PRI - M3: Search System

Group 43

João Alves - up202007614

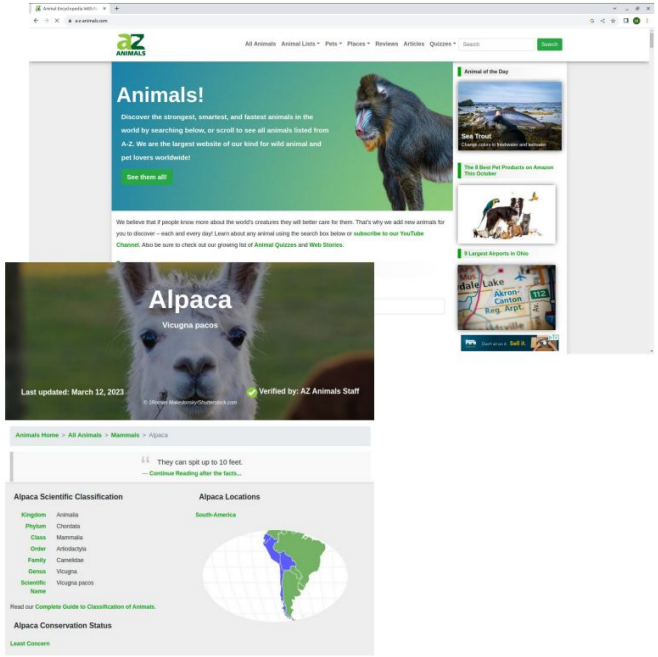
Marco André - up202004891

Mário Ferreira - up201907727

Ricardo Matos - up202007962

Introduction

- Data Source: https://a-z-animals.com/
- 2778 entries with 42 features
- Many animal specific features and many others engineered by the team during the data processing phase
- There is multiple small text fields and a main “Text” field containing all condensed text information about an animal
- For information collection and indexing we are using ***Apache Solr*** and a refined schema with many analyzers.



Global Parameters

```
"params": {
      "q": "[Query Text]",
      "defType": "edismax",
      "qf": "Name^2.5 Features^2.0 Fun_Fact^2.0
      Diet^2.0
      Text^1.5 Origin^3.5 Features^2.0 Behavior^2.0",
      "pf": "Name^2.5 Features^2.0 Fun_Fact^2.0
      Diet^2.0
      Text^1.5 Origin^3.5 Features^2.0 Behavior^2.0",
      "mm": "3<-25%",
      "ps": 5,
      "fl": "Name, score",
      "rows": "30"
}
```

The queries' text for search scenarios are:

- Energetic dog breeds suited for hunting
- North America animals that like to eat insects
- Change the color of their skin, fur or feathers for the purpose of camouflage
- Animals that walk in hierarchical groups or herds and how they deal with territory
- (NOT Birds) migrate to Mexico or migrate to America

Global Parameters for Generic Search

The team achieved great success with query tailored parameterization during the last iteration of the project. However, when using the global parameterization, the results were consistently lower for all queries and a **Mean Average Precision of 49%** was achieved.

Queries / Metrics	AvgP	P@10	R@10	F@10
Query 1	0.64	0.70	0.41	0.51
Query 2	0.32	0.2	0.22	0.21
Query 3	0.14	0.10	0.25	0.22
Query 4	0.84	0.88	0.40	0.53
Query 5	0.51	0.4	0.66	0.50

Explored Improvements

In order to improve the retrieved document relevance of the search system, the team explored the following paths:

- Taylor synonyms to be domain-specific to animals
- More Like This Suggestions for each animal entry
- Semantic Search
- Learning to Rank

The team also developed a GUI for the search system as it will be discussed later.

Domain-specific Synonyms

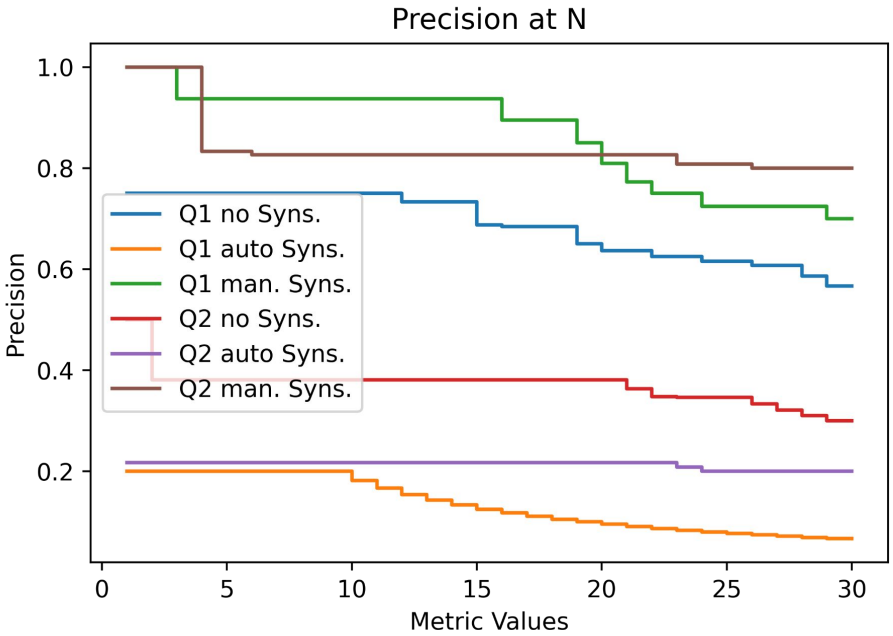
The team aimed to boost search results by **tailoring synonyms to the context of animals**. The creation of the new synonyms was done using the **nltk Python package** to establish word associations via **word-nets**.

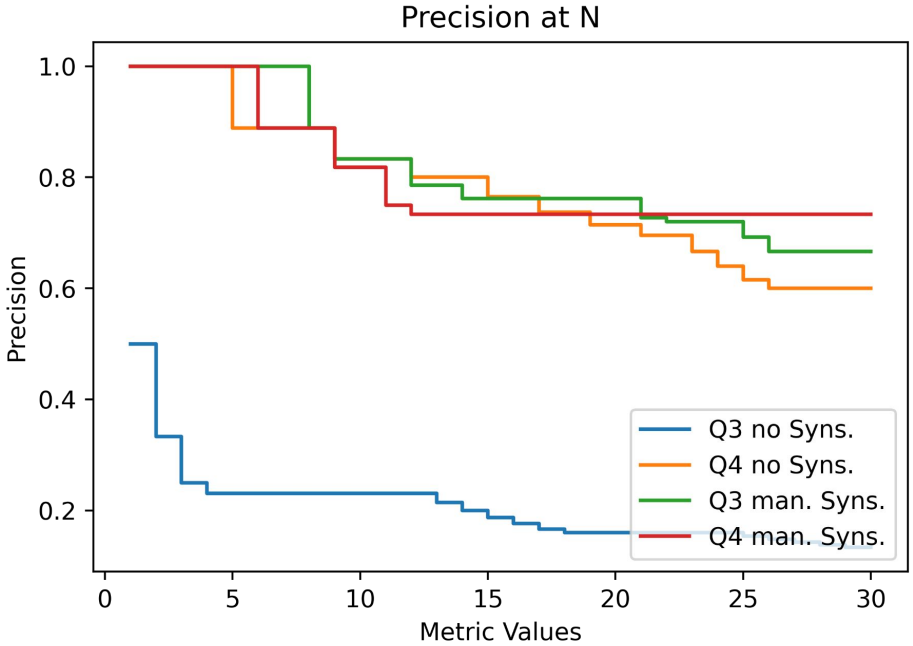
After manual review, most synonyms were correct but some captured term relations that **deviated from the original meaning** and thus **generated noise** in the results that lead to significantly **worse results than before**.

This emphasizes the need for a delicate balance in synonym-based strategies for *Solr* query optimization.

A solution would be to have a more sophisticated/manual approach for synonym generation which would provide more fine control.

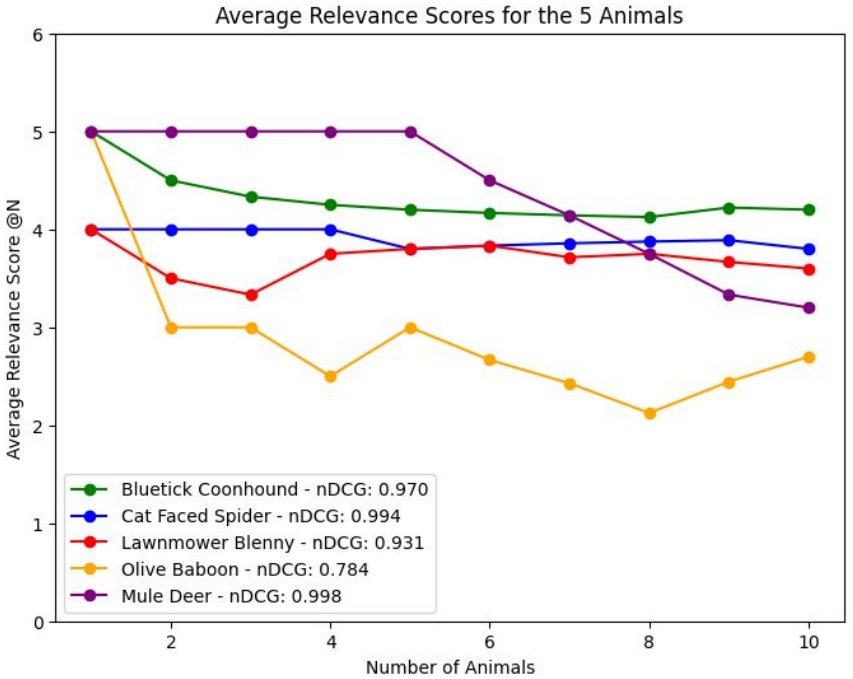
This was done for the 5 queries as proof of concept and the results clearly show noticeable improvements for some queries.


Precision at N
Precision
Metric Values
1.0
0.8
0.6
0.4
0.2
0
5
10
15
20
25
30
Q1 no Syns.
Q1 auto Syns.
Q1 man. Syns.
Q2 no Syns.
Q2 auto Syns.
Q2 man. Syns.


Precision at N
Precision
Metric Values
1.0
0.8
0.6
0.4
0.2
0
5
10
15
20
25
30
Q3 no Syns.
Q4 no Syns.
Q3 man. Syns.
Q4 man. Syns.

More Like This

- Designed to find documents similar to a given one. Can be used as recommendations.
- First animal retrieved from each query was chosen
- Parameter tuning was required
- Classify animals with a 0 to 5 scale, where 5 is the most relevant
- Analyzed top 10 results of each animal (50 in total)
- Used normalized discounted cumulative gain (nDCG) for effectiveness measurement
- Very positive results overall - 4/5 animals achieved nDCG above 0.9



Semantic Search

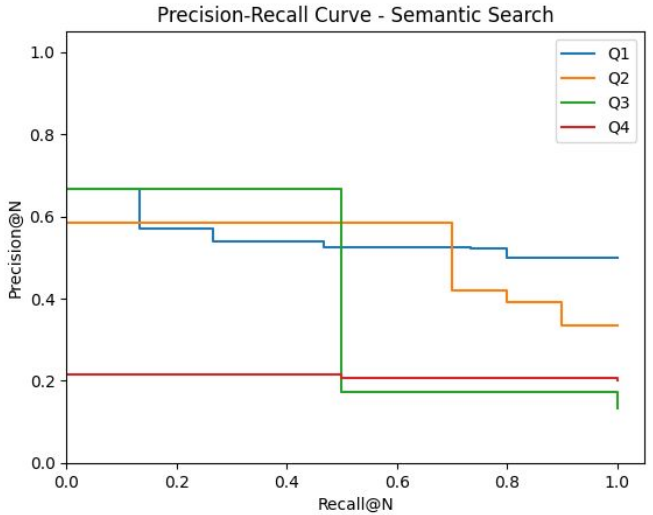
Semantic search refers to a system's ability to extract the meaning and purpose behind a query, going beyond keyword matching.

- Selected the model **all-MiniLM-L6-v2** from the Python library **sentence-transformers** to generate a dense vector for each document
- Employed the **cosine** similarity function and the **Hierarchical Navigable Small World (HNSW)** k-nearest neighbor algorithm for accurate searches

The system is now ready to receive a dense vector comprising the semantic meaning of a query and retrieves the most similar documents.

Semantic Search - Results

- Slightly better performance on queries 2 and 3
- Worse results on queries 1 and 4
- No relevant documents retrieved for query 5



Queries / Metrics	AvgP
Query 1	0.64
Query 2	0.32
Query 3	0.14
Query 4	0.84
Query 5	0.51

Search with Global Parameters

Queries / Metrics	AvgP	P@10	R@10	F@10
Query 1	0.50	0.5	0.33	0.4
Query 2	0.42	0.5	0.5	0.5
Query 3	0.22	0.2	0.5	0.29
Query 4	0.17	0.1	0.17	0.13
Query 5	0.0	0.0	0.0	0.0

Semantic Search

Learning to Rank

- **Learning to rank** is the process of developing a ranking model to predict document relevancy.
- Learning to Rank Overview:
  - **Machine learning** technique
  - Focuses on training models for automatic item ranking
  - Aims to predict the relevance order of items
- Training Objective:
  - Train **a supervised model (Linear regression)** to learn the correct ordering of items
  - Use previous defined **Features** to the model
  - Utilizes labeled examples with ground truth relevance. Manually labeled from 0 to 5, accordingly to a set of rules

Learning to Rank

- Solr LTR linear model.
- Very good results compared with the global parameters.
- MAP of **67,6%**
- Query 5 was not included in the training set as the model was not fitted for this query.

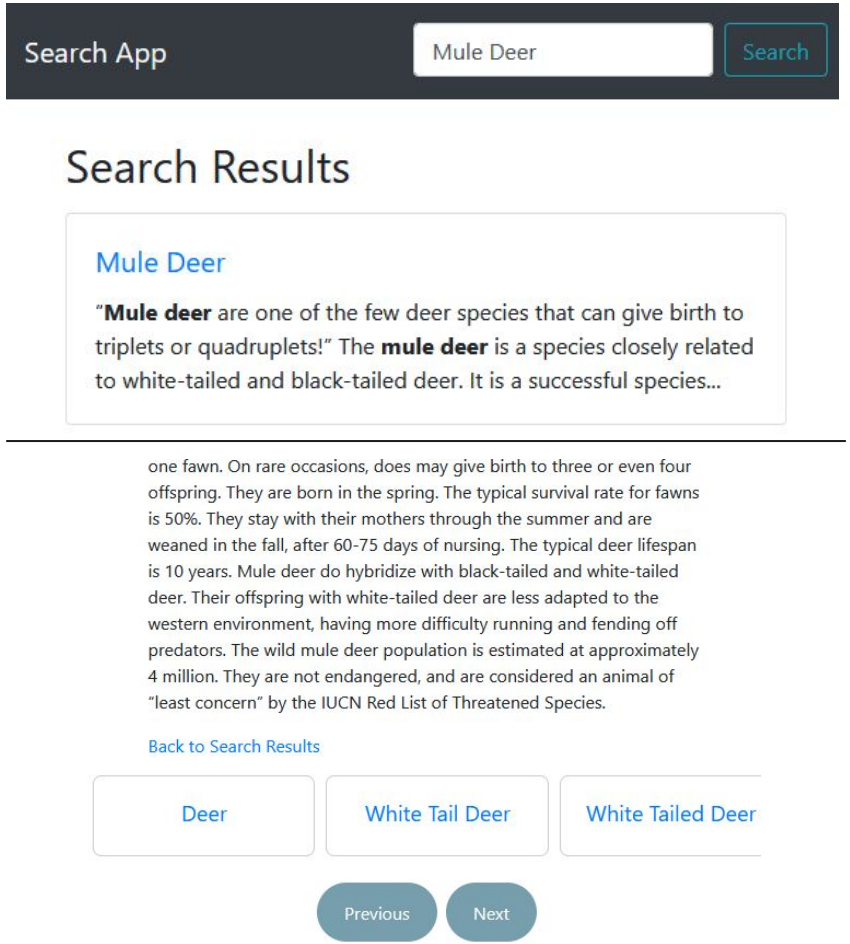
Queries / Metrics	AvgP	P@10	R@10	F@10
Query 1	0.87	0.9	0.38	0.52
Query 2	0.95	1.0	0.4	0.57
Query 3	0.47	0.2	0.2	0.2
Query 4	0.82	0.7	0.41	0.52
Query 5	0.27	0.3	0.43	0.35

Features	Description	Weight
queryMatchName	Score of matching Name field	-0.41950315
queryMatchGenus	Score of matching Genus field	0.0
queryMatchClass	Score of matching Class field	0.0
queryMatchOrigin	Score of matching Origin field	-0.42167284
queryMatchFun Fact	Score of matching the Fun Fact	0.16945207
queryMatchMigratory	Score of matching the Migratory	0.4367062
queryMatchText	Score of matching the Text field	0.08357657
originalScore	The original score	0.08357656

Table 4: LTR model and features

Graphical User Interface

- User-friendly interface with the primary aim of enhancing user experience
- Implemented snippets with highlights corresponding to the keywords in a user's query
- Introduced a "more like this" functionality to enable users to explore similar animals



Conclusion

During this final phase, the team worked hard to improve document relevance in order to achieve a better search system. This endeavour lead us through many exploration paths where some were very successful and others less so.

Nonetheless, the metrics indicate that document relevance for all tested queries increased considerably when taking into account all changes made during this project step. However, there is still room for improvement as more advanced techniques could be explored.

The outcome was the establishment of a high-quality animal information retrieval system that can provide users with relevant information about animal related topics.

End

João Alves - up202007614

Marco André - up202004891

Mário Ferreira - up201907727

Ricardo Matos - up202007962