



Hotels Search System based on their Reviews

INFORMATION PROCESSING AND
RETRIEVAL – PRI
MILESTONE #3: SEARCH SYSTEM

André Ávila – up202006767@up.pt
André Costa – up201905916@up.pt
Fábio Morais – up202008052@up.pt
Fábio Sá – up202007658@up.pt

Milestone 3



Milestone 2 Overview



Information Retrieval Improvements



Stopwords Analysis



Semantic Analysis



More Like This



Search User Interface



Final System



Conclusion and Future Work

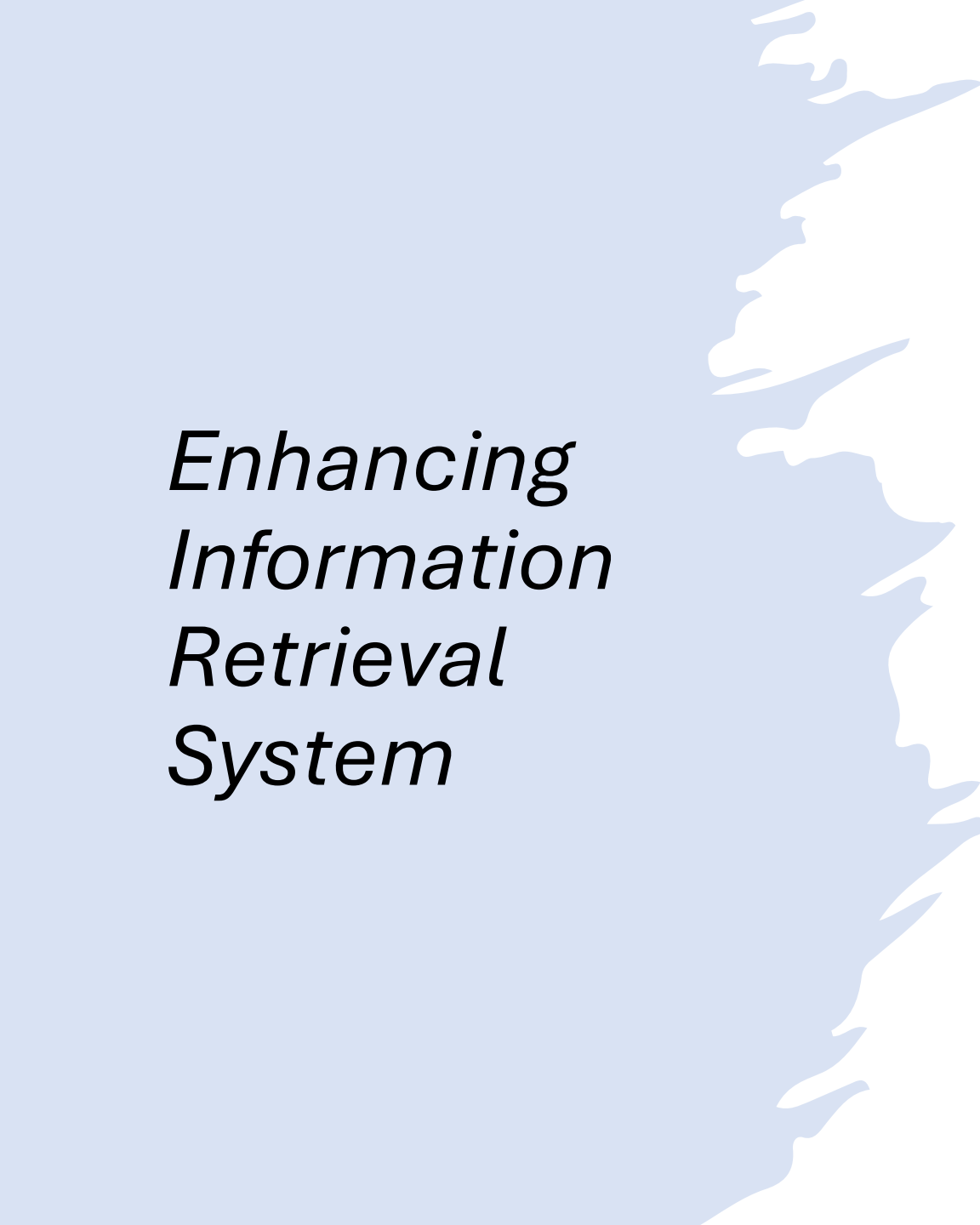
```

"reviews": [
  {
    "date": "2017-07",
    "rate": 4.8,
    "text": "Lovely hotel. I th
  },
  {
    "date": "2017-07",
    "rate": 5.0,
    "text": "Customer service was
  }
]
},
{
  "name": "Hotel Balmoral",
  "location": "Barcelona Spain",
  "average_rate": 4.33,
  "reviews": [
    {
      "date": "2017-08",
      "rate": 3.6,
      "text": "Good value

```

Milestone 2 Overview

- The previous stage introduced an initial version of the information retrieval system, evaluating it based on well-defined information needs and metrics grounded in precision and recall.
- As result we finished previous milestone with a schema that took in attendance the [ASCIIFolding](#), [LowerCase](#), [SynonymGraph](#) and [EnglishMinimalStem](#) filters.



Enhancing Information Retrieval System

The previous stage introduced an initial version of the information retrieval system, evaluating it based on well-defined information needs and metrics grounded in precision and recall.

Looking from another perspective, it also helped identify the **weaknesses** and **limitations** of the chosen approaches.

Therefore, to enhance the search engine, this phase involved the implementation and evaluation of features aimed at addressing the identified gaps:

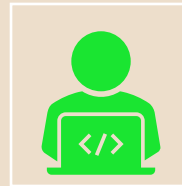
- **Stopwords** Filter
- **Semantic Search**

To explore the project's theme more focusedly, **More Like This** has been added to the list of improvements.

Stopwords



The decision to use a Stopwords Filter aimed at enhancing the exploration of **larger phrases**, allowing clients more flexibility to conduct searches with **more complex queries**.



Solr initially provides a set of predefined stop words, but a custom file was chosen, derived from a Python Library (nltk). Like the process for Synonyms, this custom stop words file was incorporated into the Solr configuration files.



In the schema, the Stopwords Filter was used both in the indexing regarding the storage of Solr and in the query analyzer regarding the search.

Stopwords *(first 2 queries)*

Considering that the query only contains **1 stopwords**, the results show a noticeable similarity even though they are not identical.

q	(good breakfast) OR (good room service)
q.op	AND
fq	{!child of="*:* -_nest_path_*"}location:"new delhi"
fl	*,[child]
sort	score desc
stopwords	true
ignoreCase	true

Rank	Previous Syst.	Improved Syst.
AvP	0.87	0.81
P@20	0.8	0.75

q	(good vegetarian) OR (good vegan)
q.op	AND
fq	{!child of="*:* -_nest_path_*"}location:*
fl	*,[child]
sort	score desc
stopwords	true
ignoreCase	true

Rank	Previous Syst.	Improved Syst.
AvP	0.55	0.21
P@20	0.5	0.4

Stopwords *(3rd query)*

In contrast, the 3rd query, which has **4 stopwords**, produced no results in the original schema without the stopwords filter. The modified schema, on the other hand, provided the expected 20 outcomes with a precision of **0.75**.

q	What are the best hotels with beach views
q.op	AND
fq	{!child of="*:* - _nest_path_*"}location:*
fl	*,[child]
sort	score desc
stopwords	true
ignoreCase	true

Rank	Previous Syst.	Improved Syst.
AvP	0	0.75
P@20	0	0.65

Observing the **MAP**, the behavior of the improved schema is significantly better when increasing the complexity of the queries.

Global	Previous Syst.	Improved Syst.
MAP	0.473	0.73

Semantic Search

Dense vector generation for the review text in the JSON data that populates Solr.



Vectorization of the query text.



Adaption of the text query into a KNN search query.



Post Request as the query field with vectorized text surpasses the URL character limit.

Semantic *(first 2 queries)*

In both queries, boosted system **outperforms** semantic search. This difference is more notable in the second query.

q	{!knn f=vector topK=20}((good breakfast) OR (good room service))
q.op	AND
fq	{!child of="*: *_nest_path_:*"}location:"new delhi"
fl	*,[child]
sort	score desc

Rank	Previous Syst.	Improved Syst.
AvP	0.87	0.81
P@20	0.8	0.75

q	{!knn f=vector topK=20}((good vegetarian) OR (good vegan))
q.op	AND
fq	{!child of="*: *_nest_path_:*"}location:*
fl	*,[child]
sort	score desc

Rank	Previous Syst.	Improved Syst.
AvP	0.55	0.21
P@20	0.5	0.4

"A very enjoyable stay and lovely staff. I wish there were more vegan options for breakfast other than just fruit",

Semantic (3rd query)

Similarly to the previous queries, boosted system **outperforms** semantic search.

q	{!knn f=v ector topK=20}{(good access to public transports)}
q.op	AND
fq	{!child of="*: *_nest_path_*"}location:"new delhi"
fl	*,[child]
sort	score desc

Rank	Previous Syst.	Improved Syst.
AvP	0.97	0.83
P@20	0.95	0.75

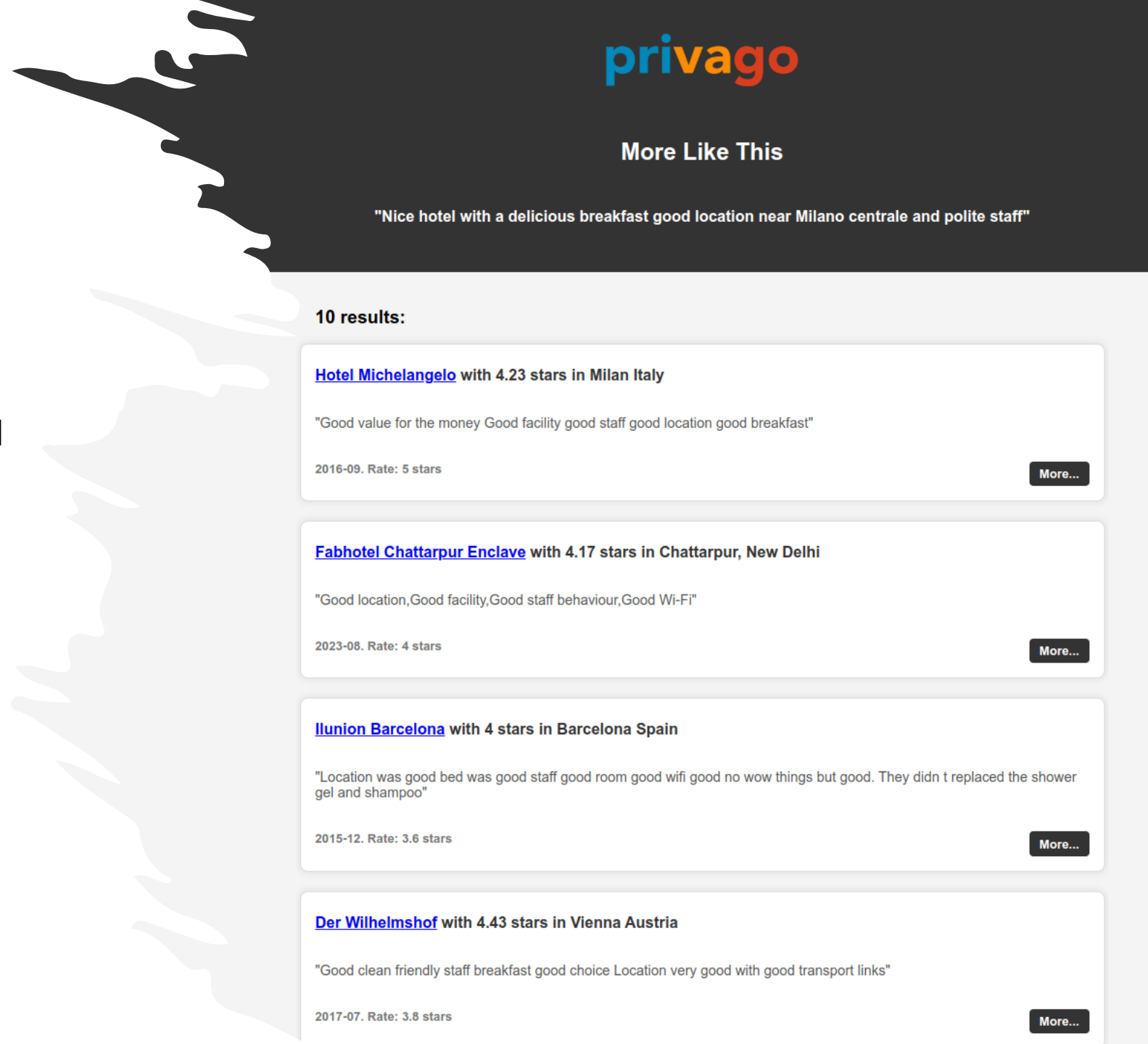
In summary, Semantic Search falls short in effectively addressing the challenge posed by negative words altering the meaning of a phrase.

Global	Previous Syst.	Improved Syst.
MAP	0.80	0.62

"Good location for easy access to the DLR and the city",

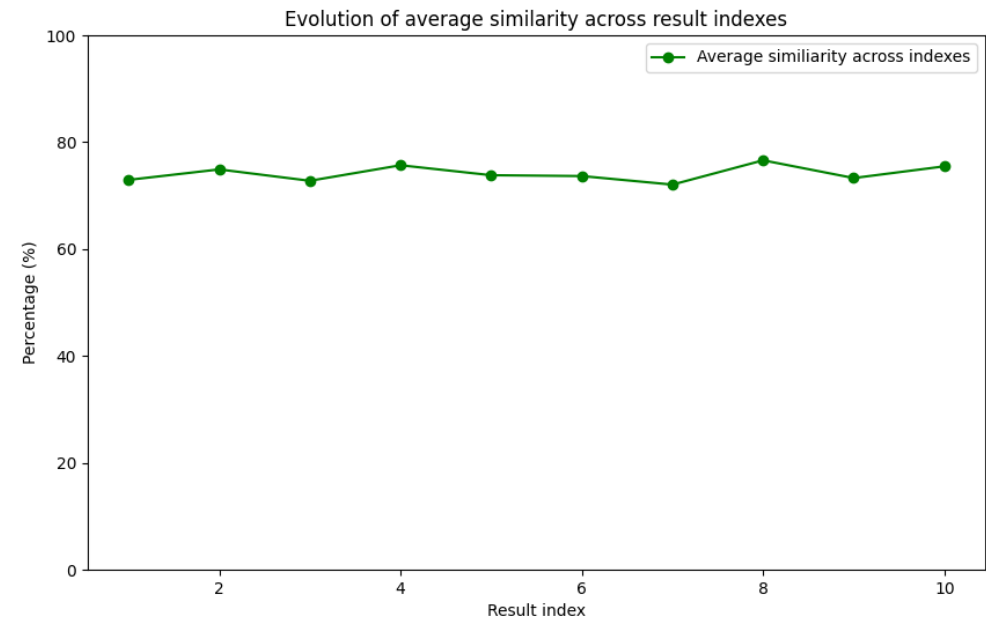
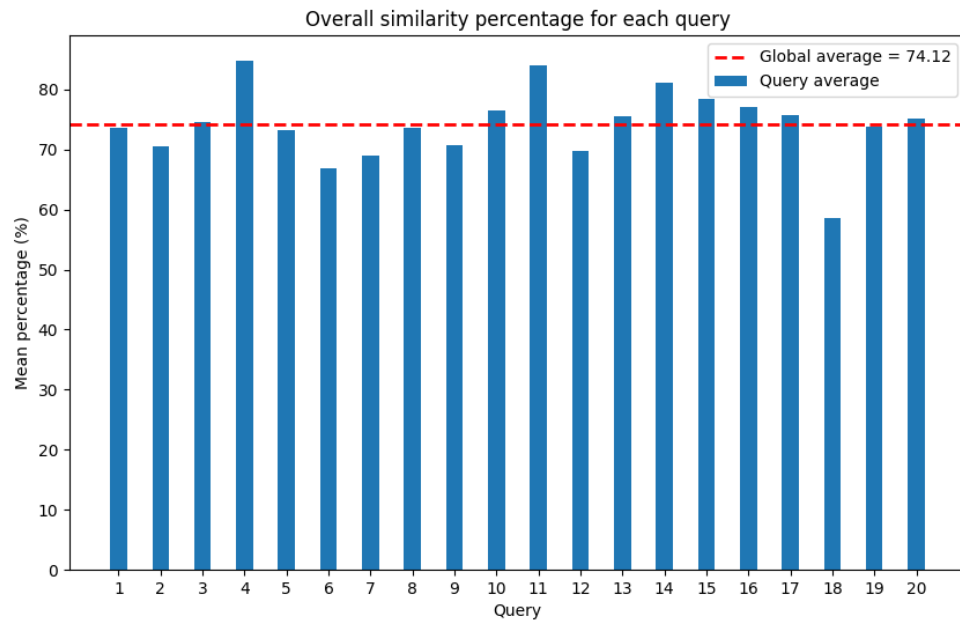
More Like This

- Empowers users to discover documents **similar to** a specified document.
 - `mlt.fl = text`
 - `mlt.mintf = 2` (terms)
 - `mlt.mindf = 5` (documents)
- We did not apply boosts to the terms or fields.

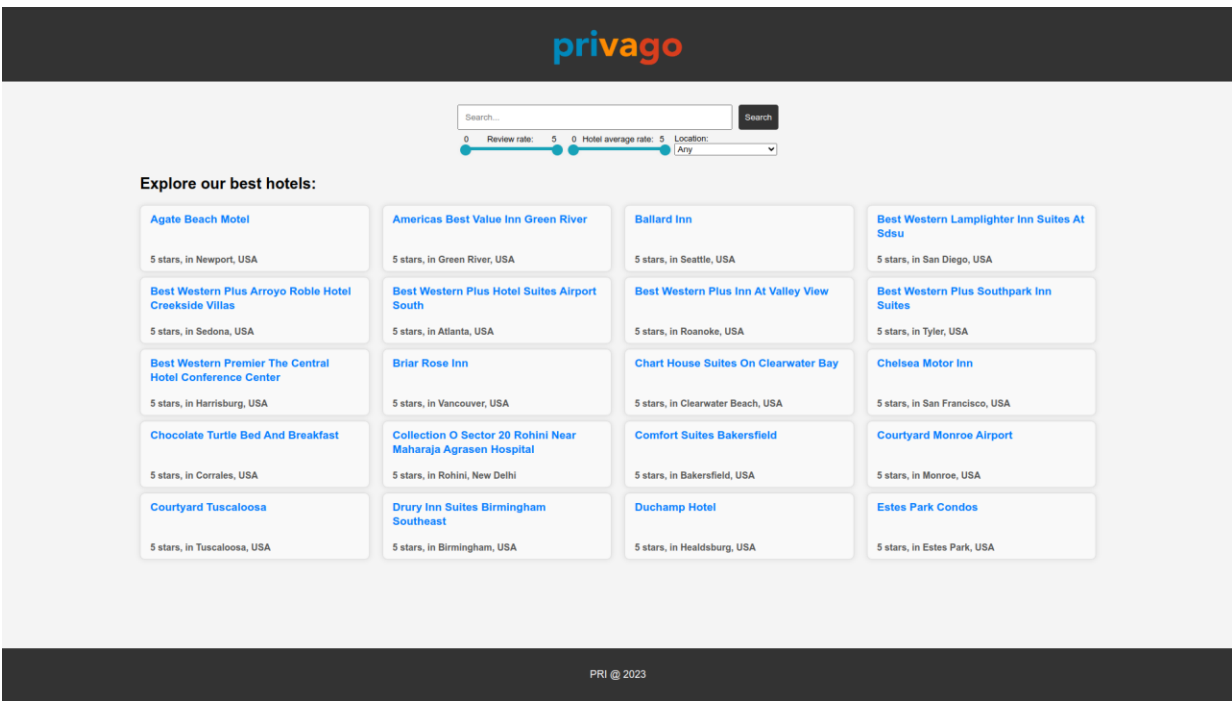


More Like This

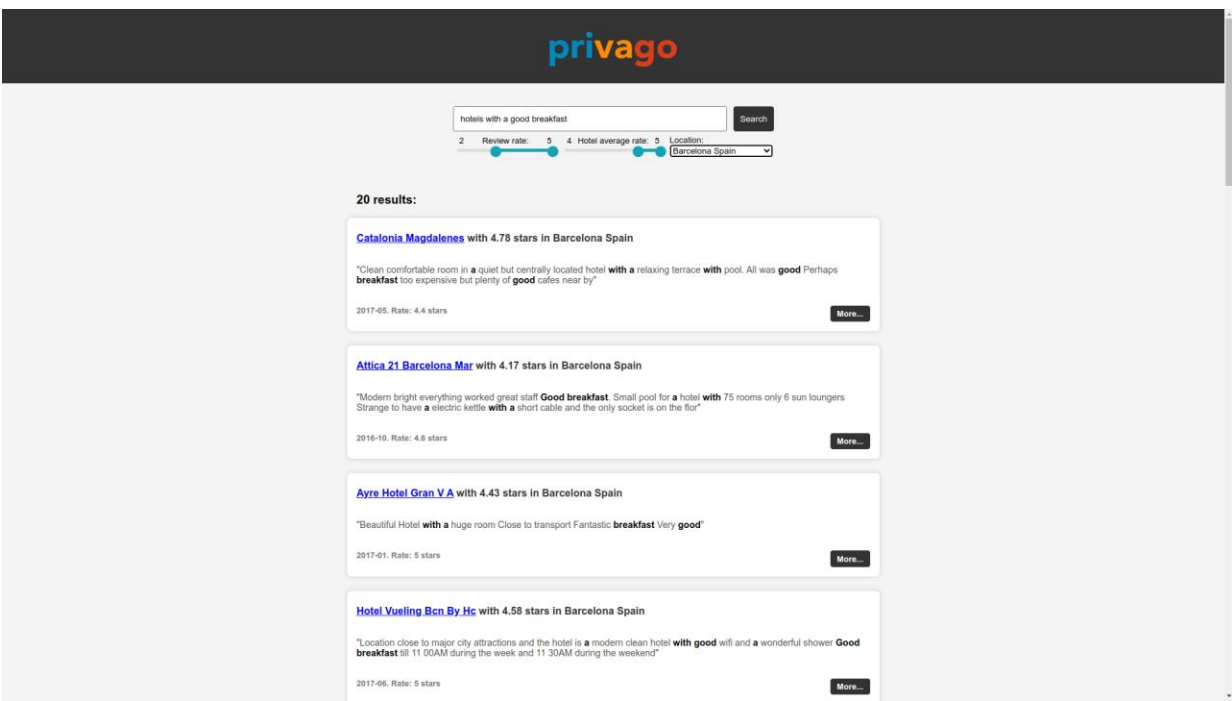
Evaluation using Python's Spacy library



Search User Interface



Home Page



Search Page



Chocolate Turtle Bed And Breakfast

5 stars, in Corrales, USA

8 reviews:

"We absolutely loved our stay at the Chocolate Turtle. Keith and Denise are delightful hosts and their BB is truly their home. And they are clearly thrilled to share it with you. The Sandia room was charming and the common areas, especially outside are lovely. Breakfast was delicious and satisfying and our location allowed us easy access to all the... More"

2015-05. Rate: 5 stars

More...

"The owners, Keith and Denise, are the exact type of people who should be running a B&B. They're friendly, full of useful (and interesting!) information about the area, and they make delicious breakfasts each and every morning. It honestly felt more like staying with friends or family members than staying in a hotel. My husband and I were visiting New... More"

2016-01. Rate: 5 stars

More...

"Our stay was great! Accommodations were wonderful and comfortable. Breakfasts were super. Hosts Denise and Kevin were friendly and sitting on the back porch or in the gazebo watching the roadrunner... More"

2015-09. Rate: 5 stars

"The Sandia room was great with it's large bed and plenty and I could go on and on for days about how beautiful it is there. Saw many beautiful birds... More"

2016-07. Rate: 5 stars

"I wanted a quiet, tasteful wedding in lovely surroundings. I got exactly that at the Chocolate Turtle. The room was beautiful with a gazebo in beautiful grounds for the ceremony. The entire place is premium New Mexico ambiance from the tasteful work and furnishings to the delightfully whimsical curios. The you step outside to the beautiful backdrop... More"

2016-06. Rate: 5 stars

More...

Search User Interface

Hotel Page

Search User Interface

An evaluation of the interface was carried out, multiple individuals were asked to try the application and then fill a form of 10 multiple choice questions that quantify certain aspects of the application's interface. Then, a score was calculated for each individual, resulting in a mean score of 81,6%. The results obtained by the forms are shown below:

User	1	2	3	4	5	6	7	8	9	10	Points
1	3	2	5	3	4	1	4	2	4	2	75
2	4	1	5	3	4	2	4	1	4	2	80
3	3	2	5	2	4	2	5	1	4	2	80
4	2	1	5	2	5	2	5	2	4	1	82,5
5	5	1	4	2	4	2	5	1	3	1	85
6	5	2	4	2	5	3	3	2	4	2	75
7	3	2	5	1	4	1	4	2	5	3	80
8	5	2	4	1	3	2	4	1	5	1	85
9	4	1	5	2	4	1	5	1	4	2	87,5
10	3	2	4	1	5	2	5	2	4	2	80
11	4	1	4	1	5	2	4	1	5	2	87,5



Final System

The final system has integrated three out of the four topics discussed throughout this presentation: **Stopwords** Filter, **More Like This**, and **Search User Interface**.

Due to the absence of improvement in the Semantic Search feature and its **independent nature** from the rest, the decision was made to exclude it from the final implementation.

The final information retrieval system has the following properties:

- **ASCII Folding**
- **LowerCase**
- **Synonyms**
- **Stopwords**
- **More Like This**

Conclusion:

Milestone Achievements & Future Work

- We have successfully accomplished all tasks set for this milestone.
- The most challenging aspect of this project was dealing with Solr's API during backend development
- User experience would improve if:
 - context-aware searches using broader NLP libraries, in addition to the features already implemented by Solr at this level, would be a strategy for a more globally applicable system with reduced bias.
 - the results of a search did not require a new page to load, due to the loading times depend on the query.



References

[Lucene](https://lucene.apache.org/core/9_9_0/index.html) 2023/12/08

[NLTK] (<https://www.nltk.org/>) 2023/12/04

[More Like This](https://solr.apache.org/guide/8_8/morelikethis.html) 2023/12/04

[Python Spacy] (<https://spacy.io/>) 2023/12/07

[Solr's API] (https://solr.apache.org/guide/8_5/client-apis.html) 2023/12/09