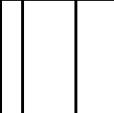


**Université**



de Strasbourg

# Recherche d'information

## Introduction

**Enzo Doyen**

# ⓘ Informations générales

- 6 séances en salle 4S04 le lundi de **15 h** à **18 h** ; 03/11, 10/11, 17/11, 24/11, 01/12 et 08/12.
- Nécessaire d'avoir un compte **Google Colab** ainsi qu'un compte **Hugging Face** pour l'évaluation et les notebooks d'exemple.

Page Moodle : <https://moodle.unistra.fr/course/view.php?id=9754>

Mot de passe : **RECHINF43**

# Objectifs

1. Comprendre les concepts fondamentaux de la recherche d'information (RI).
2. Comprendre la conception des différents systèmes de RI (recherche booléenne, vectorielle, probabiliste).
3. Implémenter des systèmes de RI et mener une évaluation qualitative de ceux-ci.
4. Créer une interface permettant une utilisation simplifiée des systèmes de RI.



# Évaluation

L'évaluation du cours se fait sur la base de deux exercices à rendre.

Date limite du 1<sup>er</sup> exercice : 01/12/2025 (23h59).

Date limite du 2<sup>e</sup> exercice : 28/12/2025 (23h59).



# Plan

- I. Recherche d'information : définition et développement
- II. Principes fondamentaux de la recherche d'information



# I. Recherche d'information : définition et développement

# Définition de la RI

## 💡 Définition :

"[...] finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)" (**Manning et al., 2008**)

# Définition de la RI

## 💡 Définition :

"[...] finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)" (**Manning et al., 2008**)

Dans ce cours, nous nous concentrerons sur la **recherche de documents textuels**.

# Exemple de RI



# **Exemples d'application de la RI**

- recherche Web (Google, Bing, etc.) ;
- recherche d'e-mails ;
- recherche de documents dans des bases de données ;
- ...

# Exemples d'application de la RI

The screenshot shows the homepage of the Université de Strasbourg library website. At the top left is a logo with three red icons: a book, a magnifying glass, and a document. To its right, the text "Les bibliothèques" and "Université de Strasbourg" is displayed. On the far right, there are links for "CONTACT", "FAQ", and "UNE PROPOSITION ?". Below this is a navigation bar with five main categories: "TROUVER UN DOCUMENT", "DÉCOUVRIR LES BIBLIOTHÈQUES", "UTILISER NOS SERVICES", "NOUS CONNAÎTRE", and "MON COMPTE". The "TROUVER UN DOCUMENT" section is currently active, indicated by a red background. Within this section, there are three tabs: "Catalogue", "Revues", and "Bases en ligne", with "Bases en ligne" being the selected tab. A search bar below these tabs contains the placeholder text "Recherche de bases de données et bases bibliographiques en ligne" and a magnifying glass icon. To the right of the search bar are links for "Recherche avancée" and "Historique des recherches". At the bottom of the page, there are links for "Nouveautés", "Liste A-Z des revues en ligne", "Numistral", "Thèses et mémoires en ligne", "univOAK", and "Recherche Data Gouv". The background features a faint network-like graphic.

Exemple de recherche d'information sur le site de la BU

# Développement de la RI

« Consider a future device ... in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory. »

(Bush, 1945)

- Appareil hypothétique (« Memex », pour « memory expansion ») imaginé par Vannevar Bush en 1945.
- Donnera plus tard naissance aux « personal knowledge bases » (Evernote, Obsidian...).

# Développement de la RI

- En 1950, terme « information retrieval » utilisé pour la première fois par Calvin Mooers.
- Système de recherche à l'aide de cartes perforées développé par IBM en 1952 (décrit par **Luhn (1958)**).
  - Une des premières implémentations du **modèle booléen**.

# Développement de la RI

- En 1976, implémentation d'un **modèle probabiliste** par **Robertson et Jones (1976)**.
- Application de « poids » aux termes en fonction de leur distribution dans les documents afin d'évaluer leur pertinence.

# Développement de la RI

- En 1992, lancement de la première édition de la *Text Retrieval Conference* (TREC), qui incite à davantage de travaux de recherche dans le domaine.
- À partir de 1994, développement d'Internet pour le grand public et des moteurs de recherche (Lycos, Yahoo! Search, AltaVista...).
  - En 1998, naissance de Google, qui propose l'algorithme PageRank pour indexer et classer les pages Web en fonction de leur popularité et de leur pertinence (**Brin et Page, 1998**).

# Développement de la RI

- Depuis les années 2010, utilisation des avancements en *machine learning* et en traitement automatique des langues pour améliorer les systèmes de RI.
- Recherches contextuelles et personnalisées, suggestions de recherche...
- Représentation des requêtes de recherche et des documents sous forme vectorielle.

# Recherche d'information et *big data*

- Phénomène de *big data* : explosion des données disponibles, notamment sur le Web.
- Comment récupérer les données les plus pertinentes dans un contexte où la quantité de données devient de plus en plus importante ?

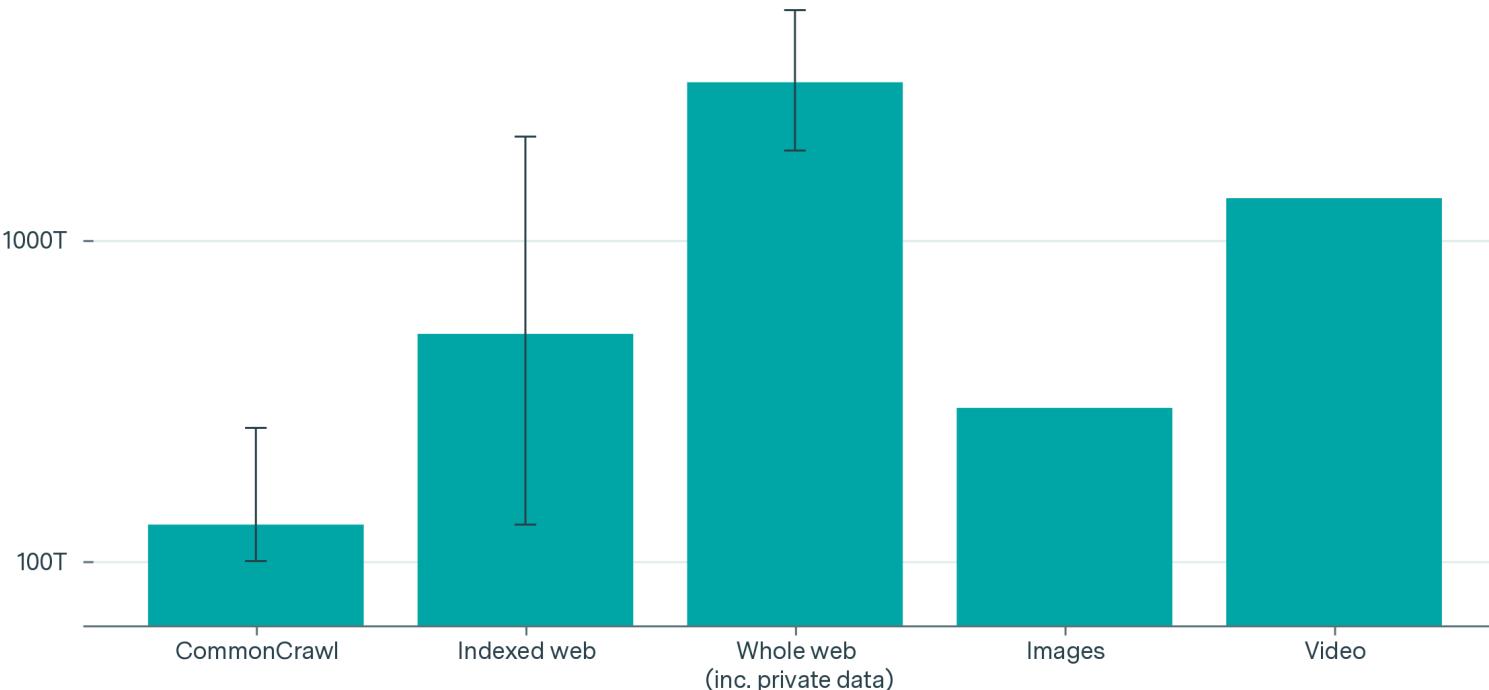
## Estimates of different stocks of data

Effective stock (number of tokens)

1000T -

1000T -

100T -



Estimation des données sur le Web | Source : <https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data>

# Recherche d'information et *big data*

- Phénomène de *big data* : explosion des données disponibles, notamment sur le Web.
- Comment récupérer les données les plus pertinentes dans un contexte où la quantité de données devient de plus en plus importante ?
- Plus récemment, l'émergence de grands modèles de langue génératifs (LLM) a conduit à une publication de masse de contenus générés par IA sur Internet.
  - Elle pose aussi des questions sur l'**utilisation future des moteurs de recherche**.

# Recherche Web à l'ère des LLM

ARTIFICIAL INTELLIGENCE

## How AI-generated text is poisoning the internet

Source : <https://www.technologyreview.com/2022/12/20/1065667/how-ai-generated-text-is-poisoning-the-internet/>

# Recherche Web à l'ère des LLM

**AI means the end of internet search  
as we've known it**

Despite fewer clicks, copyright fights, and sometimes iffy answers, AI could unlock new ways to summon all the world's knowledge.

By Mat Honan

January 6, 2025

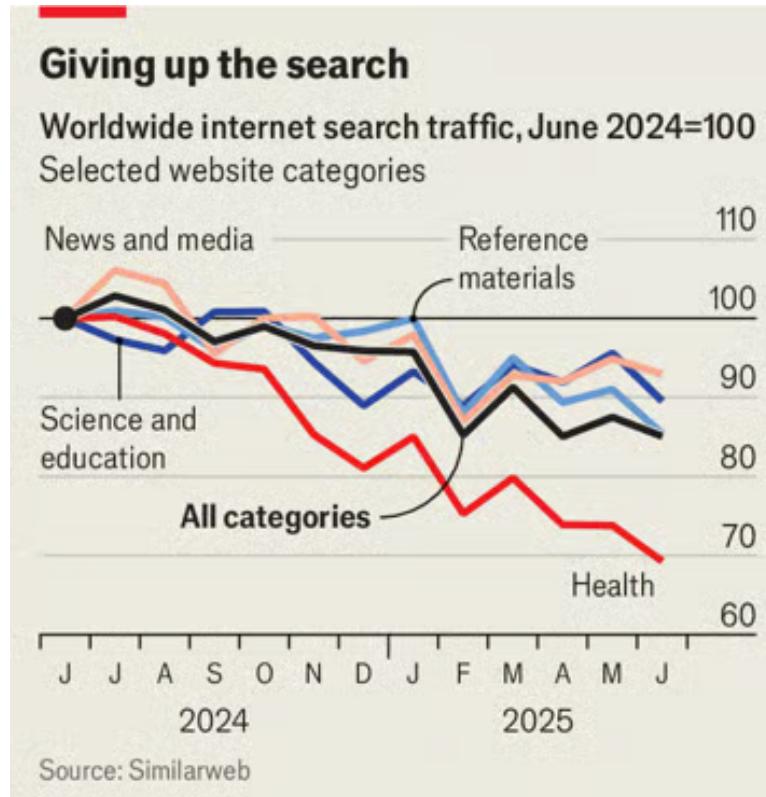
Source : <https://www.technologyreview.com/2025/01/06/1108679/ai-generative-search-internet-breakthroughs/>

# Recherche Web à l'ère des LLM

« The biggest change to the way search engines have delivered information to us since the 1990s is happening right now. No more keyword searching. No more sorting through links to click. Instead, we're **entering an era of conversational search**. Which means **instead of keywords, you use real questions, expressed in natural language**. And **instead of links, you'll increasingly be met with answers, written by generative AI and based on live information from all across the internet**, delivered the same way. »

Mat Honan, AI means the end of internet search as we've known it, *MIT Technology Review*,  
<https://www.technologyreview.com/2025/01/06/1108679/ai-generative-search-internet-breakthroughs/>

# Recherche Web à l'ère des LLM



Source : <https://www.economist.com/business/2025/07/14/ai-is-killing-the-web-can-anything-save-it>

# Recherche Web à l'ère des LLM : intégration

The screenshot shows a Google search results page for the query "explain the connection between lightning and thunder". The interface includes the Google logo, a search bar with the query, and standard navigation links like All, Images, Videos, News, Shopping, More, and Tools. A message indicates "About 88,600,000 results (0.46 seconds)". An "AI Overview" section is present, with a "Learn more" link. Below it are three interactive cards: one from the UCAR Center for Science Education and another from Wikipedia. The main content area starts with a summary: "Lightning is the most spectacular part of a thunderstorm, and it's also how thunderstorms get their name. Lightning is what causes thunder." This is followed by a section titled "Here's how lightning and thunder are connected:" with a numbered list of three points. The first point details how lightning heats the air rapidly, reaching temperatures up to 50,000°F. The second point discusses the rapid expansion of air. The third point explains how this expansion creates a sonic shock wave, resulting in thunder.

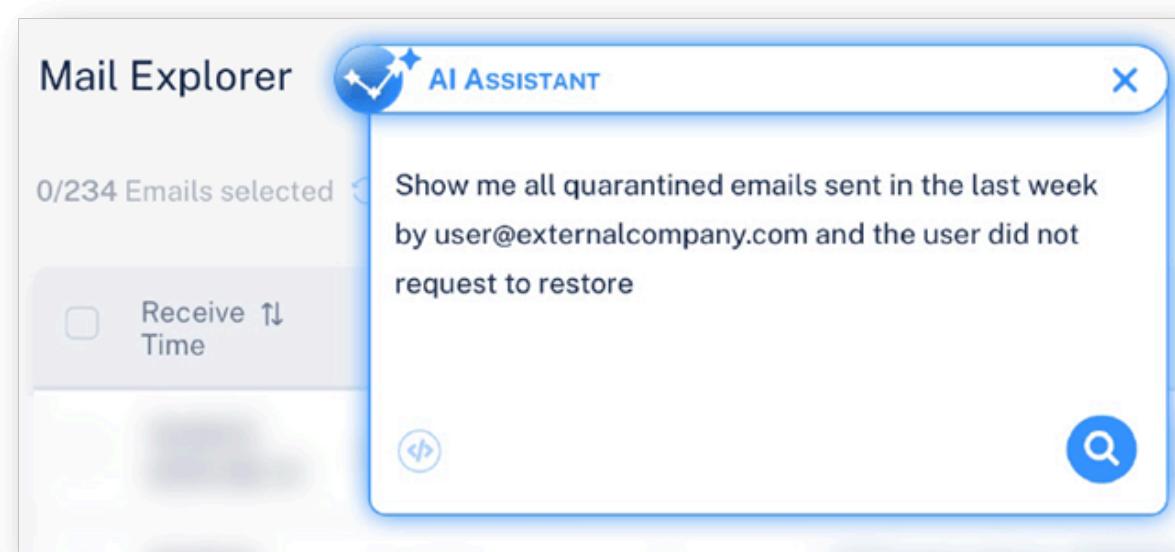
Here's how lightning and thunder are connected:

1. Lightning heats the air rapidly as it passes through it. The air temperature in the lightning channel can reach 50,000°F, which is five times hotter than the surface of the sun.
2. The sudden increase in temperature and pressure causes the air to expand rapidly.
3. The expansion of air creates a sonic shock wave, which is what we hear as thunder. ▾

The thunderclap or peal of thunder can range from a long, low rumble to a sudden, loud crack. Thunder can usually be heard from about 10 miles away from a lightning strike. ▾

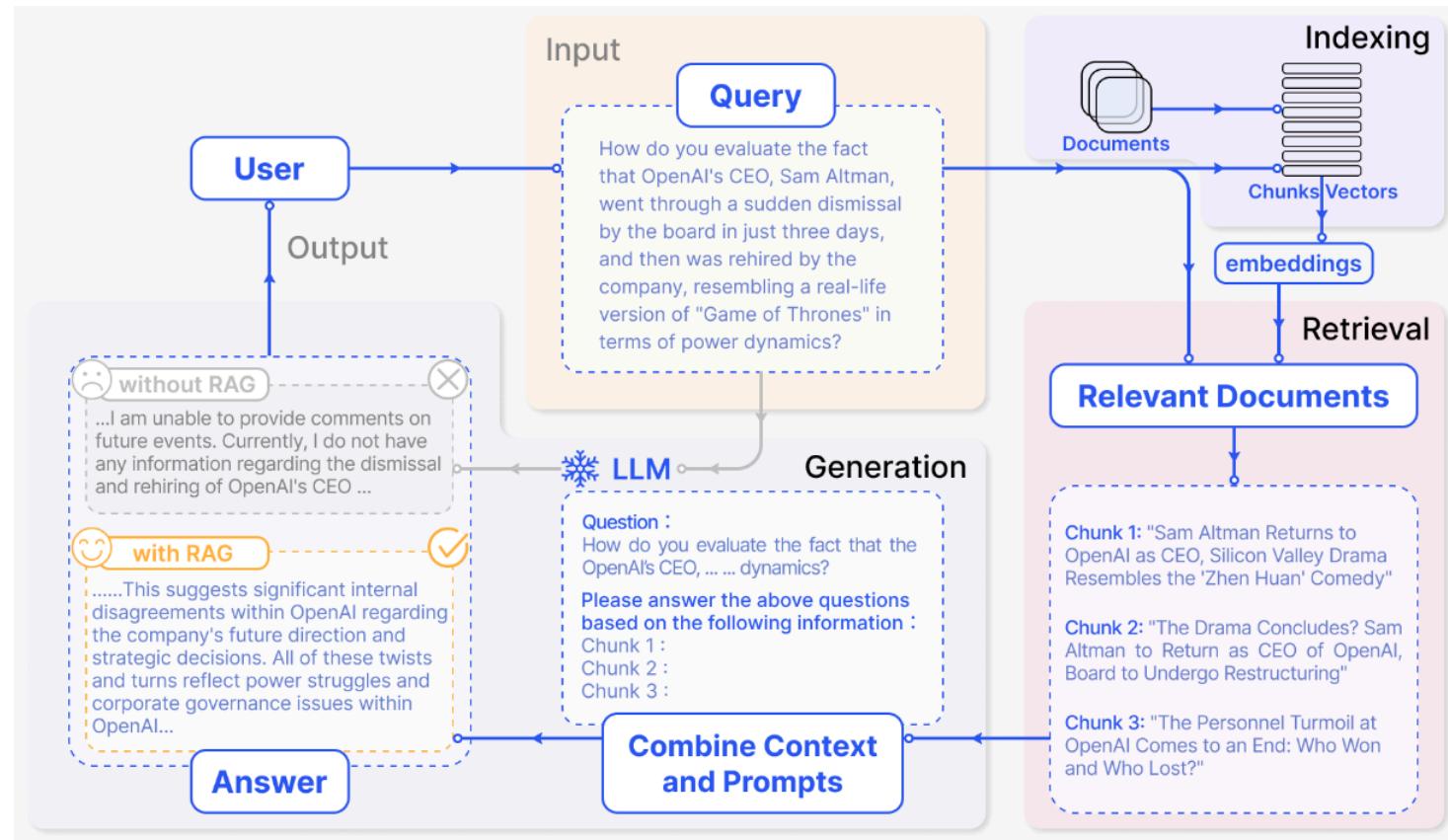
Source : <https://blog.google/products/search/generative-ai-google-search-may-2024/>

# Recherche d'e-mails à l'ère des LLM



Source : <https://emailsecurity.checkpoint.com/product-updates/ai-assistant-email-search-powered-by-generative-ai>

# Recherche d'information pour les LLM : RAG



Source : Gao et al. (2024)

The screenshot shows a digital notebook interface with a light green background. At the top left is a back arrow and the title "CS History". On the right are icons for search, refresh, and user profile. A sidebar on the left lists categories: "CS 106 class notes" (highlighted in green), "Mark I/UNIVAC history", "Lovelace excerpt", "Invention of the GUI", "Babbage intro", and a "+ New" button. At the bottom of the sidebar is an "Open notes" button. The main content area has a white background. It features a section titled "CS 106 class notes" with a sub-section "Week Two: Ada Lovelace". Below this is a text block about Ada Lovelace's life and work on the Analytical Engine. Another text block discusses her contributions to computing, mentioning her translation of an article from French to English. A third text block talks about her vision for the Analytical Engine's potential beyond calculation. On the right side of the main content area is a blue sidebar titled "Document guide" with a yellow star icon. It contains text about women in computing, specifically Ada Lovelace, Grace Hopper, and Katherine Johnson. Below this is a "Key topics" section with buttons for "STEM and Diversity" and "Alan Kay's Dynabook", followed by a button for "Katherine Johnson at NASA". At the bottom of the sidebar are two questions: "How did Grace Hopper's work on the Mark I contribute to modern computing?" and "What challenges did Johnson face as an African American woman at NASA?". At the very bottom of the sidebar are buttons for "1 Source", "Edit", and a blue upward arrow.

← CS History

CS 106 class notes

Mark I/UNIVAC history

Lovelace excerpt

Invention of the GUI

Babbage intro

+ New

Open notes

## CS 106 class notes

### Week Two: Ada Lovelace

Ada Lovelace was a British mathematician and writer who is considered to be the first computer programmer. She was born in 1815 to the poet Lord Byron and his wife Annabella Milbanke.

Showed an early aptitude for mathematics, and she was tutored by some of the leading mathematicians of her day. In 1833, she met Charles Babbage, a mathematician and inventor who was working on a machine that he called the Analytical Engine. The Analytical Engine was designed to be a programmable computer.

Contributions to Computing:

Ada Lovelace made several important contributions to the development of computing. She translated an article about the Analytical Engine from French to English. In her translation, she added a series of notes that outlined how the Analytical Engine could be used to solve mathematical problems. These notes are considered to be the first example of computer programming.

Ada Lovelace also wrote about the potential of the Analytical Engine to be used for creative purposes, such as composing music. She believed that the Analytical Engine would have a profound impact on society; one of the first people to envision the potential of computers to be used for more than just calculation.

### Document guide

Class notes containing information about women in computing, including the British mathematician Ada Lovelace, the computer scientist Grace Hopper, and the NASA mathematician Katherine Johnson. These women were all pioneers in the field of computing, and their work has helped to pave the way for women in STEM fields.

### Key topics

STEM and Diversity   Alan Kay's Dynabook  
Katherine Johnson at NASA

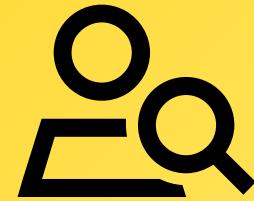
### Questions to try

How did Grace Hopper's work on the Mark I contribute to modern computing?

What challenges did Johnson face as an African American woman at NASA?

1 Source   Edit ↑

Source : <https://blog.google/technology/ai/notebooklm-google-ai/>



## **II. Principes fondamentaux de la recherche d'information**

# RI : l'utilisateur ou l'utilisatrice avant tout

- La recherche d'information est centrée sur l'utilisateur ou l'utilisatrice.
- L'objectif est de répondre à un **besoin d'information** spécifique.
- Il faut alors penser à :
  - qui est le public visé (qui va utiliser le système de RI ?) ;
  - quel est le type de documents à rechercher et le domaine ;
  - comment il/elle est susceptible de formuler sa requête ;
  - quels outils de recherche mettre à sa disposition (système de recherche avancée ? si oui, quelles options ?) ;
  - comment lui proposer les résultats les plus pertinents possibles...
    - ◆ ... quand bien même la requête n'est pas forcément formulée correctement, ou de manière exacte.

# RI : pertinence et évaluation

- Les documents indexés sont-ils pertinents par rapport à la requête ? Répondent-ils au besoin d'information exprimé ?
  - Comment évaluer cette pertinence ?
- Les documents retournés sont-ils à jour ? Fiables ?
  - Flux d'informations en continu, *fake news...*

# RI : pertinence et évaluation

La **précision** et le **rappel** sont deux mesures couramment utilisées pour évaluer la pertinence des résultats retournés par un système de RI.

	Pertinent	Non pertinent
Récupéré	TP	FP
Non récupéré	FN	TN

Tableau de contingence (adapté de **Manning et al. (2008:155)**)

$$\text{Précision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Rappel} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

# Étapes et processus d'un système de RI

1. Prétraitements des documents (nettoyage, normalisation...).
2. Indexation des documents :
  - représentation des documents sous un format exploitable par le système de RI.
3. Prétraitements de la requête.
4. Appariement de la requête avec les documents indexés :
  - recherche des documents pertinents en fonction de la requête.
5. Classement des documents pertinents.



Étapes d'un système de recherche d'information (**Amini et al., 2013**)



# Bibliographie

- Amini, M.-R., Gaussier, É., et Péan, G. (2013). *Recherche d'information : applications, modèles et algorithmes*. Eyrolles.
- Brin, S., et Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks*, 30, 107-117.
- Bush, V. (juillet 1945). As We May Think. *The Atlantic*.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., et Wang, H. (mars 2024). *Retrieval-Augmented Generation for Large Language Models: A Survey* (Numéro arXiv:2312.10997). arXiv. [10.48550/arXiv.2312.10997](https://arxiv.org/abs/2312.10997)
- Luhn, H. P. (1958). *The IBM Electronic Information Searching System*. International Business Machines Corp., Research Center.
- Manning, C. D., Raghavan, P., et Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.
- Mitra, B. (2018). *An Introduction to Neural Information Retrieval* (Numéro v.41). Now Publishers.
- Robertson, S. E., et Jones, K. S. (1976). Relevance Weighting of Search Terms. *Journal of the American Society for Information Science*, 27(3), 129-146. [10.1002/asi.4630270302](https://doi.org/10.1002/asi.4630270302)
- Zhai, C., et Massung, S. (juin 2016). *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. Association for Computing Machinery and Morgan & Claypool. [10.1145/2915031](https://doi.org/10.1145/2915031)

# ♥ Remerciements

- Pablo Ruiz Fabo pour le contenu de certaines diapositives.