

# Definition of an information retrieval system using Elasticsearch

<b>Document collaborators and roles</b>	<b>2</b>
<b>UML diagram</b>	<b>2</b>
Autocomplete-Index	2
Listings-Dev	3
<b>Indexes</b>	<b>4</b>
Index: "listings-dev"	4
Settings	4
Filters	4
Settings breakdown	4
Filters	4
Replicas	5
Shards	5
Properties of the documents in the "listings-dev" index	5
Index: "autocomplete"	6
Properties of the documents in the "autocomplete" index	6
<b>Search overview</b>	<b>6</b>
Neural Search	6
Keyword Search	7
<b>Precision vs Recall Analysis &amp; Test queries</b>	<b>7</b>
<b>Query #1</b>	<b>8</b>
<b>Query #2</b>	<b>9</b>
<b>Query #3</b>	<b>9</b>
<b>Query #4</b>	<b>10</b>
<b>Query #5</b>	<b>11</b>
<b>Query #6</b>	<b>12</b>
<b>Query #7</b>	<b>13</b>
<b>Query #8</b>	<b>14</b>
<b>Query #9</b>	<b>14</b>
<b>Query #10</b>	<b>15</b>
<b>Query #11</b>	<b>16</b>
<b>Query #12</b>	<b>17</b>
<b>Query #13</b>	<b>18</b>
<b>Query #14</b>	<b>19</b>
<b>Query #15</b>	<b>20</b>
<b>Query #16</b>	<b>21</b>
<b>Query #17</b>	<b>22</b>
<b>Query #18</b>	<b>22</b>

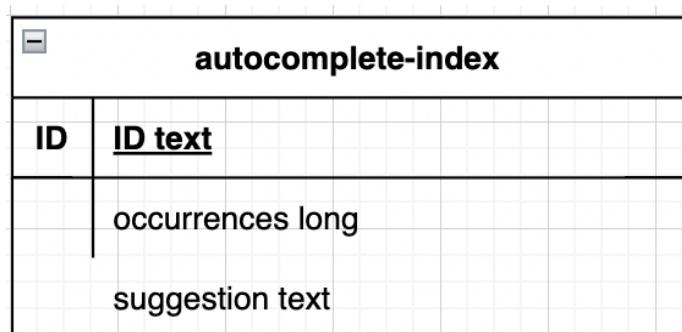
<b>Query #19</b>	<b>23</b>
<b>Query #20</b>	<b>24</b>
<b>Overall precision and recall values</b>	<b>25</b>
<b>Justification</b>	<b>25</b>
Integration of Elasticsearch into our e-commerce system	26
Potential problems or areas of opportunity regarding information management and query handling	26
Measures of mitigation for query handling	27

## Document collaborators and roles

Member	Role
Miguel Arriaga Velasco	Developer in the search team. Collaborated in the creation of the python script to create the indexes.
Stephan Guingor Falcón	Natural Language Processing module chief.
Pablo Rocha Ojeda	Developer in the search team. Collaborated in the creation of indexes.
Emilio Sibaja Villarreal	Collaborator
Ariadne Álvarez Reyes	Collaborator
Octavio Fenollosa Sosa	Collaborator

## UML diagram

### Autocomplete-Index



## Listings-Dev

ID	<u>id long</u>
	base_price long
	car_model.body_type text
	car_model.brand_id long
	car_model.id long
	car_model.spec_sheet text
	car_model.model text
	car_model.year long
	car_variant.abs text
	car_variant.air_conditioning boolean
	car_variant.car_model_id long
	car_variant.colors object
	car_variant.customizations object
	front_airbags boolean
	fuel_type boolean
	full_description text
	full_name text
	horse_power text
	is_new boolean
	liter long
	metadata object
	number_doors long
	oil_consume long
	passengers long
	rear_airbags boolean
	rim long
	rim_material text
	spin text
	tags text
	transmission text
	traction text
	used_km long
	variant_name text
	colors.hex_code text
	colors.images.img_url text
	colors.images.order text
	colors.name text
	created_at date
	description text
	description_embedding.model_id text
	description_embedding.predicted_value dense_vector
	geo_loc geo_point
	test_drive boolean

# Indexes

## Index: "listings-dev"

### Settings breakdown

#### Filters

- "lowercase": Converts all text to lowercase. This allows case-insensitive searches, ensuring that queries match indexed data regardless of capitalization.
- "asciifolding": Performs ASCII folding, which means it replaces accented characters and other special characters with their ASCII equivalents. This helps improve search accuracy by treating accented and non-accented characters as equivalent.
- "stop": Removes common words that are considered insignificant for search purposes, such as articles, prepositions, and conjunctions. These words are often excluded to reduce the index size and improve search performance.
- "custom\_stemmer": Applies advanced stemming algorithms to reduce words to their base or root form. Stemming allows search queries to match different forms of the same word, such as singular and plural forms, verb conjugations, etc. The "custom\_stemmer" likely refers to a specific implementation of a stemming algorithm.
  - "type": "snowball": This setting indicates the stemming algorithm used for the index. "Snowball" is a widely used stemming algorithm that supports multiple languages. It is designed to reduce words to their base form based on language-specific rules. In this case, the algorithm is configured for the Spanish language.
  - "language": "Spanish": Specifies the language of the indexed content. Elasticsearch uses language-specific analyzers to process text data, such as tokenizing, stemming, and applying language-specific rules. In this case, the index is configured for Spanish language content.

#### Replicas

- 3 replicas: Replicas are additional copies of index shards that provide redundancy and improve search performance by distributing data across multiple nodes. In this case, the index has three replicas, meaning each shard has three copies.

#### Shards

- 1 shard: Shards are smaller units that distribute the index across multiple nodes in a cluster, allowing for scalability and parallel processing. In this case, the index has only one shard, meaning all the data is stored on a single node.

Overall, these settings configure the "listings-dev" index to support Spanish language content with various text processing filters, including lowercase conversion, ASCII folding, stop word removal, and advanced stemming using the Snowball algorithm. The index has one shard and three replicas for redundancy and performance optimization.

## Properties of the documents in the "listings-dev" index

- "id" : The ID of the listing
- "base\_price": The starting price for the car being sold
- "car\_model.body\_type" :
- "car\_model.brand\_id" : The id of the brand selling the car
- "car\_model.id" : A unique identifier for the car model
- "car\_model.spec\_sheet" : The specification of the car in text form
- "car\_model.model" : The model of the car being sold
- "car\_model.year": The manufacturing year of the car.
- "car\_variant.abs" : A boolean representing true if the car has an abs system
- "car\_variant.air\_conditioning" : A boolean representing true if the car has air conditioning
- "car\_variant.car\_model\_id" : A unique identifier for the car model
- "car variant.colors": An array of colors the car is available in
- "car variant.customizations" : An object containing the extra features the variant may modify
- "front\_airbags" : A boolean representing true if the car has air bags
- "fuel\_type" : Type of fuel that the car uses
- "full\_description" : The description of the car being sold
- "full\_name" : The full name of the car.
- "horse\_power": The horsepower of the car
- "is\_new" : Boolean that represents the car is new or semi-new
- "liter" : Capacity in liters for fuel
- "metadata" : Object to store any additional information of the car being sold
- "number\_doors" : Number of doors in the car
- "oil\_consume" : Amount of oil consumed
- "passengers" : Number of passengers the car can have
- "rear\_airbags" : A boolean representing true if the car has rear bags
- "rim" : Size of the rim
- "rim\_material" : Material of the rim.
- "spin" : A unique identifier used to retrieve images for 3D view
- "tags" : Array of values associated with the listing
- "transmission" : The transmission type of the car.
- "traction" : Traction of the car
- "used\_km" : Number of km the car has go through
- "variant\_name" : The name of the car variant.
- "colors.hex\_code": Hex code of the color
- "colors.images.img\_url" : image url of car with the specified color
- "colors.images.order" : Order of the color
- "colors.name" : Name of the color
- "created\_at": The time when the listing was created
- "description": The description of the car being sold

- “description\_embedding.model\_id” : Id of the model that was used to create the text embeddings
- “description\_embedding.predicted\_value”: A dense vector representing the text embedding for the listing description. This property is used to calculate similarity between documents using the cosine similarity metric.
- “geo\_loc”: A field of type "geo\_point" that stores the geographic location of the listing.
- “test\_drive”: Boolean representing if the car is available for a test drive

## **Index: "autocomplete"**

### **Properties of the documents in the "autocomplete" index**

- “suggestion”: The text value of the autocomplete suggestion, the field is of type “search\_as\_you\_type” to optimize queries.
- “occurrences”: A numeric value that indicates the number of occurrences of the suggestion, in order to prioritize the most used suggestions.
- “id”: An md5 hash of the query

## **Search overview**

To create the best search experience we implemented a search endpoint that has two modes.

### **Neural Search**

The results for the users will be based on context, this is also known as semantic search, and we implemented it using text-embeddings in elastic search. The embeddings were generated by a model based on the pretrained model bert

### **Keyword Search**

This makes use of classic full text search, using elastic search. We use the user's query and full text search in different fields and rank the results based on the importance of the field and number of hits.

For both modes we can add filters on top, like ranges, or exact matches on fields, this is very important as users often want to limit the search results, we can also enable ‘near me’ searches, which looks for cars that are being sold in less than a specific distance. ( 1000KM )

We also created an autocomplete endpoint, so that the user is given suggested queries as they type. For this endpoint, we store all the previous queries in the elastic search database, as well as a score for each one of them. The more a query is used, the higher its score will be and it will have more priority in the autocomplete recommendations.

In addition if custom ner is enabled while searching, we will extract different entities from the plain text query. If the service from azure has a confidence score higher than 50% (should be higher but the model is not trained yet to give high scores) then we will add them to our filters and increase the efficiency of our results.

## Precision vs Recall Analysis & Test queries

We decided to obtain the precision and recall values for the keyword search mode, as it is going to be the main search method for the application (hard filters were also not used in the queries to fully test the natural language processing search). To do this, we made 20 test queries, created a precision vs. recall matrix for each one, as well as their overall precision value and recall value, and finally averaged the precision and recall for the 20 queries.

**It is important to keep in mind that we are only evaluating the first 10 results for each of the queries.**

The formulas used are shown below:

		Real Label		
		Positive	Negative	
Predicted Label	Positive	True Positive (TP)	False Positive (FP)	Precision = $\frac{\sum TP}{\sum TP + FP}$
	Negative	False Negative (FN)	True Negative (TN)	
		Recall = $\frac{\sum TP}{\sum TP + FN}$		Accuracy = $\frac{\sum TP + TN}{\sum TP + FP + FN + TN}$

Where:

- **True Positive (TP):** In the context of search results, a true positive represents a relevant document that is correctly retrieved by the search engine.
- **False Positive (FP):** A false positive represents a non-relevant document that is mistakenly retrieved by the search engine, indicating an incorrect positive prediction.
- **False Negative (FN):** A false negative represents a relevant document that is not retrieved by the search engine, indicating a missed relevant result.
- **True Negative (TN):** In the context of search results evaluation, true negatives are not typically considered since they refer to non-relevant documents correctly identified as such (which are not typically tracked in this context).
- **Precision:** Precision measures the proportion of retrieved documents that are actually relevant. It is calculated as  $TP / (TP + FP)$ . In the context of search results, it represents the accuracy of the retrieved documents.
- **Recall:** Recall measures the proportion of relevant documents that are successfully retrieved by the search engine. It is calculated as  $TP / (TP + FN)$ . In the context of search results, it represents the comprehensiveness or coverage of the search engine.

# Query #1

Query:

Bmw x4 del año 2020

Screenshot:

The screenshot shows a search results page for 'Bmw x4 del año 2020' on the Movu platform. The search bar at the top contains the query. To the right of the search bar are icons for a heart, a share, and a notification bell, along with a user profile picture. On the left, there are filters for 'Región o estado' (Ciudad de México, Monterrey, Puebla, Guadalajara), 'Rango de precio' (\$604,999 - \$946,999), and 'Año' (2018). The results are ordered by relevance. The first result is a red BMW X4 30i Executive from 2020 with a price of \$946,999 MX. The second result is a grey BMW X4 35i M Sport from 2018 with a price of \$744,999 MX. The third result is a dark grey BMW X4 35i M Sport from 2018 with a price of \$763,999 MX. Below these are three more cars: a grey BMW X4 28i X Line from 2018, a white BMW X4 30i X Line from 2019, and a dark grey BMW X4 35i M Sport from 2016.

Precision Value: 10%

Recall Value: 100%

Results precision vs. recall matrix:

True Positive (TP)	False Positive (FP)
1	9
False Negative (FN)	True Negative (TN)
0	0

## Query #2

Query:

quiero un coche mini-cooper azul o cafe

Screenshot:

The screenshot shows a search interface for cars. The search bar at the top contains the query "quiero un coche mini cooper azul o cafe". Below the search bar, there are two filter sections: "Región o estado" and "Rango de precio". The "Región o estado" section lists Ciudad de México (4), Cuernavaca (1), Monterrey (1), Puebla (1), Guadalajara (2), and Querétaro (1). The "Rango de precio" section has a slider set between \$304,999 and \$443,999. The main area displays six car listings, each with a thumbnail image, dealership information, model name, year, price, and a color-coded status indicator below it. The dealerships listed are Ciudad de México, Cuernavaca, Monterrey, Puebla, and Guadalajara. The models shown are Mini Cooper Chili Hatch from 2019 (\$421,999 MX), Mini Cooper Chili Hatch from 2018 (\$374,999 MX), Mini Cooper Chili Hatch from 2016 (\$304,999 MX), Mini Cooper Chili Hatch from 2018 (\$377,999 MX), Mini Cooper Chili Hatch from 2019 (\$428,999 MX), and Mini Cooper Salt Hatch from 2020 (\$443,999 MX).

Precision Value: 50%

Recall Value: 45%

Results precision vs. recall matrix:

True Positive (TP)	False Positive (FP)
5	5
False Negative (FN)	True Negative (TN)
6	0

## Query #3

**Query:**

me gustaria un coche del 2018 que cueste menos de 400 mil pesos

**Screenshot:**

The screenshot shows a search results page for Movu. The search bar at the top contains the query "me gustaria un coche del 2018 que cueste menos de 400 mil pesos". The results are ordered by Relevancia. The first three results are Mercedes-Benz vehicles from 2015, 2014, and 2022, all listed at prices above 400,000 pesos. Below them are three other vehicles: a Peugeot 2008 Feline from 2016 at \$213,999 MX, a Toyota Hilux Mid from 2018 at \$451,999 MX, and another Toyota Hilux Mid from 2019 at \$469,999 MX.

**Precision Value:** 0%

**Recall Value:** 0%

**Results precision vs. recall matrix:**

True Positive (TP)	False Positive (FP)
0	10
False Negative (FN)	True Negative (TN)
10	0

## Query #4

Query:

camioneta hyundai roja

Screenshot:

The screenshot shows a search results page for 'camioneta hyundai roja' on the movu platform. The search bar at the top contains the query. To the right of the search bar are icons for heart, basket, and notifications, along with a user profile picture. On the left, there are filters for 'Región o estado' (Cuernavaca, Puebla, Ciudad de México, Monterrey) and 'Rango de precio' (\$240,999 to \$410,999). A 'Año' filter shows 2017, 2018, and 2020. The results are ordered by Relevancia. There are six search results displayed in a grid:

Imagen	Detalles	Precio
	dealership, Cuernavaca	\$299,999 MX
	dealership, Puebla	\$240,999 MX
	dealership, Ciudad de México	\$331,999 MX
	dealership, Puebla	\$336,999 MX
	dealership, Monterrey	\$294,999 MX
	dealership, Ciudad de México	\$410,999 MX

Precision Value: 60%

Recall Value: 16%

Results precision vs. recall matrix:

True Positive (TP)	False Positive (FP)
6	4
False Negative (FN)	True Negative (TN)
32	0

## Query #5

Query:

mazda 3 negro

Screenshot:

The screenshot shows a search results page for 'mazda 3 negro' on the Movu platform. The search bar at the top contains the query 'mazda 3 negro'. Below the search bar, there are filters: 'Región o estado' (Ciudad de México, Monterrey, Guadalajara, Puebla), 'Rango de precio' (\$245,999 - \$325,999), and 'Año' (2016, 2017, 2018). The results are ordered by Relevancia. There are six car listings displayed in two rows of three:

Image	Dealer	Model	Year	Price
	dealership, Ciudad de México	Mazda 3 S Sedan	2016	\$266,999 MX
	dealership, Ciudad de México	Mazda 3 S Sedan	2016	\$253,999 MX
	dealership, Ciudad de México	Mazda 3 S Sedan	2018	\$325,999 MX
	dealership, Monterrey	Mazda 3 S Hatchback	2018	\$318,999 MX
	dealership, Ciudad de México	Mazda 3 S Hatchback	2017	\$287,999 MX
	dealership, Guadalajara	Mazda 3 S Hatchback	2018	\$324,999 MX

Precision Value: 20%

Recall Value: 1.1%

Results precision vs. recall matrix:

True Positive (TP)	False Positive (FP)
2	8
False Negative (FN)	True Negative (TN)
174	0

## Query #6

Query:

camioneta blanca

Screenshot:

The screenshot shows a search results page for 'camioneta blanca' on the Movu website. The search bar at the top contains the query 'camioneta blanca'. Below the search bar are filters for 'Región o estado' (Monterrey, Cuernavaca, Ciudad de México, Guadalajara, Puebla, Querétaro) and 'Rango de precio' (\$180,999 - \$863,999). The results are ordered by 'Relevancia'. There are six car listings displayed in two rows of three. Each listing includes a thumbnail image, the car model ('Dodge Challenger Black'), year ('2016', '2017'), location ('dealership, Monterrey'), price ('\$561,999 MX', '\$538,999 MX', '\$449,999 MX', '\$582,999 MX', '\$463,999 MX', '\$551,999 MX'), and a small icon.

Precision Value: 90%

Recall Value: 31%

Results precision vs. recall matrix:

True Positive (TP)	False Positive (FP)
9	1
False Negative (FN)	True Negative (TN)
20	0

## Query #7

Query:

dodge charger black en monterrey

Screenshot:

The screenshot shows a search results page for 'dodge charger black en monterrey' on the Movu platform. The search bar at the top contains the query. Below it, there are filters for 'Región o estado' (Monterrey, Cuernavaca, Ciudad de México, Guadalajara), 'Rango de precio' (\$180,999 - \$863,999), and 'Año' (2015-2017). The results are ordered by Relevancia. There are six cars listed in two rows of three:

Image	Location	Model	Year	Price (MX)
	dealership, Monterrey	Dodge Challenger Black ...	2017	\$449,999 MX
	dealership, Cuernavaca	Dodge Challenger Black ...	2017	\$582,999 MX
	dealership, Monterrey	Dodge Challenger Black ...	2016	\$561,999 MX
	dealership, Ciudad de México	Dodge Challenger Black ...	2015	\$463,999 MX
	dealership, Monterrey	Dodge Challenger Black ...	2016	\$538,999 MX
	dealership, Monterrey	Dodge Challenger Black ...	2016	\$551,999 MX

Precision Value: 40%

Recall Value: 100%

Results precision vs. recall matrix:

True Positive (TP)	False Positive (FP)
4	6
False Negative (FN)	True Negative (TN)
0	0

## Query #8

Query:

coche manual negro susuki

Screenshot:

The screenshot shows a search results page for 'coche manual negro susuki' on the Movu website. The search bar at the top contains the query. Below it, there are filters: 'Región o estado' (Ciudad de México, Monterrey, Cuernavaca, Puebla), 'Rango de precio' (\$205,999 - \$492,999), and 'Año'. The results are ordered by Relevancia. There are six car listings arranged in two rows of three. Each listing includes a thumbnail image, the car model, year, price, dealer location, and a star rating.

Model	Year	Price	Dealer	Rating
Suzuki Swift GLX	2017	\$251,999 MX	dealership, Ciudad de México	3 stars
Suzuki Swift GLX	2017	\$261,999 MX	dealership, Ciudad de México	4 stars
Suzuki Ciaz RS	2019	\$278,999 MX	dealership, Monterrey	3 stars
Suzuki Jimny GLX	2021	\$491,999 MX	dealership, Monterrey	3 stars
Suzuki Jimny GLX	2021	\$492,999 MX	dealership, Ciudad de México	4 stars
Suzuki Ciaz RS	2018	\$251,999 MX	dealership, Ciudad de México	2 stars

Precision Value: 30%

Recall Value: 0.24%

Results precision vs. recall matrix:

True Positive (TP)	False Positive (FP)
3	7
False Negative (FN)	True Negative (TN)
121	0

## Query #9

Query:

mini cooper azul

Screenshot:

The screenshot shows a search results page for 'mini cooper azul' on the movu website. The search bar at the top contains the query. Below it, there are filters for 'Región o estado' (Ciudad de México, Cuernavaca, Monterrey, Puebla, Guadalajara, Querétaro) and 'Rango de precio' (\$304,999 - \$443,999). The results are ordered by Relevancia and show six Mini Cooper Chili Hatchback models. Each result includes a thumbnail image, dealership information, year, and price.

Dealership	Model	Year	Price
dealership, Ciudad de México	Mini Cooper Chili Hatch...	2019	\$421,999 MX
dealership, Ciudad de México	Mini Cooper Chili Hatch...	2018	\$374,999 MX
dealership, Cuernavaca	Mini Cooper Chili Hatch...	2016	\$304,999 MX
dealership, Monterrey	Mini Cooper Chili Hatch...	2018	\$377,999 MX
dealership, Puebla	Mini Cooper Chili Hatch...	2019	\$428,999 MX
dealership, Guadalajara	Mini Cooper Salt Hatchb...	2020	\$443,999 MX

Precision Value: 100%

Recall Value: 33%

Results precision vs. recall matrix:

True Positive (TP)	False Positive (FP)
10	0
False Negative (FN)	True Negative (TN)
5	0

## Query #10

Query:

coche negro automatico deportivo

Screenshot:

The screenshot shows a search results page for 'coche negro automatico deportivo' on the Movu website. The search bar at the top contains the query. Below it, there are filters for 'Región o estado' (Ciudad de México, Monterrey, Querétaro, Puebla), 'Rango de precio' (\$240,999 - \$621,999), and 'Año'. The results are ordered by Relevancia and show six car listings:

Imagen	Detalles	Precio
	dealership, Ciudad de México Infiniti QX80 - 2017	\$621,999 MX
	dealership, Monterrey Mg5 Excite 2022	\$340,999 MX
	dealership, Monterrey Mg5 Elegance 2022	\$333,999 MX
	dealership, Querétaro Nissan Kicks Advance 2019	\$334,999 MX
	dealership, Ciudad de México Kia Soul LX 2020	\$322,999 MX
	dealership, Puebla Kia Stinger EX 2018	\$517,999 MX

Precision Value: 0%

Recall Value: 0%

Results precision vs. recall matrix:

True Positive (TP)	False Positive (FP)
0	10
False Negative (FN)	True Negative (TN)
0	0

## Query #11

Query:

Sedan azul mazda 2018

Screenshot:

The screenshot shows a search results page for a blue Mazda 3 Sedan from 2018. The search bar at the top contains the query 'Sedan azul mazda 2018'. The interface includes filters for 'Región o estado' (Querétaro, Ciudad de México, Puebla), 'Rango de precio' (\$243,999 - \$322,999), and 'Año' (2016-2018). The results are ordered by 'Relevancia'. Each listing includes a thumbnail image of the car, its location (dealership name and city), the model name ('Mazda 3 S Sedan' or 'Mazda 3 I Sedan'), the year ('2018' or '2016'), the price ('\$316,999 MX', '\$315,999 MX', '\$259,999 MX', '\$247,999 MX', '\$272,999 MX', '\$243,999 MX'), and a set of three circular icons below the price.

Precision Value: 40%

Recall Value: 1.3%

Results precision vs. recall matrix:

True Positive (TP)	False Positive (FP)
4	6
False Negative (FN)	True Negative (TN)
303	0

## Query #12

Query:

audi del 2015 gris

Screenshot:

The screenshot shows a search results page for 'audi del 2015 gris' on the Movu website. The search bar at the top contains the query. Below it, there are filters for 'Región o estado' (Ciudad de México, Guadalajara, Monterrey), 'Rango de precio' (\$318,999 - \$1,063,999), and 'Año' (2018). The results are ordered by Relevancia. Five cars are listed:

Imagen	Detalles	Precio
	dealership, Ciudad de México Audi A4 Dynamic 2018	\$445,999 MX
	dealership, Ciudad de México Audi A4 Dynamic 2019	\$526,999 MX
	dealership, Guadalajara Audi S6 S6 2018	\$1,063,999 MX
	dealership, Monterrey Audi A4 Sport 2016	\$318,999 MX
	dealership, Monterrey Audi A4 Dynamic 2018	\$453,999 MX
	dealership, Ciudad de México Audi Q3 Select 2018	\$492,999 MX

Precision Value: 0%

Recall Value: 0%

Results precision vs. recall matrix:

True Positive (TP)	False Positive (FP)
0	10
False Negative (FN)	True Negative (TN)
10	0

## Query #13

Query:

tesla

Screenshot:

The screenshot shows a search results page for 'tesla' on the Movu platform. The search bar at the top contains the query 'tesla'. Below the search bar are several filters: 'Región o estado' set to 'Guadalajara', 'Rango de precio' set to '\$ 1,009,999 - \$ 1,009,999', and 'Año' set to '2022'. The main result is a Tesla Model 3 Standard ... from a dealership in Guadalajara, priced at '\$1,009,999 MX'. The car is shown in a thumbnail image.

Precision Value: 100%

Recall Value: 100%

Results precision vs. recall matrix:

True Positive (TP)	False Positive (FP)
1	0
False Negative (FN)	True Negative (TN)
0	0

## Query #14

Query:

Fiat 500 en querétaro

Screenshot:

The screenshot shows a search interface for Fiat 500 cars in Querétaro. The search bar at the top contains the query "fiat 500 en queretaro". Below the search bar are filter options: "Región o estado" (Querétaro, Puebla, Ciudad de México, Cuernavaca), "Rango de precio" (\$191,999 to \$478,999), and "Año" (2014 to 2019). The results are ordered by Relevancia. There are six cars displayed in a grid:

Imagen	Localización	Tipo	Año	Precio
	dealership, Querétaro	Fiat 500 Pop	2016	\$246,999 MX
	dealership, Querétaro	Fiat 500 Sporting Hatch...	2016	\$248,999 MX
	dealership, Puebla	Fiat 500 Abarth Hatchba...	2019	\$478,999 MX
	dealership, Querétaro	Fiat 500 Trendy Hatchba...	2015	\$202,999 MX
	dealership, Ciudad de México	Fiat 500 Sport Hatchback	2014	\$191,999 MX
	dealership, Querétaro	Fiat 500 Trendy Hatchba...	2015	\$198,999 MX

Precision Value: 50%

Recall Value: 100%

Results precision vs. recall matrix:

True Positive (TP)	False Positive (FP)
5	5
False Negative (FN)	True Negative (TN)
0	0

## Query #15

Query:

coche nuevo de 2 puertas

Screenshot:

The screenshot shows a search results page for 'coche nuevo de 2 puertas' on the movu website. The search bar at the top contains the query. Below it, there are filters: 'Región o estado' (Guadalajara, Ciudad de México), 'Rango de precio' (\$225,999 - \$439,999), and 'Año' (2016-2018). The results are ordered by Relevancia. The first three results are BMW Serie 2 220i Coupé (2016, \$338,999 MX, dealership, Guadalajara), BMW Serie 2 220i Coupé (2017, \$439,999 MX, dealership, Ciudad de México), and BMW Serie 2 220i Coupé (2017, \$419,999 MX, dealership, Ciudad de México). The fourth result is Volkswagen Jetta A6 2 (2017, \$248,999 MX, dealership, Ciudad de México), the fifth is Volkswagen Jetta A6 2 (2017, \$244,999 MX, dealership, Ciudad de México), and the sixth is Volkswagen Jetta A6 2 (2018, \$282,999 MX, dealership, Ciudad de México).

Precision Value: 0%

Recall Value: 0%

Results precision vs. recall matrix:

True Positive (TP)	False Positive (FP)
0	10
False Negative (FN)	True Negative (TN)
15	0

## Query #16

Query:

jetta en la cdmx

Screenshot:

Precision Value: 60%

Recall Value: 3.1 %

Results precision vs. recall matrix:

True Positive (TP)	False Positive (FP)
6	4
False Negative (FN)	True Negative (TN)
190	0

## Query #17

Query:

Hatchback blanco

Screenshot:

The screenshot shows a search results page for 'hatchback blanco'. At the top, there's a search bar with the query 'hatchback blanco' and a microphone icon. Below it, there are filters: 'Región o estado' (Guadalajara, Monterrey, Puebla, Cuernavaca, Ciudad de México) and 'Rango de precio' (\$174,999 - \$863,999). The results are ordered by Relevancia. There are six cars displayed in a grid:

Imagen	Detalles	Precio
	dealership, Guadalajara Smart Fortwo Black And ... 2014	\$180,999 MX
	dealership, Monterrey Dodge Challenger Black ... 2016	\$561,999 MX
	dealership, Monterrey Dodge Challenger Black ... 2016	\$538,999 MX
	dealership, Puebla Ford Bronco Sport Outer ... 2022	\$760,999 MX
	dealership, Monterrey Dodge Challenger Black ... 2017	\$449,999 MX
	dealership, Cuernavaca Dodge Challenger Black ... 2017	\$582,999 MX

Precision Value: 0%

Recall Value: 0%

Results precision vs. recall matrix:

True Positive (TP)	False Positive (FP)
0	10
False Negative (FN)	True Negative (TN)
549	0

## Query #18

Query:

Mazda 2 Hatchback

Screenshot:

The screenshot shows a search results page for 'MAZDA2 HATCHBACK' on the movu website. The interface includes a header with 'INGRESA' and 'REGISTRO' buttons, a search bar, and a dropdown menu for sorting by 'Relevancia'. On the left, there are filters for 'Región o estado' (Ciudad de México, Puebla, Monterrey, Guadalajara) and 'Rango de precio' (\$235,999 - \$324,999). Below these filters, there are three rows of car listings. Each listing includes a thumbnail image, the car model, year, location ('dealership'), price (\$318,999 MX, \$315,999 MX, \$319,999 MX, \$263,999 MX, \$249,999 MX), and a small icon indicating the number of reviews (e.g., 1, 2, 3).

Precision Value: 10%

Recall Value: 14%

Results precision vs. recall matrix:

True Positive (TP)	False Positive (FP)
1	9
False Negative (FN)	True Negative (TN)
6	0

## Query #19

Query:

marca nissan color azul

Screenshot:

The screenshot shows a search results page for "marca nissan color azul". The results are filtered by region (Ciudad de México, Monterrey, Puebla, Guadalajara), price range (\$140,999 - \$280,999), and year (2017-2019). The results are ordered by relevance. There are 8 blue Nissan March vehicles displayed, each with a thumbnail, model name, year, location, and price.

Thumbnail	Model	Year	Location	Price
	Nissan March Advance	2018	dealership, Ciudad de México	\$226,999 MX
	Nissan March Advance	2018	dealership, Monterrey	\$206,999 MX
	Nissan March Advance	2019	dealership, Ciudad de México	\$235,999 MX
	Nissan March Advance	2017	dealership, Monterrey	\$195,999 MX
	Nissan March Advance	2018	dealership, Monterrey	\$216,999 MX
	Nissan March Sense	2018	dealership, Puebla	\$198,999 MX

Precision Value: 50%

Recall Value: 0.98%

Results precision vs. recall matrix:

True Positive (TP)	False Positive (FP)
5	5
False Negative (FN)	True Negative (TN)
505	0

## Query #20

Query:

auto manual

Screenshot:

The screenshot shows a search results page for 'auto manual' on the Movu platform. The search bar at the top contains the query 'auto manual'. Below the search bar are filters for 'Región o estado' (Ciudad de México, Guadalajara, Querétaro, Monterrey, Puebla), 'Rango de precio' (\$181,999 - \$384,999), and 'Año' (2016, 2022). The results are ordered by Relevancia. The first result is a Smart car from a dealership in Ciudad de México for \$257,999 MX. The second result is an Mg5 Excite from a dealership in Guadalajara for \$320,999 MX. The third result is an Mg5 Elegance from a dealership in Ciudad de México for \$334,999 MX. The fourth result is a Nissan Sentra Sense from a dealership in Ciudad de México for \$257,999 MX. The fifth result is a Seat Toledo Style from a dealership in Querétaro for \$320,999 MX. The sixth result is a Nissan March Sense from a dealership in Ciudad de México for \$334,999 MX.

Precision Value: 100%

Recall Value: 0.0033%

Results precision vs. recall matrix:

True Positive (TP)	False Positive (FP)
10	0
False Negative (FN)	True Negative (TN)
3058	0

## Overall precision and recall values

Averaging the results of the 20 queries, we obtained the following values:

Overall precision: 40.5%

Overall recall: 27.28%

## Justification

Our data schema in Elasticsearch has been designed to enhance the precision and recall of the information retrieval system, achieving a balance between the two. Here's how our data schema contributes to obtaining more precise documents while maintaining a good recall:

- Descriptive Embeddings: The "description\_embedding.predicted\_value" property in the "listings-dev" index represents a dense vector that captures the textual representation of the listing description. By utilizing text embeddings and cosine similarity, we can calculate the similarity between documents. This approach allows us to retrieve more precise results by considering the semantic context of the search query.
- Keyword Search: In addition to neural search, we have implemented classic full-text search using Elasticsearch. By performing keyword-based searches across different fields, we leverage the power of Elasticsearch's search capabilities. This helps us retrieve relevant documents based on the user's query and rank them using the importance of the field and the number of hits. It improves precision by matching exact keywords and terms.
- Filters and Ranges: We provide users with the ability to apply filters, such as ranges or exact matches on specific fields, to refine their search results. By incorporating these filters, users can narrow down their search criteria and obtain more precise documents that meet their specific requirements.
- Autocomplete and User Behavior: We have implemented an auto-complete feature that suggests queries to users as they type. This feature not only improves the user experience but also contributes to precision by offering frequently used and highly scored queries. Storing previous queries and their respective scores in Elasticsearch allows us to prioritize popular queries and provide more accurate suggestions.
- Index settings: the settings that configure the "listings-dev" index are made to support Spanish language content with various text processing filters, including lowercase conversion, ASCII folding, stop word removal, and advanced stemming using the Snowball algorithm. The index has one shard and three replicas for redundancy and performance optimization.

By combining these strategies, we aim to strike a balance between precision and recall in our information retrieval system. Users can expect precise results based on the semantic context of their search query while ensuring a good recall rate by leveraging the power of full-text search and user behavior analysis.

## Integration of Elasticsearch into our e-commerce system

To incorporate Elasticsearch into our e-commerce system, we will integrate it as the primary search engine. Elasticsearch will handle the search queries from users and provide highly relevant and efficient results.

## Potential problems or areas of opportunity regarding information management and query handling

- Scalability: As our e-commerce platform grows and handles an increasing amount of data, scalability might become a concern. The size of the Elasticsearch cluster and the performance of search queries can be potential challenges. To mitigate this, we can implement sharding and replication strategies, optimize the Elasticsearch cluster configuration, and utilize hardware resources effectively.
- Relevance and Ranking: Ensuring the relevance and ranking of search results is an ongoing challenge. We may encounter situations where certain documents are ranked inaccurately or irrelevantly. Continuous monitoring, performance analysis, and feedback from users will be crucial in fine-tuning the relevance and ranking algorithms. Regular updates and improvements to our data schema and relevance models can help address these challenges.
- Query Understanding: Understanding user queries accurately and extracting relevant entities from plain text can be a complex task. Incorporating natural language processing techniques and leveraging external services, such as Azure's custom NER, can enhance the query understanding process. Regularly updating and fine-tuning these services based on user feedback and data analysis can further improve query understanding and accuracy.

## Measures of mitigation for query handling

- Continuous Monitoring and Analysis: Regular monitoring of search performance metrics, user feedback, and query logs will help identify any potential issues or areas for improvement. Analyzing search patterns, user behavior, and query performance will enable us to make data-driven decisions to enhance precision and recall.
- A/B Testing: Conducting A/B testing with different ranking algorithms, relevance models, or query processing techniques can provide insights into the effectiveness of different approaches. By comparing the performance of different variants, we can identify the most effective strategies and continuously iterate on them.
- User Feedback and Iterative Improvements: Actively soliciting and analyzing user feedback on search results and suggestions will help us understand user expectations and address any shortcomings. Incorporating user feedback into our iterative improvement process will enable us to enhance the precision and recall of the search system over time.
- Regular Maintenance and Updates: Keeping Elasticsearch and its associated components up to date with the latest versions, security patches, and performance enhancements will ensure a stable and efficient search system. Regularly reviewing and updating the data schema, relevance models, and indexing strategies will also contribute to improved query handling and information management.

By addressing these potential challenges and implementing the mitigation measures mentioned above, we aim to optimize the performance of our e-commerce search system, enhance precision and recall, and provide an excellent user experience.