Vector Space Model

Zahra Younes Pour Langaroudi

Part1

Introduction

Project Overview

Goal:

- To build a complete IR system from scratch based on the Vector Space Model

Dataset:

- Cranfield Collection (1,398 aeronautical abstracts)

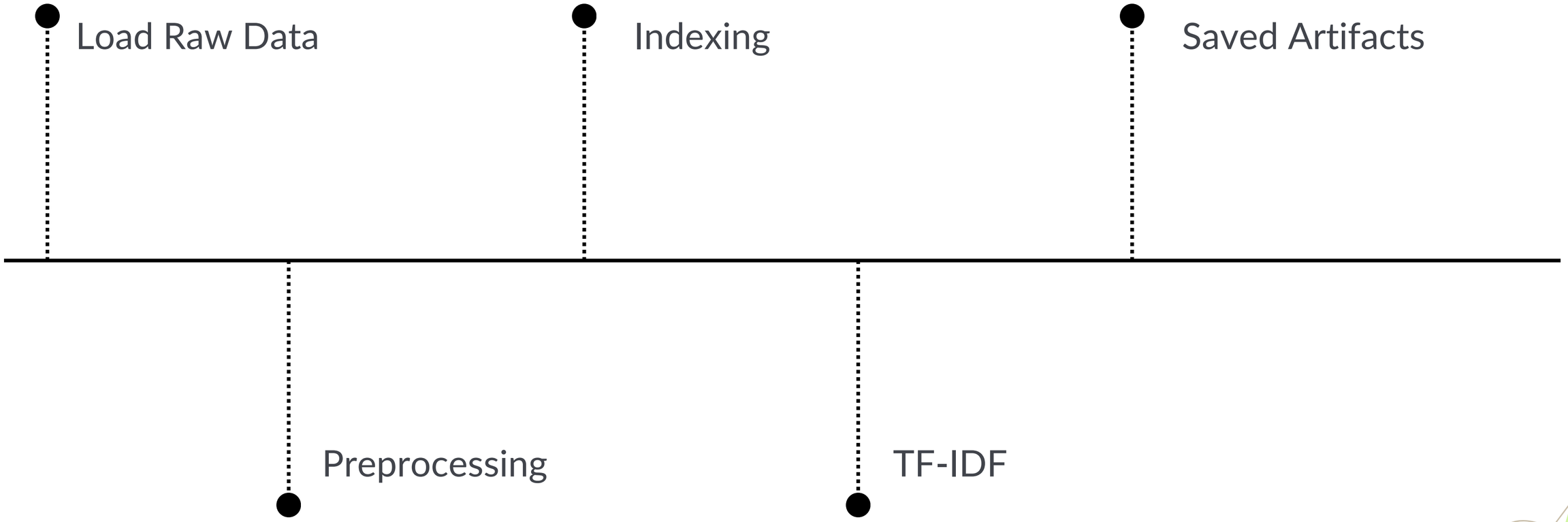
Core Features:

- TF-IDF based vector representation
- Cosine Similarity for ranking
- Pseudo-Relevance Feedback (Rocchio Algorithm)
- Systematic evaluation using Precision & Recall

Part2

System Architecture & Methodology

Indexing Pipeline (One-time process)



Evaluation Or User Query Pipeline

- Load Artifacts
- Load Relevance Judgment
- Load Cranfield Queries Or User Query

- Preprocessing
- Create Query Vector

- Rank Documents
- Initial Rank
- Final Rank with Feedback (Optional)

User Query:

- Retrieved 10 Highest Documents

Evaluate:

- Calculate Metrics such as Precision, Recall, F1

Preprocessing & Weighting

Preprocessing:

- **Tokenization:** Extracting words
- **Stop-Word Removal:** Using the standard NLTK list
- **Stemming:** Using PorterStemmer to reduce words to their root form

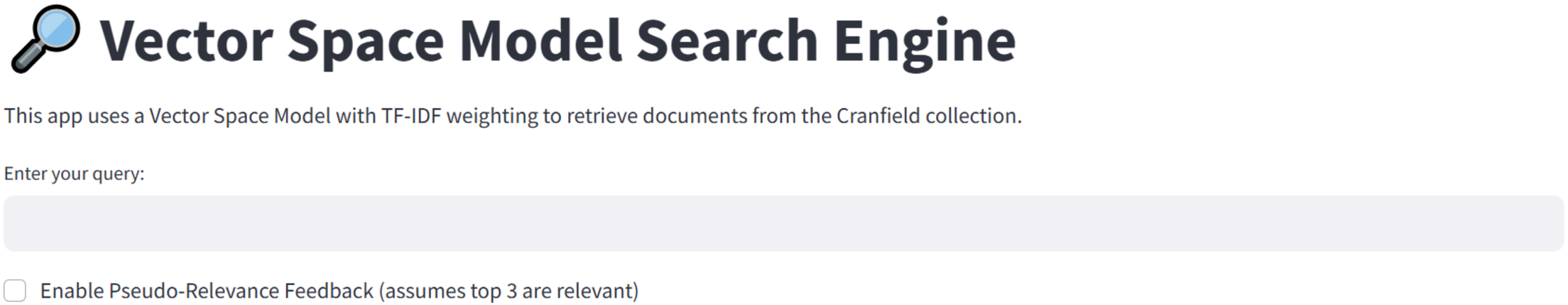
Weighting Scheme: TF-IDF

- **Term Frequency (TF):** Used a raw count of term occurrences
- **Inverse Document Frequency (IDF):** Standard log(N/df) to measure term importance

Part3

Live Demonstration

Live Demo


🔎 Vector Space Model Search Engine
This app uses a Vector Space Model with TF-IDF weighting to retrieve documents from the Cranfield collection.
Enter your query:
Enable Pseudo-Relevance Feedback (assumes top 3 are relevant)

Part4

Evaluation & Results

System Performance & Challenges

Metric(@K =20) For 225 Queries	Baseline Score	Score with Feedback
Mean Precision	0.0049	0.0049
Mean Recall	0.0029	0.0029
Mean F1-Score	0.0035	0.0035

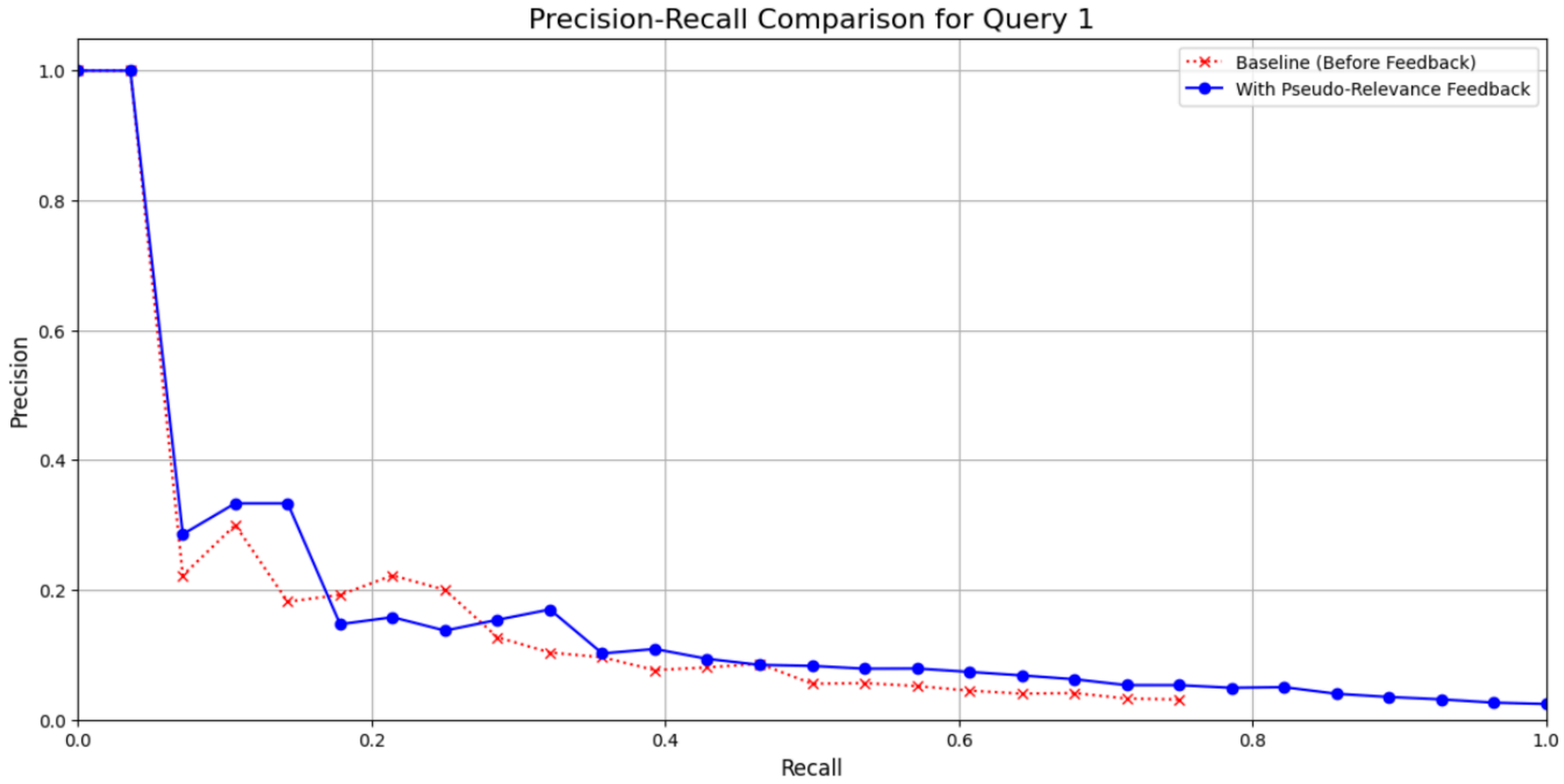
Key Challenge: Uneven relevance judgments in the Cranfield dataset.

System Performance & Challenges

Metric(@K =20) For Query 1	Baseline Score	Score with Feedback
Mean Precision	0.15	0.20
Mean Recall	0.10	0.14
Mean F1-Score	0.12	0.17

Finding: When initial results are good, as in this case, pseudo-relevance feedback successfully improves performance.

P-R Curve



Analysis: For Query 1, feedback provides a small but consistent improvement in precision at most recall levels.

Part5
Conclusion

Conclusion & Future Work

Achievements:

- Successfully built a complete Vector Space Model IR system from scratch.
- Scientifically evaluated the baseline model's performance and challenges.
- Analyzed the limitations of pseudo-relevance feedback on this baseline.
- Built a user-friendly web interface.

Future Work:

- Implement a more advanced weighting (like lnc.ltc)
- Experiment with true user relevance feedback.

Thank You & Questions