338.513 euros. Así ha sido siempre y así será esta vez. Aunque sí que hay algo que es distinto en el nuevo Aventador. Como es lógico, es más rápido que su predecesor el Lamborghini Murcielago, tiene más potencia y un diseño aún más espectacular, brutal y arrebatador al mismo tiempo. (…)

Figura 7. Ejemplo de documento indexado por el rastreador

La etiqueta *<a>* define un hiperenlace, el cual es usado para enlazar una página con otra. El atributo *href* de la etiqueta indica el link de destino.

```
<BODY>
...some text...
<P>You'll find a lot more in <A href="chapter2.html"
title="Go to chapter two">chapter two</A>.
<A href="./chapter2.html"
title="Get chapter two.">chapter two</A>.
See also this <A href="../images/forest.gif"
title="GIF image of enchanted forest">map of
the enchanted forest.</A>
</BODY>
```

Figura 8. Ejemplo de hiperenlaces en documento HTML
(Fuente: http://www.w3.org/TR/html4/struct/links.html)

Los enlaces extraídos (URLs) son añadidos a la frontera URL, que en todo momento se compone de las direcciones URL cuyas páginas correspondientes aún no han sido recolectadas por la araña. A medida que las páginas son obtenidas, las correspondientes URLs son borradas de la frontera y añadidas a una lista auxiliar de URLs cuyas páginas ya han sido obtenidas. Inicialmente la frontera URL contiene el conjunto semilla. La frontera URL sigue una estructura de cola *FIFO* (*First In, First Out*) cogiendo el rastreador la URL con más antigüedad en la frontera. La mayoría de los rastreadores dirigidos priorizan las URLs de las páginas a descargar teniendo en cuenta factores como la relevancia que el clasificador (binario, *Naïve Bayes* u otro) asoció a la página de origen (página padre), el texto del enlace o el texto en las proximidades del enlace, entre otros.

No todos los enlaces de una página son extraídos. Cada URL extraída pasa una serie de pruebas para determinar si debe ser añadida a la frontera URL. Se ha creado un filtro URL para excluir enlaces según el tipo de documento al que conducen en base al dominio de la URL. En este filtro son desechados enlaces orientados a un tipo de documento diferente al de hipertexto como formatos de documento portatiles (*.pdf*), imágenes (*.jpg*, *.jpeg*, *.gif*, *.png*) , archivos de áudio (*.mp3*, *.mp4*) o archivos de vídeo (*.avi*, *.mov*), entre otros. El filtro ha sido construido mediante una expresión regular obtenida del rastreador *crawler4j*[3] de código abierto. Se decidió además complementar el filtro para que deseche también aquellas URLs según las palabras contenidas en ella que hagan pensar que se tratan de direcciones de correo electrónico (símbolo @), descargas (*/download/*) o código *javascript* (*javascript:*). También se debe de evitar duplicados en la frontera. La URL extraída es normalizada para obtener la URL canónica de la página (URL única de la página). Todas las URLs son añadidas en la frontera en su forma canónica. Si la URL está ya en la frontera o su página ha sido ya obtenida entonces no es añadida.

[3] https://code.google.com/p/crawler4j/

El número de páginas a recolectar por el rastreador viene dado por la profundidad de búsqueda indicada por el usuario del sistema y la cual es asociada a cada una de las URLs que constituyen el conjunto semilla. Una vez alcanzada dicha profundidad no se añadirán más URLs a la frontera. El rastreador finaliza su labor cuando la frontera URL es vaciada por completo.

En base a la información descrita en párrafos anteriores, la arquitectura del rastreador queda definida como se muestra en la siguiente ilustración:

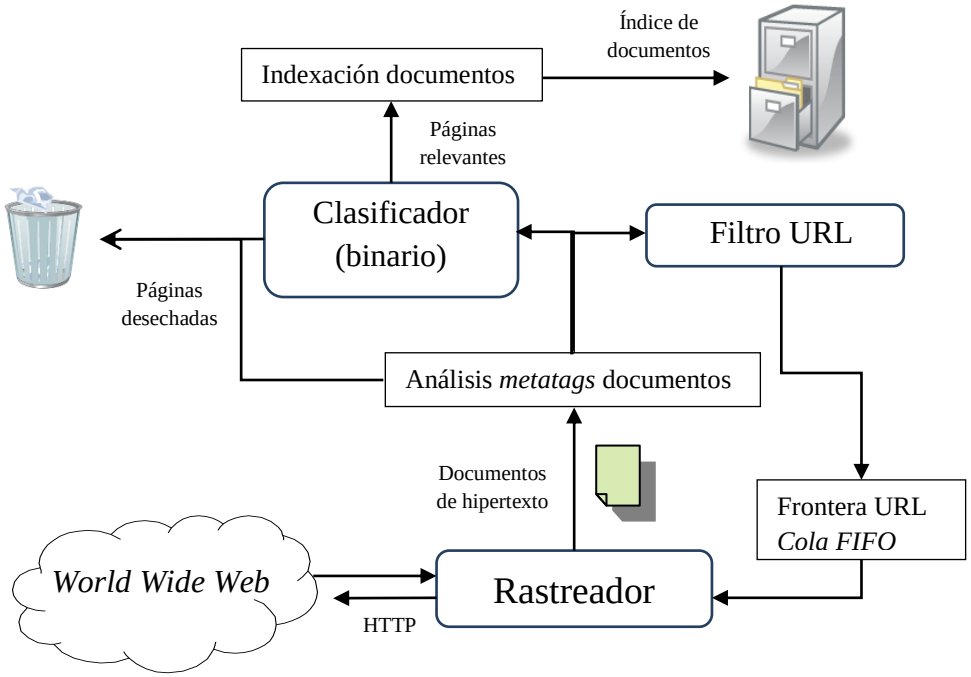


Figura 9. Arquitectura del rastreador dirigido

3.3.1. Evaluación

Se ha establecido la evaluación del rastreador en tres métricas: velocidad de recolección, relevancia (precisión) y cobertura (cobertura) de los recursos descubiertos.

Es importante medir la velocidad en la cual las páginas relevantes para el clasificador binario son adquiridas, y como, efectivamente, las páginas irrelevantes son descartadas. Esta velocidad de recolección debe ser alta, o de otro modo el rastreador emplearía bastante tiempo meramente descartando páginas irrelevantes (Soumen Chakrabarti, Martin van den Berg, and Byron Dom, 1999). Debido a que la carga de la red es un factor que influye en la velocidad de recolección y que varía enormemente a lo largo del día, se ha decidido no presentar en los resultados el tiempo como el tiempo que emplea el rastreador en realizar su tarea, sino como el número total de URLs extraídas en la operación de rastreo.

Cogemos el mismo conjunto semilla de URLs que el utilizado en la realización de los experimentos y lanzamos dos ejecuciones. La primera ejecución aplica una profundidad de búsqueda igual a uno sobre cada URL del conjunto semilla, mientras que la segunda aplica una profundidad de dos.

Profundidad de búsqueda	Páginas recolectadas	URLs recuperadas	Velocidad de recolección
1	305	497	**0,614**
2	5301	11561	**0,459**

Tabla 2. Velocidad de recolección del rastreador para diferentes niveles de búsqueda

De este resultado podemos extraer que, aproximádamente, de cada dos URLs que extrae nuestro rastreador , una de ellas es irrelevante para el clasificador. También se observa cómo el rastreador diverge según aumenta la profundidad de búsqueda.

Los resultados mostrados en la tabla 2. fueron extraídos del fichero de traza generado por el rastreador en su ejecución. En este fichero se pueden identificar las direcciones URLs de todas las páginas que el rastreador intentó recolectar y el resultado final de esta operación. Los resultados han sido clasificados en diferentes categorías dependiendo del éxito o no éxito de la operación y del tipo de error localizado.

Resultado	Descripción
OK	La página ha sido marcada como relevante por el clasificador y añadida al índice con éxito.
Casos de error creados	
Non-relevant page error	La página ha sido marcada como no relevante por el clasificador.
Language error	El idioma identificado en la página difiere al deseado.
Keywords error	Las palabras clave identificadas en la página difieren a las asociadas al tema de búsqueda.
Content not found error	No se ha encontrado texto a extraer de la página.
HTTP Protocol Error Codes [4]	
Client Error 4xx	Se aplica a los casos en los que el cliente parece haber errado.
Server Error 5xx	Indica los casos en que el servidor es consciente de que ha cometido un error o es incapaz de realizar la solicitud.

Tabla 3. Listado de posibles resultados en la recolección de una página

La precisión y la cobertura son otros dos indicadores del rendimiento de un rastreador dirigido. La precisión puede ser definida como la fracción de páginas recolectadas de la Web que son relevantes para el propósito final del proyecto, mientras que la rellemada es definida como la fracción de páginas relevantes de la Web que son recolectadas.

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved})$$

Figura 10. Cálculo precisión

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(\text{retrieved}|\text{relevant})$$

Figura 11. Cálculo cobertura

[4] http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html

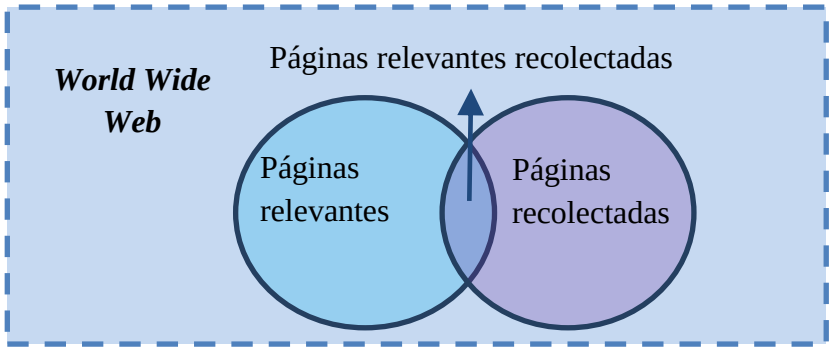


Figura 12. Concepto de precisión y cobertura en World Wide Web

Es difícil medir la cobertura para un rastreador dirigido al no tener una buena noción de qué es una "buena cobertura" para un tema de búsqueda (Soumen Chakrabarti, Martin van den Berg, and Byron Dom, 1999). En el presente proyecto ha bastado con una profundidad de búsqueda de igual a dos para conseguir el número de documentos suficientes para su consecución.

Lo ideal sería juzgar la precisión del rastreador por inspección humana, sin embargo, al recuperar el rastreador miles de páginas esto es imposible. Como en última instancia es el procesador lingüístico el que determina qué documentos de la colección son relevantes para el propósito del proyecto, por medio de un umbral de decisión fijado y el léxico construído por el usuario del sistema, éste será el clasificador automático que trabaje sobre la colección de documentos obtenida por el rastreador.

La precisión que asigne el procesador al rastreador variará significativamente según el número de términos que conformen el léxico. Realizamos dos evaluaciones, la primera empleando en el procesador lingüístico el mismo léxico y umbral que en la realización del experimento grande. La segunda, y dado que el tema de búsqueda del rastreador eran artículos de coches, modificamos el umbral, ignorando los sentimientos, marcando como relevante todo aquel documento que contenga como mínimo una de las marcas automovilísticas del léxico.

	Documentos colección	Documentos relevantes	Precisión
Evaluación #1	5301	543	**0,10**
Evaluación #2		4609	**0,87**

Tabla 4. Precisión del rastreador para diferentes umbrales de decisión

3.4. Procesado lingüístico

Debido a que los documentos no pueden ser interpretados directamente por el algoritmo, se hace necesaria una representación compacta de cualquier documento.

En el campo de recuperación de la información la representación más utilizada es el modelo de espacio vectorial. Este modelo parte del modelo "bag of words" en el que un texto puede ser representado como un vector de palabras que aparecen en él. Las palabras del texto que se consideran apropiadas para ser características del vector constituyen el léxico.

Primero se construye un léxico donde cada término es una característica. Qué palabras son consideradas apropiadas y por tanto deben formar el léxico es uno de los desafios que presenta este modelo. En el presente proyecto se ha considerado que el léxico esté formado por los siguientes diccionarios: 1) Marcas, 2) Sentimientos, donde los términos que los conforman son dictaminados por el usuario del sistema.

Cada documento es entonces representado mediante un vector de pesos, di= {w1j, w2j, w3j, … , wTj}, donde T es el conjunto de términos. Este vector es denominado vector documento. El peso asignado a cada posición i del vector debe representar la importancia del término i en el documento j. En el modelo "bag of words" la ordenación exacta de las palabras en un documento es ignorado, no así el número de apariciones de cada palabra (Christopher D. Manning, et al., 2008). La representación del vector documento no sería posible sin la previa construcción del índice invertido, que permite una localización rápida de los documentos. Con el pre-procesado de texto realizado por Apache Lucene en la indexación de documentos, la frecuencia de aparición de los términos del léxico en cada documento puede ser obtenida.

En el presente proyecto se ha considerado asignar a cada término un peso que depende del número de apariciones del término en el documento. Este sistema de ponderación es conocido como *term frecuency*, pero no es el único. La asignación de pesos puede ser clasificada como local, si en la asignación solo es tenido en cuenta el presente documento, o global, si lo es el conjunto de documentos.

A continuación se detallan diferentes sistemas para determinar el peso de cada posición en el vector documento:

- Asignación de pesos local:
    - Binaria: En cada posición del vector se indica la presencia (1) o no presencia (0) en el documento del término correspondiente a esa posición.

- Frecuencia de aparición: Es el sistema empleado en el presente proyecto. En cada posición del vector se indica el número de apariciones en el documento del término correspondiente a esa posición.

- Asignación de pesos global:

  - TF-IDF (Frecuencia de aparición – Frecuencia inversa del documento): Tiene en cuenta la frecuencia de aparición general del término. Esta medida es útil para determinar qué palabras del "bag of words" son más importantes y por tanto consideradas apropiadas para formar el léxico de términos. La frecuencia de aparición de la palabra en un documento es normalizada por el número total de palabras en él (TF). Para reducir la importancia de aquellas palabras que son muy comunes en el conjunto de documentos se establece que la importancia de cada palabra es inversamente proporcional al número de documentos donde aparece. El objetivo es que la relevancia de estas palabras sea mínima y por tanto no deban ser consideradas características de ningún vector.

Si se ponen los vectores documento en las columnas de una matriz se obtiene la matriz término por documento, representación necesaria para iniciar el algoritmo *Latent Semantic Analysis*. De la misma manera, las filas de esta matriz pueden ser vistas como vectores característicos de los términos. El número de filas corresponde siempre al número de términos, es decir, al tamaño del léxico.

D/T	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7
t1	4	1	6	1	0	8	1
t2	1	6	0	2	6	0	4
t3	0	5	3	0	3	7	0

Figura 13. Ejemplo matriz término por documento empleando frecuencia de palabra

Como no todos los documentos son relevantes para el propósito final del proyecto, el sistema establece un umbral de decisión que determina qué vectores documento forman la matriz término por documento. Si las frecuencias de aparición de los términos del léxico en el documento no superan el umbral fijado, el documento es marcado como no relevante y su vector documento es, por tanto, desechado. Por ejemplo, se podría implementar un umbral que garantice que todos los documentos de la matriz tienen al menos en su texto un término del diccionario de marcas y otro término presente en el diccionarios de sentimientos.

3.5. Procesado semántico

La representación del vector espacio sufre de su incapacidad de afrontar los dos problemas clásicos derivados en los lenguajes naturales: sinonimia y polisemia (Christopher D. Manning, et al., 2008). Si dos documentos contienen diferentes palabras pero con significados similares, sus vectores característicos pueden llegar a ser muy diferentes.

"The Lord of the Rings is a **fantastic movie**"

"The Lord of the Rings is an **excellent film**"

Cada una de estas dos parejas de sinónimos puede ser representada como una característica simple. Utilizar una base de datos léxica como *WordNet*, que contiene relaciones entre palabras similares, podría solucionar esta limitación. Sin embargo, esta solución sería muy costosa en recursos y por tanto ha sido desechada.

Latent Semantic Analysis (LSA) es el algoritmo utilizado por el sistema y tiene el objetivo de capturar la similitud en significado, asumiendo que palabras que aparecen en contextos similares a menudo estarán relacionadas. Este algoritmo usa la *Singular Value Descomposition (SVD)* de la matriz término por documento para analizar la relación estadística entre las palabras del conjunto de documentos (Turney, P., & Littman, M., 2003). En *SVD*, una matriz rectangular es descompuesta en el producto de otras tres matrices que, cuando son multiplicadas matricialmente, la matriz original es reconstruída (Landauer, T. K., Foltz, P. W., & Laham, D., 1998).

La descomposición es escrita:

$$A = U\Sigma V^T,$$

Donde U es la matriz ortogonal $t \times t$ que tiene los valores singulares izquierdos de A como sus columnas, V es la matriz ortogonal $d \times d$ que tiene los valores singulares derechos de A como sus columnas (donde V^t es la matriz traspuesta de V) y Σ es la matriz diagonal $t \times d$ que tiene los valores singulares de A en orden decreciente a lo largo de su diagonal (Michael W. Berry, et al., 1999).

La dimensionalidad de la solución puede ser reducida simplemente eliminando coeficientes de la matriz diagonal, comenzando con los más pequeños (Landauer, T. K., Foltz, P. W., & Laham, D., 1998). Cómo elegir el rango que proporciona un desarrollo óptimo del *LSA* para cualquier colección de datos sigue siendo una pregunta abierta y es normalmente determinado empíricamente (Michael W. Berry, et al., 1999). La reconstrucción basada en una dimensionalidad reducida será aproximada a la matriz original. El paso de reducción de la dimensionalidad colapsa las matrices de *SVD* de tal

modo que palabras que ocurren en varios contextos aparecezcan con una mayor o menor frecuencia estimada, y algunas que no aparecen originariamente, ahora aparezcan, aunque sea fracionalmente (Landauer, T. K., Foltz, P. W., & Laham, D., 1998). Una menor frecuencia estimada puede reflejar el hecho de que esta palabra no es esperada en este contexto y por lo tanto debería ser cuantificada como no importante en la caracterización del vector documento.

Una vez que las relaciones indirectas son inferidas por la *SVD* de la matriz, las relaciones entre los términos son entonces imputadas por correlaciones. El modelo de espacio vectorial no es solo un mecanismo para comparar búsquedas (querys) con documentos o documentos con documentos, sino también puede ser usado para comparar términos con términos (Michael W. Berry, et al., 1999). En este modelo la manera estándar de cuantificar la similitud es mediante la correlación coseno que viene determinada por la medida del coseno del ángulo entre dos vectores representativos. La razón principal es que dos documentos con contenido muy similar pueden tener un vector representativo diferente porque un documento sea mucho mas largo que el otro. Así, la distribución relativa de los términos puede ser idéntica en los dos documentos, pero las frecuencias de los términos absolutas de uno puede ser mucho más grandes (Christopher D. Manning, et al., 2008). Los sistemas de recuperación de la información calculan las similitudes coseno entre el vector consulta (query) y cada uno de los vectores documentos de la colección, ordenan las puntuaciones resultantes y seleccionan los documentos con mayor puntuación para esta consulta (claro ejemplo es el motor de búsqueda *Google* y su página de resultados).

En el presente proyecto se ha considerado dar otro enfoque distinto al modelo de espacio vectorial para cuantificar la similitud entre los términos del léxico. La correlación aquí utilizada es la correlación de *Spearman*, basada en la medida de la fuerza de la relación monotónica entre pares de datos. Así pues, las filas de la matriz término por documento dejan de ser vistas como vectores representativos de los términos, ahora son solo conjuntos de datos. Para entender la correlación de *Spearman* es necesario conocer que una función monotónica es una función que o bien nunca aumenta o bien nunca disminuye a medida que aumenta la variable independiente.

En ambas correlaciones, una correlación entre dos términos cercana a 1 indica que los términos ocurren en contextos de similar significado, por tanto son considerados términos similares. Los términos similares del léxico son entonces agrupados en grupos de términos semánticamente independientes en la fase de presentación del sistema. Una correlación entre dos términos cercana a -1 indica que los términos no solo no son similares, sino que son opuestos. Por último, una correlación cercana a 0 indica una correlación débil. El resultado de calcular la correlación entre cada par de términos es la obtención de la matriz de similitudes, que origina la fase de presentación de resultados del sistema, siendo una matriz cuadrada $t \times t$ con valores comprendidos entre -1 y 1 y con todos los valores de su diagonal principal iguales a 1.

3.6. Presentación de los resultados

El proceso de agrupar términos acorde a su contenido relacionado en grupos semánticos independientes de términos es conocido como agrupación (Michael W. Berry, et al., 1999). La entrada clave para un algoritmo de agrupación es la medida distancia (Christopher D. Manning, et al., 2008). La matriz de similitudes es transformada en una matriz de distancias siguiendo el siguiente algoritmo: *dist(i,j)=1-sim(i,j)*, con valores comprendidos entre 0 y 2, siendo iguales a 0 los valores de su diagonal principal.

La presentación del resultado es entonces llevada al espacio semántico. Este espacio semántico puede ser visto como un mapa conceptual donde analizar y entender las percepciones del consumidor sobre distintos productos. En el presente proyecto se ha empleado el escalado multidimensional (en inglés, *multidimensional scaling MDS*) utilizado habitualmente en marketing para la visualización y explotación de datos. Aplicando el escalado multidimensional sobre la matriz de distancias se obtiene una configuración de puntos en 2 dimensiones, donde cada uno de estos puntos representa un término. Los puntos son dibujados en el espacio visualizando cómo los sentimientos se agrupan en grupos semánticos independientes y las relaciones de similitud existentes entre estos grupos y las marcas de estudio.

Capítulo 4

Experimentos y resultados

Índice Capítulo 4

4.1. Planteamiento

De acuerdo al diseño y desarrollo del sistema comentado en el apartado anterior, se va a detallar el planteamiento seguido en la realización del experimento. El principal objetivo del experimento reside en la visualización de las relaciones de similitud existentes entre las principales marcas automovilísticas del mercado y los sentimientos comúnmente empleados en este sector.

En primer lugar fue necesario contar con un conjunto de documentos con los que desarrollar el experimento. En nuestro caso era esencial que este corpus estuviese formado por artículos de coches escritos en español. Sin embargo, como no se encontró ningún recurso de la naturaleza requerida para este idioma se optó por la generación de un corpus propio. Este corpus debía estar formado por cientos (incluso miles) de documentos válidos para los intereses del proyecto por lo que se descartó su construcción manual al ser la carga de trabajo demasiado excesiva. Finalmente, se decidió que la recolección de documentos fuera automatizada, por lo que se desarrolló el rastreador web descrito en el **apartado 3.3**. La construcción de este rastreador otorga grandes beneficios al usuario final al permitirle generar tantos corpus de documentos como él quiera en un tiempo razonable.

La identificación de páginas interesantes desde donde iniciar la operación de rastreo fue el siguiente paso. La búsqueda se centró en encontrar en la Web sitios especializados en el mundo del motor y la actualidad del automóvil que contasen con un número elevado de artículos de coches. También interesaba que en estos sitios se recogiesen repetidamente en sus contenidos expresiones sentimentales hacia la marca automovilística sobre la que se escribe. Otros aspectos tenidos en cuenta son el *PageRank*[5] del sitio que, de forma númerica, indica su relevancia dentro de la Web o el número de usuarios únicos[6]. La evaluación de la precisión del rastreador detallada en el **apartado 3.3.1** es una buena forma de determinar si las páginas seleccionadas para iniciar el proceso de rastreo fueron las más adecuadas.

SITIOS DE REVISTAS ESPECIALIZADAS	PUBLICACIONES DIGITALES
▪ Autobild ▪ Marca Motor	▪ DiarioMotor ▪ 8000vueltas ▪ Auto10

Figura 14. Sitios web elegidos para iniciar la recolección de documentos

[5] http://www.whatsmypr.net/
[6] http://www.infocoches.com/ranking-coches/

Con el fin de reforzar las relaciones semánticas entre sentimientos se optó por la posibilidad de construir corpus de apoyo. Estos corpus de apoyo estarían formados por textos que contuvieran gran fuerza sentimental y se utilizarían en los experimentos junto al corpus principal para la construcción de la matriz término por documento. Se decidió generar uno de esos corpus sobre el sitio web *Traveler*[7] empleando el rastreador web. Este sitio es una revista de viajes viajes online que cuenta con numerosas guías de viaje y artículos sobre visitas turísticas de todas partes del mundo, que recoge un gran número de expresiones sentimentales positivas en sus contenidos.

La construcción de un léxico se considera vital para la extracción de la información a realizar durante el procesado lingüístico de los documentos detallado en el **apartado 3.4**. Este léxico puede ser dividido en distintas categorías o diccionarios. En el presente proyecto se decidió que el léxico estuviera constituido por un diccionario de marcas y un diccionario de sentimientos. Los términos que constituyen estos diccionarios son seleccionados manualmente y harán referencia a las marcas automovilísticas y sentimientos que se desean detectar en el texto no estructurado de cada documento en el procesamiento lingüístico. Cabe destacar que en muchos de los artículos de opinión se recogen comentarios introducidos por los usuarios del sitio. Esto significa que los textos de los documentos del corpus pueden contener faltas de ortografia sobre las que nada puede hacer el procesador lingüístico (sería necesario el uso de un corrector ortográfico para solventar estas faltas).

Se decidió dejar construído un diccionario completo de marcas con las 58 marcas automovilísticas más importantes del mercado[8]. Sobre este diccionario serían construídos pequeños subconjuntos, donde cada subcojunto constituiría el diccionario de marcas en cada una de las fases de las que se compone el experimento. Nuestra principal base de sentimientos será extraída del *JDPA Sentiment Corpus for the Automotive Domain*, guía de cómo se deben anotar las entidades y las estructuras sentimentales que acompañan a estas entidades, con un corpus formado por 457 documentos que contienen opiniones sobre coches en inglés extraídas de sitios como *MySpace*, siendo anotadas manualmente sobre estas opiniones: 1) las entidades y sus relaciones y 2) las expresiones sentimentales hacia estas entidades. Estas expresiones sentimentales son etiquetadas en 4 valores potenciales de polaridad: positiva, negativa, neutral y mixta (Kessler et al., 2010). Al igual que con las marcas, sobre esta base de sentimientos fueron construídos pequeños subconjuntos, donde cada subcojunto constituiría el diccionario de sentimientos en cada una de las fases de las que se compone el experimento.

Al final del procesado lingüístico cada documento del corpus quedaría representado como un vector de pesos:

Doc_1= {tf(Marca_1, Doc_1) , …, tf(Marca_M, Doc_1), tf(Sentimiento_1, Doc_1), …, tf(Sentimiento_S, Doc_1)};

[7] http://www.traveler.es/
[8] http://www.autobild.es/coches

Siendo *M* el número total de marcas automovilísticas del diccionario de marcas, *S* el número total de sentimientos del diccionario de sentimientos y cada posición del vector documento el número de apariciones en el documento del término correspondiente a esa posición.

No todos los documentos del corpus serán relevantes para el propósito del proyecto, siendo el umbral de decisión establecido en el procesador lingüístico quien determine qué vectores documentos conforman la matriz término por documento. Dependiendo del tipo de corpus, principal o de apoyo, se decidió establecer diferentes umbrales de decisión.

CORPUS PRINCIPAL	**CORPUS DE APOYO**
▪ 1 o + marcas automovilísticas del diccionario marcas presentes en el texto del documento. ▪ Igual a 2 o mayor el número de apariciones de sentimientos del diccionario de sentimientos en el texto del documento.	▪ 2 o + sentimientos del diccionario de sentimientos presentes en el texto del documento.

Figura 15. Umbrales de decisión establecidos en cada tipo de corpus

Los vectores documento que superen el umbral de decisión fijado serían las columnas de la matriz término por documento a construir.

D/T	Documento_1	…	Documento_N
Sentimiento_1	tf(s1,doc1)	…	tf(s1,docN)
.	.		.
.	.	…	.
.	.		.
Sentimiento_S	tf(sS,doc1)	…	tf(sS,docN)
Marca_1	tf(m1,doc1)	…	tf(m1,docN)
.	.		.
.	.	…	.
.	.		.
Marca_M	tf(mM,doc1)	…	tf(mM,docN)

Figura 16. Representación final de matriz término por documento

El procesador lingüístico construye la matriz término por documento correspondiente de un corpus. Si se quiere que esta matriz esté compuesta por los documentos de varios corpus bastaría con procesar cada uno de ellos por separado y después concatenar las matriz término por documento resultantes.

En el **apartado 3.5.** se mencionaron los problemas que presenta la representación del vector espacio con la sinonima y polisemia de palabras, donde los vectores característicos de dos documentos pueden llegar a ser muy diferentes aunque contengan términos con significados similares. Para capturar esta similitud en significado, y evitar tener que utilizar bases de datos léxicas como *WordNet* que complementen al léxico construido, se decidió aplicar el algoritmo *Latent Semantic Analysis (LSA)* sobre la matriz término por documento. Posteriormente, mediante la correlación de *Spearman*, las relaciones entre los términos serían imputadas obteniendo la matriz de similitudes. Dado que un término será idéntico a si mismo, la diagonal principal de esta matriz es rellenada automáticamente con 1's.

T/T	Sentimiento_1	…	Sentimiento_N	Sentimiento_1	…	Sentimiento_N
Sentimiento_1	1	…	[-1,1]	[-1,1]	…	[-1,1]
.	.		.	.		.
.	.	…	.	.	…	.
.	.		.	.		.
Sentimiento_S	[-1,1]	…	1	[-1,1]	…	[-1,1]
Marca_1	[-1,1]	…	[-1,1]	1	…	[-1,1]
.	.		.	.		.
.	.	…	.	.	…	.
.	.		.	.		.
Marca_M	[-1,1]	…	[-1,1]	[-1,1]	…	1

Figura 17. Representación de matriz de correlaciones

Se pensó que la manera más adecuada de presentar los resultados del experimento sería el espacio semántico como se explica en el **apartado 3.6.**. En este espacio se visualizaría la similitud existente entre marcas y sentimientos. Previamente, la matriz de similitudes tendría que ser transformada en una matriz de distancias donde se aplicaría un escalado multidimensional para obtener una configuración de puntos en 2 dimensiones (coordenadas X, coordenadas Y). Cada punto representaría la localización de un término en el espacio semántico.

T/T	Sentimiento_1	…	Sentimiento_N	Sentimiento_1	…	Sentimiento_N
Sentimiento_1	0	…	[0,2]	[0,2]	…	[0,2]
.	.		.	.		.
.	.	…	.	.	…	.
.	.		.	.		.
Sentimiento_S	[0,2]	…	0	[0,2]	…	[0,2]
Marca_1	[0,2]	…	[0,2]	0	…	[0,2]
.	.		.	.		.
.	.	…	.	.	…	.
.	.		.	.		.
Marca_M	[0,2]	…	[0,2]	[0,2]	…	0

Figura 18. Representación de matriz de distancias

4.2. Metodología

El corpus principal completo construído para la realización del experimento consta de un total de más de 5.000 documentos. A su vez, el corpus auxiliar consta de no más de 400 documentos.

La estructura del corpus empleado en el experimento puede ser vista en la siguiente figura:

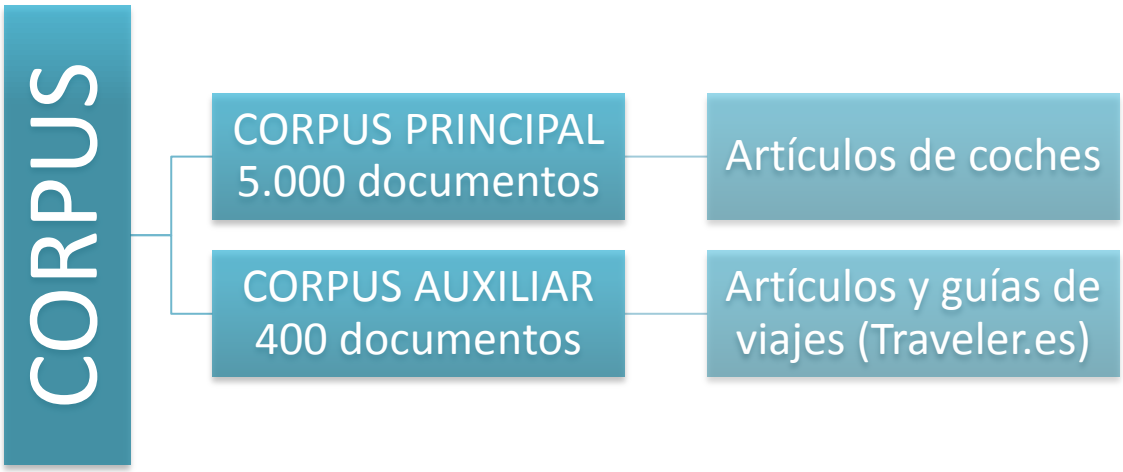


Figura 19. Estructura del corpus empleada en la realización del experimento.

Los sentimientos que constituyen el diccionario de términos del léxico serán extraídos del *JDPA Sentiment Corpus*. La polaridad de estos sentimientos será únicamente positiva. La transcripción al castellano ha sido realizada mediante *Google Translator*. Recordamos que las expresiones sentimentales han sido anotadas manualmente en el *JDPA Sentiment Corpus* con esta clasificación por sus autores en función del contexto de la oración. Por tanto, tendremos que ser meticulosos dado el momento de seleccionar los sentimientos que formarán parte del diccionario de no escoger términos que puedan resultar ambiguos.

La estructura del léxico empleada en el experimento puede ser vista en la siguiente figura:

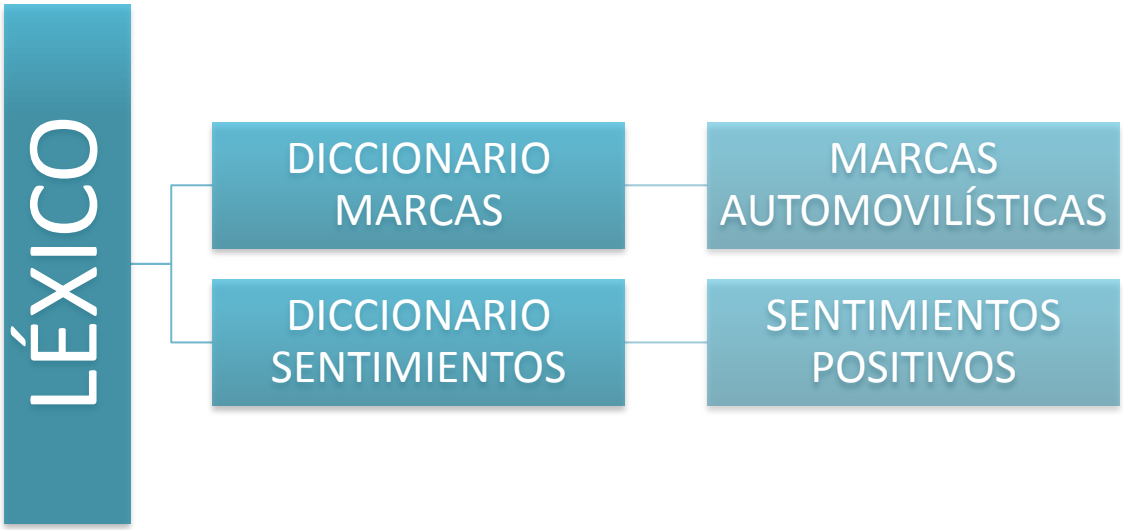


Figura 20. Estructura del léxico empleada en la realización del experimento.

El experimento estará dividido en las tres fases que se describen a continuación. Debido a la complejidad de integración de diferentes tecnologías de búsqueda, procesado y análisis de texto, se ha decidido comenzar con experimentos simples que permitan una exploración manual y una interpretación segmentada de las diferentes etapas, con el objetivo de ver cómo encaja las diferentes partes y cómo se comporta el sistema completo. Posteriormente generalizaremos a casos más grandes incrementando el número de términos pertenecientes al léxico y por consiguiente el número de documentos que forman la matriz término por documento.

1) Experimento pequeño

Se decidió partir de un experimento pequeño que constase de un reducido conjunto de documentos del corpus. Se seleccionó manualmente cinco documentos del corpus principal y otros cinco documentos del corpus auxiliar. La matriz término por documento resultante será por tanto visualizable, permitiendo determinar si los resultados del experimento son o no son razonables en comparación a los vectores documento de la matriz. El número de términos del diccionario de marcas será de cinco y el de sentimientos de seis. Este experimento también servirá de guía para comprender mejor el funcionamiento del algoritmo *Latent Semantic Analysis (LSA).*

2) Experimento intermedio

En este experimento ya no se limita el número de documentos a usar del corpus, teniendo un total de 456 documentos del corpus principal y 15 documentos del corpus auxiliar. El número de términos del léxico también incrementa, teniendo diez marcas y

ocho sentimientos. En este segundo experimento se visualizan los primeros grupos semánticos independientes de términos.

3) Experimento grande

Se aumenta el número de marcas del léxico de diez a treinta. Como consecuencia, también incrementa el número de documentos a usar del corpus principal. Se decidió no aumentar el número de documentos a usar del corpus auxiliar debido a su nulo impacto en los resultados.

La estructura del experimento realizado puede ser vista en la siguiente figura:

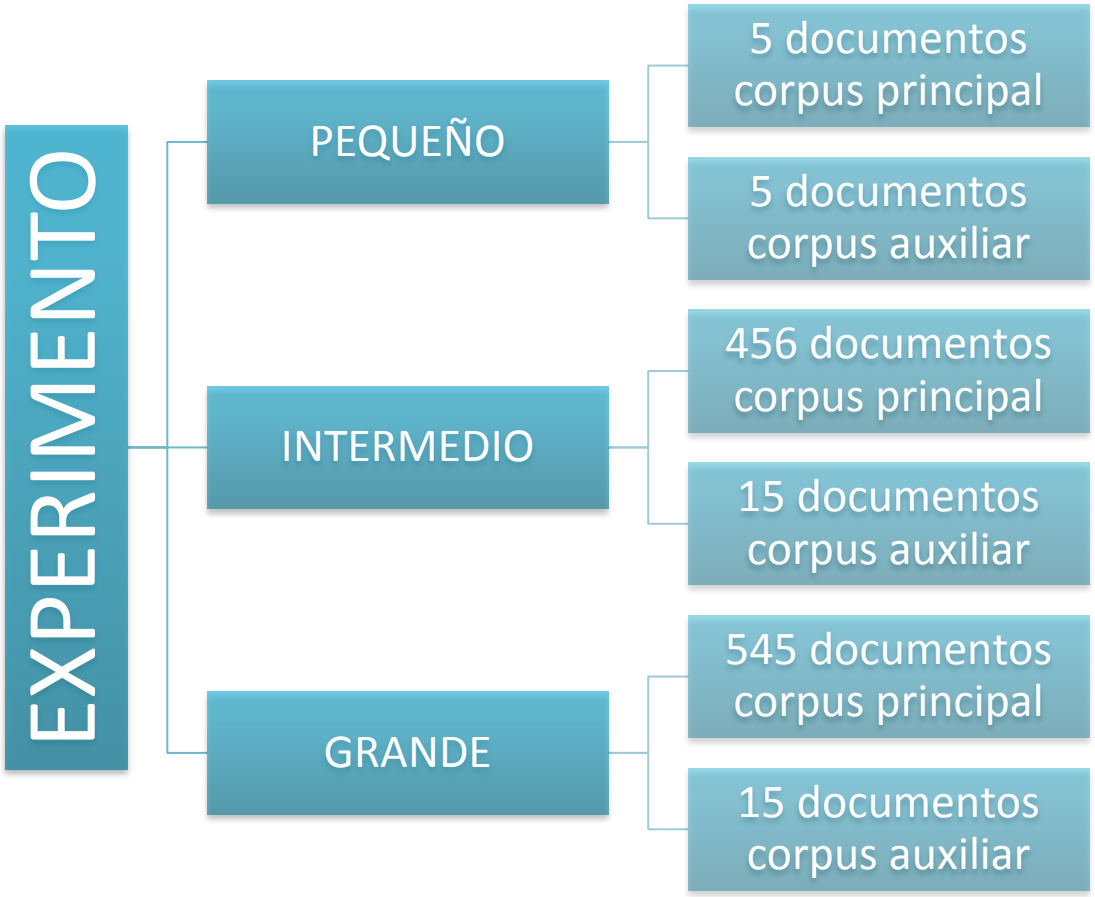


Figura 21. Estructura del experimento realizado

4.3. Resultados

4.3.1. Experimento pequeño

Se construye matriz término por documento seleccionando manualmente del corpus principal 5 documentos y otros 5 documentos del corpus auxiliar, formando una matriz 11x10.

	CORPUS PRINCIPAL					CORPUS AUXILIAR				
T/D	Doc_1 (id_424)	Doc_2 (id_1895)	Doc_3 (id_3263)	Doc_4 (id_4278)	Doc_5 (id_5982)	Doc_6 (id_63)	Doc_7 (id_70)	Doc_8 (id_129)	Doc_9 (id_142)	Doc_10 (id_297)
cómodo	0	1	0	2	1	1	0	0	1	0
divertido	0	0	0	0	1	0	0	0	2	0
bonito	0	0	3	0	1	0	0	1	0	1
suave	0	1	0	0	0	0	1	0	0	0
agradable	2	1	0	0	0	1	1	0	0	0
precioso	0	0	2	0	0	0	0	1	0	3
bmw	0	0	1	0	14	0	0	0	0	0
fiat	0	0	3	12	0	0	0	0	0	0
lamborghini	0	0	3	0	0	0	0	0	0	0
skoda	0	9	0	0	0	0	0	0	0	0
renault	7	0	0	0	0	0	0	0	0	0

Tabla 5. Matriz término por documento 11x10 experimento pequeño

La *Singular Value Descomposition (SVD)* de la matriz da origen a la matriz diagonal.

	1	2	3	4	5	6	7	8	9	10
1	14,2	0	0	0	0	0	0	0	0	0
2	0	12,5	0	0	0	0	0	0	0	0
3	0	0	9,17	0	0	0	0	0	0	0
4	0	0	0	7,28	0	0	0	0	0	0
5	0	0	0	0	5,16	0	0	0	0	0
6	0	0	0	0	0	2,39	0	0	0	0
7	0	0	0	0	0	0	2,19	0	0	0
8	0	0	0	0	0	0	0	1,6	0	0
9	0	0	0	0	0	0	0	0	0,92	0
10	0	0	0	0	0	0	0	0	0	0,48
11	0	0	0	0	0	0	0	0	0	0

Tabla 6. Matriz diagonal 11x10 con los valores singulares de la matriz término por documento del experimento pequeño

Dado que el número de dimensiones retenidas en *LSA* es un tema empírico (Landauer, T. K., Foltz, P. W., & Laham, D., 1998), determinaremos qué número de coeficientes hemos de eliminar observando si los resultados son o no son razonables en comparación a los vectores documento de la matriz anteriormente generada. El número de coeficientes a eliminar elegido puede darnos una pequeña pista de cómo actuar cuando se tenga que tomar esta misma decisión en el experimento intermedio.

Se ha decidido que el mejor resultado se obtiene cuando la dimensionalidad de la solución es reducida a cinco, eliminando los cinco coeficientes más bajos de la matriz diagonal. A continuación se muestra la matriz término por documento reconstruida.

	CORPUS PRINCIPAL					CORPUS AUXILIAR				
T/D	Doc_1 (id_424)	Doc_2 (id_1895)	Doc_3 (id_3263)	Doc_4 (id_4278)	Doc_5 (id_5982)	Doc_6 (id_63)	Doc_7 (id_70)	Doc_8 (id_129)	Doc_9 (id_142)	Doc_10 (id_297)
cómodo	0,04	1,03	0,12	1,99	1,01	0,07	0,03	-0,12	0,07	-0,25
divertido	0,00	0,02	-0,02	0,03	1,03	0,01	0,00	-0,03	0,02	-0,06
bonito	0,00	0,00	2,78	0,06	1,02	-0,05	0,00	0,74	-0,07	1,51
suave	0,04	1,02	0,00	0,00	0,00	0,03	0,03	0,00	0,01	0,00
agradable	2,07	1,04	-0,02	0,02	0,01	0,11	0,11	0,00	0,01	-0,01
precioso	0,00	0,01	2,83	-0,21	-0,06	-0,07	0,00	0,79	-0,10	1,62
bmw	0,00	0,00	1,02	-0,01	13,99	0,07	0,00	0,03	0,22	-0,03
fiat	-0,01	-0,01	3,01	11,99	0,00	0,16	0,00	0,01	0,16	0,00
lamborghini	0,00	0,00	2,17	0,21	0,06	-0,04	0,00	**0,58**	-0,07	**1,18**
skoda	-0,02	8,99	-0,01	0,00	0,00	0,21	0,21	0,01	0,11	0,03
renault	6,98	-0,01	0,00	0,00	0,00	0,27	0,27	0,00	0,00	0,00

Tabla 7. Matriz término por documento 11x10 reconstruída experimento pequeño

Aplicando el modelo de energías, el porcentaje de energía de la matriz empleada en el *Singular Value Descomposition* es calculado.

	1	2	3	4	5	6	7	8	9	10
1	201,36	0	0	0	0	0	0	0	0	0
2	0	157,25	0	0	0	0	0	0	0	0
3	0	0	84,089	0	0	0	0	0	0	0
4	0	0	0	52,998	0	0	0	0	0	0
5	0	0	0	0	26,626	0	0	0	0	0
6	0	0	0	0	0	5,7121	0	0	0	0
7	0	0	0	0	0	0	4,7961	0	0	0
8	0	0	0	0	0	0	0	2,56	0	0
9	0	0	0	0	0	0	0	0	0,8464	0
10	0	0	0	0	0	0	0	0	0	0,2304
11	0	0	0	0	0	0	0	0	0	0
					97,3%					100%
				E_5=	522,75				E_{10}=	537

Tabla 8. Porcentaje de energía a emplear en la matriz diagonal del experimento pequeño

Percepción sentimental de marcas de coches
MATRIZ DE DISTANCIAS

T/T	cómodo	divertido	bonito	suave	agradable	precioso	bmw	fiat	lamborghini	skoda	renault
cómodo	0,00	0,17	1,05	1,19	0,80	1,44	0,82	0,67	1,18	1,07	1,65
divertido	0,17	0,00	1,36	0,99	0,61	1,75	0,78	0,85	1,50	0,86	1,59
bonito	1,05	1,36	0,00	1,68	1,70	0,35	1,01	1,02	0,05	1,61	1,08
suave	1,19	0,99	**1,68**	0,00	0,17	1,09	1,43	1,61	1,67	0,44	0,78
agradable	0,80	0,61	**1,70**	**0,17**	0,00	1,44	1,28	1,42	1,71	0,70	0,86
precioso	1,44	1,75	**0,35**	**1,09**	**1,44**	0,00	1,08	1,36	0,36	1,16	0,76
bmw	0,82	0,78	1,01	1,43	1,28	1,08	0,00	0,69	1,13	1,24	0,92
fiat	0,67	0,85	1,02	1,61	1,42	1,36	0,69	0,00	0,88	1,08	1,18
lamborghini	1,18	1,50	**0,05**	**1,67**	**1,71**	**0,36**	1,13	0,88	0,00	1,59	0,97
skoda	1,07	0,86	1,61	0,44	0,70	1,16	1,24	1,08	1,59	0,00	1,21
renault	1,65	1,59	1,08	0,78	0,86	0,76	0,92	1,18	0,97	1,21	0,00

Tabla 9. Matriz de distancias 11x11 experimento pequeño

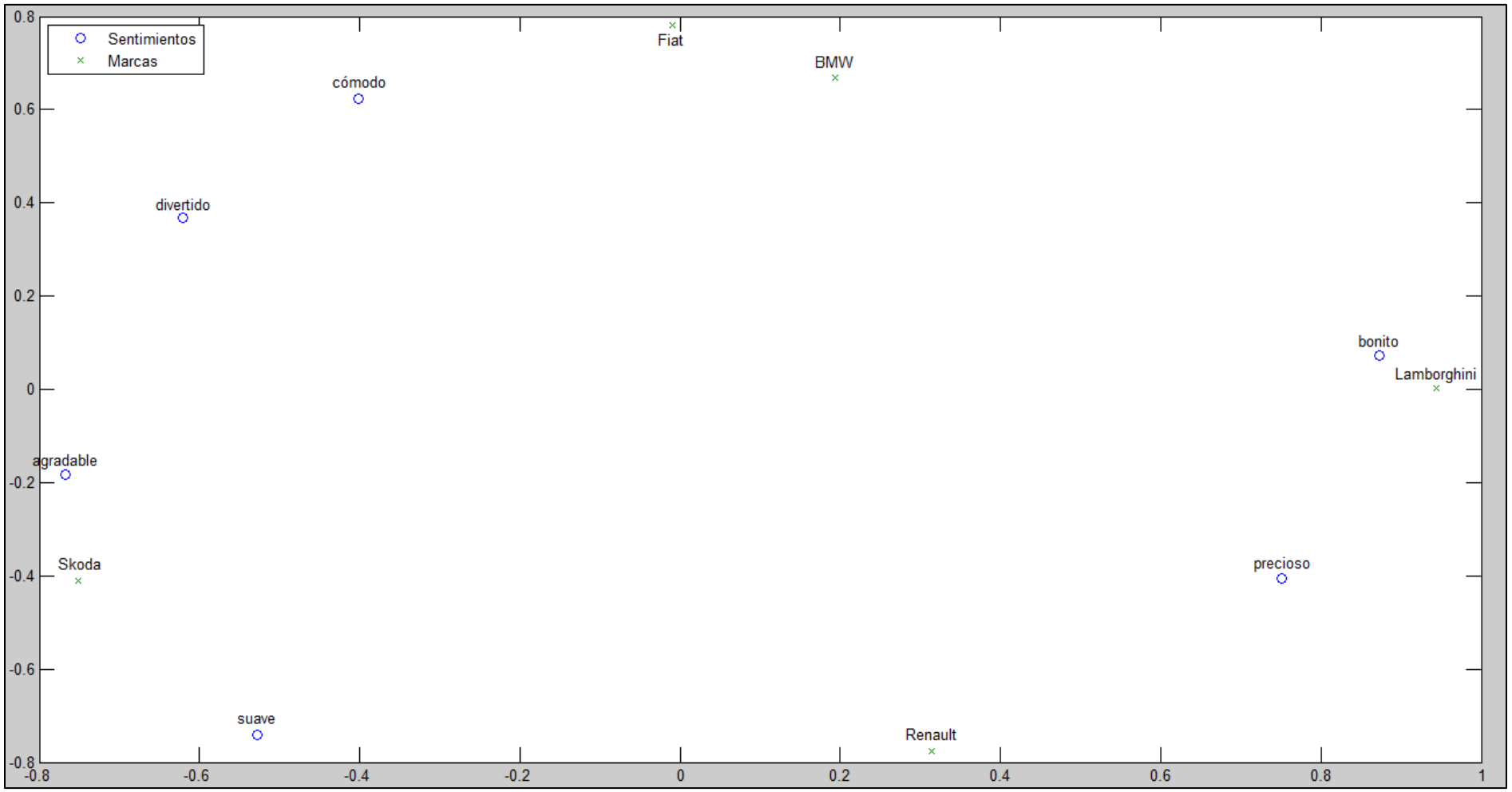


Figura 22. Posicionamiento sentimental relativo 2D experimento pequeño

En la solución se observa que algunas distancias entre marcas y sentimientos son más cortas que otras. Para entender mejor el funcionamiento del algoritmo *Latent Semantic Analysis* nos centramos en las celdas sombreadas en la matriz término por documento reconstruida y en la matriz de distancias:

- Los sentimientos *bonito* y *precioso* aparecen simultáneamente en tres de los diez documentos de la matriz término por documento original (nº 3, 8 y 10).
- Dado que *Lamborghini* solo ocurre en el documento nº 3, el algoritmo considera a este fabricante como propio de este contexto, de ahí la corta distancia con los sentimientos *bonito* y, algo mayor, *precioso*.
- Por el contrario, los sentimientos *suave* y *agradable* no ocurren en ninguno de los documentos donde sí aparecen los sentimientos *bonito* y *precioso*, por lo que el algoritmo no solo no considera estos sentimientos como similares, sino que va más lejos y los considera como términos de contextos opuestos, de ahí la alta distancia entre estos dos pares de sentimientos.
- Debido a que el algoritmo ha clasificado al fabricante Lamborghini como propio del contexto de los sentimientos *bonito* y *precioso*, también le otorga una distancia grande con respecto al par de sentimientos *suave* y *agradable*.

4.3.2. Experimento intermedio

El número de términos que conforman el léxico es aumentado en este segundo experimento. En total tendremos dieciocho términos, de los cuales diez serán marcas automovilísticas y ocho sentimientos.

El diccionario de marcas estará entonces compuesto de las siguientes diez marcas automovilísticas del sector:

1. *Audi*
2. *BMW*
3. *Mercedes*
4. *Lamborghini*
5. *Porsche*
6. *Volkswagen*
7. *Renault*
8. *Skoda*
9. *Kia*
10. *Opel*

Y el diccionario de sentimientos por los siguientes ocho sentimientos relacionados con coches:

1. *cómodo*
2. *divertido*
3. *bonito*
4. *suave*
5. *agradable*
6. *precioso*
7. *elegante*
8. *confortable*

El número de documentos del corpus principal que supera el umbral de decisión es de 456 documentos. A continuación se muestran las frecuencias totales de aparición, en orden decreciente, de cada uno de los términos del léxico en estos documentos:

1. *audi (Marca): 1351*
2. *bmw (Marca): 1269*
3. *mercedes (Marca): 975*
4. *porsche (Marca): 965*
5. *bonito (Sentimiento): 454*
6. *renault (Marca): 433*
7. *volkswagen (Marca): 352*
8. *opel (Marca): 324*
9. *cómodo (Sentimiento): 233*
10. *divertido (Sentimiento): 206*
11. *suave (Sentimiento): 203*
12. *precioso (Sentimiento): 193*
13. *elegante (Sentimiento): 176*
14. *agradable (Sentimiento): 164*
15. *lamborghini (Marca): 157*
16. *confortable (Sentimiento): 102*
17. *kia (Marca): 94*
18. *skoda (Marca): 83*

Se ha decidido que el número de documentos a utilizar del corpus auxiliar sea de 15 documentos. Todos ellos habrán superado el umbral de decisión establecido para esta colección de documentos.

Se construye la matriz término por documento uniendo las matrices obtenidas en el corpus principal y en el corpus auxiliar. El resultado es una matriz término por documento *18 x 471* que, dado su enorme tamaño, imposibilita su visualización en el presente documento.

La *Singular Value Descomposition (SVD)* de la matriz da origen a la matriz diagonal donde el número de coeficientes a suprimir desencadenará un desarrollo óptimo o no óptimo del algoritmo *Latent Semantic Analysis (LSA)*.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	...	471
1	202,94																			0
2		139,08																		0
3			127,98																	0
4				106,23																0
5					62,63															0
6						51,65														0
7							44,02													0
8								34,01												0
9									28,13											0
10										25,37										0
11											24,43									0
12												22,48								0
13													21,3							0
14														18,03						0
15															17,68					0
16																15,09				0
17																	13,03			0
18																		12,2		0

Tabla 10. Matriz diagonal 18x471 con los valores singulares de la matriz término por documento del experimento intermedio

Si aplicásemos el porcentaje de energía empleado en la matriz diagonal del experimento pequeño el número de valores singulares a mantener estaría comprendido entre diez y once. Se decidió, empíricamente mediante pruebas, que la mejor solución es obtenida manteniendo los primeros diez valores singulares de la matriz, eliminando los ocho coeficientes más bajos.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	...	471
1	41184,64																			0
2		19343,25																		0
3			16378,88																	0
4				11284,81																0
5					3922,52															0
6						2667,72														0
7							1937,76													0
8								1156,68												0
9									791,30											0
10										643,64										0
11											596,82									0
12												505,35								0
13													453,69							0
14														325,08						0
15															312,58					0
16																227,71				0
17																	169,78			0
18																		148,84	...	0
										97,29%								100%		
										E_{10}= 9,93E+08								E_{18}= 1,02E+09		

Tabla 11. Porcentaje de energía a emplear en la matriz diagonal del experimento intermedio

Los resultados obtenidos en el experimento intermedio son mostrados a continuación:

Percepción sentimental de marcas de coches
MATRIZ DE DISTANCIAS

T/T	cómodo	divertido	bonito	suave	agradable	precioso	elegante	confortable	Audi	BMW	Mercedes	Lamborghini	Porsche	Volkswagen	Renault	Skoda	Kia	Opel
cómodo	0,00	0,39	1,32	0,08	0,25	1,10	1,07	0,21	0,90	0,92	0,84	1,15	1,01	0,77	0,89	0,53	0,71	0,65
divertido	0,39	0,00	0,83	0,49	0,46	0,59	0,77	0,54	0,81	0,74	0,92	0,91	0,82	1,15	0,70	1,00	0,87	0,90
bonito	1,32	0,83	0,00	1,41	1,34	0,17	0,26	1,38	0,89	0,91	1,06	0,63	0,88	1,33	1,11	1,57	1,22	1,21
suave	0,08	0,49	1,41	0,00	0,16	1,15	1,20	0,26	1,00	0,96	0,77	1,15	1,12	0,76	0,79	0,49	0,72	0,65
agradable	0,25	0,46	1,34	0,16	0,00	1,05	1,05	0,37	0,95	0,84	0,76	1,20	1,13	0,68	0,70	0,44	0,51	0,87
precioso	1,10	0,59	0,17	1,15	1,05	0,00	0,41	1,25	1,00	0,95	1,09	0,72	0,76	1,23	0,97	1,47	1,17	1,22
elegante	1,07	0,77	0,26	1,20	1,05	0,41	0,00	1,03	0,71	0,65	0,90	0,67	0,95	1,09	1,22	1,34	1,12	1,25
confortable	0,21	0,54	1,38	0,26	0,37	1,25	1,03	0,00	0,62	0,79	0,80	1,06	1,09	0,73	0,89	0,72	0,92	0,80
Audi	0,90	0,81	0,89	1,00	0,95	1,00	0,71	0,62	0,00	0,50	0,71	0,90	0,90	0,92	1,03	1,11	1,02	0,92
BMW	0,92	0,74	0,91	0,96	0,84	0,95	0,65	0,79	0,50	0,00	0,65	0,99	0,86	0,98	0,96	1,07	1,03	0,99
Mercedes	0,84	0,92	1,06	0,77	0,76	1,09	0,90	0,80	0,71	0,65	0,00	1,00	1,00	0,84	1,00	0,88	0,95	0,79
Lamborghini	1,15	0,91	0,63	1,15	1,20	0,72	0,67	1,06	0,90	0,99	1,00	0,00	1,09	1,17	1,22	1,35	1,34	1,29
Porsche	1,01	0,82	0,88	1,12	1,13	0,76	0,95	1,09	0,90	0,86	1,00	1,09	0,00	1,05	0,87	1,08	1,11	0,97
Volkswagen	0,77	1,15	1,33	0,76	0,68	1,23	1,09	0,73	0,92	0,98	0,84	1,17	1,05	0,00	0,96	0,57	0,67	0,80
Renault	0,89	0,70	1,11	0,79	0,70	0,97	1,22	0,89	1,03	0,96	1,00	1,22	0,87	0,96	0,00	0,91	0,84	0,74
Skoda	0,53	1,00	1,57	0,49	0,44	1,47	1,34	0,72	1,11	1,07	0,88	1,35	1,08	0,57	0,91	0,00	0,20	0,68
Kia	0,71	0,87	1,22	0,72	0,51	1,17	1,12	0,92	1,02	1,03	0,95	1,34	1,11	0,67	0,84	0,20	0,00	0,73
Opel	0,65	0,90	1,21	0,65	0,87	1,22	1,25	0,80	0,92	0,99	0,79	1,29	0,97	0,80	0,74	0,68	0,73	0,00

Tabla 12. Matriz de distancias 18x18 experimento intermedio

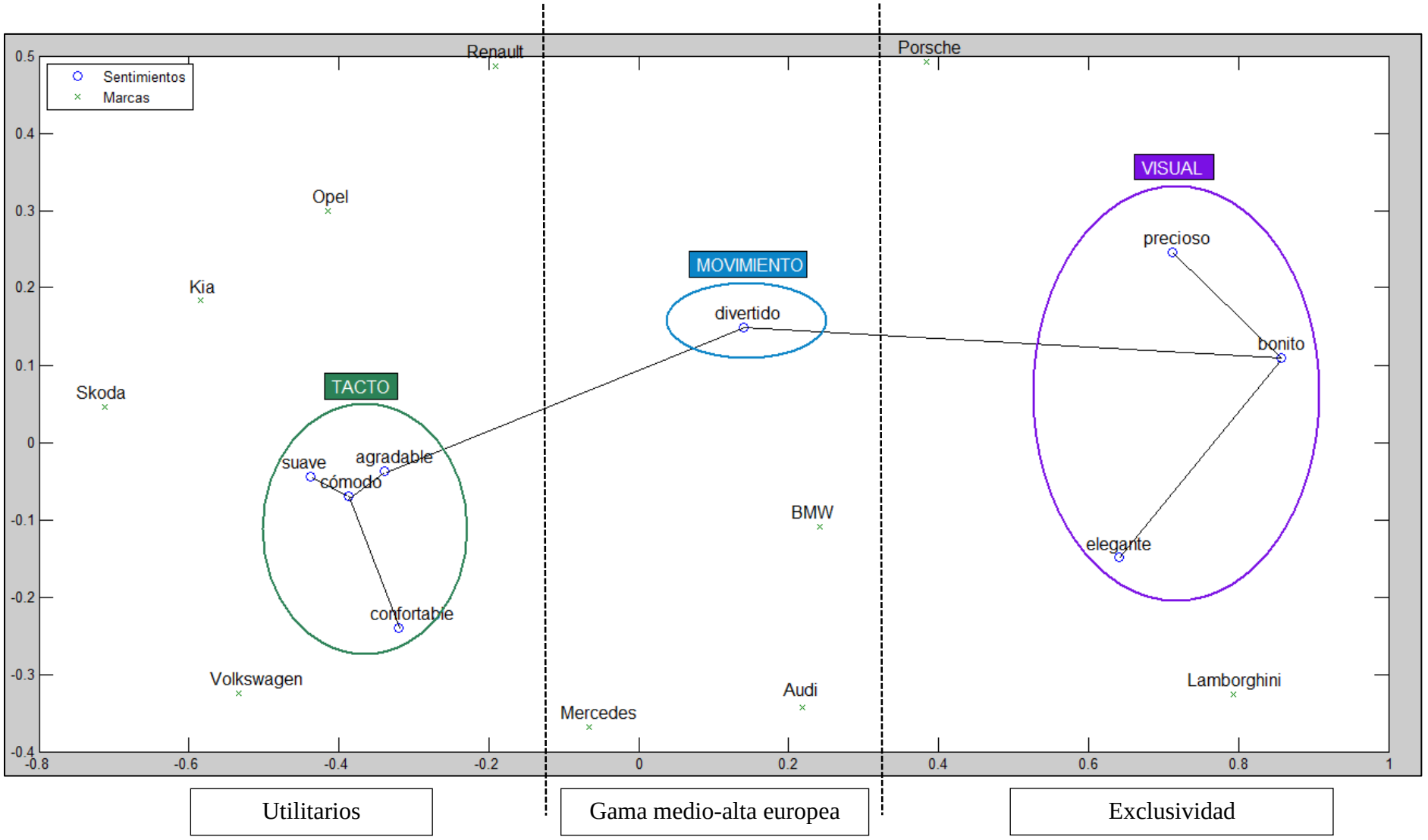


Figura 23. Posicionamiento sentimental relativo 2D experimento intermedio

En la solución se observa que algunos sentimientos tienden a estar muy próximos, formándose los primeros grupos semánticos independientes de sentimientos. Se han identificado los siguientes tres grupos:

1) Los sentimientos *cómodo*, *suave*, *agradable* y *confortable* se agrupan constituyendo un grupo semántico que perfectamente podría ser considerado como grupo de sentimientos relacionados con el tacto.
2) Los sentimientos *bonito*, *precioso* y *elegante* se agrupan constituyendo un grupo semántico de sentimientos relacionados con lo visual.
3) El sentimiento *divertido* se ha decidido que sea categorizado en un grupo semántico de sentimientos relacionados con el movimiento.

TACTO	VISUAL	MOVIMIENTO
• *cómodo* • *suave* • *agradable* • *confortable*	• *bonito* • *precioso* • *elegante*	• *divertido*

Figura 24. Grupos semánticos independientes de sentimientos identificados

No sólo los sentimientos se agrupan. También se observa que hay marcas que tienden a estar próximas entre sí, aunque quizás no estén tan próximas como sucede con los sentimientos. Estas marcas pueden ser agrupadas en grupos semánticos independientes al igual que como se hizo con los sentimientos. Se han identificado los siguientes tres grupos según el segmento de mercado de la marca:

UTILITARIOS	EXCLUSIVIDAD	GAMA MEDIO-ALTA
• *Renault* • *Opel* • *Kia* • *Skoda* • *Volkswagen*	• *Lamborghini* • *Porsche*	• *Audi* • *BMW* • *Mercedes*

Figura 25. Grupos semánticos independientes de marcas identificados

Antes de interpretar los resultados de este o de posteriores experimentos hay que tener en cuenta las puntualizaciones a continuación descritas:

- El posicionamiento relativo (2D) aquí realizado sólo será una aproximación a la solución del problema, la cual habrá que completar con los datos ofrecidos en la matriz de distancias.
- En el texto de los artículos recogidos se tiende a potenciar los puntos fuertes que ofrece un determinado modelo de coche con el fin de aumentar su atractivo hacia el lector (y así ganar lectores). El lector buscará principalmente en un utilitario comodidad y confort, mientras que en modelos de mayor gama buscará un status que lo diferencie del resto de vehículos (y de conductores) en aspectos relacionados con la imagen del mismo.

4.3.3. Experimento grande

En este último experimento se sigue aumentando el número de términos que conforman el léxico, incrementando aún más la dimensionalidad del problema. En total tendremos treinta y ocho términos, de los cuales treinta serán marcas automovilísticas y los mismos ocho sentimientos utilizados en el experimento anterior.

El diccionario de marcas estará entonces compuesto de las siguientes treinta marcas automovilísticas del sector:

1. *Abarth*
2. *Audi*
3. *BMW*
4. *Chevrolet*
5. *Dacia*
6. *Ferrari*
7. *Fiat*
8. *Ford*
9. *Honda*
10. *Hyundai*
11. *Jaguar*
12. *Kia*
13. *Lamborghini*
14. *Lancia*
15. *Lexus*
16. *Maserati*
17. *Mazda*
18. *Mercedes*
19. *Mitsubishi*
20. *Nissan*
21. *Opel*
22. *Peugeot*
23. *Porsche*
24. *Renault*
25. *Seat*
26. *Skoda*
27. *Subaru*
28. *Toyota*
29. *Volkswagen*
30. *Volvo*

Y el diccionario de sentimientos por los siguientes ocho sentimientos relacionados con coches:

1. *cómodo*
2. *divertido*
3. *bonito*
4. *suave*
5. *agradable*
6. *precioso*
7. *elegante*
8. *confortable*

El número de documentos del corpus principal que supera el umbral de decisión es de 565 documentos. A continuación se muestran las frecuencias totales de aparición, en orden decreciente, de cada uno de los términos del léxico en estos documentos:

1. *audi (Marca): 1351*
2. *bmw (Marca): 1269*
3. *mercedes (Marca): 975*
4. *porsche (Marca): 965*
5. *ferrari (Marca): 739*
6. *mitsubishi (Marca): 653*
7. *ford (Marca): 501*
8. *bonito (Sentimiento): 488*
9. *toyota (Marca): 454*
10. *renault (Marca): 433*
11. *honda (Marca): 390*
12. *seat (Marca): 380*
13. *nissan (Marca): 355*
14. *volkswagen (Marca): 352*
15. *opel (Marca): 324*
16. *lexus (Marca): 320*
17. *mazda (Marca): 318*
18. *cómodo (Sentimiento): 274*
19. *peugeot (Marca): 261*
20. *fiat (Marca): 249*

21. suave (Sentimiento): 241
22. divertido (Sentimiento): 225
23. dacia (Marca): 221
24. subaru (Marca): 215
25. elegante (Sentimiento): 206
26. precioso (Sentimiento): 205
27. chevrolet (Marca): 202
28. jaguar (Marca): 197
29. agradable (Sentimiento): 189
30. volvo (Marca): 167
31. lamborghini (Marca): 157
32. confortable (Sentimiento): 120
33. maserati (Marca): 99
34. hyundai (Marca): 95
35. kia (Marca): 94
36. skoda (Marca): 83
37. abarth (Marca): 64
38. lancia (Marca): 64

Se ha decidido no aumentar el número de documentos a utilizar del corpus auxiliar, manteniendo los 15 documentos utilizados en el experimento anterior. La razón fue que la adhesión de documentos de este corpus auxiliar no produjo impacto alguno en la representación obtenida en el espacio semántico, reduciendo las distancias entre sentimientos del mismo grupo únicamente en una o dos décimas que resultaron ser insignificantes en el resultado del experimento.

Se construye la matriz término por documento uniendo las matrices obtenidas en el corpus principal y en el corpus auxiliar. El resultado es una matriz término por documento *38 x 560* que, dado su enorme tamaño, imposibilita su visualización en el presente documento.

La *Singular Value Descomposition (SVD)* de la matriz da origen a la matriz diagonal donde el número de coeficientes a suprimir desencadenará un desarrollo óptimo o no óptimo del algoritmo *Latent Semantic Analysis (LSA).*

Se decidió, empíricamente mediante pruebas, que la mejor solución es obtenida manteniendo los primeros doce valores singulares de la matriz, eliminando los veintiséis coeficientes más bajos. El porcentaje de energía empleado en la matriz diagonal será entonces del 83,30%, muy por debajo del 97,29% de energía del experimento intermedio.

El porcentaje de energía de la matriz diagonal a emplear así como los resultados obtenidos en el experimento grande son mostrados a continuación:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	...	560
1	216,38																																							0
2		156,87																																						0
3			134,81																																					0
4				108,93																																				0
5					107,41																																			0
6						97,97																																		0
7							92,83																																	0
8								86,99																																0
9									78,49																															0
10										76,24																														0
11											71,74																													0
12												66,27																												0
13													64,57																											0
14														54,51																										0
15															52,88																									0
16																51,55																								0
17																	48,70																							0
18																		45,99																						0
19																			44,53																					0
20																				42,20																				0
21																					40,22																			0
22																						37,01																		0
23																							34,32																	0
24																								32,23																0
25																									30,37															0
26																										28,45														0
27																											25,79													0
28																												24,98												0
29																													23,40											0
30																														21,24										0
31																															19,38									0
32																																17,94								0
33																																	17,35							0
34																																		17,05						0
35																																			15,56					0
36																																				14,34				0
37																																					13,84			0
38																																						12,22		0

Tabla 13. Matriz diagonal 38x560 con los valores singulares de la matriz término por documento del experimento grande

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	...	560
1	46822,18																																							0
2		24609,19																																						0
3			18172,51																																					0
4				11864,98																																				0
5					11537,36																																				0
6						9597,63																																			0
7							8618,24																																		0
8								7566,49																																	0
9									6161,00																																0
10										5812,30																															0
11											5147,01																														0
12												4391,96																													0
13													4168,68																												0
14														2971,76																											0
15															2796,63																										0
16																2657,79																									0
17																	2371,42																								0
18																		2115,51																							0
19																			1983,05																						0
20																				1780,57																					0
21																					1617,59																				0
22																						1369,58																			0
23																							1178,09																		0
24																								1039,09																	0
25																									922,37																0
26																										809,34															0
27																											665,10														0
28																												624,21													0
29																													547,68												0
30																														451,10											0
31																															375,70										0
32																																321,75									0
33																																	300,85								0
34																																		290,79							0
35																																			242,08						0
36																																				205,54					0
37																																					191,47				0
38																																						149,41			0
											83,30%																											100%			
											E_{12}= 1,60E+09																											E_{38}= 1,92E+09			

Tabla 14. Porcentaje de energía a emplear en la matriz diagonal del experimento grande

Percepción sentimental de marcas de coches
MATRIZ DE DISTANCIAS

T/T	cómodo	divertido	bonito	suave	agradable	precioso	elegante	confortable	Abarth	Audi	BMW	Chevrolet	Dacia	Ferrari	Fiat	Ford	Honda	Hyundai	Jaguar	Kia	Lamborghini	Lancia	Lexus	Maserati	Mazda	Mercedes	Mitsubishi	Nissan	Opel	Peugeot	Porsche	Renault	Seat	Skoda	Subaru	Toyota	Volkswagen	Volvo
cómodo	0,00	0,47	0,47	0,22	0,17	0,61	0,63	0,32	0,38	0,80	0,77	0,63	0,88	0,90	0,87	1,13	0,78	0,27	0,62	0,46	0,65	0,57	0,96	0,80	0,51	1,05	0,95	0,81	0,51	0,37	0,97	0,63	0,99	0,56	0,67	0,87	0,34	0,78
divertido	0,47	0,00	0,44	0,80	0,60	0,48	0,58	0,98	0,42	0,89	0,61	0,56	0,80	0,66	0,63	0,89	0,57	0,52	0,44	0,47	0,64	0,66	1,03	0,71	0,51	1,00	1,24	0,68	0,62	0,89	0,75	0,44	0,93	1,06	0,31	0,56	0,81	0,86
bonito	0,47	0,44	0,00	0,66	0,59	0,15	0,18	0,79	0,60	0,62	0,39	0,71	0,81	0,80	0,93	1,15	0,86	0,54	0,48	0,40	0,67	0,54	0,99	0,59	0,95	0,70	1,34	0,96	0,58	0,91	0,74	0,63	1,02	0,97	0,75	0,87	0,76	0,43
suave	0,22	0,80	0,66	0,00	0,19	0,65	0,85	0,27	0,74	1,00	0,94	0,57	0,71	1,17	0,90	0,93	0,98	0,28	0,62	0,56	0,70	0,60	0,95	0,82	0,62	0,93	0,89	0,94	0,35	0,19	1,12	0,55	0,97	0,34	0,91	0,95	0,27	0,75
agradable	0,17	0,60	0,59	0,19	0,00	0,63	0,74	0,36	0,64	0,95	0,86	0,78	0,61	1,13	0,71	1,04	0,91	0,13	0,65	0,35	0,94	0,80	1,07	1,03	0,66	1,00	0,87	0,93	0,37	0,20	1,20	0,40	0,89	0,40	0,66	0,75	0,37	0,63
precioso	0,61	0,48	0,15	0,65	0,63	0,00	0,33	0,94	0,58	0,80	0,46	0,60	0,70	0,88	0,77	0,99	0,82	0,67	0,43	0,57	0,68	0,57	1,01	0,58	0,87	0,69	1,25	0,98	0,61	0,90	0,65	0,55	0,86	0,90	0,82	0,90	0,85	0,64
elegante	0,63	0,58	0,18	0,85	0,74	0,33	0,00	0,83	0,58	0,51	0,27	0,70	0,94	0,73	0,78	0,99	0,87	0,72	0,38	0,53	0,65	0,40	0,94	0,45	0,95	0,58	1,32	1,05	0,75	0,98	0,82	0,82	0,76	0,97	0,73	0,97	0,86	0,47
confortable	0,32	0,98	0,79	0,27	0,36	0,94	0,83	0,00	0,93	0,64	0,87	0,90	0,79	1,14	1,09	0,93	1,12	0,37	0,87	0,48	0,89	0,67	0,99	0,90	0,85	0,98	0,81	0,97	0,53	0,20	1,16	0,75	1,02	0,39	0,94	0,99	0,19	0,62
Abarth	0,38	0,42	0,60	0,74	0,64	0,58	0,58	0,93	0,00	0,88	0,76	0,53	1,28	0,62	0,69	1,17	0,57	0,77	0,52	0,91	0,51	0,52	1,01	0,66	0,37	1,06	1,00	0,88	0,92	0,96	0,68	0,92	0,82	0,98	0,62	0,87	0,82	1,21
Audi	0,80	0,89	0,62	1,00	0,95	0,80	0,51	0,64	0,88	0,00	0,63	1,00	0,92	0,90	0,99	1,02	1,05	0,91	0,89	0,68	0,85	0,69	0,91	0,74	1,13	0,68	1,05	0,89	0,89	0,98	0,83	0,98	0,88	0,93	0,95	0,99	0,77	0,63
BMW	0,77	0,61	0,39	0,94	0,86	0,46	0,27	0,87	0,76	0,63	0,00	0,88	0,89	0,75	0,93	0,91	0,76	0,81	0,61	0,62	0,82	0,67	0,97	0,60	0,93	0,67	1,25	1,04	0,94	0,96	0,77	0,86	0,91	1,06	0,71	0,97	0,97	0,54
Chevrolet	0,63	0,56	0,71	0,57	0,78	0,60	0,70	0,90	0,53	1,00	0,88	0,00	1,09	0,76	0,66	0,56	0,75	0,71	0,19	0,90	0,15	0,20	0,88	0,27	0,38	0,71	1,15	0,82	0,53	0,83	0,77	0,79	0,81	0,80	0,61	0,82	0,70	1,05
Dacia	0,88	0,80	0,81	0,71	0,61	0,70	0,94	0,79	1,28	0,92	0,89	1,09	0,00	1,24	0,79	0,77	1,06	0,60	0,94	0,41	1,27	1,13	1,10	1,15	1,17	0,92	1,01	0,85	0,56	0,50	1,15	0,25	0,90	0,61	0,96	0,88	0,77	0,66
Ferrari	0,90	0,66	0,80	1,17	1,13	0,88	0,73	1,14	0,62	0,90	0,75	0,76	1,24	0,00	1,07	1,04	0,84	1,10	0,73	1,04	0,49	0,70	1,15	0,50	0,77	0,89	1,12	1,07	1,18	1,19	0,67	1,10	1,11	1,20	0,80	1,02	1,18	1,14
Fiat	0,87	0,63	0,93	0,90	0,71	0,77	0,78	1,09	0,69	0,99	0,93	0,66	0,79	1,07	0,00	0,53	0,70	0,83	0,53	0,78	0,89	0,72	0,85	0,93	0,85	0,88	1,07	0,73	0,57	0,83	0,96	0,54	0,16	0,63	0,75	0,69	0,78	1,05
Ford	1,13	0,89	1,15	0,93	1,04	0,99	0,99	0,93	1,17	1,02	0,91	0,56	0,77	1,04	0,53	0,00	0,87	0,95	0,63	0,89	0,77	0,66	0,85	0,68	0,83	0,77	0,92	0,80	0,67	0,76	0,97	0,73	0,55	0,70	0,82	0,83	0,78	0,95
Honda	0,78	0,57	0,86	0,98	0,91	0,82	0,87	1,12	0,57	1,05	0,76	0,75	1,06	0,84	0,70	0,87	0,00	0,88	0,88	0,92	0,79	0,87	0,96	0,92	0,70	0,95	1,06	0,78	0,94	1,05	0,76	1,00	0,84	1,01	0,63	0,78	0,99	1,07
Hyundai	0,27	0,52	0,54	0,28	0,13	0,67	0,72	0,37	0,77	0,91	0,81	0,71	0,60	1,10	0,83	0,95	0,88	0,00	0,58	0,20	0,88	0,73	1,15	0,94	0,62	1,00	1,02	0,97	0,26	0,27	1,24	0,37	1,08	0,54	0,49	0,70	0,36	0,51
Jaguar	0,62	0,44	0,48	0,62	0,65	0,43	0,38	0,87	0,52	0,89	0,61	0,19	0,94	0,73	0,53	0,63	0,88	0,58	0,00	0,64	0,35	0,18	0,98	0,27	0,50	0,67	1,29	0,96	0,47	0,77	0,88	0,58	0,68	0,79	0,54	0,84	0,72	0,81
Kia	0,46	0,47	0,40	0,56	0,35	0,57	0,53	0,48	0,91	0,68	0,62	0,90	0,41	1,04	0,78	0,89	0,92	0,20	0,64	0,00	0,99	0,76	1,15	0,94	0,94	0,98	1,19	0,96	0,30	0,46	1,13	0,24	1,00	0,70	0,59	0,69	0,43	0,41
Lamborghini	0,65	0,64	0,67	0,70	0,94	0,68	0,65	0,89	0,51	0,85	0,82	0,15	1,27	0,49	0,89	0,77	0,79	0,88	0,35	0,99	0,00	0,19	0,88	0,16	0,51	0,70	1,18	0,88	0,74	0,99	0,61	1,01	0,94	0,94	0,74	0,93	0,78	1,08
Lancia	0,57	0,66	0,54	0,60	0,80	0,57	0,40	0,67	0,52	0,69	0,67	0,20	1,13	0,70	0,72	0,66	0,87	0,73	0,18	0,76	0,19	0,00	0,92	0,14	0,63	0,66	1,24	0,90	0,54	0,78	0,74	0,87	0,68	0,69	0,76	0,98	0,54	0,86
Lexus	0,96	1,03	0,99	0,95	1,07	1,01	0,94	0,99	1,01	0,91	0,97	0,88	1,10	1,15	0,85	0,85	0,96	1,15	0,98	1,15	0,88	0,92	0,00	0,94	1,04	0,90	1,01	0,61	1,05	1,09	1,00	1,15	0,77	0,99	1,16	1,07	1,00	0,96
Maserati	0,80	0,71	0,59	0,82	1,03	0,58	0,45	0,90	0,66	0,74	0,60	0,27	1,15	0,50	0,93	0,68	0,92	0,94	0,27	0,94	0,16	0,14	0,94	0,00	0,67	0,53	1,23	0,97	0,82	0,98	0,63	1,02	0,85	0,92	0,79	1,08	0,89	0,86
Mazda	0,51	0,51	0,95	0,62	0,66	0,87	0,95	0,85	0,37	1,13	0,93	0,38	1,17	0,77	0,85	0,83	0,70	0,62	0,50	0,94	0,51	0,63	1,04	0,67	0,00	1,06	0,86	0,97	0,87	0,77	0,92	0,84	1,09	1,01	0,43	0,85	0,82	1,21
Mercedes	1,05	1,00	0,70	0,93	1,00	0,69	0,58	0,98	1,06	0,68	0,67	0,71	0,92	0,89	0,88	0,77	0,95	1,00	0,67	0,98	0,70	0,66	0,90	0,53	1,06	0,00	1,04	0,96	0,88	0,99	0,77	1,03	0,77	0,85	0,92	0,98	1,06	0,49
Mitsubishi	0,95	1,24	1,34	0,89	0,87	1,25	1,32	0,81	1,00	1,05	1,25	1,15	1,01	1,12	1,07	0,92	1,06	1,02	1,29	1,19	1,18	1,24	1,01	1,23	0,86	1,04	0,00	1,01	1,15	0,76	1,03	1,13	0,98	0,82	1,05	0,92	0,98	1,12
Nissan	0,81	0,68	0,96	0,94	0,93	0,98	1,05	0,97	0,88	0,89	1,04	0,82	0,85	1,07	0,73	0,80	0,78	0,97	0,96	0,96	0,88	0,90	0,61	0,97	0,97	0,96	1,01	0,00	0,78	0,94	0,81	0,84	0,71	0,77	0,93	0,79	0,80	0,91
Opel	0,51	0,62	0,58	0,35	0,37	0,61	0,75	0,53	0,92	0,89	0,94	0,53	0,56	1,18	0,57	0,67	0,94	0,26	0,47	0,30	0,74	0,54	1,05	0,82	0,87	0,88	1,15	0,78	0,00	0,39	1,13	0,29	0,77	0,38	0,76	0,70	0,24	0,61
Peugeot	0,37	0,89	0,91	0,19	0,20	0,90	0,98	0,20	0,96	0,98	0,96	0,83	0,50	1,19	0,83	0,76	1,05	0,27	0,77	0,46	0,99	0,78	1,09	0,98	0,77	0,99	0,76	0,94	0,39	0,00	1,26	0,47	0,86	0,20	0,87	0,92	0,26	0,71
Porsche	0,97	0,75	0,74	1,12	1,20	0,65	0,82	1,16	0,68	0,83	0,77	0,77	1,15	0,67	0,96	0,97	0,76	1,24	0,88	1,13	0,61	0,74	1,00	0,63	0,92	0,77	1,03	0,81	1,13	1,26	0,00	1,09	0,89	1,13	0,87	0,79	1,11	1,07
Renault	0,63	0,44	0,63	0,55	0,40	0,55	0,82	0,75	0,92	0,98	0,86	0,79	0,25	1,10	0,54	0,73	1,00	0,37	0,58	0,24	1,01	0,87	1,15	1,02	0,84	1,03	1,13	0,84	0,29	0,47	1,09	0,00	0,85	0,62	0,70	0,66	0,54	0,72
Seat	0,99	0,93	1,02	0,97	0,89	0,86	0,76	1,02	0,82	0,88	0,91	0,81	0,90	1,11	0,16	0,55	0,84	1,08	0,68	1,00	0,94	0,68	0,77	0,85	1,09	0,77	0,98	0,71	0,77	0,86	0,89	0,85	0,00	0,50	1,02	0,89	0,81	1,04
Skoda	0,56	1,06	0,97	0,34	0,40	0,90	0,97	0,39	0,98	0,93	1,06	0,80	0,61	1,20	0,63	0,70	1,01	0,54	0,79	0,70	0,94	0,69	0,99	0,92	1,01	0,85	0,82	0,77	0,38	0,20	1,13	0,62	0,50	0,00	1,10	1,01	0,28	0,80
Subaru	0,67	0,31	0,75	0,91	0,66	0,82	0,73	0,94	0,62	0,95	0,71	0,61	0,96	0,80	0,75	0,82	0,63	0,49	0,54	0,59	0,74	0,76	1,16	0,79	0,43	0,92	1,05	0,93	0,76	0,87	0,87	0,70	1,02	1,10	0,00	0,35	0,92	0,79
Toyota	0,87	0,56	0,87	0,95	0,75	0,90	0,97	0,99	0,87	0,99	0,97	0,82	0,88	1,02	0,69	0,83	0,78	0,70	0,84	0,69	0,93	0,98	1,07	1,08	0,85	0,98	0,92	0,79	0,70	0,92	0,79	0,66	0,89	1,01	0,35	0,00	0,86	0,87
Volkswagen	0,34	0,81	0,76	0,27	0,37	0,85	0,86	0,19	0,82	0,77	0,97	0,70	0,77	1,18	0,78	0,78	0,99	0,36	0,72	0,43	0,78	0,54	1,00	0,89	0,82	1,06	0,98	0,80	0,24	0,26	1,11	0,54	0,81	0,28	0,92	0,86	0,00	0,76
Volvo	0,78	0,86	0,43	0,75	0,63	0,64	0,47	0,62	1,21	0,63	0,54	1,05	0,66	1,14	1,05	0,95	1,07	0,51	0,81	0,41	1,08	0,86	0,96	0,86	1,21	0,49	1,12	0,91	0,61	0,71	1,07	0,72	1,04	0,80	0,79	0,87	0,76	0,00

Tabla 15. Matriz de distancias 38x38 experimento grande

Percepción sentimental de marcas de coches
POSICIONAMIENTO RELATIVO (2D)

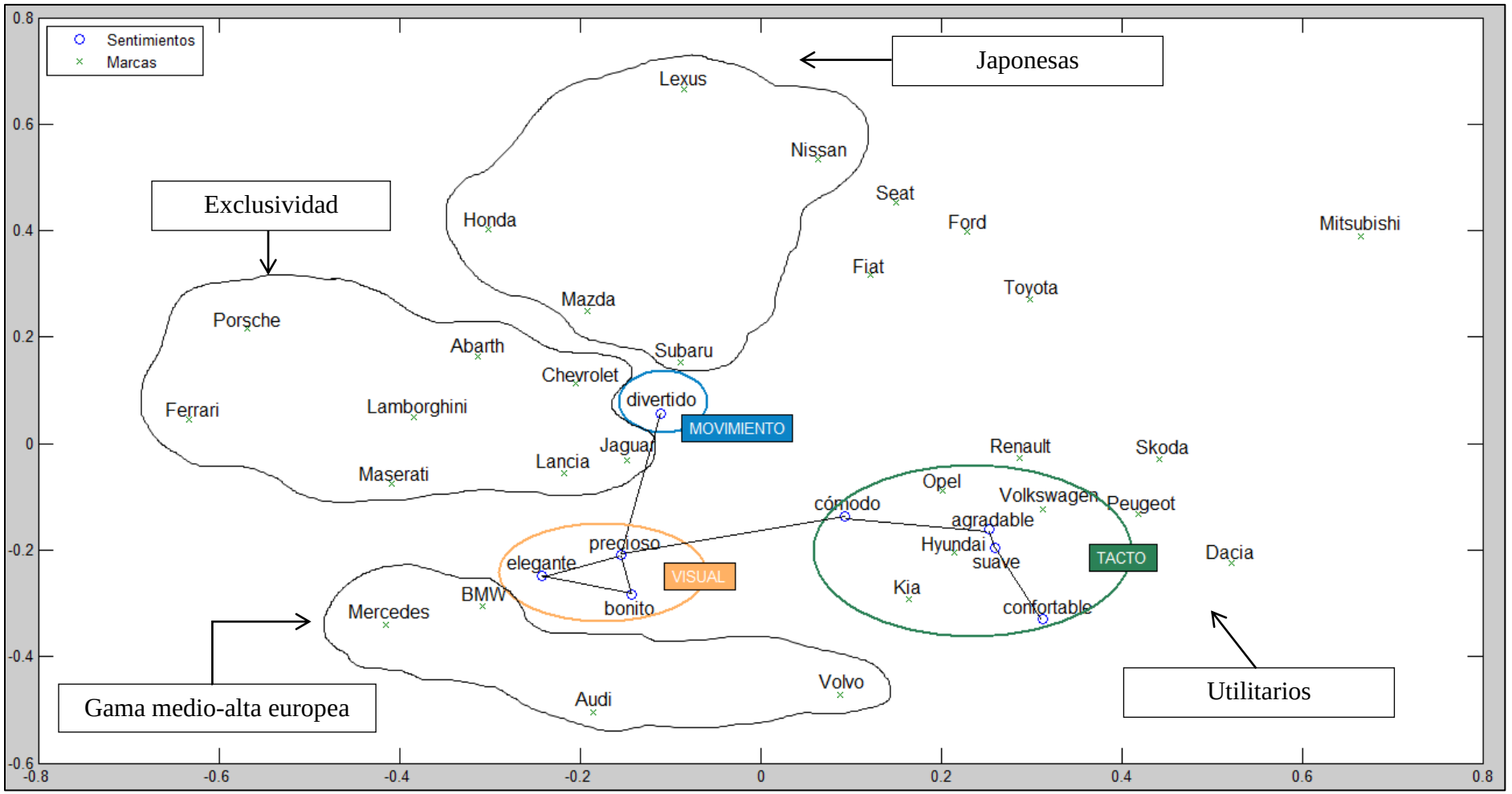


Figura 26. Posicionamiento sentimental relativo 2D experimento grande

En la solución se observa que cuando se incrementa la dimensión del problema los sentimientos siguen agrupándose de forma adecuada. Este hecho tiene sentido ya que se presupone que la sinonimia existente entre los sentimientos no debe verse afectada por la nueva dimensión del problema.

La inclusión de nuevas marcas al problema ha reforzado los grupos semánticos independientes de marcas que se habían identificado en el experimento anterior.

UTILITARIOS	EXCLUSIVIDAD	GAMA MEDIO-ALTA
• *Renault* • *Opel* • *Kia* • *Skoda* • *Volkswagen*	• *Lamborghini* • *Porsche*	• *Audi* • *BMW* • *Mercedes*
• *Peugeot* • *Hyundai* • *Dacia*	• *Jaguar* • *Lancia* • *Maserati* • *Chevrolet* • *Abarth* • *Porsche* • *Ferrari*	• *Volvo*

Figura 27. Grupos semánticos independientes de sentimientos identificados reforzados

Además de estos tres grupos semánticos se ha identificado en este experimento grande un nuevo grupo de marcas: *Lexus*, *Nissan*, *Honda*, *Mazda* y *Subaru*. Todas estas marcas tienen en común que son marcas niponas y dado que el sistema las sitúa próximas en el espacio semántico hace pensar que actúan sobre el mismo segmento de mercado.

Al contrario que los cuatro grupos semánticos mencionados anteriormente, hay otros grupos identificados en este experimento que no son fácilmente interpretables. Un claro ejemplo serían las marcas *Fiat*, *Ford* o *Seat*, las cuales el sistema ha sacado fuera del grupo de utilitarios. Este hecho podría significar que estas tres marcas actúan sobre un segmento de mercado diferente que el de los utilitarios y deberían ser por tanto categorizados en otro grupo, sin embargo, como hemos mencionado al principio de este párrafo, no es fácil sacar una interpretación de este grupo.

Un segundo grupo que no es fácilmente interpretable estaría formado por *Toyota* y *Mitsubishi*. Estas dos marcas, sobre todo la última, disponen de una gran variedad de vehículos 4x4, por lo que se explicaría su proximidad en el espacio semántico y, sobre todo, el por qué aparecen tan alejadas del resto de grupos, en especial *Mitsubishi*. La inclusión de algún sentimiento relacionado con los vehículos 4x4 como *robusto* o *dominante* y el comportamiento de estas marcas hacía estos sentimientos haría más fácil la interpretación.

JAPONESAS	NO INTERPRETABLE 1	NO INTERPRETABLE 2
• *Lexus* • *Nissan* • *Honda* • *Mazda* • *Subaru*	• *Seat* • *Ford* • *Fiat*	• *Toyota* • *Mitsubishi*

Figura 28. Nuevos grupos semánticos independientes de sentimientos identificados en experimento grande

4.4. Consideraciones sobre la selección de autovalores en LSA

Durante el desarrollo de los experimentos se ha concluido que no hay regla con el modelo de energías, decidiéndose empíricamente mediante pruebas la mejor solución del problema. La toma de la decisión ha sido fruto de la guía en base a la agrupación de los sentimientos en el espacio semántico. En caso de que los sentimientos se agrupen en grupos semánticos correctos, es decir, se observa que en cada grupo existe sinonimia entre los sentimientos que lo conforman, entonces se considerará que la selección de los autovalores ha sido acertada.

La acción de identificar el rango de autovalores en el que los sentimientos se agrupan de forma consistente podría ser automatizada en base a las distancias entre sentimientos reflejadas en la matriz de distancias. A continuación se visualizan las matrices diagonales tanto del experimento intermedio como del experimento grande indicando, en cada una de ellas, en qué corte de valores singulares los sentimientos comienzan a agruparse de forma consistente.

Tabla 16. Matriz diagonal 18x471del experimento intermedio dividida en cortes en función del posicionamiento de los puntos en el espacio

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	...	471
1	41184,64																			0
2		19343,25																		0
3			16378,88																	0
4				11284,81																0
5					3922,52															0
6						2667,72														0
7							1937,76													0
8								1156,68												0
9									791,30											0
10										643,64										0
11											596,82									0
12												505,35								0
13													453,69							0
14														325,08						0
15															312,58					0
16																227,71				0
17																	169,78			0
18																		148,84	...	0

Puntos en el espacio se solapan

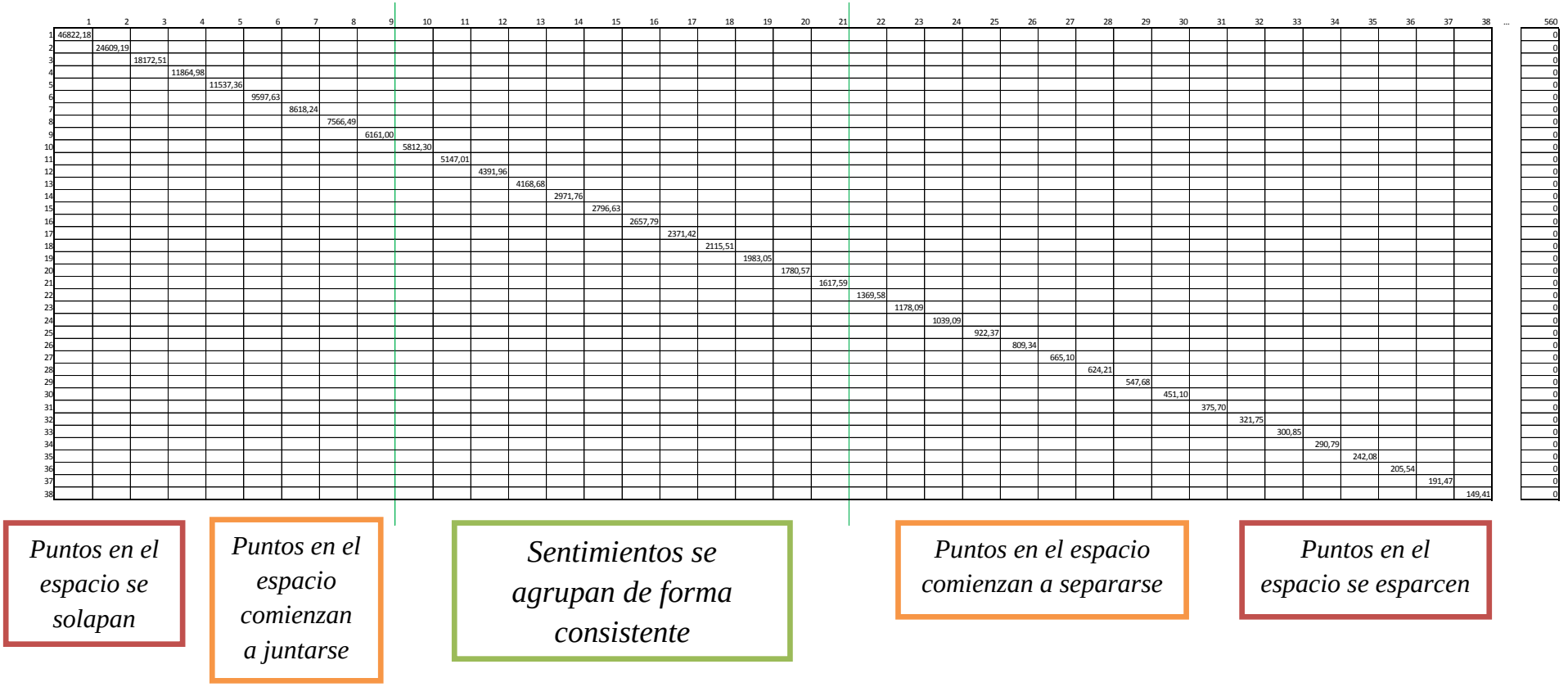
Puntos en el espacio comienzan a juntarse

Sentimientos se agrupan de forma consistente

Puntos en el espacio comienzan a separarse

Puntos en el espacio se esparcen

Tabla 17. Matriz diagonal 38x560 del experimento grande dividida en cortes en función del posicionamiento de los puntos en el espacio



Puntos en el espacio se solapan

Puntos en el espacio comienzan a juntarse

Sentimientos se agrupan de forma consistente

Puntos en el espacio comienzan a separarse

Puntos en el espacio se esparcen

Capítulo 5

Conclusiones y líneas futuras

Índice Capítulo 5

5.1. Conclusiones

El trabajo aquí realizado constituye un primer acercamiento al diseño de sistemas que analicen la similitud existente entre marcas y sentimientos presentes online. Como requisito de diseño se impuso la simplicidad y flexibilidad del sistema, dejando el procesado de textos más dependiente del idioma como trabajo futuro en función del compromiso complejidad/prestaciones alcanzado.

Como en todo sistema de estas características, el primer paso consistió en construir una colección de documentos lo suficientemente grande para garantizar un buen desarrollo de los experimentos realizados. El rastreador web aquí construido otorga una gran flexibilidad (y sobretodo comodidad) en la construcción de esta colección. La ausencia de un modelo conceptual en la Web provoca que los documentos de hipertexto que recolecta el rastreador difieran frecuentemente en el modo en que representan la información. La extracción de texto como de enlaces de estos documentos por tanto fue realizada de la manera más homogénea posible, con el propósito de que el rastreador se amoldase al mayor número de páginas web posible.

La escasez de sentimientos con polaridad negativa en los documentos recolectados determinó que estos no fueran tenidos en cuenta en la realización de los experimentos. En gran parte de los artículos recolectados se potencia el atractivo de una marca hacia el lector, recogiéndose sentimientos negativos únicamente en algunos de los comentarios de los usuarios. Estos comentarios propiciaban la diversidad de opiniones, generando un debate entre los usuarios en los que terminaban mencionándose diferentes marcas.

Se decidió utilizar el algoritmo *Latent Semantic Analysis (LSA)* para capturar la similitud de significados. La dificultad con este algoritmo residió en la decisión de qué rango de autovalores proporcionaba un resultado óptimo del algoritmo, el cual tuvo que ser determinado empíricamente mediante pruebas. Como un resultado de interés, tanto académico como de aplicación, se ha encontrado que la consistencia semántica de los adjetivos es guía suficiente para la determinación del espectro elegido, aspecto que puede ser automatizado. La efectividad de este algoritmo parece demostrada en los resultados mostrados en los experimentos realizados, en los que se observa que los sentimientos se agrupan en grupos semánticos independientes. Estos grupos parecen completamente lógicos en base a la similitud en significado de los sentimientos que los conforman, permitiendo su clasificación semántica. De igual manera, aunque con una similitud en significado algo menor, las marcas automovilísticas pueden ser identificadas en diferentes grupos semánticos, dependiendo de la gama y las prestaciones de sus modelos. Una vez identificados los grupos, la similitud entre grupos de marcas y grupos de sentimientos parecen razonables, entendiendo que los grupos de marcas estarán próximos a aquellos grupos de sentimientos que potencian su atractivo hacia el lector. Además, tanto el léxico como el umbral de decisión empleados en el

procesado lingüístico de los documentos de la colección juegan un papel importante en la consecución de los experimentos, aportando gran valor a los resultados obtenidos.

Desde el punto económico y de marketing, se presenta el reto de decidir qué sentimientos son los más acordes para satisfacer la necesidad de investigación de mercado de una compañía. Además, los resultados presentados por el sistema arrojan preguntas en base al marketing en función de la razón por la cual las marcas de estudio son representadas de tal manera en el espacio semántico. Los motivos tendrán que ser averiguados por los responsables de esta área con el objetivo de identificar el porqué de estas debilidades o fortalezas en el posicionamiento relativo de unas marcas frente a otras, con el fin de mejorar su estrategia de marketing y ganar ventaja competitiva en el sector.

5.2. Líneas Futuras

A consecuencia del ámbito del proyecto llevado a cabo existen numerosas líneas de trabajo que surgen como ideas durante su implementación. A continuación se detallan las posibles líneas futuras de trabajo a las que da origen el presente proyecto.

5.2.1. Análisis sintáctico

El aumento de funcionalidades permitiría en un futuro mejorar los resultados siendo el primer enfoque mejorar el alcance del procesador lingüístico.

- Identificar el sujeto sobre el que va dirigido el sentimiento permitiría una mejor representación de las marcas en el espacio semántico y evitaría los problemas sufridos con los sentimientos con polaridad negativa.

- Además del uso de diccionarios, la construcción de conjuntos de reglas permitiría solucionar las dificultades en la detección de aquellas entidades que pueden aparecer en diferentes formas. Estas reglas aplicarían patrones de expresiones regulares a las entidades del léxico. Un claro ejemplo sería el fabricante de automóviles *Volkswagen* que con frecuencia aparece en los textos de la forma *VW*. La regla que solventaría el problema presentado en el ejemplo sería *Volkswagen => VW*.

- Otra mejora del procesador sería la desambiguación basada en heurísticos que solucionaría el problema de la ambigüedad en la clasificación de entidades del léxico. Con el fin de posibilitar esta desambiguación semántica en la extracción

de información sería necesario emplear herramientas de análisis morfosintáctico y sintáctico como las que utiliza *Daedalus*. Un ejemplo de desambiguación sería el término *Mercedes* que dependiendo del contexto puede hacer referencia al fabricante de automóviles o a una persona. Los heurísticos para establecer la clasificación de esta entidad podrían ser los siguientes:

1. *Mercedes* clasificado como fabricante si en el texto aparece la palabra coche.
2. Preposición + *Mercedes*, por ejemplo: 'a *Mercedes*', entonces *Mercedes* clasificado como nombre de persona.
3. Artículo + *Mercedes*, por ejemplo: 'el *Mercedes*', entonces *Mercedes* clasificado como fabricante.

5.2.2. Redes sociales

Otra línea de ampliación sería la extensión del trabajo a comunidades electrónicas (foros de discusión y redes sociales). El problema que aquí se presenta es que en estos sitios las reglas de gramática y ortografía son frecuentemente ignoradas por sus usuarios (faltas de ortografía o uso de letras repetidas entre otros) dificultando la detención de sentimientos. El uso de correctores ortográficos durante el procesado lingüístico sería necesario para solventar el problema.

5.2.3. Usabilidad

Por último, y con vistas a negocio, sería importante trasladar este proyecto a un proyecto software listo para ser comercializado a compañías. Este software debería atender todos los requisitos de usabilidad necesarios para que los usuarios puedan trabajar de la manera más fácil posible con el sistema.

Capítulo 6

Anexo I: Herramientas desarrolladas ad hoc

Índice Capítulo 6

6.1. Introducción

En este anexo se recogen todos los aspectos relacionados con la implementación de cada uno de los módulos que conforman el sistema desarrollado, así como las herramientas de programación y lenguajes utilizados.

6.2. Metodología de la programación

Se ha elegido la orientación a objetos como paradigma de la programación para las herramientas de recolección de documentos y procesamiento lingüístico nombradas como *Mi Araña Web* y *Mi Procesador*. Ambas herramientas son aplicaciones de escritorio en lenguaje *Java* siendo *NetBeans* el entorno de desarrollo utilizado. Para su elección han sido consideradas las siguientes ventajas:

- IDE libre y gratuito sin restricciones de uso.
- Fácil de usar, cómodo y de calidad.
- Asistente para la creación de interfaces sencillo de usar, permitiendo crear interfaces de usuario por medio de ventanas.

La funcionalidad de ambas herramientas reside en la tecnología facilitada por *Apache Lucene*. *Lucene* es una API flexible que permite añadir capacidades de indexación y búsqueda a cualquier sistema que se esté desarrollando. Las principales ventajas para su elección son descritas a continuación:

- Indexación incremental. Se pueden añadir documentos a un índice ya creado con anterioridad.
- Contenido etiquetado. *Lucene* permite dividir el contenido de los documentos en campos y así poder realizar consultas con un mayor contenido semántico, buscando términos en los distintos campos del documento. También se puede optar a tener campos almacenados pero no indexados como metadatos asociados al documento.
- Técnicas de indexación. Eliminación de palabras poco representativas del documento, o *stopwords*, como *"a"*, *"el"*, *"la"*, etcétera., reduciendo el tamaño del índice y el tiempo de indexación.
- Multiplataforma.

6.3. Herramienta Mi Araña Web

La herramienta *Mi Araña Web* permite recolectar artículos de coches de la Web para la obtención de una colección de documentos sobre la cual poder realizar el procesado lingüístico.

Para mostrar el modelado de la herramienta se utiliza un diagrama de clases. En este diagrama se muestran las relaciones entre las clases más importantes de la herramienta *Mi Araña Web*, siendo omitidas las clases de menor importancia para simplificar el diseño y su comprensión.

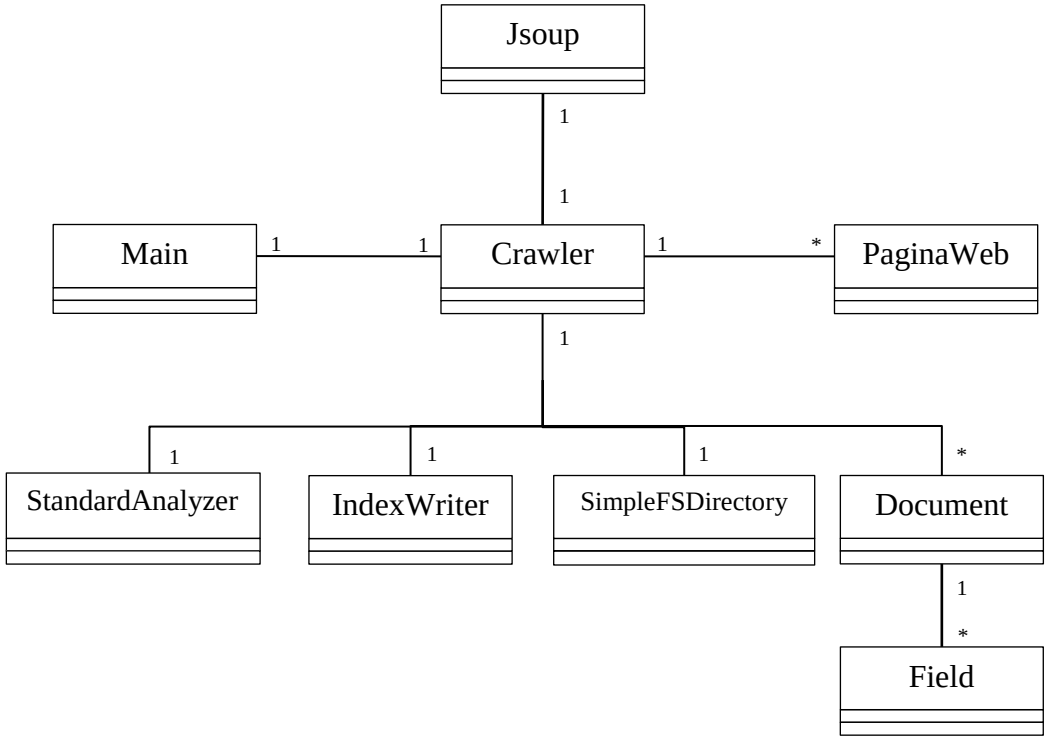


Figura 29. Diagrama de clases de la herramienta Mi Araña Web

A continuación se describen las clases visualizadas en el diagrama anterior:

IndexWriter

La clase *IndexWriter* se usa tanto para la creación del índice como para su mantenimiento. Cuando se crea el objeto de esta clase, al constructor se le pasan los siguientes cuatro parámetros:

- El primer parámetro representa la ruta donde será creado el índice (*Directory*).
- El segundo parámetro es el analizador de texto (*Analyzer*).
- El tercer parámetro pregunta el tipo de operación que se desea realizar en el índice. En nuestro caso queremos crear el índice por lo que pasamos *true* como valor del parámetro. Si hubiésemos querido añadir documentos a un índice ya creado con anterioridad pasaríamos el valor *false*.
- El cuarto parámetro limita el tamaño de los campos de los documentos a indexar. Hemos considerado no limitar su tamaño, por lo que se pasa *UNLIMITED* como valor.

Para añadir un documento al índice basta con invocar el método *addDocument*. Siempre que se termine de usar el *IndexWriter* se optimiza y se cierra invocando a los métodos *optimize* y *close*.

Document

La clase *Document* se usa para crear el documento a indexar. Un documento está formado por un conjunto de campos (*Fields*). Al crear el objeto de esta clase el constructor no solicita ningún parámetro de entrada, construyendo un nuevo documento sin ningún campo. Cada vez que se quiera añadir un campo al documento bastará con invocar el método *add*, con el nuevo campo como parámetro de entrada.

Field

La clase *Field* se usa para añadir un nuevo campo a un documento. Cuando se crea el objeto de esta clase, al constructor se le pasan los siguientes parámetros:

- El primer parámetro representa el nombre del campo.
- El segundo parámetro representa el valor del campo.
- El tercer parámetro pregunta si se desea que el valor del campo sea almacenado en el índice pudiendo ser entonces los datos originales del campo recuperados. Posibles valores: *YES* o *NO*.
- El cuarto parámetro pregunta si el campo será indexado para las búsquedas. Posibles valores:
    1. *ANALYZED*: el valor del campo será "*tokenizado*" previamente a que sea indexado, siendo convertido en una secuencia de "*tokens*" o trozos de texto.
    2. *NO*: en el caso en que no se quiera indexar el campo.

3. *NOT_ANALYZED*: No pasa por el filtro (*Analyzer*).

StandardAnalyzer

Esta clase sirve como filtro cuando se indexa texto. La clase *Analyzer* se usa para preprocesar el texto de entrada convirtiendo éste en una secuencia de "*tokens*".

- El texto del campo a indexar es segmentado en unidades (palabras)
- Los signos de puntuación localizados en el texto son suprimidos.
- Las letras mayúsculas son convertidas a minúsculas.

Estos analizadores también suprimen del índice las palabras poco representativas del documento como *"a"*, *"el"*, *"la"*, etcétera., conocidas como *stopwords*, reduciendo considerablemente el tamaño del mismo. El analizador aquí utilizado es *StandardAnalyzer*, cuya lista de *stopwords*, por desgracia, está pensada únicamente para la lengua inglesa. Otros analizadores emplean algoritmos de extracción de raíces[9] (*stemming*), o de eliminación de sufijos, orientados a obtener un único término a partir de diferentes palabras que constituyen, esencialmente, variaciones morfológicas con un mismo significado. El resultado de aplicar el algoritmo sobre el campo a indexar será el lema de cada una de las palabras, permitiendo trabajar únicamente con las raices lingüísticas de las mismas.

Lista de *analyzers*: *BrazilianAnalyzer, ChineseAnalyzer, CJKAnalyzer, CzechAnalyzer, DutchAnalyzer, FrenchAnalyzer, GermanAnalyzer, GreekAnalyzer, KeywordAnalyzer, PatternAnalyzer, PerFieldAnalyzerWrapper, QueryAutoStopWordAnalyzer, RussianAnalyzer, ShingleAnalyzerWrapper, SimpleAnalyzer, SnowballAnalyzer, SpanishAnalyzer,* ***StandardAnalyzer****, StopAnalyzer, ThaiAnalyzer y WhitespaceAnalyzer.*

SimpleFSDirectory

La clase *SimpleFSDirectory* almacena ficheros del índice en el fichero del sistema.

Jsoup

La clase *Jsoup* se usa para crear una nueva conexión a una URL. El método *connect* permitirá obtener y analizar el documento HTML de la página de destino.

En el diagrama de clases se han incluído dentro de la clase *Jsoup* las clases *Document*, *Element* y *Elements*.

[9] http://snowball.tartarus.org/

1. La clase *Document* será el documento HTML de la página de destino.
2. Las clases *Element* y *Elements* permiten manipular elementos contenidos en el documento HTML de la página (*Document*), permitiendo sus métodos obtener la información con la que rellenar los campos de los documentos a añadir al índice y extraer las URLs contenidas en la página.

PaginaWeb

La clase *PaginaWeb* se usa para almacenar información previa de la página que el rastreador tiene pensado en un futuro recolectar.

Crawler

La clase *Crawler* se usa para realizar la recolección de documentos, siendo el objeto de esta clase equivalente al rastreador web descrito en el apartado 3.3. Esta clase contiene todos los métodos necesarios para crear el índice de documentos, recolectar páginas web (obteniendo tanto el texto como las URLs contenidas en ellas) y añadir documentos al índice.

A continuación se muestran los atributos y métodos de la clase y se escribe una breve descripción de aquellos métodos más importantes:

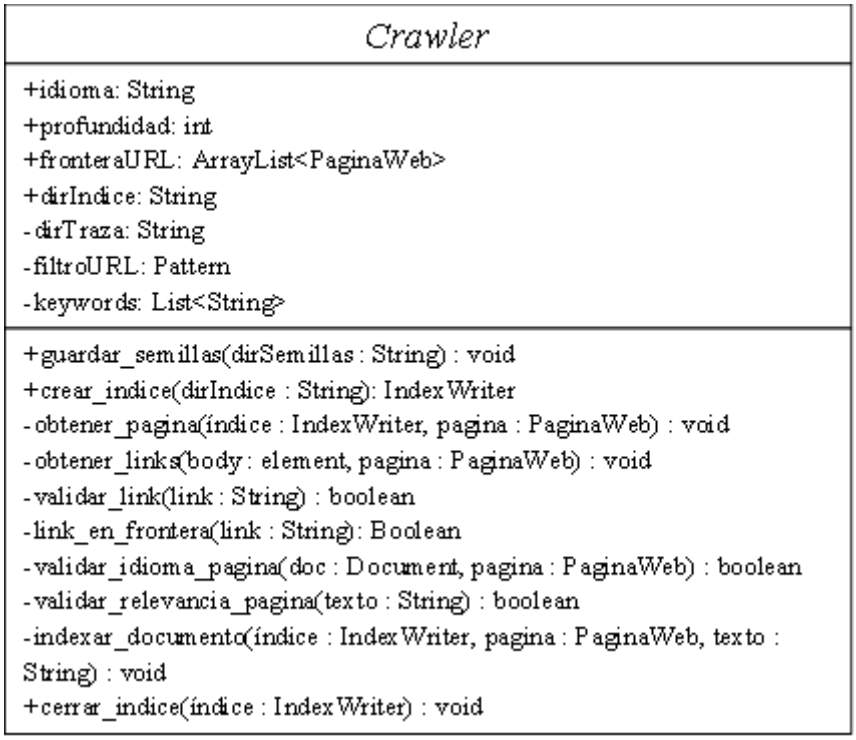


Figura 30. Atributos y métodos de la clase Crawler

guardar_semillas

Este método lee las URLs semillas contenidas en el fichero de texto. Por cada URL leída se crea un objeto de la clase *PaginaWeb* el cual es añadido a la frontera URL. Los parámetros a pasar al constructor serán:

- URL semilla.
- "SEMILLA"
- Profundidad de búsqueda= 0.

crear_indice

Este método crea el índice de documentos construyendo un objeto de la clase *IndexWriter* que será almacenado en el directorio especificado por parámetro.

obtener_pagina

Este método invoca a los métodos correspondientes para realizar el proceso de recolección completo de la página.

obtener_links

Este método invoca al método *getElementsByTag* del objeto *Elements* para extraer todos los textos ancla *<a>* de los hiperenlaces contenidos en el cuerpo del documento HTML la página. Por cada hiperenlace extraído se sigue la siguiente operación:

1. Se obtiene la URL canónica de la página de destino invocando al método *attr* del objeto *Element*.
2. Se pasa la URL por el filtro URL creado invocando al método *validar_link* del objeto *Crawler*.
3. Se comprueba que la URL extraída no esté ya en frontera invocando al método *link_en_frontera* del objeto *Crawler*.
4. Si la URL supera los filtros se crea un objeto de la clase *PaginaWeb* pasándole al constructor los siguientes parámetros:
   - URL extraída (en su forma canónica).
   - URL de la página que se está recolectando en ese momento y que será la página padre de la URL extraída.
   - Profundidad de búsqueda que se alcanzaría al obtener la página a la que apunta la URL extraida (Profundidad página actual + 1).

   Una vez que el objeto es creado es añadido a la frontera URL, que será un *ArrayList* de objetos de esta clase.

validar_idioma_pagina

Este método identifica el idioma de la página y valida que sea igual al deseado por el rastreador. Invoca al método *getElementsByTag* del objeto *Document* para

obtener la etiqueta <*html* > del documento HTML de la página donde está (o puede estar) localizado el idioma de la página.

validar_relevancia_pagina

Este método busca las palabras claves asociadas a la temática de búsqueda del rastreador en el texto extraído de la página para determinar la relevancia de la misma.

indexar_documento

Este método crea el documento a añadir al índice creando un objeto de la clase *Document*, le añade los correspondientes campos creando objetos de la clase *Field* e invocando al método *add* del objeto *Document* creado y, finalmente, lo indexa invocando método *addDocument* del objeto *IndexWriter*.

cerrar_indice

Este método invoca a los métodos *optimize* y *close* del objeto *IndexWriter* para optimizar y cerrar el índice de documentos.

6.4. Herramienta Mi Procesador

La herramienta *Mi procesador* permite obtener una representación compacta de los documentos que forman la colección para que puedan ser interpretados por el algoritmo empleado en el procesamiento semántico.

Para mostrar el modelado de la herramienta se utiliza un diagrama de clases. En este diagrama se muestran las relaciones entre las clases más importantes de la herramienta *Mi procesador*, siendo omitidas las clases de menor importancia para simplificar el diseño y su comprensión.

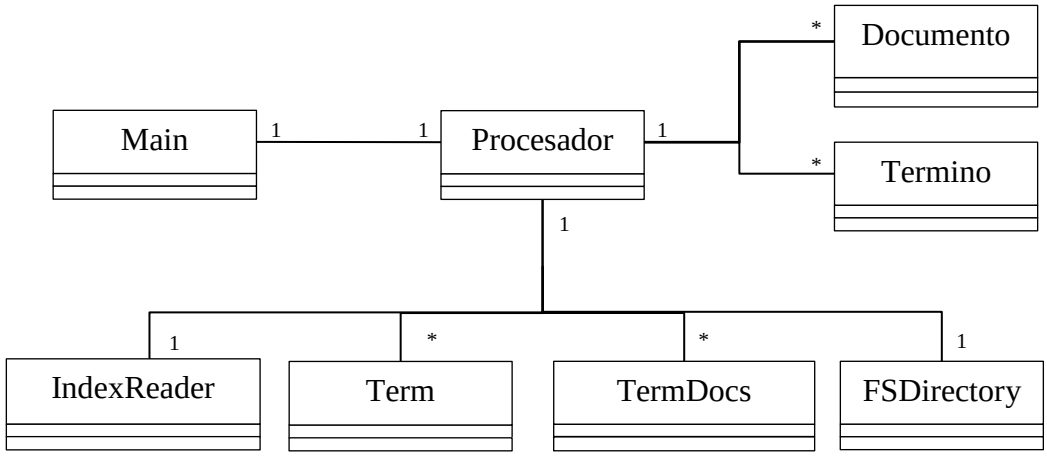


Figura 31. Diagrama de clases de la herramienta Mi Procesador

A continuación se describen las clases visualizadas en el diagrama anterior:

IndexReader

La clase IndexReader se usa para acceder al índice. Para abrir el índice se invoca al método *open*, al que se le pasan los siguientes dos parámetros:

- El primer parámetro representa el directorio donde el índice está almacenado en el sistema (*Directory*).
- El segundo parámetro pregunta el modo en el que se desea acceder el índice. En nuestro caso queremos acceder al índice en modo solo lectura por lo que pasamos *true* como valor del parámetro. Si hubiésemos querer permitir modificaciones sobre el índice al que se va acceder pasaríamos como valor *false*.

Term

La clase *Term* se usa para definir el término a buscar en el índice. Un término representa una palabra del texto siendo éste la unidad de búsqueda. Cuando se crea el objeto de esta clase, al constructor se le pasan los siguientes dos parámetros:

- El texto de la palabra.
- El nombre del campo en el que el texto se produjo.

TermDocs

*TermDoc*s no es una clase propiamente dicha, sino una interfaz con una colección de métodos. *TermDocs* proporciona una interfaz para enumerar las parejas <documento, frecuencia> para un término. La parte de documento nombra a cada documento que contiene el término (los documentos son indicados por números) y la parte de frecuencia da el número de veces que el término aparece en cada documento.

FSDirectory

La clase *FSDirectory* es utilizada para abrir el directorio que contiene los ficheros generados en la creación del indice.

Documento

La clase *Documento* se usa para almacenar toda la información obtenida en el procesamiento de un documento y que será la base para obtener una representación compacta del documento que sea interpretable por el algoritmo empleado en el procesamiento semántico. Cuando se crea el objeto de esta clase, al constructor se le pasan los siguientes parámetros:

- Número de documento en el índice.
- Texto del campo URL del documento.

Termino

La clase *Termino* se usa para almacenar la información obtenida de un término en el procesamiento de los documentos del índice y que se verá reflejada en los ficheros de salida de la herramienta. Cuando se crea el objeto de esta clase, al constructor se le pasan los siguientes parámetros:

- Texto del término.
- Clasificación del término (dependiendo del diccionario al que pertenezca).

Procesador

La clase *Procesador* se usa para realizar la representación de los documentos tal y como se describió en el apartado 3.4. Esta clase contiene todos los métodos necesarios para leer el índice de documentos creado por el rastreador y procesar los documentos en base al léxico y los umbrales de decisión establecidos por el usuario.

A continuación se muestran los atributos y métodos de la clase y se escribe una breve descripción de aquellos métodos más importantes:

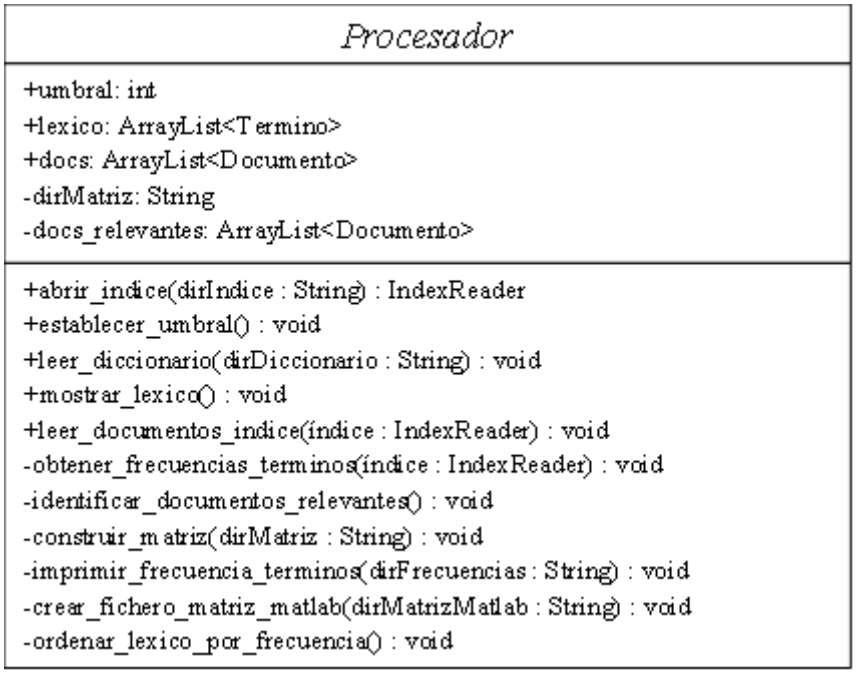


Figura 32. Atributos y métodos de la clase Procesador

abrir_indice

Este método abre el índice de documentos invocando al método *open* de la clase *IndexReader*. Previamente habrá que crear un objeto de la clase *FSDirectory*, invocando al método *open* de la susodicha clase. La razón es que es necesario abrir el índice

utilizando la misma implementación que se usó en la herramienta anterior en el almacenado de los ficheros del índice en el sistema.

establecer_umbral

Este método establece el umbral a superar por un documento para ser marcado como relevante en su representación.

leer_diccionario

Este método almacena en un array los términos que conforman el léxico. El directorio que contiene el léxico es facilitado por el usuario del sistema y estará compuesto de uno o varios ficheros que son los diccionarios en los que se divide el léxico. La lectura de cada fichero se realiza empleando la norma *ISO 8859-1* que define la codificación del alfabeto latino evitando problemas con tíldes y caracteres especiales. Por cada término leído en el interior de estos ficheros se crea un objeto de la clase *Termino* que es añadido al *ArrayList termino*.

leer_documentos_indice

Este método recorre el índice de documentos creando un objeto *Documento* por cada documento leído que es añadido al *ArrayList docs*. El tamaño del índice viene determinado por el método *maxDoc* del objeto *IndexReader*.

obtener_frecuencias_terminos

Este método obtiene por cada documento el número de apariciones de los términos del léxico. Para lograr las frecuencias de aparición se crea un objeto de la clase *Term* por cada término del léxico, siendo *Texto* el campo de búsqueda. Con el método *TermDocs* del objeto *IndexReader* se crea un objeto *TermDocs* con la enumeración de documentos donde se encuentra presente el término y el número de apariciones correspondiente. Esta información es recogida en los objetos *Documento* afectados.

identificar_documentos_relevantes

Este metodo recorre el *ArrayList docs* evaluando todos los objetos *Documento* almacenados en él. Si el objeto *Documento* supera el umbral de decisión establecido es marcado como relevante y añadido al *ArrayList docs_relevantes*.

construir_matriz

Este método dibuja la matriz término por documento en un fichero de texto dentro del directorio de salida especificado por el usuario. Esta matriz unicamente será contruida con los vectores documento de los documentos contenidos en el *ArrayList docs_relevantes*.

6.5. Procesador semántico

El procesado semántico es íntegramente realizado en *MATLAB*, herramienta que permite, entre otras funcionalidades, la manipulación de matrices y la representación de datos.

Las funciones empleadas en la realización del procesado semántico y la presentación de los resultados se describen a continuación.

Función	Descripción	Parámetros entrada	Parámetros salida
lsa.m	Función desarrollada que aplica el algoritmo *Latent Semantic Analysis* sobre la matriz término por documento: *SVD* de la matriz más correlación de *Spearman*.	- Matriz término por documento. - Número de valores singulares a eliminar en *SVD*.	- Matriz de correlaciones.
distancias.m	Función desarrollada que obtiene la matriz de distancias a partir de la matriz de correlaciones.	- Matriz de correlaciones.	- Matriz de distancias.
mdscale.m	Función definida por *MATLAB* que realiza el escalado multidimensional sobre una matriz dada.	- Matriz de distancias. - Dimensión *d* (2 en nuestro caso).	-Configuración de *n* puntos en 2 dimensiones.
plot.m	Función definida por *MATLAB* que construye una gráfica lineal en 2D.	-Configuración de *n* puntos en 2 dimensiones.	-Representación gráfica del espacio semántico.
horzcat.m (*opcional*)	Función definida por *MATLAB* que concatena arrays y matrices horizontalmente.	- Matrices término por documento a fusionar.	- Matriz término por documento.
energia.m (*opcional*)	Función desarrollada que obtiene la energía de la matriz diagonal empleada en *SVD*.	- Matriz diagonal. - Número de valores singulares a eliminar en *SVD*.	- Energía empleada en *SVD* (%).
singulares.m (*opcional*)	Función desarrollada que a partir de una energía dada calcula el número de valores singulares a mantener en la matriz diagonal.	- Matriz diagonal. - Energía (0._).	- Número de valores singulares a mantener en matriz diagonal.

Tabla 18. Funciones empleadas en el procesado semántico

Capítulo 7

Anexo II: Manual de usuario

Índice Capítulo 7

7.1. Introducción

Este manual ha sido elaborado con la intención de ofrecer la información necesaria para el uso del sistema desarrollado. Por cada una de las herramientas que lo conforman se sigue la siguiente estructura:

- Interfaz de usuario.
- Parámetros de entrada y salida.
- Pasos a seguir.

7.2. Mi Araña Web

A continuación se muestra la interfaz de usuario de la herramienta *Mi Araña Web* y se especifican los parámetros de entrada y salida.

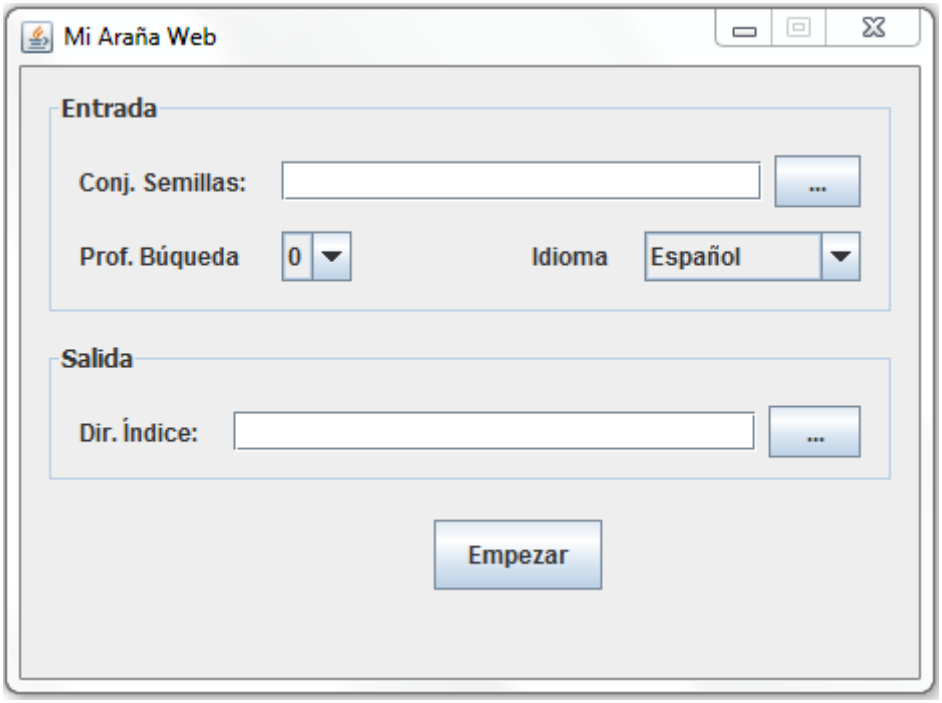


Figura 33. Interfaz herramienta Mi Araña Web

Campo	Descripción
Conj. Semillas	Fichero de texto con el conjunto de semillas.
Prof. Búsqueda	Profundidad de búsqueda de la araña web, a elegir: 0 (ninguna profundidad, opción por defecto), 1, 2, 3 o 4.
Idioma	Idioma de las páginas a recolectar, a elegir: Español (opción por defecto) o Inglés.
Dir.Índice	Directorio de salida donde guardar el índice Lucene generado por la araña web.

Tabla 19. Parámetros de entrada y salida de la herramienta Mi Araña Web

Selección del fichero con el conjunto de URLs semillas

El fichero de texto contendrá las URLs de las páginas desde donde se quiere que la Araña Web comience con el proceso de recolección. Válida una única URL por línea de texto.

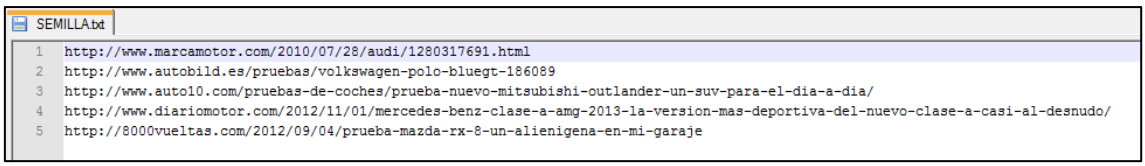


Figura 34. Fichero SEMILLA.txt con el conjunto de URLs semilla

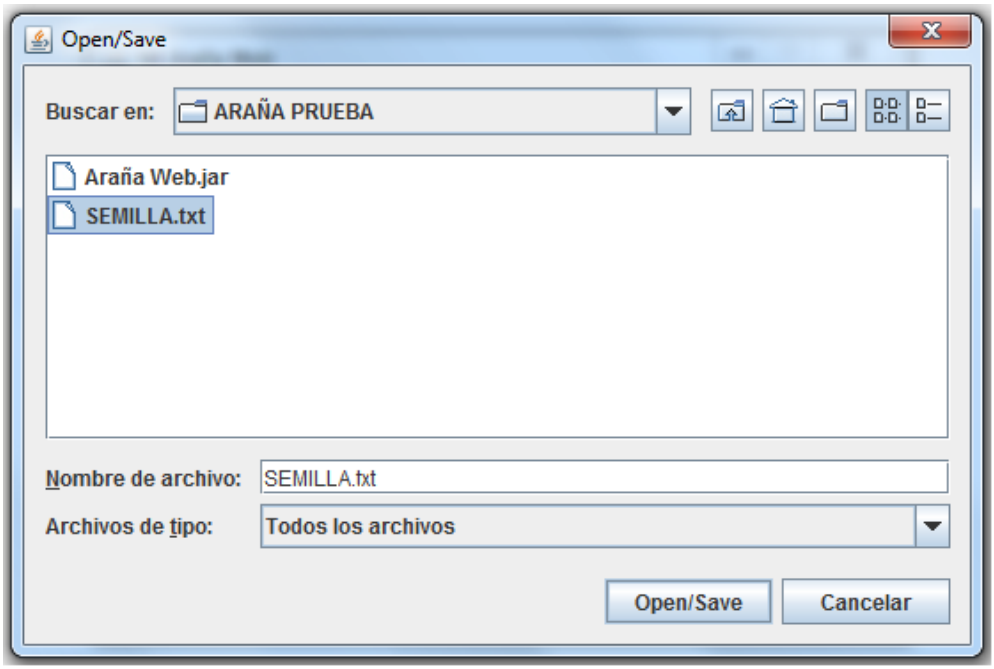


Figura 35. Selección desde la herramienta Mi Araña Web del fichero SEMILLA.txt

Selección de la profundidad de búsqueda

Si se deja la opción por defecto (opción 0) la araña solo recolectará las páginas de las URLs del conjunto semilla.

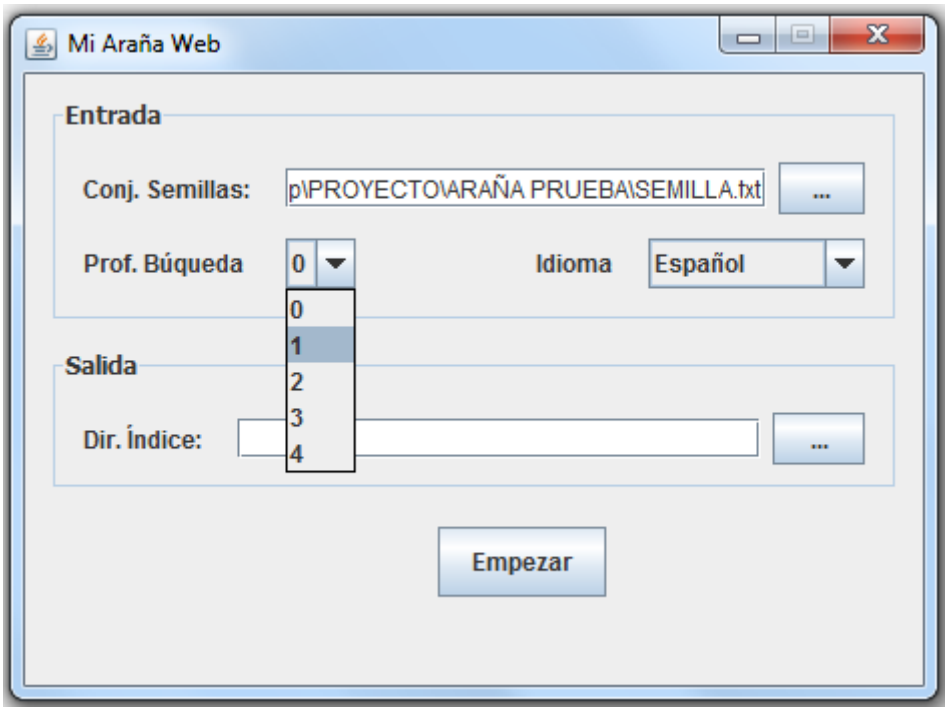


Figura 36. Selección desde la herramienta Mi Araña Web de la profundidad de búsqueda

Selección de idioma

Las páginas recolectadas por la araña web cuyo idioma identificado sea diferente al seleccionado no serán indexadas. Identificación mediante el atributo *lang* de HTML. Dos opciones disponibles: Español (opción por defecto) e Inglés.

Creación del directorio salida del índice Lucene

En *Nombre de archivo* escribir el nombre deseado del directorio. Pulsar tecla *Enter* o seleccionar opción *Open/Save* para su posterior creación.

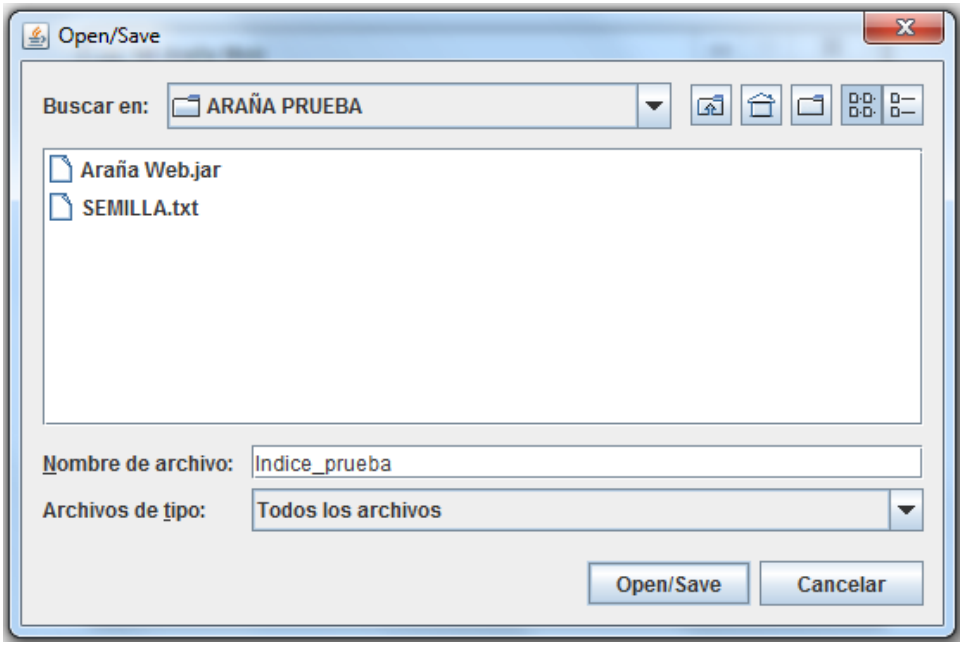


Figura 37. Creación directorio de salida desde la herramienta Mi Araña Web donde guardar el índice

Ejecución de la Araña Web

Una vez completados los formularios de *Entrada* y *Salida* pulsar botón *Empezar*. Una barra de progreso indicará el progreso de la ejecución. Si se cierra la ventana de *Progreso* durante la ejecución se aborta el proceso.

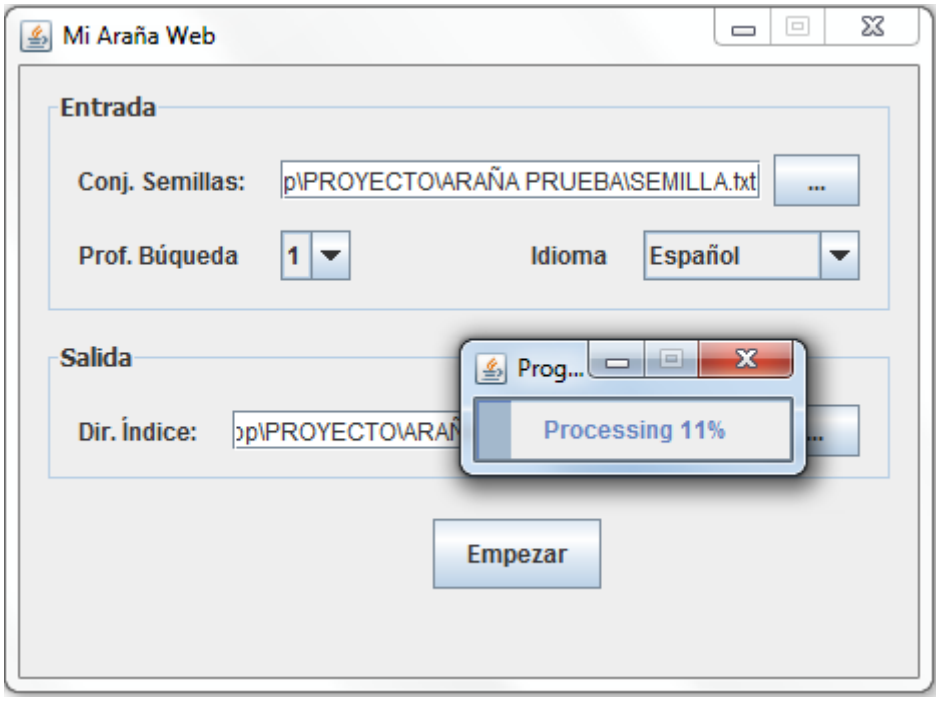


Figura 38. Ejecución de la Araña Web desde la herramienta Mi Araña Web

Resultados

Terminado el proceso son generados en el directorio de salida el índice Lucene y el fichero de texto con la traza de ejecución de la herramienta.

Nombre de fichero traza: traza_paginas_araña

Formato fichero traza: URL + Título página + Resultado: OK (página indexada), Otro (página no indexada + detalle del error) + Idioma página + Padre página + Profundidad de búsqueda en la que página fue recolectada + Número de Links (URLs) extraídas de la página.

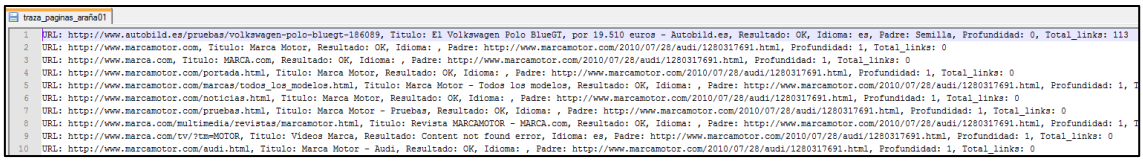


Figura 39. Ejemplo traza de ejecución generada por la herramienta Mi Araña Web

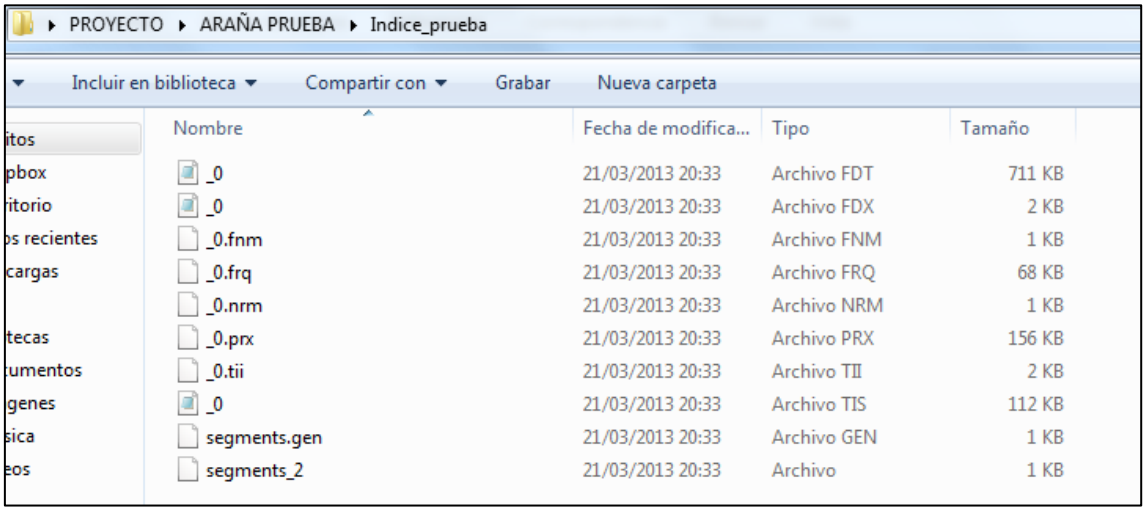


Figura 40. Directorio con el índice Lucene generado

7.3. Mi Procesador

A continuación se muestra la interfaz de usuario de la herramienta *Mi Procesador* y se especifican los parámetros de entrada y salida.

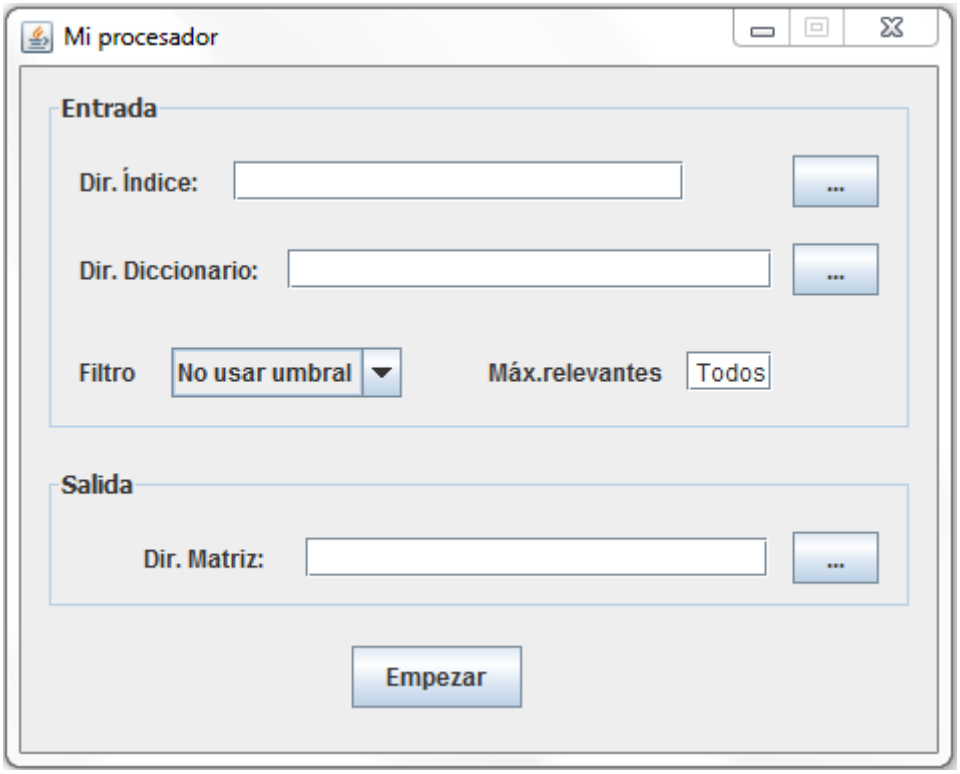


Figura 41. Interfaz herramienta Mi Procesador

Campo	Descripción
Dir.Índice	Directorio donde se guarda el índice Lucene sobre el que trabajar.
Dir.Diccionario	Directorio donde se guarda el léxico (diccionarios) a utilizar.
Filtro	Umbral a utilizar en la selección de documentos relevantes, a elegir: No usar umbral (opción por defecto), Umbral 1 o Umbral 2.
Máx.relevantes	Máximo de documentos relevantes deseados, a elegir: Todos (Todos los documentos marcados como relevantes) o valor numérico.
Dir.Matriz	Directorio donde guardar matriz término por documento generada por el procesador.

Tabla 20. Parámetros de entrada y salida de la herramienta Mi Procesador

Selección del directorio del índice Lucene

Seleccionar *Open/Save* sobre alguno de los ficheros contenidos en directorio índice para seleccionar directorio.

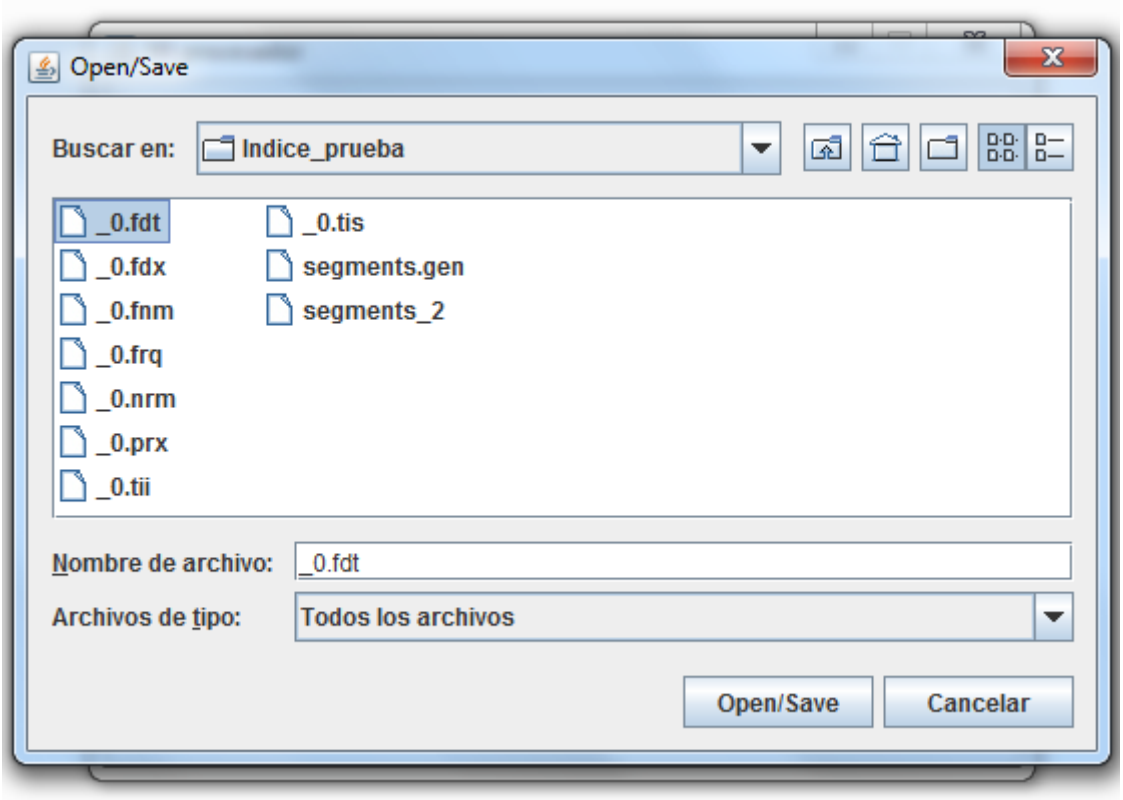


Figura 42. Selección desde la herramienta Mi Procesador del directorio con índice Lucene

Selección del directorio del léxico de términos

Seleccionar *Open/Save* sobre alguno de los ficheros contenidos en el directorio del léxico para seleccionar directorio. El directorio contendrá uno o más ficheros de texto, siendo cada fichero uno de los diccionarios que conforman el léxico. Válido un único término por línea de texto. Es indiferente si los términos son escritos en mayúsculas o minúsculas.

Para la clasificación de términos nombrar los ficheros de texto de la siguiente manera:

1. _MARCAS: Cada término del fichero es clasificado como Marca.
2. _SENTIMIENTOS: Cada término del fichero es clasificado como Sentimiento.
3. _ALEATORIOS: Cada término del fichero es clasificado como Aleatorio.
4. “OTRO”: Términos sin clasificación.

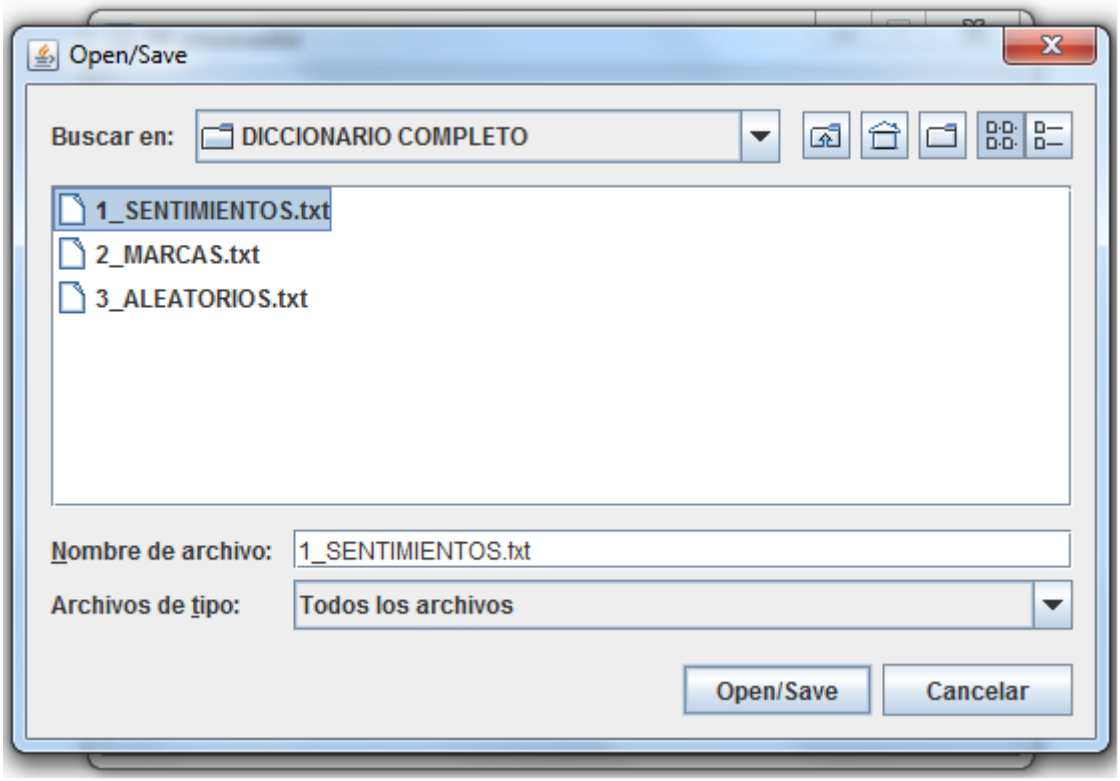


Figura 43. Selección desde la herramienta Mi Procesador del directorio con diccionarios del léxico

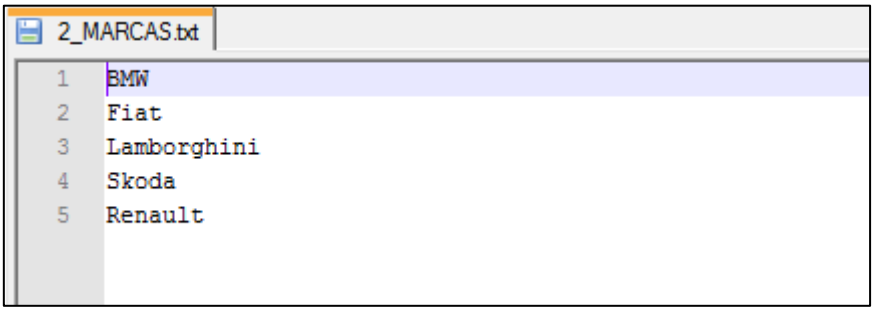


Figura 44. Ejemplo diccionario de términos

Selección de filtro/umbral

Requisito a superar por un documento del índice para ser marcado como relevante.

Umbral	Descripción
No usar umbral	Todos los documentos del índice son marcados como relevantes
Umbral 1	2 o + sentimientos presentes en el texto del documento.
Umbral 2	1 o + marcas presentes en el texto del documento. Igual a 2 o mayor el número de apariciones de sentimientos en el texto del documento.

Tabla 21. Umbrales de decisión a elegir en la herramienta Mi Procesador

Número máximo de documentos relevantes deseados

Ofrece limitar el número de documentos relevantes deseados en la matriz término por documento. Si se deja *Todos* (opción por defecto), o se escribe algo diferente a un número, todos los documentos marcados como relevantes durante el procesado formarán parte de la matriz. Si se quiere limitar este número entonces indicar número máximo.

Creación del directorio salida de la matriz término por documento

En *Nombre de archivo* escribir el nombre deseado del directorio. Pulsar tecla *Enter* o seleccionar opción *Open/Save* para su posterior creación.

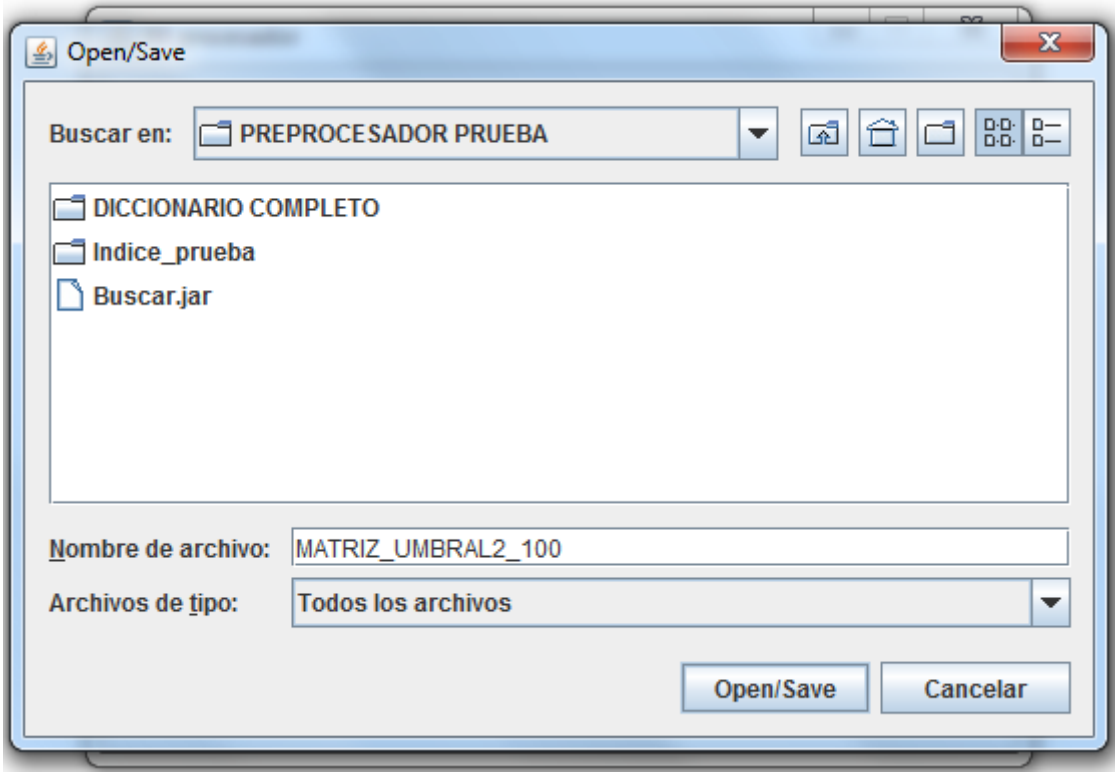


Figura 45. Creación directorio de salida desde la herramienta Mi Procesador donde guardar la matriz

Ejecución del procesador

Una vez completados los formularios de *Entrada* y *Salida* pulsar botón *Empezar*. Una barra de progreso indicará el progreso de la ejecución. Si se cierra la ventana de *Progreso* durante la ejecución se aborta el proceso.

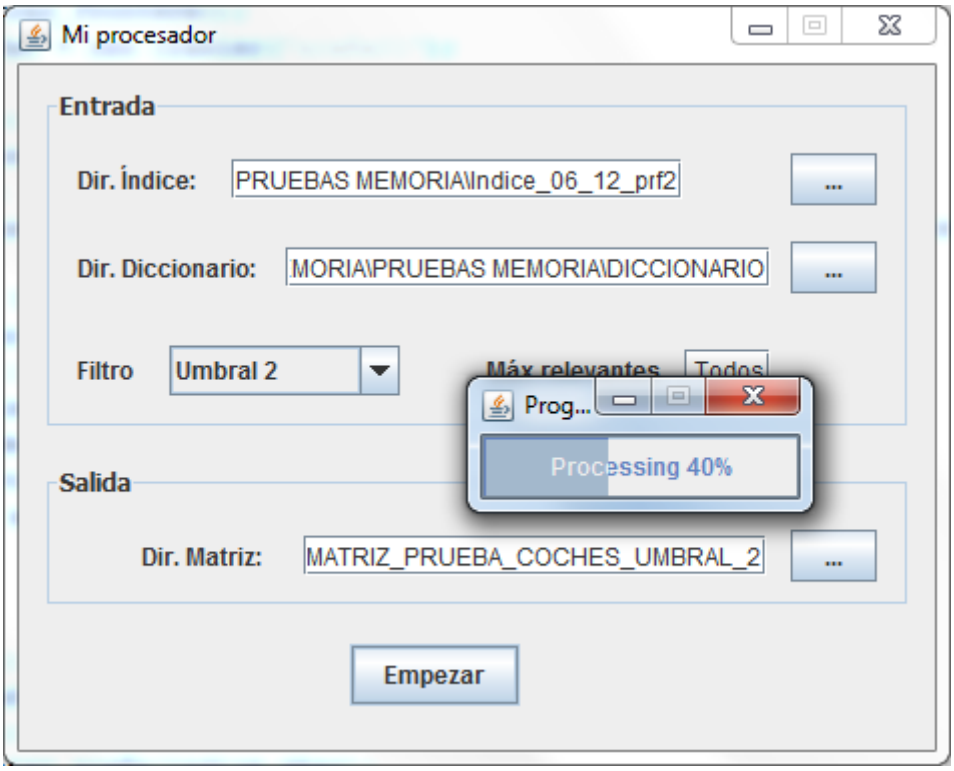


Figura 46. Ejecución del procesador desde la herramienta Mi Procesador

Resultados

Terminado el proceso son generados en el directorio de salida los siguientes ficheros:

Fichero	Descripción
Matriz.m	Fichero *MATLAB* con matriz término por documento. Se incluye la siguiente información: Fecha de inicio del procesado, usuario del equipo y umbral de decisión empleado.
Matriz.txt	Fichero de texto con matriz término por documento.
Termino_frecuencia.txt	Fichero de texto con los términos del léxico y el número de apariciones totales de cada uno de ellos en los documentos que conforman la matriz (ordenados descendentemente).
Urls_documentos_relevantes.txt	Fichero de texto con los campos URL de los documentos marcados como relevantes en el procesado.

Tabla 22. Ficheros generados por la herramienta Mi Procesador en el directorio de salida

matriz.m

```
%Sat May 25 19:37:28 CEST 2013
%Fernando
%Umbral 2
M=[0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 2, 1, 0,
1, 1, 2, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 2, 3, 0, 0, 0, 2, 0, 0, 2, 6, 1, 2,
0, 0, 0, 0, 0, 0, 0, 1, 1, 8, 6, 2, 2, 0, 1, 1, 2, 1, 2, 0, 0, 0, 1, 2, 0, 1,
1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0,
0, 1, 0, 2, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 2, 0, 0, 0, 0, 1, 0, 1, 1, 0,
0, 0, 0, 0, 0, 0, 0, 1, 2, 3, 0, 0, 0, 0, 0, 2, 0, 2, 3, 0, 1, 1, 0, 0, 0, 0,
1, 3, 1, 1, 0, 0, 21, 0, 18, 6, 2, 3, 0, 2, 1, 1, 0, 19, 2, 1, 2, 0, 2, 1, 0,
0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4, 0, 0, 0, 0, 0, 0, 2, 2, 0, 5, 0, 0,
0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 3, 1, 0, 1, 1, 0, 0, 0, 0, 4, 0, 0, 8, 0, 0, 0, 0, 0, 0, 2, 6, 1, 1, 3, 0,
```

Figura 47. Ejemplo fichero matriz.m

termino_frecuencia.txt

```
bmw (Marca): 1035
renault (Marca): 419
bonito (Sentimiento): 345
fiat (Marca): 224
divertido (Sentimiento): 171
cómodo (Sentimiento): 158
precioso (Sentimiento): 145
suave (Sentimiento): 142
lamborghini (Marca): 115
agradable (Sentimiento): 113
skoda (Marca): 72
```

Figura 48. Ejemplo fichero termino_frecuencia.txt

urls_documentos_relevantes.txt

```
http://8000vueltas.com/2012/09/04/prueba-mazda-rx-8-un-alienigena-en-mi-garaje
http://www.marcamotor.com/pruebas.html
http://www.autobild.es/pruebas/vw-polo-gti-car%C3%A1cter-agresivo
http://www.autobild.es/pruebas/nuevo-seat-leon-20-tdi-184-fr-190736
http://www.auto10.com/tecnologia-coches/cuantos-cv-necesito/#comment-32073
http://www.auto10.com/tecnologia-coches/cuantos-cv-necesito/#comment-31873
http://www.diariomotor.com/
http://www.diariomotor.com/2012/11/06/chevrolet-corvette-gs-a-prueba-i-un-deportivo-a-la-americana/
http://www.diariomotor.com/2012/05/13/bmw-presenta-al-serie-1-de-3-puertas-y-nos-muestra-al-m135i-de-produccion/
http://www.diariomotor.com/2012/03/05/mercedes-clase-a-llega-el-nuevo-compacto-de-la-flecha-plateada/
http://www.diariomotor.com/2012/07/10/mercedes-clase-a-presentacion-y-prueba-en-eslovenia-i-diseno-y-deportividad-las-claves-del-nuevo-compacto-de-mercedes/
http://www.diariomotor.com/2012/10/04/mercedes-clase-a-presentacion-y-prueba-en-burgos-i-algo-ha-pasado-en-mercedes/
http://www.diariomotor.com/2012/11/15/quiere-seguir-siendo-el-rey-asi-podria-ser-el-nuevo-bugatti-veyron-de-1-600-cv/
http://8000vueltas.com/2012/08/30/mg-b-los-comienzos-de-un-deportivo-de-hace-50-anos-justos
http://8000vueltas.com/categoria/pruebas
```

Figura 49. Ejemplo fichero urls_documentos_relevantes.txt

Glosario

Blog	Es un sitio web periódicamente actualizado que recopila cronológicamente textos o artículos de uno o varios autores, apareciendo primero el más reciente, donde el autor conserva siempre la libertad de dejar publicado lo que crea pertinente.
Deflate	Es un algoritmo de compresión de datos sin pérdidas que usa una combinación del algoritmo LZ77 y la codificación Huffman.
Emoción	Alteración del ánimo intensa y pasajera, agradable o penosa, que va acompañada de cierta conmoción somática.
gzip	Abreviatura de GNU ZIP, un software libre GNU que reemplaza al programa compress de UNIX
Herramienta	Elemento elaborado con el objetivo de hacer más sencilla una determinada actividad o labor mecánica.
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
ISO_8859-1	Norma de la ISO que define la codificación del alfabeto latino, incluyendo los diacríticos (como letras acentuadas, ñ, ç), y letras especiales (como ß, Ø).
Java	Es un lenguaje de programación de propósito general, concurrente, orientado a objetos y basado en clases que fue diseñado específicamente para tener tan pocas dependencias de implementación como fuera posible.
Javascript	Es un lenguaje de programación que permite a los desarrolladores crear acciones en sus páginas web. Tiene la ventaja de ser incorporado en cualquier página web, puede ser ejecutado sin la necesidad de instalar otro programa para ser visualizado.
Librería	Es un conjunto de subprogramas utilizados para desarrollar software. Las bibliotecas contienen código y datos, que proporcionan servicios a programas independientes.
Marketing online	Es el estudio de las técnicas del uso de Internet para publicitar y vender productos y servicios.

PDF	Portable Document Format
Red social	Son estructuras sociales compuestas de grupos de personas, las cuales están conectadas por uno o varios tipos de relaciones, tales como amistad, parentesco, intereses comunes o que comparten conocimientos.
Sentimiento	Se refiere tanto a un estado de ánimo como a una emoción conceptualizada que determina el estado de ánimo. En análisis de sentimientos se refiere a los aspectos relacionados con las emociones de los mensajes.
Servidor	Es un nodo que forma parte de una red y que provee servicios a otros nodos denominados clientes. El término servidor se utiliza tanto para referirse al ordenador físico como al tipo de software que funciona en el equipo.
Sitio Web	Conjunto de páginas web que están relacionadas entre sí, por lo general porque se ingresan desde un mismo dominio.
URL	Uniform Resource Locator
Usuario	Es quien usa ordinariamente algo.
UTF-8	8-bit Unicode Transformation Format
Web 2.0	Es la transición que se ha dado de aplicaciones tradicionales hacia aplicaciones que funcionan a través del web enfocado al usuario final. Se trata de aplicaciones que generen colaboración y de servicios que reemplacen las aplicaciones de escritorio.
Web semántica	La Web Semántica es una Web extendida, dotada de mayor significado en la que cualquier usuario en Internet podrá encontrar respuestas a sus preguntas de forma más rápida y sencilla gracias a una información mejor definida.
WWW	World Wide Web
XML	Extensible Mark-up Language

Referencias

1. Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008.

2. Landauer, T. K., Foltz, P. W., & Laham, D. (1998). *Introduction to Latent Semantic Analysis*. Discourse Processes, 25, 259-284.

3. Michael W. Berry, Zlatko Drmac, and Elizabeth R. Jessup (1999). *Matrices, Vector Spaces, and Information Retrieval*. Society for Industrial and Applied Mathematics.

4. Soumen Chakrabarti, Martin van den Berg, and Byron Dom (1999). *Focused crawling: a new approach to topic-specific Web resource discovery*. WWW '99 Proceedings of the eighth international conference on World Wide Web.

5. R.Baeza-Yates, and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

6. Manuel Álvarez Díaz, *Arquitectura para Crawling Dirigido de Información Contenida en la Web Oculta*, Tesis doctoral, Universidad de Coruña, Diciembre de 2007.

7. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). *Sentiment strength detection in short informal text*. Journal of the American Society for Information Science and Technology, 61(12), 2544–2558. Copyright © 2010 (American Society for Information Science and Technology)

8. Jacinto Esteban Feliú, *Iniciación a la tecnología Lucene y aplicación*, Proyecto fin de carrera, Universidad Carlos III de Madrid (UC3M), Junio de 2011.

9. *Lucene 3.0.3 core API. [Acceso: 10 de Julio de 2013]* http://lucene.apache.org/core/3_0_3/api/core/index.html

10. *Jsoup 1.7.3-SNAPSHOT API. [Acceso: 10 de Julio de 2013]* http://jsoup.org/apidocs/

11. *Header Field Definitions (W3C). [Acceso: 10 de Julio de 2013]* http://www.w3.org/Protocols/rfc2616/rfc2616-sec14.html

12. *Status Code Definitions (W3C). [Acceso: 10 de Julio de 2013]* http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html

13. Jason S. Kessler, Miriam Eckert, Lyndsie Clark, and Nicolas Nicolov. *The ICWSM 2010 JDPA Sentiment Corpus for the Automotive Domain*. 4th International AAAI Conference on Weblogs and Social Media Data Challange Workshop (ICWSM-DCW 2010), 2010. Washington, D.C.

14. *IDC Forecasts Business Analytics Software Market to Continue on Its Strong Growth Trajectory Through 2017***.** IDC Press Release. *[Acceso: 10 de Julio de 2013]* Obtenido de: http://www.idc.com/getdoc.jsp?containerId=prUS24194613

15. *¿Cómo influye internet en nuestro proceso de decisión de compra? [Acceso: 10 de Julio de 2013]* Obtenido de: http://www.puromarketing.com/47/13924/como-influye-internet-nuestro-proceso-decision-compra.html

16. *Las opiniones, críticas y recomendaciones online ejercen una mayor influencia que nunca. [Acceso: 10 de Julio de 2013]* Obtenido de: http://www.puromarketing.com/88/10761/opiniones-criticas-recomendaciones-online-ejercen-mayor-influencia-nunca.html

17. *No podemos evaluar emociones y sentimientos de la misma manera que evaluamos o medimos ingresos. [Acceso: 10 de Julio de 2013]* Obtenido de: http://www.puromarketing.com/42/13743/podemos-evaluar-emociones-sentimientos-misma-manera-evaluamos-medimos-ingresos.html

18. *Adelantarse a la crisis de reputación. [Acceso: 10 de Julio de 2013]* Obtenido de: http://www.activainternet.es/adelantarse-crisis-reputacion/