Universidad Tecnológica Nacional
DICIEMBRE 2020 / Año 18- Nº 39

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.39.195-206.2020 - ISSN 1666-6933

Enfoque combinado de *Word2Vec* y *2-grams* para la recuperación de avisos clasificados inmobiliarios semánticamente relacionados

Combined approach of Word2Vec and 2-grams for the retrieval of semantically related real estate ads

Presentación: 15/10/2020

Aprobación: 01/12/2020

José Federico Medrano
Universidad Nacional de Jujuy - Argentina
jfmedrano@fi.unju.edu.ar

Resumen

La publicación de avisos clasificados de inmuebles se ha convertido en el medio de publicidad preferido tanto para particulares como empresas inmobiliarias. Esto ha provocado un crecimiento importante en la cantidad de avisos, tornando difícil la búsqueda un inmueble adecuado, mucho más si la búsqueda es en una gran ciudad. Este trabajo propone un enfoque basado en técnicas de minería de textos y procesamiento del lenguaje natural para la recuperación de avisos clasificados semánticamente relacionados. Para tal propósito se recolectaron los avisos publicados por el sitio web lavoz.com.ar, mediante un *scraper*. El título y la descripción de estos avisos fueron empleados para conformar un corpus textual modelado mediante *Word2Vec,* evaluando la similitud por medio de *Word Mover's Distance*. El empleo de 2-grams (bigramas) frente a otros esquemas de agrupación de términos ofrecieron los mejores resultados comparando los resultados con búsquedas sintácticas.

Palabras claves: Búsqueda semántica, Word2Vec, Procesamiento del Lenguaje Natural, Minería de Textos.

Abstract

The publication of real estate ads has become the preferred advertising medium for both individuals and real estate companies. This has caused a significant growth in the number of ads, making it difficult to search for a suitable property, much more so if the search is in a big city. This work proposes an approach based on text mining and natural language processing techniques for the recovery of semantically related classified ads. For this purpose, the ads

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.39.195-206.2020 - ISSN 1666-6933

published by the lavoz.com.ar website were collected, through a scraper. The title and description of these ads were used to form a textual corpus modeled using Word2Vec, evaluating the similarity using Word Mover's Distance. The use of 2-grams compared to other term grouping schemes offered the best results comparing the results with syntactic searches.

Keywords: Semantic search, Word2Vec, Natural Language Processing, Text Mining.

Introducción

Los clasificados inmobiliarios constituyen una parte integral del mercado inmobiliario. Este tipo de avisos cumplen el rol de informar, entregando una descripción concisa sobre una unidad que puede ser un terreno, un departamento o una casa disponible para la venta o alquiler. Durante años el único modo de publicitar estos avisos era por medios impresos como periódicos, pero paulatinamente el medio de publicación se fue mudando a la web. Como un mecanismo para aumentar la visibilidad, tanto las inmobiliarias como los particulares publican los anuncios en portales web de bajo costo o gratuitos.

Al respecto, existen periódicos que se volvieron muy populares no solo por ofrecer publicidad gratuita de este tipo de avisos, sino también por ser referentes en la materia. Tal es el caso del periódico digital Cordobés La Voz[1] , que posee todo un portal muy bien estructurado para la publicación de avisos clasificados (Clasificados La Voz [2]). Los cinco rubros que ofrece son: Inmuebles, Vehículos, Empleos, Servicios, Rurales, Productos y Servicios. De estos, el rubro que más destaca por poseer el mayor número son los avisos clasificados de Inmuebles superando los 56.000 registros en octubre de 2020.

Si bien este portal posee cierta estructuración en la carga de datos, para luego poder filtrar las búsquedas, el campo Descripción del inmueble requiere un estudio más profundo puesto que es un campo de texto no estructurado (libre), donde la descripción textual del aviso es muy rica en información y aporta muchos más datos sobre el inmueble que el resto de campos paramétricos (cantidad de dormitorios, cantidad de baños, ubicación, entre otros). Los campos textuales son un buen ejemplo de información no estructurada, que es una de las formas más simples de datos que se pueden generar en la mayoría de los escenarios. El texto no estructurado es procesado y percibido fácilmente por los humanos, pero es mucho más difícil de entender para las máquinas (Feldman & Sanger, 2006).

Desde la página inicial del portal de clasificados se puede acceder a la opción “Inmuebles”, y luego de una búsqueda inicial se ofrecen una serie de filtros, a saber: Operación (Venta, Alquiler), Ubicación, Precio, Tipo de unidad, Cantidad de dormitorios, Tipo de vendedor, Tipo de Barrio, Apto crédito/escritura, A estrenar, Matriculado y Ubicación del departamento (para el caso de los departamentos). Pero no existen filtros para el campo Descripción, ni para indicar cuestiones tales como si el inmueble cuenta con aire acondicionado, comodidades, tipo de piso, cocina, lugares de cercanía, entre muchísimos otros. Ejemplo de descripciones de avisos se observan en la Tabla 1, donde se marcan en negrita algunas palabras relevantes que se pierden o no son tenidas en cuenta al momento de revisar un listado de avisos, el problema aumenta cuando se trata de ciudades grandes donde existen cientos y miles de avisos.

1 https://www.lavoz.com.ar/

2 https://clasificados.lavoz.com.ar/

Revista Tecnología y Ciencia
DOI:https://doi.org/10.33414/rtyc.39.195-206.2020 - ISSN 1666-6933

Título	Descripción
Nueva Córdoba-inversor-1 Dormitorio	Bvd Illia 654, piso 11 externo, depto. de 1 dormitorio en **piso parquet**, **placard** con interiores armados, laterales de yeso, ventana con persiana, living comedor en **piso parquet**. salida **balcon**, cocina separda equipad completa (**bajomesada, alacema, mesada de granito, bacha doble**) baño revestido en cermico, vanitory, **superficie total 40 metros cuadrados**.
Bº Alta Cordoba a 1 Cuarda de La Plaza - Piscina y Asadores	U$S 70.000 BARRIO ALTA CORDOBA UBICACIÓN: Antonio del Viso Nº 696 CONSTA DE: Cocina separada con **alacenas** y **bajo mesada, calefón,** artefacto de cocina , instalación para lavarropas, living comedor amplio con **calefactor** y salida al **balcón** , antebaño con mueble bajo mesada y baño con ducha.Un dormitorio con **placard** con interiores y **aire acondicionado**. todas las aberturas de aluminio con cortinas de enrollar. Dos sectores de terrazas uno de ellos con solárium , piscina y asadores – SUPERFICIE **CUBIERTA: 47 mts.2** ANTIGÜEDAD: 5 años
Amplio Depto Terminado y Con Financiación (b° Nueva Córdoba)	Ubicado en Vélez Sarsfield, esquina Perú, 7° piso frente. Amplio departamento en excelente ubicación. Ideal **para estudiante** a **metros de ciudad universitaria**.Living/comedor muy amplio**Cocina separada** toda equipadaDormitorio amplio con **placard** e interioresBaño completo**Balcón**Excelente vista Ubicación estratégica posesión inmediataVER!!!

Tabla 1: Ejemplos de avisos Clasificados

En este contexto sería deseable contar con opciones de búsqueda de tipo textual que permitan múltiples combinaciones de términos para poder afinar y mejorar la búsqueda de avisos, es por ello que las búsquedas semánticas se han convertido en la opción más empleada en diversos ámbitos, para mejorar la recuperación de videos (Aytar, Shah, & Luo, 2008; Andonian, y otros, 2020), para la búsqueda de resúmenes biomédicos en bases bibliográficas de ciencias de la salud (Soto, Przybyła, & Ananiadou, 2019), para la detección de similitud semántica en textos médicos (Zheng, y otros, 2019), en los motores de búsqueda (Fang, y otros, 2020), para la traducción automática (Biçici & Way, 2016), para la búsqueda de respuestas (Limbasiya & Agrawal, 2019; Zhou, Zhou, He, & Wu, 2016), entre otras aplicaciones.

Este trabajo propone un esquema combinando la técnica de modelado Word2Vec y el empleo de 2-grams (bigramas) para recuperar avisos clasificados inmobiliarios semánticamente relacionados a partir de un conjunto grande de registros. Si bien no existe un esquema base con el cual comparar y evaluar objetivamente los resultados obtenidos puesto que la calidad de un resultado en este contexto es muy subjetiva, se realizó una comparación entre distintas consultas y agrupaciones de términos para evaluar la cantidad de documentos relevantes recuperados con cada esquema a evaluar.

Metodología

Recolección de datos

La minería de texto o *Knowledge Discovery from Text* (KDT), un concepto introducido inicialmente por (Feldman & Dagan, 1995), se refiere al proceso de extracción de información de alta calidad de un texto o conjunto de textos. El uso de métodos automáticos de minería de texto y procesamiento del lenguaje natural (PLN) han ganado una creciente atención en la comunidad académica como mecanismo para analizar textos no estructurados. El objetivo

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.39.195-206.2020 - ISSN 1666-6933

primario de estas técnicas es encontrar nuevo conocimiento a partir de las relaciones que se da entre las palabras, términos, y oraciones que forman el conjunto textual objeto de análisis. Este conocimiento puede luego clasificarse utilizando reglas o patrones específicos al traducir los datos sin procesar en información útil. Si este mismo concepto se traslada a la extracción de datos relevantes de diferentes páginas web, se habla de *Web Data Mining*.

El portal de "Clasificados La Voz" posee algo más de 56.000 avisos clasificados de inmuebles es su totalidad (sin distinción de ubicación ni tipo de inmuebles). Como este portal es un sitio publicitario, no posee un mecanismo que permita recolectar o exportar los resultados de las búsquedas de forma automática. Por esta razón se diseñó un proceso de *web scraping* (una de las técnicas más difundidas y empleadas de *Web Data Mining* según (Glez-Peña, Lourenço, López-Fernández, Reboiro-Jato, & Fdez-Riverola, 2014)) para recolectar el conjunto total de avisos clasificados, el cual fue implementado en el lenguaje *Python* utilizando la librería *BeautifulSoup*. Se recolectaron un total de 56.755 registros en octubre de 2020.

Como en este enorme conjunto de registros recolectados existen avisos de todo el país, se decidió tomar una muestra de los mismos para llevar a cabo esta aproximación, ya que es entendible que en distintos lugares o regiones, el modo de publicar avisos y las relaciones que se pueden encontrar entre los términos analizados puede variar. Así mismo, la descripción de un aviso de un terreno o local varía enormemente con la descripción de una casa o departamento. Teniendo esto en mente se seleccionaron todos los avisos que pertenecen a la región del centro de Córdoba. Se filtraron los registros que pertenecen específicamente a anuncios del: "Centro de Córdoba", "Nueva Córdoba", "Alta Córdoba", "Barrio General Paz" y "Barrio Alberdi", con año de publicación 2020, dando un total de 15.853 registros. Luego de un proceso de limpieza y normalización de datos de forma automática, para extraer las características de cada registro y poder clasificarlos según el tipo de inmueble, según la condición (venta o alquiler), y según la moneda, se obtuvo el detalle que se indica en la Tabla 2. Se decidió analizar en una primera instancia el conjunto de registros que posee el mayor número de registros (departamentos para la venta con un total de 8.081), los resultados obtenidos podrán luego replicarse a los demás conjuntos.

Tipo de Inmueble	Condición	Moneda	Cantidad de Registros
Casa	Alquiler	Dolar	5
Casa	Venta	Dolar	516
Casa	Alquiler	Peso	83
Casa	Venta	Peso	114
Cochera	Alquiler	Dolar	1
Cochera	Venta	Dolar	375
Cochera	Alquiler	Peso	103
Cochera	Venta	Peso	47
Departamento	Alquiler	Dolar	20
Departamento	***Venta***	***Dolar***	***6624***
Departamento	Alquiler	Peso	3473
Departamento	***Venta***	***Peso***	***1457***
Local	Alquiler	Dolar	5

Revista Tecnología y Ciencia
DOI:https://doi.org/10.33414/rtyc.39.195-206.2020 - ISSN 1666-6933

Local	**Venta**	**Dolar**	**384**
Local	**Alquiler**	**Peso**	**667**
Local	**Venta**	**Peso**	**98**
Oficina	**Alquiler**	**Dolar**	**9**
Oficina	**Venta**	**Dolar**	**340**
Oficina	**Alquiler**	**Peso**	**485**
Oficina	**Venta**	**Peso**	**59**
Terreno	**Venta**	**Dolar**	**127**
Terreno	**Alquiler**	**Peso**	**4**
Terreno	**Venta**	**Peso**	**34**

Tabla 2: Resumen de cantidad de registros por tipo de aviso inmobiliario

Análisis textual

Como se mencionó, el campo descripción es un campo muy rico de información, por ello en este trabajo se decidió emplear tanto el título como la descripción de los avisos para conformar el conjunto textual a analizar. En una recorrida previa del conjunto de avisos de departamentos del centro de córdoba se obtuvo el histograma de frecuencia de la longitud de cada registro (titulo + descripción) que se aprecia en la Figura 1, en términos generales, cada registro posee una longitud aceptable (menos de 700 caracteres, aproximadamente 140 palabras).

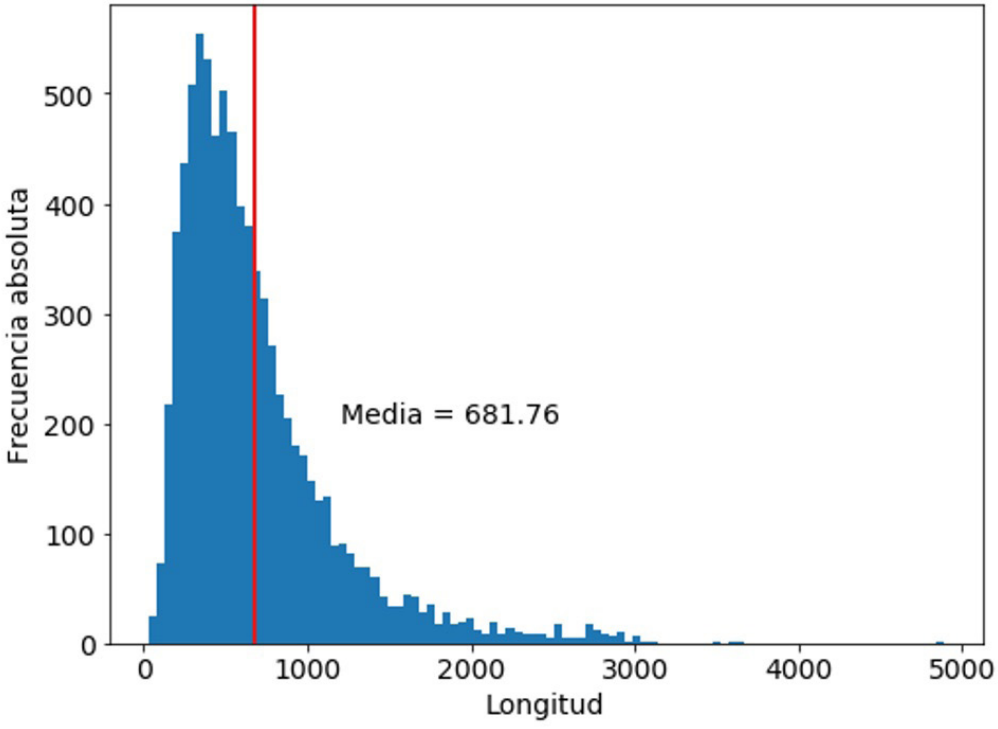

Figura 1: Histograma de la longitud del conjunto textual

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.39.195-206.2020 - ISSN 1666-6933

En este mismo sentido, resulta atractivo no solo analizar la frecuencia de uso de términos en los avisos, sino también la coocurrencia de términos, y mucho más la coocurrencia de combinaciones de términos (bigramas, trigramas o cuatrigramas). Luego de algunas pruebas se comprobó que existen diversos bigramas muy empleados en las descripciones de los avisos, estos tienen que ver con detalles o características del inmueble ofrecido, como ser la cantidad de dormitorios, el tipo de piso, el tipo de cocina, la proximidad a algún lugar, el tamaño de alguna zona del departamento, entre muchos otros. La Tabla 3 presenta el listado de los 30 bigramas más frecuentemente empleados al momento de describir un aviso en el conjunto de datos seleccionado, en dicha tabla se excluyeron previamente las palabras vacías (*stopwords*) del español, puesto que no aportan datos relevantes al análisis.

Bigrama	Frecuencia	Bigrama	Frecuencia
1 dormitorio	4525	aire acondicionado	956
nueva córdoba	4219	alta córdoba	946
2 dormitorios	3455	vendo departamento	906
living comedor	3390	gral paz	854
bajo mesada	2400	comedor amplio	830
cocina separada	2197	2 baños	804
departamento 1	2084	ubicado calle	784
general paz	1757	planta baja	767
baño completo	1444	3 dormitorios	750
departamento 2	1277	1 baño	743
venta departamento	1272	cocina integrada	718
comedor cocina	1140	hermoso departamento	674
dormitorio placard	1110	cocina comedor	645
departamento venta	1088	pisos porcelanato	609
muebles bajo	972	salida balcón	584

Tabla 3: Listado de la frecuencia de los 30 bigramas más empleados

Todos estos detalles permanecen ocultos al usuario que realiza la búsqueda, salvo que ingrese aviso por aviso y proceda a leer dicha descripción, para un número pequeños de avisos esto es una tarea sencilla, pero cuando se trata de decenas, centenas o miles de avisos es impracticable realizarlo de forma manual. Por estas razones sería deseable contar con un instrumento de búsqueda que permitiera recuperar avisos con algún tipo de relación según un conjunto de términos o frase que describiera "mejor" al inmueble buscado.

Preprocesamiento del texto

El preprocesamiento es uno de los componentes clave en muchas de las tareas de minería de texto, para este trabajo consistió en las siguientes tareas: corrección, normalización, tokenización, filtrado y lematización. La corrección se realizó para completar abreviaturas comunes que fueron detectadas en las descripciones de los avisos, por ejemplo: "Nva Cba" corresponde a "Nueva Córdoba", "dpto" y "depto" corresponden a "departamento", "dorm" corresponde a "dormitorio", "oport" corresponde a "oportunidad", entre los más frecuentes.

Revista Tecnología y Ciencia
DOI:https://doi.org/10.33414/rtyc.39.195-206.2020 - ISSN 1666-6933

La normalización consistió en pasar todo el texto a minúsculas, eliminar los signos de puntuación, eliminar las tildes de las palabras acentuadas, y eliminar las cadenas vacías. Esto es necesario no solo para reducir la cantidad de términos a procesar y por ende reducir los cálculos, sino también para evitar que dos palabras iguales sean tratadas como distintas por la inclusión de un punto, guion o acento. Por su parte, la tokenización es la tarea de dividir una secuencia de caracteres en pedazos (palabras / frases) llamados *tokens*. La lista de *tokens* luego se usa para un procesamiento posterior (Webster & Kit, 1992). El filtrado generalmente se realiza sobre los documentos para eliminar algunas de las palabras. Un filtrado común es la eliminación de *stopwords*. Las stopwords son las palabras que aparecen con frecuencia en el texto sin tener mucha información de contenido (por ejemplo, preposiciones, conjunciones, etc.). Del mismo modo, las palabras que aparecen con bastante frecuencia en el texto y que se dice que tienen poca información para distinguir diferentes documentos y también las palabras que aparecen muy raramente tampoco tienen una relevancia significativa y pueden eliminarse de los documentos (Silva & Ribeiro, 2003). Por último, la lematización es la tarea que considera el análisis morfológico de las palabras, es decir, agrupar las diversas formas flexionadas de una palabra para que puedan analizarse como un solo ítem.

Modelado

Como en muchas áreas del PLN, en los últimos años la mayoría de los trabajos ha incorporado técnicas basadas en *Word Embeddings* como mecanismo para mejorar los resultados (Chiruzzo, Etcheverry, & Rosá, 2020). *Word2Vec* es un modelo de *Word Embeddings* (incrustación de palabras en español) que emplea redes neuronales para representar palabras de un texto en un espacio vectorial (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Este esquema de representación ha obtenido mejores resultados que su predecesor *bag-of-words* (Bojanowski, Grave, Joulin, & Mikolov, 2017), aunque en la actualidad sigue demostrando buenos resultados cuando se trabaja con textos muy cortos (Yang, Huang, & Ofoghi, 2019).

El título y la descripción de los 8.081 avisos de venta de departamentos fueron empleados para conformar un corpus textual modelado mediante Word2Vec, evaluando la similitud por medio del esquema *Word Mover's Distance* (WMD) (Kusner, Sun, Kolkin, & Weinberger, 2015). WMD mide la distancia total que las incrustaciones de palabras de dos textos (para este caso la tupla *título* + *descripción*) deben viajar para convertirse en idénticas. WMD aprovecha los resultados de las técnicas de incrustación avanzada como Word2Vec, que aprende representaciones semánticamente significativas de palabras a partir de coocurrencias locales en oraciones. Las técnicas basadas en distancia de edición como la distancia de *Hamming* (para cadenas de la misma longitud) o la generalización de ésta, la distancia de *Levenshtein* (Levenshtein, 1966), no resultan efectivas para realizar búsquedas semánticas puesto que uno busca con un conjunto muy reducido de términos (de dos a cuatro) y la comparación se realiza con registros de hasta cientos de términos, en estos casos los valores de distancia de *Hamming* o *Levenshtein* no aportan resultados coherentes al problema en cuestión.

Para facilitar las tareas de procesamiento de los datos de este trabajo se decidió emplear las bibliotecas de Python *NLTK* (Bird, Klein, & Loper, 2009) para todo lo que respecta al PLN y *Gensim* (Řehůřek & Sojka, 2010) la cual posee implementación del modelo Word2Vec y WMD.

Resultados

Tanto las implementaciones de los modelos como de las funciones de similitud empleados, facilitan enormemente las etapas de prueba y verificación de los enfoques planteados.

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.39.195-206.2020 - ISSN 1666-6933

Posibilitan realizar pruebas simultaneas con diferentes combinaciones de grupos de términos, es decir, empleando unigramas, bigramas, trigramas o cuatrigramas. Corroborando en cada caso cual es el esquema que mejor se ajusta al problema y sobre todo al conjunto de datos, ya que el conocimiento que "aprenden" las técnicas de PLN proviene del conjunto de datos procesado.

Para una búsqueda con los siguientes términos: Consulta= *"piso parque amplio balcón ciudad universitaria"* algunos de los resultados más relevantes se observan en la Tabla 4 ordenados de mayor a menor relevancia, según la puntuación de la función de similitud WMD empleando bigramas.

#	Similitud	Aviso
1	0.5326	vende departamento nueva córdoba frente **ciudad universitaria** departamento 1 dormitorio sobre calle richarson al 50 frente ciudad universitaria cocina integrada al livin/comedor - baño completo - dormitorio con placar grande - **balcon** - luminoso con excelente vista al parque las tejas - escritura inmediata piso 5 externo con **balcón grande - piso de parquet** plastificado se vende amoblado o sin amoblar
2	0.5293	**nueva córdoba** a estrenar 1 dormitorio **amplio balcon** 1 dormitorio living cocina baño **balcon** cerrado con carpinteria de aluminio **pisos de parquet** amplio placardcompletamente reciclado!!!
3	0.5286	oportunidad unica frente buen pastor luminoso con **balcón** excelente ubicacion frente a **buen pastor independencia e hipolito yrigoyen** piso alto muy luminoso **balcon amplio** todo el **piso de parquet** aire acondicionado cocina completa baño equipado completopiso alto con vista a las sierras de córdobaedificio de categoriaa 2 cuadras shopping patio olmos y media cuadra sanatorio allendeplaza españa **ciudad universitaria** y parque sarmiento
4	0.5181	nueva córdoba frente sanatorio allende orientacion frente tres dormitorios con placares tres baños uno zonificado living comedor amplio con salida a **balcon** cerrado y de ampla dimensiones cocina comedor amplia habitacion y baño servicio
5	0.5176	nueva córdoba impecable !!! excelente departamento muy bien ubicado en la ciudad de córdoba en una de las **calles que salen al parque sarmiento** a metros de la terminal de omnibus la uidad cuena con un dormitorio cocina separada living comedor baño grande habitacion com placard **pisos de parquet** el departamento esta muy bien iluminado en piso alto

Tabla 4: Top 5 de resultados para una búsqueda con 6 términos

En negrita se pueden observar los términos que guardan alguna relación con la búsqueda realizada (la existencia de los términos buscados dentro del texto analizado), o que a primera vista pueden coincidir de algún modo con los términos o frase de búsqueda. En este sentido, las búsquedas sintácticas no entregan la mayor cantidad de resultados puesto que la falta de algún carácter o los errores ortográficos producen resultados distintos, es decir, al buscar el término "Córdoba" es distinto a "córdoba" (en minúscula), "cordoba" (en minúscula y sin acento), "Crodoba" (caracteres intercambiados y sin acento) y "Cordoba" (sin acento) por ejemplo. Sin embargo, las búsquedas semánticas y sobre todo las relaciones semánticas entre el texto, pasan por alto este tipo de errores u omisiones, y pueden seguir entregando buenos resultados (Sarker & Gonzalez-Hernandez, 2018).

Por ejemplo, los resultados 4 y 5 de la tabla anterior no poseen dentro de la descripción

Revista Tecnología y Ciencia
DOI:https://doi.org/10.33414/rtyc.39.195-206.2020 - ISSN 1666-6933

los términos "Ciudad Universitaria", a pesar de ello ambos avisos guardan una estrecha relación con la búsqueda realizada ya que estos inmuebles se encuentran muy cercanos a Ciudad Universitaria, entre 600 y 700 metros de distancia. Pero este conocimiento inducido no es casual, ya que proviene de otros avisos que incorporan algunas indicaciones de la ubicación del "Sanatorio Allende" y "El parque Sarmiento", señalando la proximidad con "Ciudad Universitaria". Ambos resultados mediante una búsqueda sintáctica habrían sido descartados por la ausencia de los términos buscados. Es aquí donde cobran real importancia el empleo de técnicas de búsqueda de resultados basados en la semántica del contenido (Medrano, 2018).

Algunos de los malos resultados encontrados, es decir, registros que luego de una revisión manual no eran relevantes a la búsqueda realizada, tienen que ver en su mayoría con avisos muy cortos, es decir, avisos que poseen entre el título y la descripción no más de 100 caracteres (entre 10 y 13 palabras aproximadamente). Estos avisos poco o nada descriptivos con la sola existencia de uno de los términos buscados ya entregan una "buena puntuación" aunque en realidad sea un aviso que no está nada relacionado con la búsqueda realizada. Por ejemplo el siguiente aviso: "*Nueva Córdoba* (Título) –M*onoambiente 30m2 cerca ciudad universitaria ambrosio olmos* (Descripción)".

Evaluación

Si bien no existe un esquema base con el cual comparar objetivamente los resultados obtenidos a lo largo de las pruebas realizadas, el esquema que se empleó fue seleccionar una serie de consultas con términos representativos de la ubicación, característica, servicio, comodidad o proximidad del inmueble y definir los 20 resultados más adecuados (agrupándolos por longitud de registro) de acuerdo a cada una de las 5 consultas, para luego comparar los resultados entregados con el modelo propuesto y con las distintas variantes de conjuntos de términos (n-gramas) adoptando la métrica de evaluación exhaustividad (*recall*) (Baeza-Yates & Ribeiro-Neto, 1999). La sensibilidad o exhaustividad es la fracción de instancias o elementos relevantes recuperados por una consulta:

$$Recall = \frac{Elementos \quad Relevantes \quad Recuperados}{Total \quad de \quad Elementos \quad Recuperados} \tag{1}$$

Por cada consulta se obtuvieron los primeros 20 resultados y en cada caso se computaron los elementos relevantes recuperados contra los 20 registros definidos inicialmente. La Figura 2 resume el promedio de exhaustividad calculado para cada una de las alternativas empleadas teniendo en cuenta la longitud de los registros. Se puede observar que a medida que el tamaño de los registros aumenta el esquema que entrega los mayores valores de exhaustividad es el 2-gramas (agrupaciones de 2 términos). Sin embargo para registros pequeños (menos de 300 caracteres) el esquema que entrega los mejores resultados es el de 1-grama (un único término).

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.39.195-206.2020 - ISSN 1666-6933

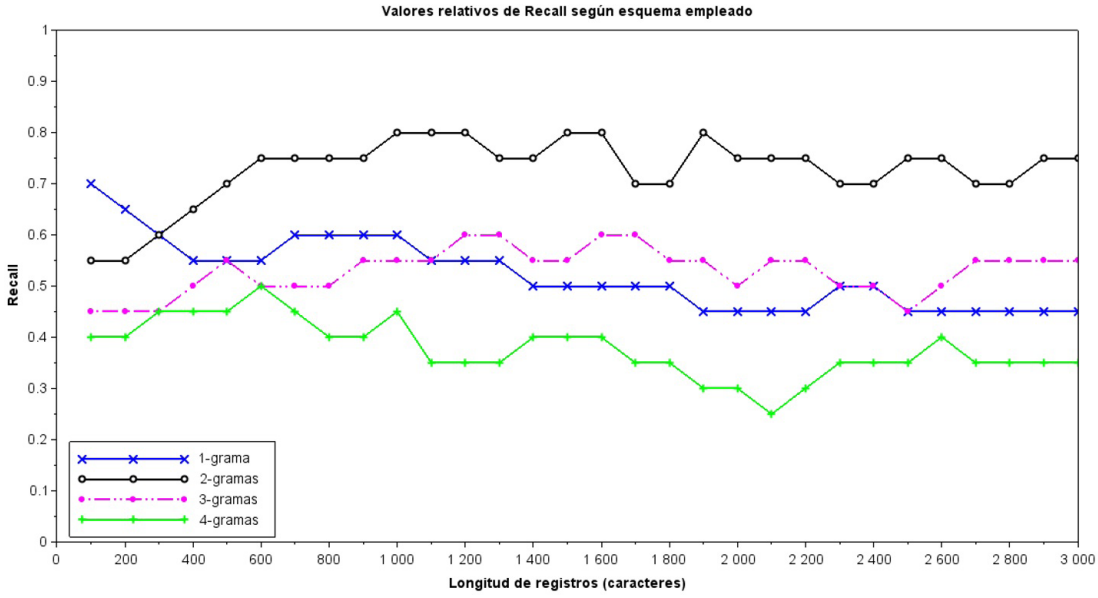

Figura 2: Valores promedio de recall según esquema n-grama empleado y tamaño de los registros

Conclusiones

A partir de la recolección de un gran conjunto de avisos clasificados de inmuebles, se llevó a cabo un experimento basado en la recuperación de avisos comenzando con una búsqueda inicial empleando términos no paramétricos analizando textos no estructurados. Los resultados indican que el enfoque propuesto es prometedor y que el uso de modelos basados en word embedding entrenados con bigramas parece funcionar mejor que un modelo unigrama, trigrama o cuatrigrama a medida que aumenta el tamaño de los títulos y descripciones de los avisos.

En este artículo se describió un prototipo destinado a la tarea de encontrar, en un gran corpus de clasificados, un conjunto de avisos que están relacionados semánticamente a una frase o conjunto de términos. Distintas consultas fueron empleadas para evaluar la capacidad del proceso expuesto para encontrar la mayor cantidad de avisos relevantes a estas. Las comparaciones realizadas sugieren que la calidad del modelo planteado aumenta con el empleo de n-gramas de palabras de diferentes órdenes (1–3). Con descripciones cortas (entre 100 y 300 caracteres) el mejor esquema es el empleo de unigramas tradicionales, sin embargo como la media del tamaño de registros supera los 600 caracteres, entrenar el modelo con bigramas resulta la mejor elección.

El enfoque que se plantea puede ser claramente mejorado a partir del empleo de evaluaciones o sugerencias colaborativas en base a opiniones de los propios usuarios. Además podría restringirse el tamaño mínimo de las descripciones a 300 caracteres para obtener un detalle más rico en información y de esta forma aumentar la calidad de los resultados recuperados. Del mismo modo, con la sugerencia de ciertos términos se podría mejorar el detalle del inmueble.

Revista Tecnología y Ciencia
DOI:https://doi.org/10.33414/rtyc.39.195-206.2020 - ISSN 1666-6933

Referencias

Andonian, A., Fosco, C., Monfort, M., Lee, A., Feris, R., Vondrick, C., & Oliva, A. (2020). We Have So Much In Common: Modeling Semantic Relational Set Abstractions in Videos. arXiv preprint arXiv:2008.05596.

Aytar, Y., Shah, M., & Luo, J. (2008). Utilizing semantic word similarity measures for video retrieval. 2008 IEEE Conference on Computer Vision and Pattern Recognition, (págs. 1-8).

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern Information Retrieval.

Biçici, E., & Way, A. (2016). Referential translation machines for predicting semantic similarity. language resources and evaluation, 50(4), 793-819.

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5(1), 135-146.

Chiruzzo, L., Etcheverry, M., & Rosá, A. (2020). Sentiment Analysis in Spanish Tweets: Some Experiments with Focus on Neutral Tweets. Proces. del Leng. Natural, 64, 109-116.

Fang, K., Zhao, L., Shen, Z., Wang, R., Zhou, R., & Fan, L. (2020). Beyond Lexical: A Semantic Retrieval Framework for Textual SearchEngine. arXiv preprint arXiv:2008.03917.

Feldman, R., & Dagan, I. (1995). Knowledge discovery in Textual Databases (KDT). KDD'95 Proceedings of the First International Conference on Knowledge Discovery and Data Mining, (págs. 112-117).

Feldman, R., & Sanger, J. (2006). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2014). Web scraping technologies in an API world. Briefings in Bioinformatics, 15(5), 788-797.

Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015). From Word Embeddings To Document Distances. Proceedings of The 32nd International Conference on Machine Learning, (págs. 957-966).

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, 10(8), 707–710.

Limbasiya, N., & Agrawal, P. (2019). Semantic Textual Similarity and Factorization Machine Model for Retrieval of Question-Answering. International Conference on Advances in Computing and Data Sciences, 195-206.

Medrano, J. (2018). Filtrado basado en contenido para artículos académicos en repositorios institucionales. XXIV Congreso Argentino de Ciencias de la Computación (La Plata, 2018).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems 26, (págs. 3111-3119).

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.39.195-206.2020 - ISSN 1666-6933

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora.

Sarker, A., & Gonzalez-Hernandez, G. (2018). An unsupervised and customizable misspelling generator for mining noisy health-related text sources. Journal of Biomedical Informatics, 88, 98-107.

Silva, C., & Ribeiro, B. (2003). The importance of stop word removal on recall values in text categorization. Proceedings of the International Joint Conference on Neural Networks, 2003., 3, págs. 1661-1666.

Soto, A., Przybyła, P., & Ananiadou, S. (2019). Thalia: semantic search engine for biomedical abstracts. Bioinformatics, 35(10), 1799-1801.

Webster, J., & Kit, C. (1992). Tokenization as the initial phase in NLP. COLING '92 Proceedings of the 14th conference on Computational linguistics - Volume 4, (págs. 1106-1110).

Yang, S., Huang, G., & Ofoghi, B. (2019). Short Text Similarity Measurement Using Context from Bag of Word Pairs and Word Co-occurrence. international conference on the digital society, 1179, 221-231.

Zheng, T., Gao, Y., Wang, F., Fan, C., Fu, X., Li, M., . . . Ma, H. (2019). Detection of medical text semantic similarity based on convolutional neural network. BMC Medical Informatics and Decision Making, 19(1), 1-11.

Zhou, G., Zhou, Y., He, T., & Wu, W. (2016). Learning semantic representation with neural networks for community question answering retrieval. Knowledge Based Systems, 93, 75-83.