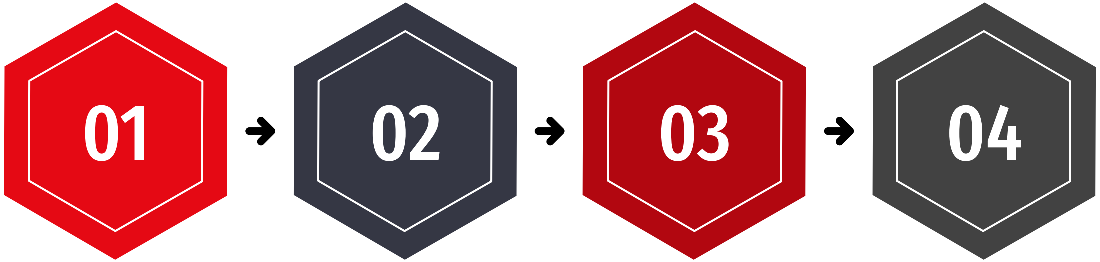
Netflix Titles and IMDb Reviews

Improving the system

PRI Group 2136
Ricardo Fontão, up201806317
Telmo Baptista, up201806554
Tiago Silva, up201806516

Review Language Identification

Language Identification - Pipelines



Foreach processor

Can only run a single processor per entry

Nested pipeline

Allow multiple processors per entry

Language identifier

Inference processor with the lang_ident_model_1

Selects top 3 candidates

New fields

language: “en”
en: “Review here”

Language Identification - Indexing

Per-field approach

Easily integrated into previous work
Allows multi-language queries
Duplicate fields for each language used

Per-index approach

Extra index per language supported
Avoids duplication of data

Language Identification - Languages Present

Language	Prevalence
English	Majority of reviews
Spanish	Over 1k reviews
French	500 reviews
German	170 reviews
Portuguese	168 reviews
Others	All under 150 reviews each

Language Identification - System Performance

	Total index time	Index time per doc
W/o Language detection	120 sec	13 ms
W/ Language detection	57 min	420 ms

38 x Slower ●

Increasing the Elastic container’s RAM had no effect

Search Need IV

In Spanish:

Search for horror movies

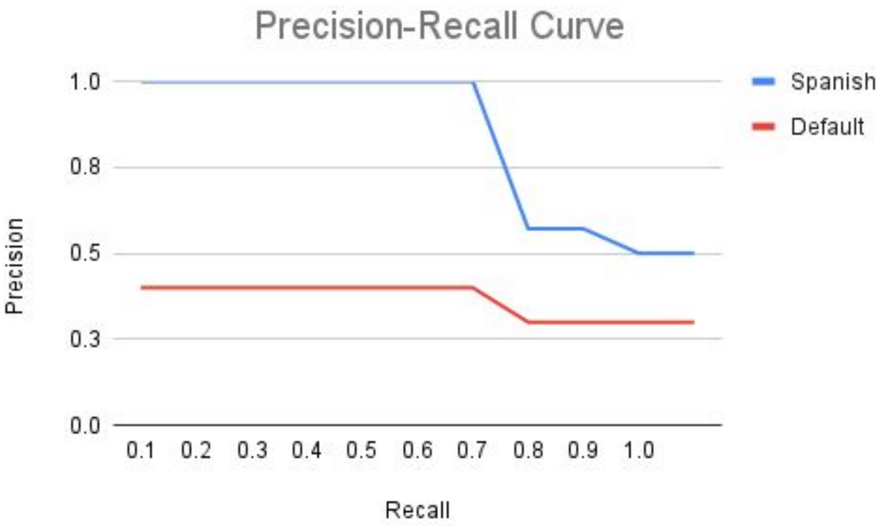
Search Need IV

es: Película horror

Experiments performed:

- **Spanish**:
  es (reviews in Spanish)
- **Default**:
  review_details

	Result	P@10	R@10	AP
Spanish	R R R N N N R N N R	0.50	0.50	0.81
Default	N N R N R N N N N R	0.30	0.75	0.34



Search Need V

In Spanish:

Search for movies with aliens

Search Need V

es: Película extraterrestres

Experiments performed:

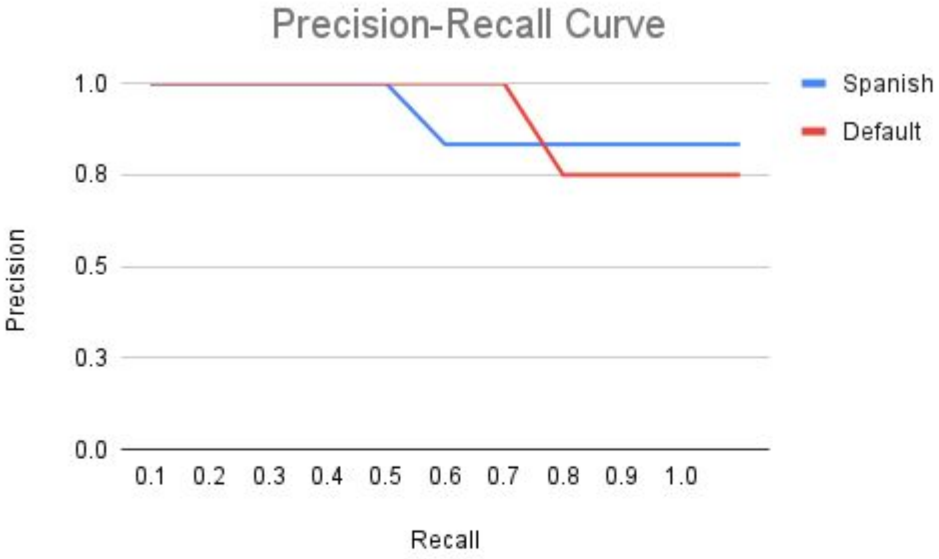
- **Spanish**:

  es (reviews in Spanish)

- **Default**:

  review_details

	Result	P@10	R@10	AP
Spanish	R R N R R R N N N N	0.50	0.83	0.88
Default	R R N R N N N N N N	0.30	0.75	0.92



Synonyms

Synonyms - Basics

Why use them?

Synonym token filters allow more variety and accuracy to our search

When are they useful?

Bollywood ↴
Hindi cinema, Indian movies

How to handle them?

Elasticsearch provides ways to customize analyzers

How to set it up?

Provide a synonym filter to the custom analyzer

How does it work?

Elasticsearch maps synonyms to the inverted index to be used when querying the index

Synonyms - Implementation

	Scalability	Results
Wordnet	Highly scalable	Apparent randomness

English words are grouped into cognitive synonyms (synsets)
Can cover the majority of the words used in the documents stored

Manual Listing	Not scalable	Acceptable, but limited scope

Covers a small portion of all the variance of words present in the documents

Search Need VI

Search for documentaries covering topics about pollution on our planet

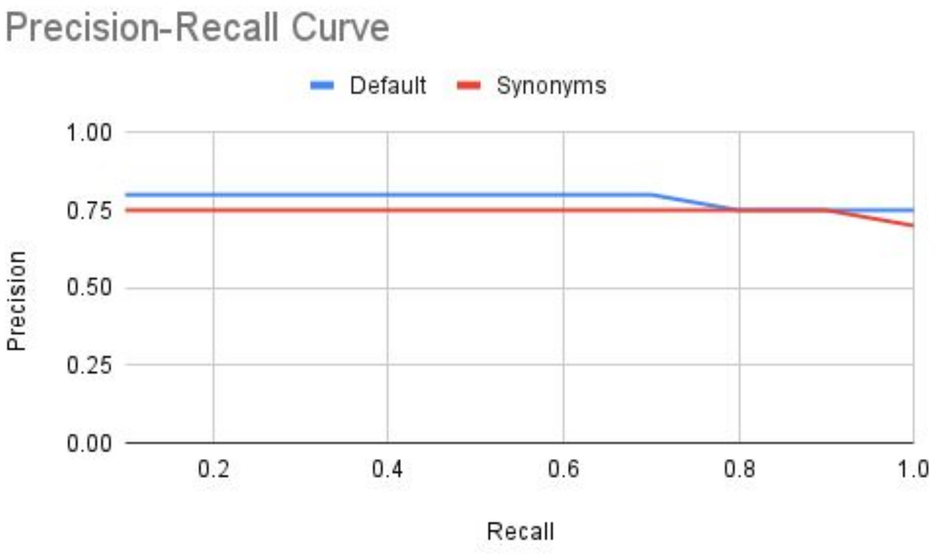
Search Need VI

documentaries about pollution

Query Text:
Documentaries about pollution

- **Default**:
  Without synonyms filter
- **Synonym Search**:
  With synonyms filter

	Result	P@10	R@10	AP
Default	N R R R R N R R N N	0.60	0.75	0.69
Synonym	N R R N R R R R N R	0.70	0.875	0.65



Search Need VII

Search for sci-fi movies featuring mechas

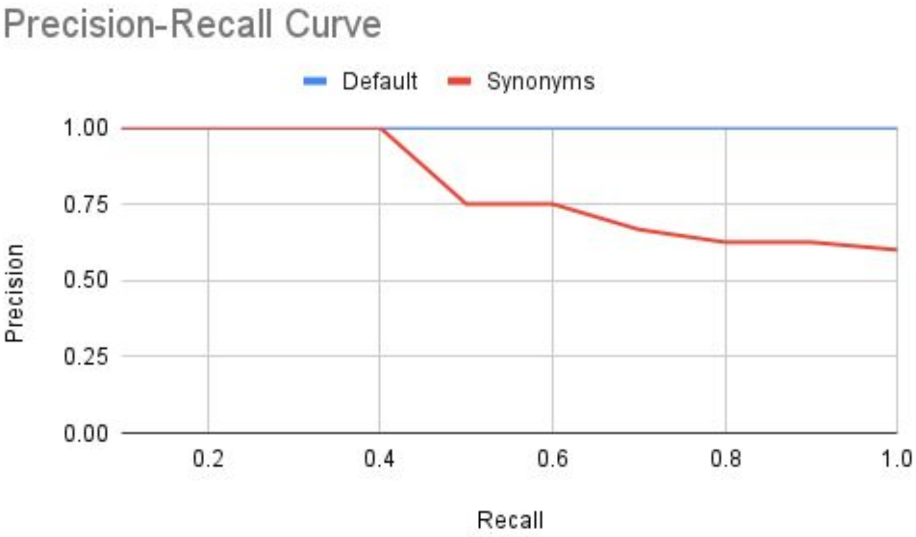
Search Need VII

scifi and mecha movies

Query Text:
scifi and mecha movies

- **Default**:
  Without synonyms filter
- **Synonym Search**:
  With synonyms filter

	Result	P@10	R@10	AP
Default	R R R R R R R R R R	1.00	0.90	1.00
Synonym	R R N R N R N R N R	0.60	0.85	0.77



Future work

Translate fields such as genre

Integrate Wordnet properly for lexical analysis

Create a Search User Interface