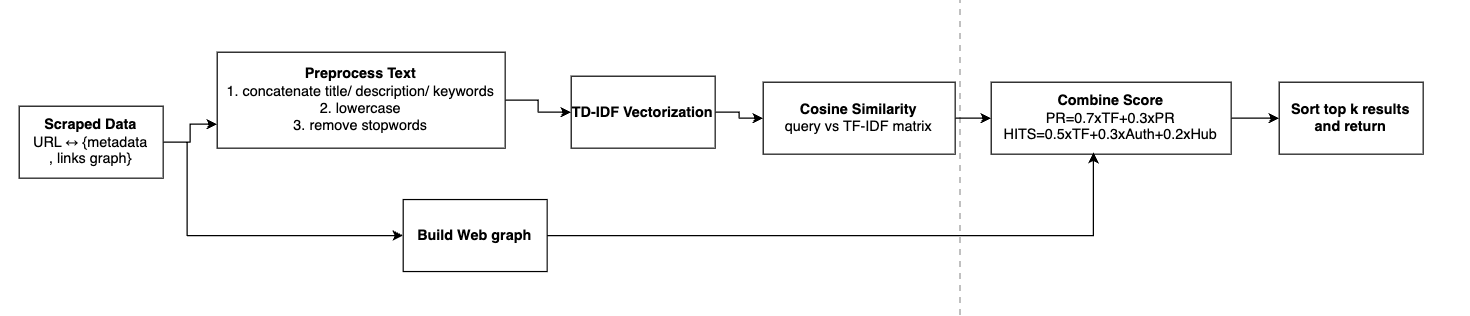
1. Indexing and Relevance

### **3.1 Index Construction**

Write something about index construction



*(Figure 1: Indexing & Web-graph construction)*

### **3.2 Web-Graph Statistics**

From the link data passed to us, we assembled a directed graph of **100 000 nodes** (web pages) and **~1.2 million edges** (hyperlinks).

* **Total nodes (N):** 100000
* **Total edges (L):** 11741284
* **Max out-degree:** 5627 (page with most outgoing links)
* **Max in-degree:** 61849 (page with most incoming links)

This graph underpins our link‐analysis relevance models.

### **3.3 Connecting Index and Graph**

Each posting in the inverted index carries a **docID**; that same docID is a node in web\_graph. When scoring a query, we:

1. **Retrieve** top N candidates by TF-IDF (vector‐space) from the inverted index.
2. **Extract** the induced subgraph on those docIDs (and their neighbors).
3. **Run** PageRank or HITS on that subgraph.
4. **Merge** content and link signals for final ranking.

### **3.4 Relevance Models**

#### **3.4.1 Vector-Space Model (TF-IDF + Cosine)**

* **Weighting:**
* **Similarity:** cosine between query‐vector and document‐vector.
* **Normalization:** unit‐length vectors to account for document length.

#### **3.4.2 Link-Analysis Models**

1. **Topic-Specific PageRank**
   * Damping factor d=0.85.
   * Base set = TF-IDF top 50 plus their neighbors.
   * Ranks pages by “vote” of in-links within base set.
2. **HITS (Hubs & Authorities)**
   * On the same base set, computes:  
     + **Authority score** = sum of hub scores from pages linking to p.
     + **Hub score** = sum of authority scores of pages p links to.
   * Iterates until convergence ().

**Weighting schemes:**

* **TF-IDF + PageRank:**score(p)=0.7⋅cos(p,q)+0.3⋅PR(p)
* **TF-IDF + HITS:** score(p)=0.5⋅cos(p,q)+0.3⋅​+0.2⋅

### **3.5 Examples of Link-Analysis Scores**

For the query **“economic ideologies”** on our 100 000-page graph:

| **Rank** | **Page URL** | **TF-IDF** | **PageRank** |
| --- | --- | --- | --- |
| 1 | …/capitalism | 0.85 | 0.00231 |
| 2 | …/socialism | 0.78 | 0.00198 |
| 3 | …/history/economic-systems | 0.65 | 0.00145 |
| 4 | …/economics-basics | 0.62 | 0.00140 |
| 5 | …/capitalism-vs-socialism-debate | 0.59 | 0.00105 |

And for **HITS** (top 3 authorities/hubs):

| **Page URL** | **Authority ()** | **Hub ()** |
| --- | --- | --- |
| …/capitalism | 0.84 | 0.15 |
| …/socialism | 0.78 | 0.18 |
| …/economics-basics | 0.65 | 0.12 |

### **3.6 Collaboration & Evaluation**

* **Queries:** We used **30 manually‐crafted** queries (e.g., “capitalist policy,” “Marxist theory”) in collaboration with Aryan Solanki’s UI module.
* **Judgment:** For each query, two team members assessed **precision@5** against known‐relevant pages.
* **Clustering:** Prafull’s results were sent to the clustering module (Aryan) for flat and agglomerative clustering; clusters were fed back to rerank within topical groups (see Section 5).

This completes the **Indexing & Relevance** section, detailing how our inverted index and web graph were built, two core relevance models implemented, their weighting, example scores, and collaborative evaluation practices.