

Introduction to Information Retrieval

Text Preprocessing and Text Laws

Marta Arias, José Luis Balcázar,
Ramon Ferrer-i-Cancho, Ricard Gavaldà
Department of Computer Science, UPC

Grau en Enginyeria Informàtica, UPC

September 08, 2025

Information Retrieval

From *library science* to *web search*

- ▶ **(1960-70s)**: Initially restricted to librarians and government agencies.
- ▶ **Today**: we all rely heavily on it (web search!).

The Web

- ▶ **Web 1.0 (1990s)**: Static, read-only, informational.
- ▶ **Web 2.0 (2000s)**: Interactive, social, participatory.
- ▶ **Web 3.0 (2010s)**: Decentralized, data-driven.
- ▶ **(2020s)**: AI-augmented?

Web Search as Comprehensive Computing

Algorithms, data structures, computer architecture, networking, logic, discrete mathematics, interface design, user modelling, databases, software engineering, programming languages, multimedia technology, image and sound processing, data mining, artificial intelligence, ...

- ▶ **Challenge:** search billions of pages and return useful results within tenths of a second.

IR vs Databases

Database Queries

- ▶ Require both data tuples and schema.
- ▶ To retrieve: need to know both the item and its location.

Information Retrieval

- ▶ We may not know:
 - ▶ Where the information is.
 - ▶ Whether it exists.
 - ▶ The exact query to express it.
- ▶ Queries may be vague or ambiguous.
- ▶ “Too literal” answers can be unhelpful.

User Expectations

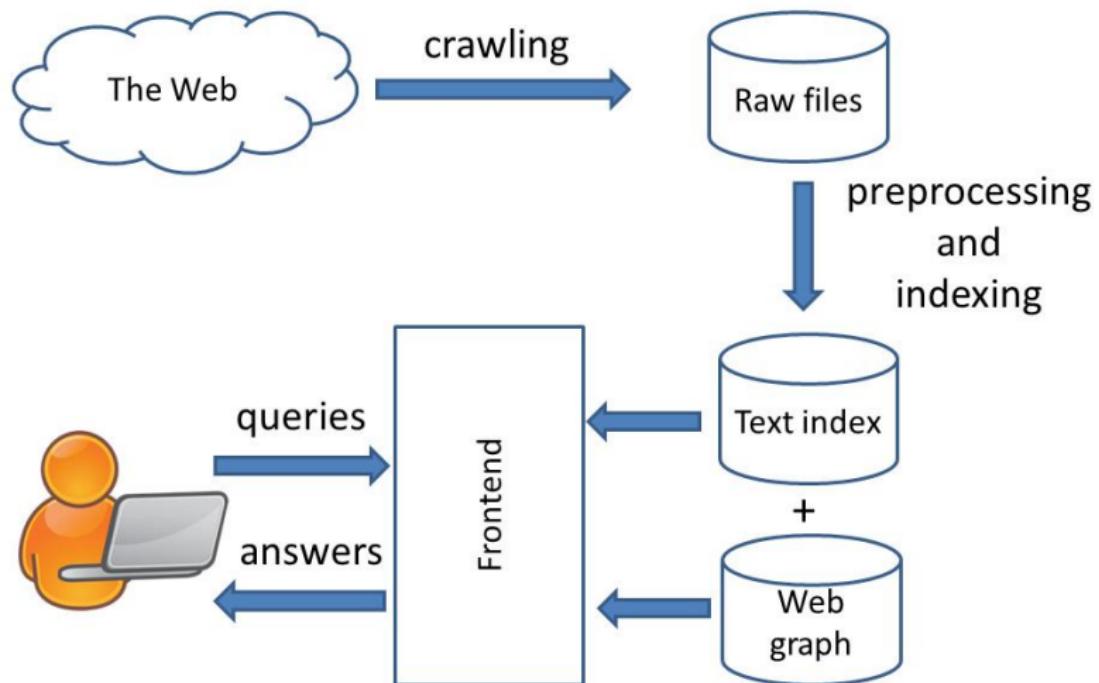
- ▶ Often we don't know exactly what we want.
- ▶ Retrieval systems rely on large document collections.
- ▶ **Assessing relevance is complex!**
 - ▶ Heuristics required.
 - ▶ No strict "keys"—only keywords.
- ▶ **Large** collections allow better answers however present its own challenges in management and search – we are in the era of *Big Data*

The IR process

Focus for first half of the course

- ▶ Retrieving (hyper)text documents from the web
 - ▶ Hypertext documents contain **terms** and **links**
 - ▶ Users issue **queries** to look for documents
 - ▶ Queries typically formed by terms as well

The IR process, cont.



The IR process, cont.

Offline

- ▶ Crawling
- ▶ **Preprocessing** (next!)
- ▶ Indexing

Goal: prepare compact data structures for *fast* online queries.

Online

- ▶ Query input
- ▶ Retrieve candidate documents
- ▶ Rank results
- ▶ Present answer (may depend on context, e.g., location, ads).

Preprocessing

Term Extraction

- ▶ **Parsing** (HTML, XML).
- ▶ **Tokenization** (split character sequences into *tokens* “words”).
- ▶ **Enriching** (metadata, POS tags, synonyms).
- ▶ **Lemmatization** or **stemming** (normalize *tokens*).

Tokenization

Join consecutive characters into “words”: use spaces and punctuation to mark their borders.

- ▶ Special tokens: IPs, emails, URLs.
- ▶ Hyphens: *afro-american* vs *state-of-the-art*.
- ▶ Case sensitivity: “Windows” vs “windows”.
- ▶ Named Entity Recognition (NER).

Stopword Removal

- ▶ Remove frequent, low-value words:
 - ▶ Articles, prepositions, common verbs.
- ▶ Reduces index size by up to 40%.
- ▶ But: sometimes words like *may*, *will* are meaningful.
- ▶ Current trend: keep all words, filter later by relevance.

Case folding

Move everything into lower case, so search is case-independent

Enriching

- ▶ Add semantic or structural info:
 - ▶ Synonyms (*gun* → *weapon*).
 - ▶ Related words (*laptop* → *portable computer*).
 - ▶ Categories (*fencing* → *sport*).
 - ▶ POS tagging.
- ▶ Step further: Word Sense Disambiguation (WSD).

Lemmatization vs Stemming

- ▶ **Stemming:** cut suffixes (fast, crude).
- ▶ **Lemmatization:** reduce to linguistic root (accurate, slower).
- ▶ Examples:
 - ▶ *swim, swimming, swimmer* → *swim* (stemming).
 - ▶ *are, is, am* → *be* (lemmatization).

Text Laws

Word Frequencies

- ▶ Some words very frequent, others rare.
- ▶ Fundamental questions:
 - ▶ How many unique words appear?
 - ▶ How much more frequent are common words?
- ▶ Natural language follows precise **statistical laws**.

Heavy-tailed Distributions

In many natural and artificial phenomena, the probability distribution “decreases slowly” compared to Gaussians or exponentials.

- ▶ Rare items still matter:
 - ▶ Zipf’s law for texts.
 - ▶ Popularity of websites.
 - ▶ Wealth distribution.
 - ▶ Earthquake intensity.

Zipf-Mandelbrot Law

- ▶ Frequency of i -th most common word:

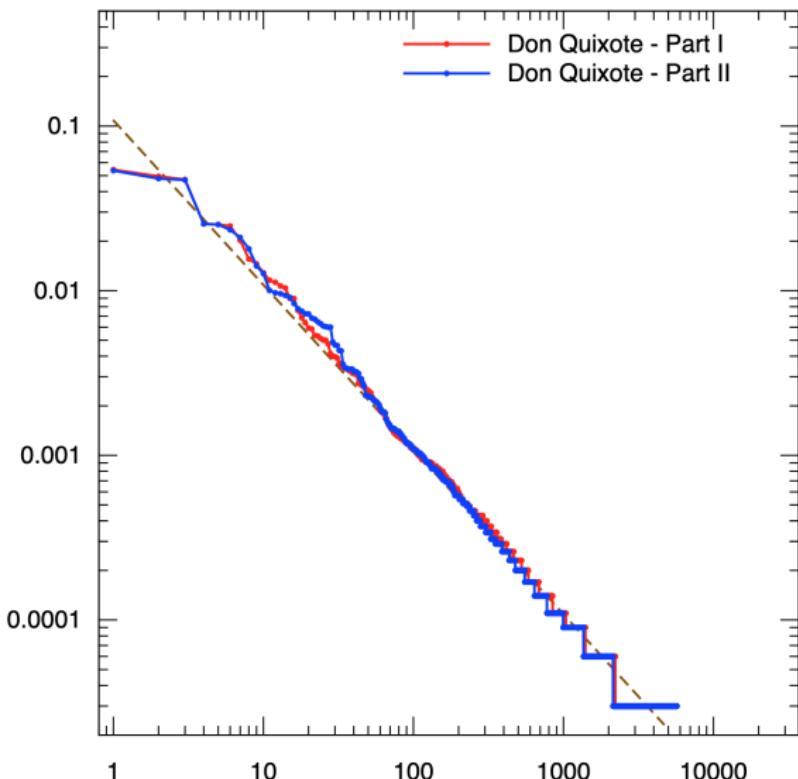
$$f(i) \approx \frac{c}{(i + b)^a}$$

- ▶ Special case (Zipf's law, $b = 0, a = 1$): $f(i) \propto \frac{1}{i}$.
- ▶ Empirical estimates based on natural text: exponent a varies around 1.

Detecting Power Laws

- ▶ Rank words by frequency.
- ▶ Plot rank vs frequency (log-log scale).
- ▶ Straight line consistent with power law.

Example: Don Quijote shows Zipf's law clearly.



Vocabulary Growth

- ▶ Longer texts → larger lexicon.
- ▶ But growth slows down.
- ▶ *Don Quijote*:
 - ▶ First 2,500 words → ~1,100 unique words.
 - ▶ Entire book (383k words) → <40k unique words.

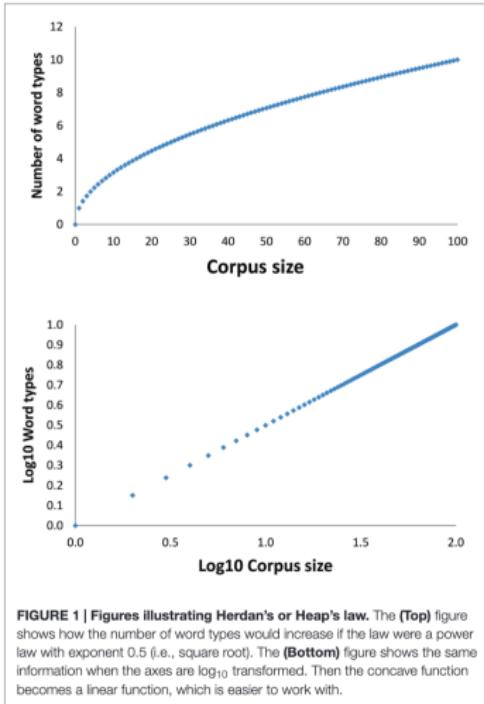
Heaps' (Herdan's) Law

- ▶ Number of distinct words grows polynomially with text size:

$$d = k \cdot N^\beta$$

- ▶ β between 0.5–0.8 depending on language.
- ▶ For Don Quijote: $\beta \approx 0.806$.

Example: Don Quijote



Harmonic Series

- ▶ Appears in text statistics.
- ▶ Harmonic series diverges:

$$\sum_{i=1}^N \frac{1}{i} \approx \ln(N) + \gamma$$

- ▶ But generalized series converge for $\alpha > 1$:

$$\sum_{i=1}^{\infty} \frac{1}{i^\alpha} = \zeta(\alpha)$$

Conclusion

- ▶ Information Retrieval requires preprocessing and statistical insight.
- ▶ Texts obey robust empirical laws (Zipf, Heaps).
- ▶ Effective IR combines:
 - ▶ Linguistic processing,
 - ▶ Efficient data structures,
 - ▶ Mathematical modeling of language.

Harmonic Series

What is the Harmonic Series?

- ▶ Defined as the infinite series:

$$H_N = \sum_{i=1}^N \frac{1}{i}$$

- ▶ Example (first few terms):

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \dots$$

- ▶ Appears in:

- ▶ Algorithm analysis (e.g., average-case QuickSort, hashing).
- ▶ Probability and statistics.
- ▶ Information retrieval and text laws.
- ▶ Number theory.

Divergence of the Harmonic Series

- ▶ Even though the terms get smaller, the series **diverges**.
- ▶ Intuition:
 - ▶ Group terms:

$$1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \cdots + \frac{1}{8}\right) + \cdots$$

- ▶ Each group $\geq 1/2$
- ▶ Therefore, the sum grows without bound as $N \rightarrow \infty$.

Growth Rate

- ▶ Divergence is **very slow**.
- ▶ For large N :

$$H_N \approx \ln(N) + \gamma$$

where:

- ▶ $\ln(N)$ = natural logarithm.
- ▶ $\gamma \approx 0.5772\dots$ is **Euler's constant**.

Euler–Mascheroni Constant γ

- ▶ Defined as the limiting difference between harmonic numbers and the natural log:

$$\gamma = \lim_{N \rightarrow \infty} (H_N - \ln(N))$$

- ▶ Approximate value:
 $\gamma \approx 0.5772156649\dots$
- ▶ Still mysterious:
 - ▶ Unknown if γ is rational or irrational.
 - ▶ Appears across number theory, analysis, and algorithms.

Example Values

- ▶ $H_{10} = 2.928968\dots$
- ▶ $H_{100} = 5.187377\dots$
- ▶ $H_{1000} = 7.485471\dots$
- ▶ Compare with $\ln(N) + \gamma$: very close!

$$H_{1000} - \ln(1000) \approx 0.5772$$

Generalized Harmonic Series

Definition

- ▶ The **generalized harmonic series** of order α is:

$$H_{N,\alpha} = \sum_{i=1}^N \frac{1}{i^\alpha}$$

- ▶ Special cases:
 - ▶ $H_{N,1} = H_N$ (the ordinary harmonic series).
 - ▶ $H_{N,2} = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \dots$

Convergence Behavior

- ▶ Depends on α :

$$\sum_{i=1}^{\infty} \frac{1}{i^\alpha}$$

- ▶ If $\alpha \leq 1$: diverges.
- ▶ If $\alpha > 1$: converges (finite sum).

Connection to Riemann Zeta Function

- As $N \rightarrow \infty$, the series tends to the **Riemann zeta function**:

$$\zeta(\alpha) = \sum_{i=1}^{\infty} \frac{1}{i^\alpha}$$

- Examples:

- $\zeta(2) = \frac{\pi^2}{6} \approx 1.6449$
- $\zeta(3) \approx 1.2021$ (Apéry's constant)
- $\zeta(4) = \frac{\pi^4}{90} \approx 1.0823$