Vector Space Model

Jose Carlos Archundia Adriano

Carlos Andres Reyes Evangelista

Erick Siordia Nagaya

Problem Description

Data collections need a system to quickly find specific data

Search Engines

Document Ranking

Vector Space Model

Retrieval System Evaluation

Precision

Recall

Problem Description: In practice

CRANFIELD COLLECTION OF ARTICLES

TEST QUERIES

ARTICLES RELEVANCE

What is the Vector Space Model?

"Vector Space Model is an information retrieval model described as a simple and intuitively appealing framework for implementing term weighting, ranking and relevance feedback"

Bruce Croft et al. (2015)

$$idf_t = \log_{10}(\frac{N}{df_t})$$

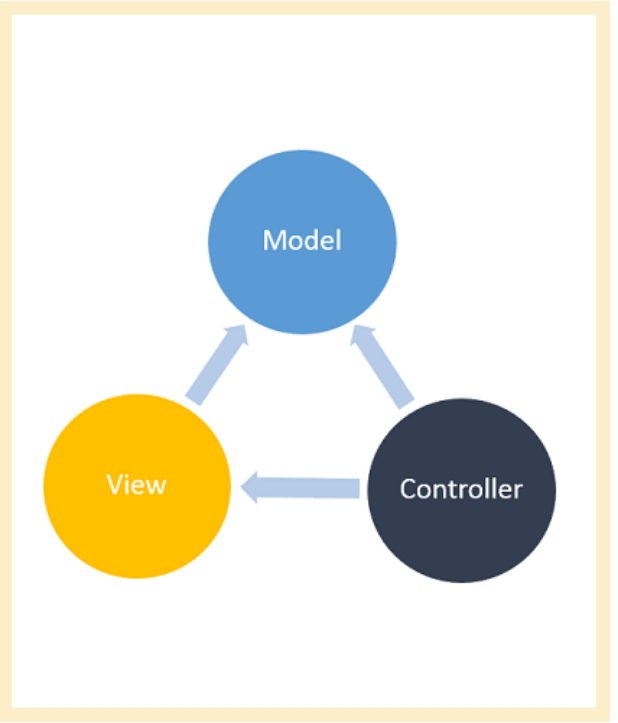
Objective

Achieve a fully-functional implementation of the Vector Space Model

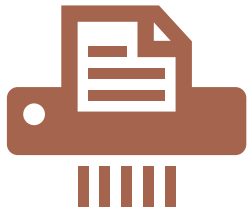
Parse a test collection to let the user choose among a set of pre-processed queries and to compute and display the most relevant results for that given query

Provide an option to select 10 random queries, compute its precision and recall and graph its average.

Methodology and Technologies



Model



Reader

- Load from disk the articles, queries and relevant samples to memory.

- Pre-analysis tasks to organize the data for its further usage.

VectorSpaceModel

- Generation of documents and query weighed vectors.

- Computation of the similarity coefficient for the articles vectors.

- Creation of ranked results list for a given query.

- Selection of random queries for the evaluation of the system.

View

Home

First screen

Searchbar

Results

Searchbar

Ranked result list

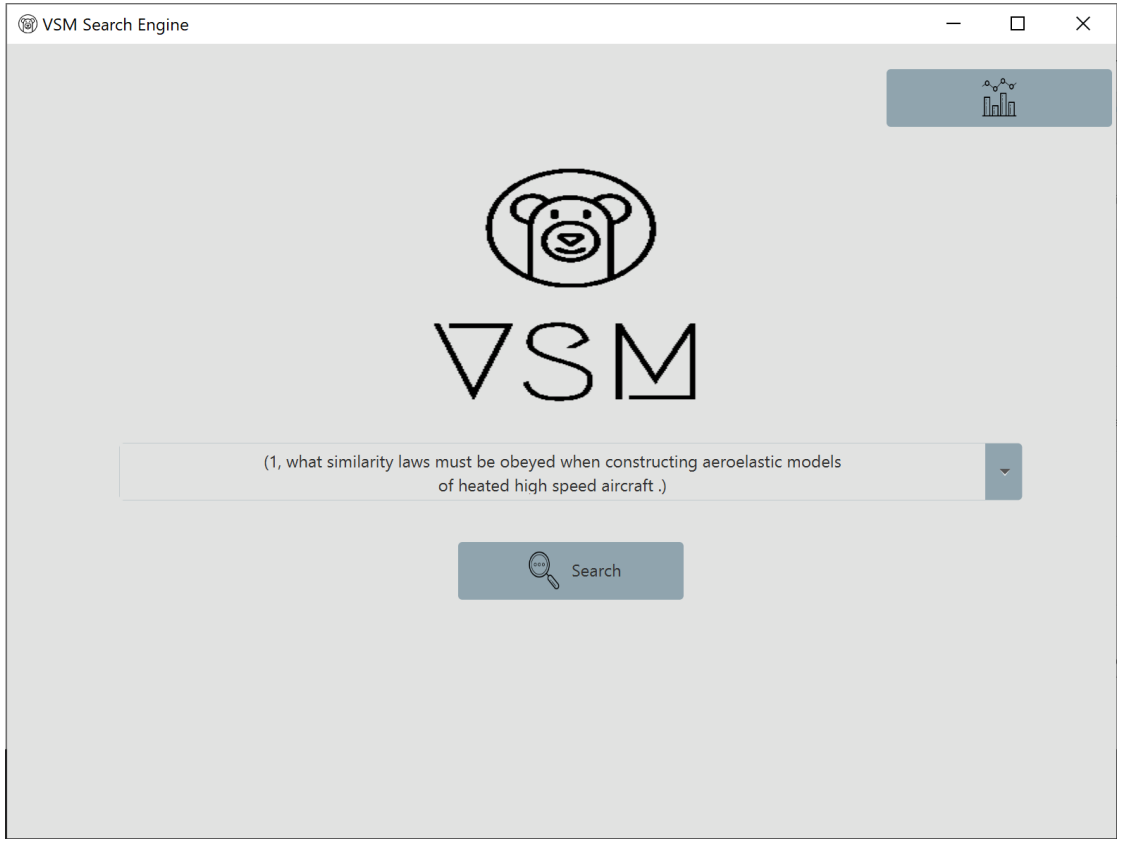
Evaluation button

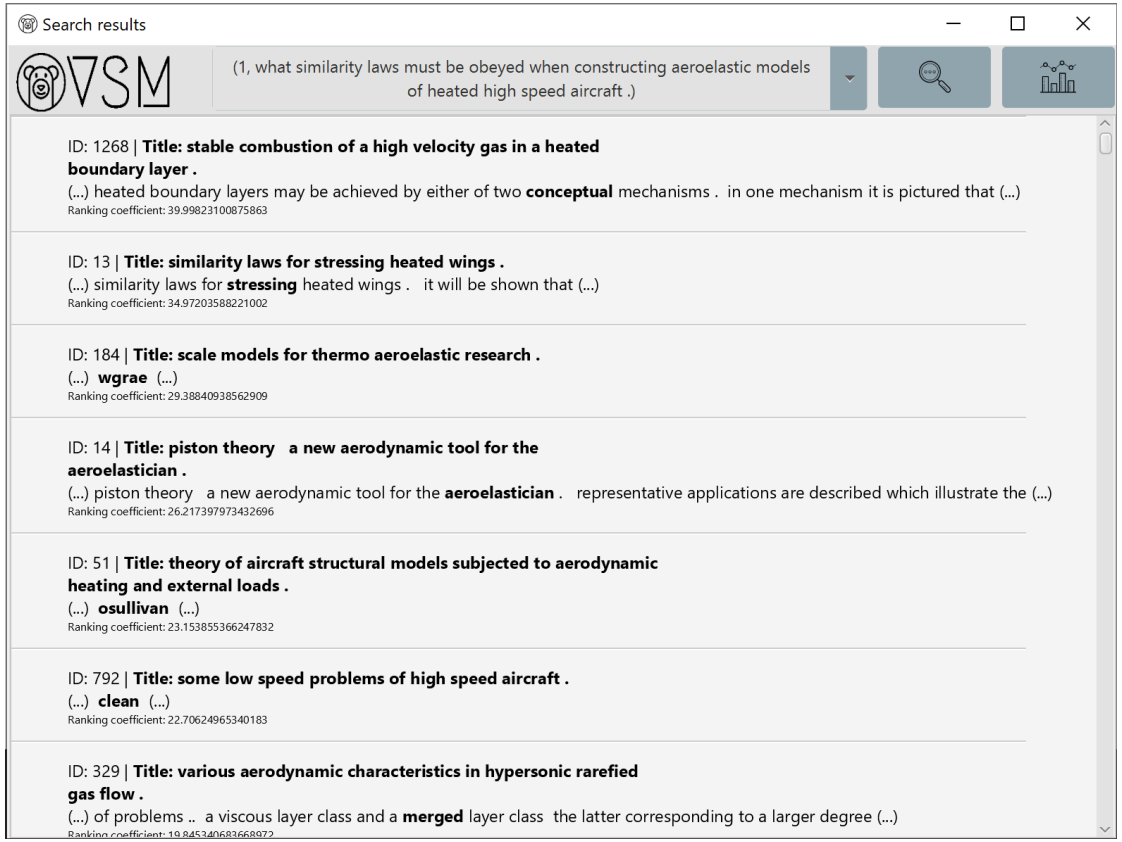
Chart

Presentation of precission/recall data

Results

Results


VSM Search Engine
(1, what similarity laws must be obeyed when constructing aeroelastic models of heated high speed aircraft .)
Search


Search results
(1, what similarity laws must be obeyed when constructing aeroelastic models of heated high speed aircraft .)
ID: 1268 | Title: stable combustion of a high velocity gas in a heated boundary layer .
(...) heated boundary layers may be achieved by either of two conceptual mechanisms . in one mechanism it is pictured that (...)
Ranking coefficient: 39.99823100875863
ID: 13 | Title: similarity laws for stressing heated wings .
(...) similarity laws for stressing heated wings . it will be shown that (...)
Ranking coefficient: 34.97203588221002
ID: 184 | Title: scale models for thermo aeroelastic research .
(...) wgrae (...)
Ranking coefficient: 29.38840938562909
ID: 14 | Title: piston theory a new aerodynamic tool for the aeroelastician .
(...) piston theory a new aerodynamic tool for the aeroelastician . representative applications are described which illustrate the (...)
Ranking coefficient: 26.217397973432696
ID: 51 | Title: theory of aircraft structural models subjected to aerodynamic heating and external loads .
(...) osullivan (...)
Ranking coefficient: 23.153855366247832
ID: 792 | Title: some low speed problems of high speed aircraft .
(...) clean (...)
Ranking coefficient: 22.70624965340183
ID: 329 | Title: various aerodynamic characteristics in hypersonic rarefied gas flow .
(...) of problems .. a viscous layer class and a merged layer class the latter corresponding to a larger degree (...)

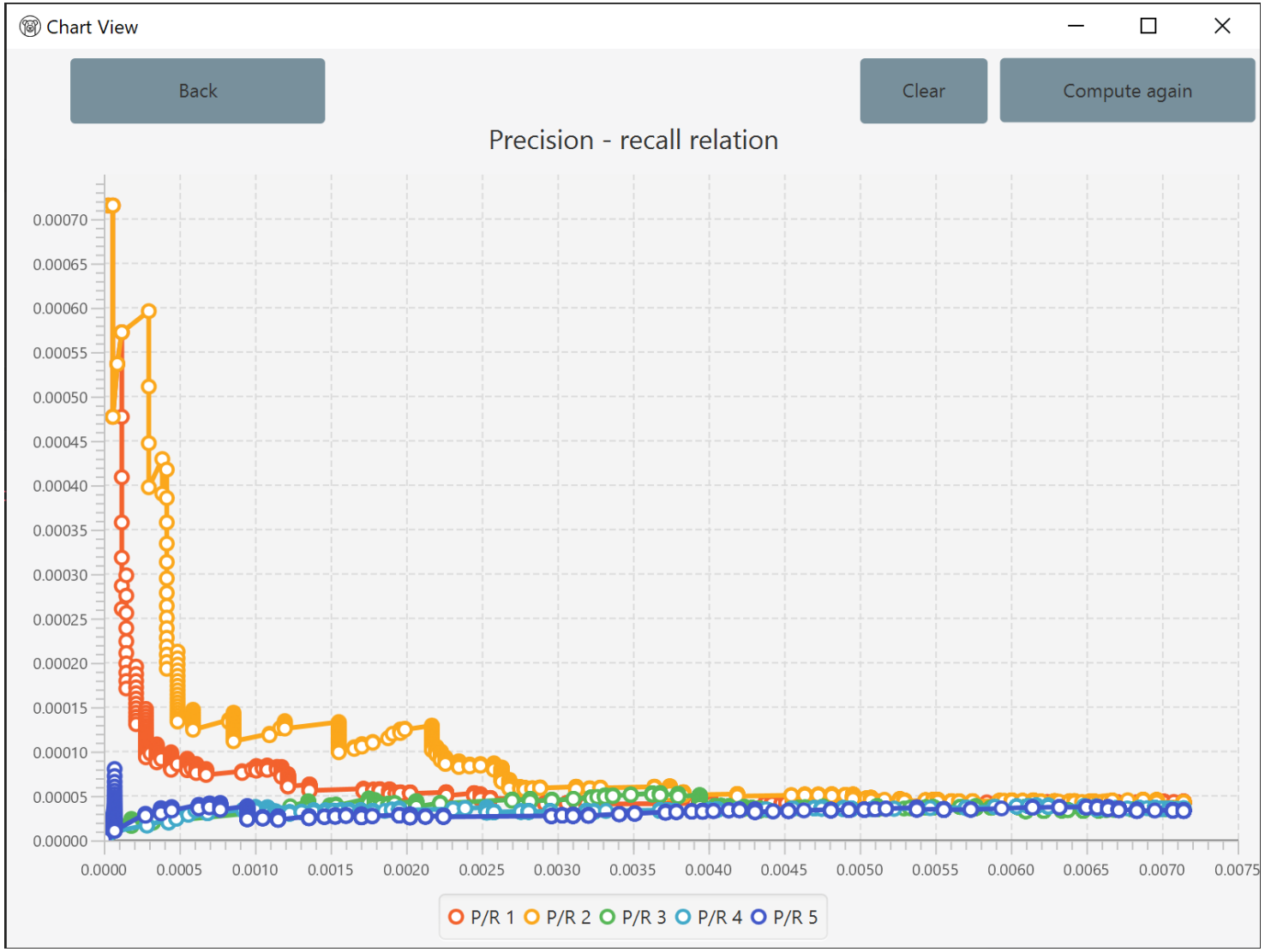


Chart View
Back
Clear
Compute again
Precision - recall relation
0.00070
0.00065
0.00060
0.00055
0.00050
0.00045
0.00040
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
0.0030
0.0035
0.0040
0.0045
0.0050
0.0055
0.0060
0.0065
0.0070
0.0075
P/R 1
P/R 2
P/R 3
P/R 4
P/R 5

Conclusions

- Challenges
  - Responsive UI
  - Kotlin-specific features and tools
  - Dependency management (Maven)
  - IntelliJ IDEA
- Lessons
  - Team work
  - Version Control System
  - Communication
  - GUI development