Group A1 - IR Assignment

Team

- MADHUMITHA M - 2020MT12048
- PAVAN KUMAR VANNEMREDDY - 2020MT12393
- PAVITHRA B - 2020MT12247
- SAKTHI SARAVANAN S - 2020MT12198

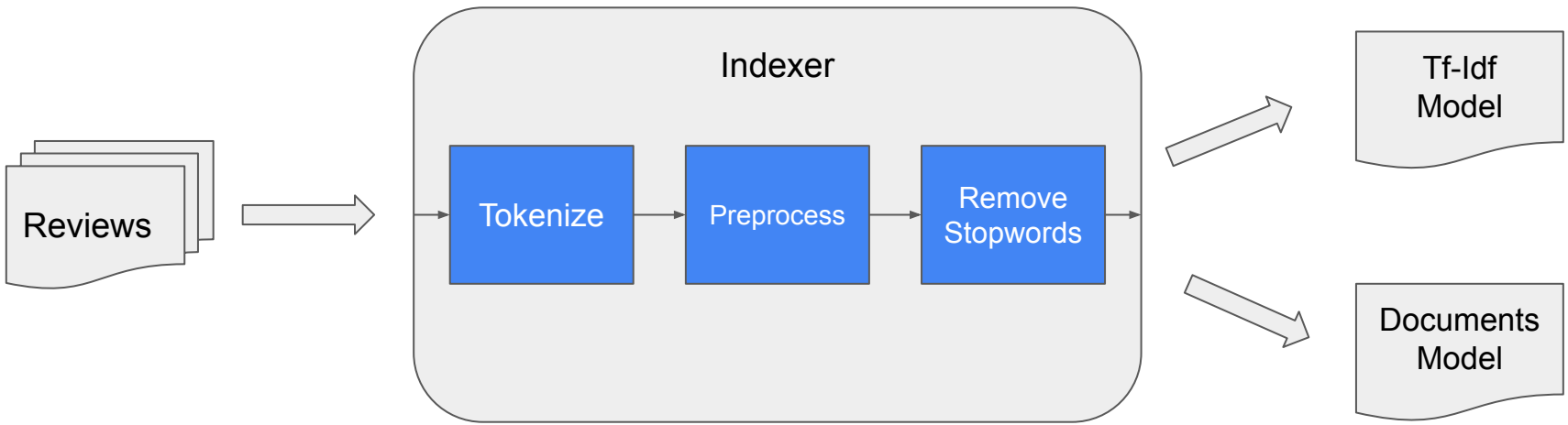
Summary

- **Problem Statement**: Build a Search Engine that caters to a particular domain
- **Dataset Used**: Amazon Product Reviews (https://www.kaggle.com/datafiniti/consumer-reviews-of-amazon-products)
- **Scoring Approach**: TF-IDF
- **Ranking Approach**: Cosine similarity
- **Tools**:
  - Language: Java
  - Build Tool: Maven
  - Ser/De: Kryo
  - Apache Commons Math for Vector operations

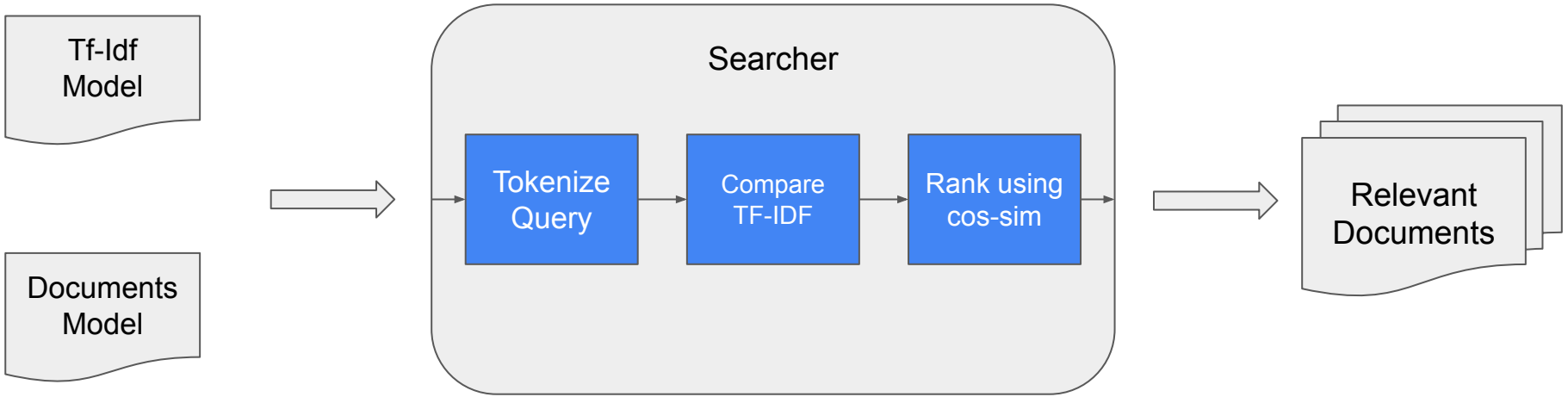
Major Data Structures (Custom Built)

- **Tfldf** - TF Table contains vocab, tfldf vectors and document frequencies
- **Counter** - It helps generate count for each token from a list
- **OutputRecord** - Helps store the docId, cosine similarity score and actual document

Indexing Workflow



Searching Workflow



Demo - Indexing

```
ir-assignment >java -cp target/ir_assignment-1.0-SNAPSHOT.jar edu.bits.wilp.ir_assignment.index.Indexer
11:36:00.408 [main] INFO  e.b.wilp.ir_assignment.index.Indexer - Opening datasets/sample.csv for reading
11:36:00.433 [main] INFO  e.b.wilp.ir_assignment.index.Indexer - Files read, starting to compute DF
11:36:01.148 [main] INFO  e.b.wilp.ir_assignment.index.Indexer - DF Complete
11:36:01.148 [main] INFO  e.b.wilp.ir_assignment.index.Indexer - Starting Tf-Idf calculations
11:36:01.424 [main] INFO  e.b.wilp.ir_assignment.index.Indexer - Tf-Idf calculations, done. Writing the model file: output.bin
11:36:01.602 [main] INFO  e.b.wilp.ir_assignment.index.Indexer - Persisting the documents: documents.bin
11:36:01.618 [main] INFO  e.b.wilp.ir_assignment.index.Indexer - Indexing Complete
```

Demo - Searching - “love tablet”

```
ir-assignment >java -cp target/ir_assignment-1.0-SNAPSHOT.jar edu.bits.wilp.ir_assignment.search.Searcher "love tablet"
11:37:22.536 [main] INFO  e.b.w.ir_assignment.search.Searcher - Loading D SparseMatrix
11:37:22.628 [main] INFO  e.b.w.ir_assignment.search.Searcher - D SparseMatrix Loaded
11:37:22.659 [main] INFO  e.b.w.ir_assignment.search.Searcher - Searching for query: love tablet
11:37:22.706 [main] INFO  e.b.w.ir_assignment.search.Searcher - Computed Query Vectors
11:37:22.706 [main] INFO  e.b.w.ir_assignment.search.Searcher - Searching across documents
11:37:23.554 [main] INFO  e.b.w.ir_assignment.search.Searcher - Computed cosine-sim across documents
Precision: 0.546, Recall: 0.0184
{"docId":463,"cosineSim":0.816370774647198,"document":"Love the tablet! Easy to use and easy to take with"}
{"docId":570,"cosineSim":0.7069005655369549,"document":"I love the amazon tablet very much and love to use"}
{"docId":137,"cosineSim":0.631736058859506,"document":"Good tablet. Wife loves it and would recommend...."}
{"docId":231,"cosineSim":0.6308758405740864,"document":"I got this tablet on sale for my wife and she loves it"}
{"docId":698,"cosineSim":0.618752056932315,"document":"We have had no issues with this tablet. Love it!TY"}
{"docId":761,"cosineSim":0.5772570526908305,"document":"I love to read and this tablet work great for my needs."}
{"docId":680,"cosineSim":0.5771066450965584,"document":"This is a great tablet with great features. I love my kindle."}
{"docId":566,"cosineSim":0.5768396132505973,"document":"great tablet to replace kindle. Love the features."}
{"docId":755,"cosineSim":0.5764721594546459,"document":"No problems with tablet, kids love it and the size is great"}
{"docId":508,"cosineSim":0.5761544749214773,"document":"My daughter love this tablet! Easy to use and carry"}
ir-assignment >
```

Citations

- Stopwords used are from https://www.ranks.nl/stopwords
  - We use the very large stop word list.
- TF-IDF from scratch in python on a real-world dataset.