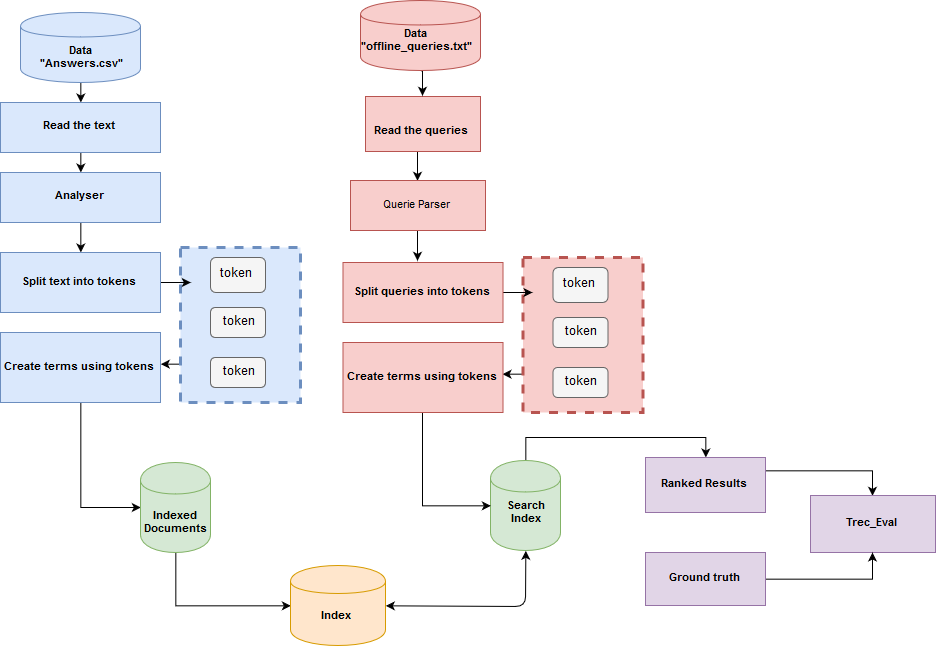
**Information retrieval**

**Lab 1.1- Building and Searching the Index**

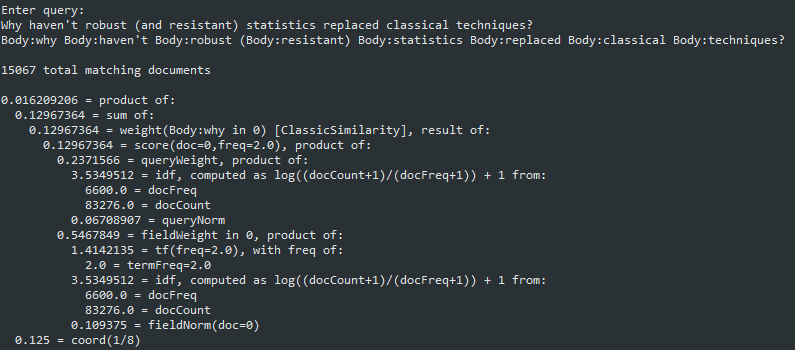
**Data workflow**



Data Workflow

**Understanding the search results**

Utilizámos a classe *Explanation* do *Lucene framework* que descreve sucintamente os cálculos realizados na avaliação dos documentos e das *queries* e obtivemos o seguinte output:



Lucene Explanation Example

Na figura anterior conseguimos visualizar que antes de ser feito qualquer cálculo, cada *query* sofre um *parsing* onde é subdivida em *tokens*. Depois dos *tokens* terem sido gerados é feita a pesquisa nos documentos e obtemos o número de documentos onde os mesmos estão presentes.

A avaliação de cada *token* varia consoante o seu número de ocorrências nos documentos em que estamos a procurar.

Neste excerto de output temos a ocorrência de apenas um dos tokens em oito possíveis, nomeadamente do token “*why*”, com uma avaliação de **0.016209206**. Este valor é obtido através cálculo do *cosine similarity*, que resulta do produto entre peso de cada token no determinado documento (neste caso de **0.12967364**) e o fator *coord* (divisão entre o número de *tokens* encontrados no documento pelo número de *tokens* criados pelo *parser*) que toma o valor **0.125**.

O peso de cada token (*weight*) é obtido através do produto entre o *queryWeight* (**0.2371566**) e o *fieldWeight* (**0.5467849**).

O valor do *queryWeigh*t resulta do produto entre o *idf* (**3.5349512**) e *queryNorm* (**0.06708907**) que representam respetivamente a rareza do token em cada documento e o fator normalizador que torna possível duas queries comparáveis.

O peso do campo (fieldWeight) resultado do produto entre tf (**1.4142135**), idf (**3.5349512**) e o fieldform (**0.109375**).

**Precision vs Recall**

Através dos valores obtidos no trec\_eval produzimos o seguinte gráfico precision-recall:

[Dizer o que representa cada eixo e analisar]

Rank Precision:

Através dos valores obtidos no trec\_eval produzimos o seguinte gráfico Rank Precision:

[Dizer o que representa cada eixo e analisar]

Segundo os valores obtidos através do trec eval criamos o seguinte gráfico de precisão e revocação, onde a precisão representa a capacidade de o sistema nos retornar os documentos mais relevante e a revocação a capacidade de o sistema encontrar dentro da coleção os documentos mais relevantes.

Ao observamos este gráfico conseguimos verificar que existe uma área significativa por baixo da curva que nos indica que existe uma elevada taxa de precisão isto, é poucos falsos positivos.

Precision-Recall is a useful measure of success of prediction when the classes are very imbalanced. In information retrieval, precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned.

The precision-recall curve shows the tradeoff between precision and recall for different threshold. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall).

A system with high recall but low precision returns many results, but most of its predicted labels are incorrect when compared to the training labels. A system with high precision but low recall is just the opposite, returning very few results, but most of its predicted labels are correct when compared to the training labels. An ideal system with high precision and high recall will return many results, with all results labeled correctly.

Cosine distance retriavel mode

O TF-IDF com o cosine distance retriavel mode serve para comparar a semelhança entre dois documentos, que é feita através da comparação de termos existentes em cada entre documentos; Varia em [0,1], isto é se o cos do angulo resultante dos dos dois documentos estiver proximo de 1, os mesmo serao muito semelhantes;

O output gerado pela aplicação dá-nos a[ print do output] frequência de cada termo e em cada documento, com isto pode calcular o cos x ;;;;