Text Mining & Bioinformatics

Patrick Ruch
Bibliomics and Text Mining – BiTeM
HEG-HES-SO & SIB
patrick.ruch@hesge.ch

Swiss Institute of Bioinformatics

Hes·so
Haute Ecole Spécialisée de Suisse occidentale
University of Applied Sciences Western Switzerland

Overview

- Introduction and objectives
- Metrics
- Words
- Tasks
- Methodologies
- Text Categorization

Objectives

- Introduce how text mining can support bioinformatics tasks
- Explain how text mining operate with biological entities and the « biological » ecosystem
- Stimulate your interest into a satellite - yet very lifeful - bioscience field

Text Data Mining

- Text Mining is like Data Mining but works with textual contents
- … So any statistical analysis can be performed with text mining provided the content is available in text ?
- Answer: **Jein !**

Disciplines

- Natural language processing, computational linguistics (+)
- Machine learning / data mining (++)
- **Information retrieval (+++)**

Common application fields

- Information retrieval
- Biocuration support tools → tools to maintain KB
- Biological modelling, e.g. biotic interactions

User level tasks

- Search – Foundations
- Triage (i.e., binary classification)
- Keyword assignment (i.e., multi-class classification)
- (Named-)Entity recognition
- Extract passages or more complex entities (e.g. protein protein interactions)

User level tasks

- Summarization
- Retrieval-augmented summarization/generation
- ChatBots

Metrics

- Precision
- Recall
- Other metrics…

Evaluation

- Like most data mining tasks, information retrieval and text mining tasks are assessed using two dimensions metrics

Precision

- Given 5 relevant documents in a collection for a given query, a search engine returns **10** documents, including **3**, which are pertinent

- P = 3/10 = 0.30 or 30%

Recall

- Given **5** relevant documents in a collection for a given query, a search engine returns 10 documents, including **3**, which are pertinent

- Recall = 3/5 = 0.60 or 60%

Precision

- Given 8 relevant documents in a collection for a given query, a search engine returns 10 documents, including 8, which are pertinent

- Please compute the precision ?

Recall

- Given 8 relevant documents in a collection for a given query, a search engine returns 10 documents, including 8, which are pertinent

- Please compute the recall ?

Recall vs. Precision

- Precision is usually regarded as more important because redundancy is (usually) high in large collections...

- Exceptions are numerous
  - Looking for allergies of patients
  - Looking for rare variants
  - Looking for known items
  - [...]

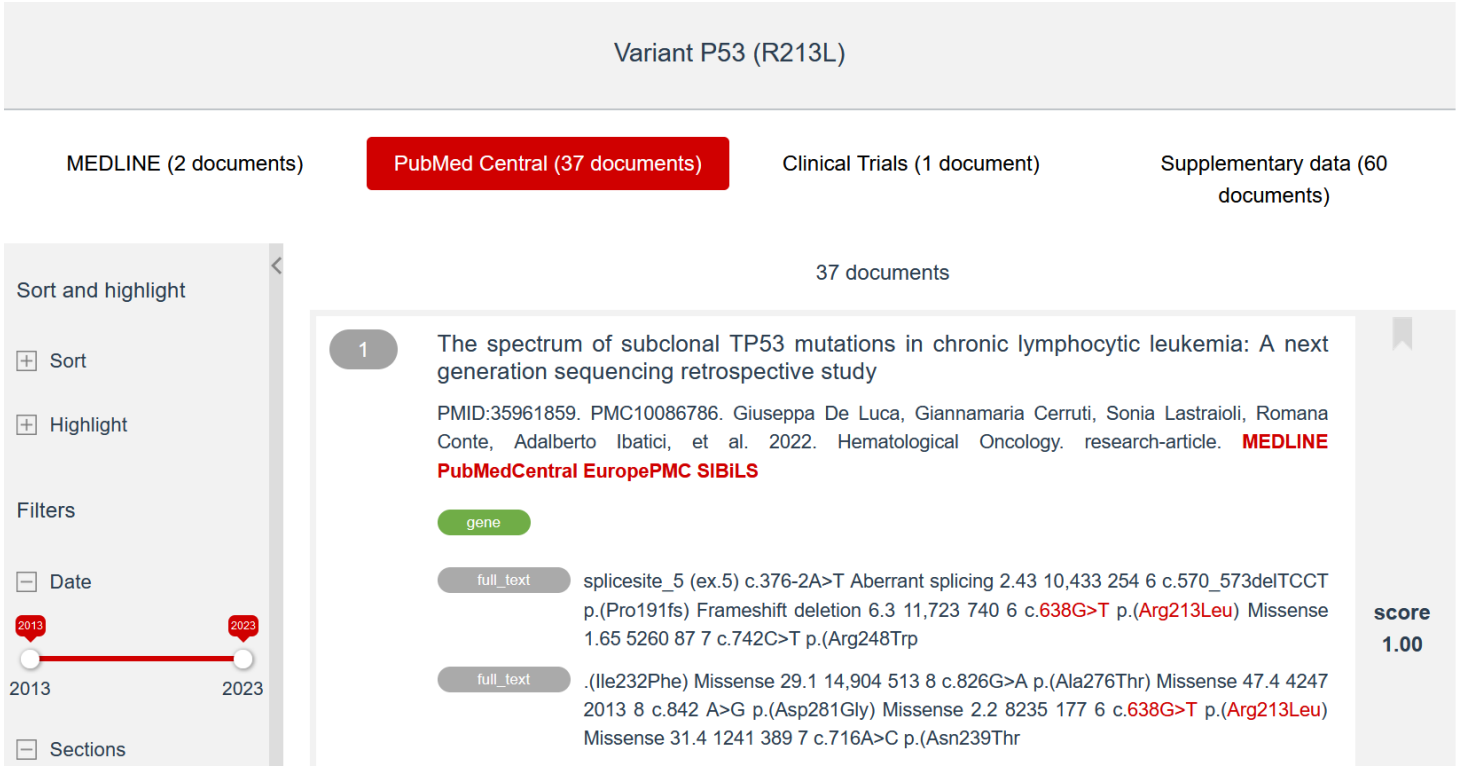
Exemple

- Rare variants

https://variomes.text-analytics.ch/

- SynVar (expansion engine)

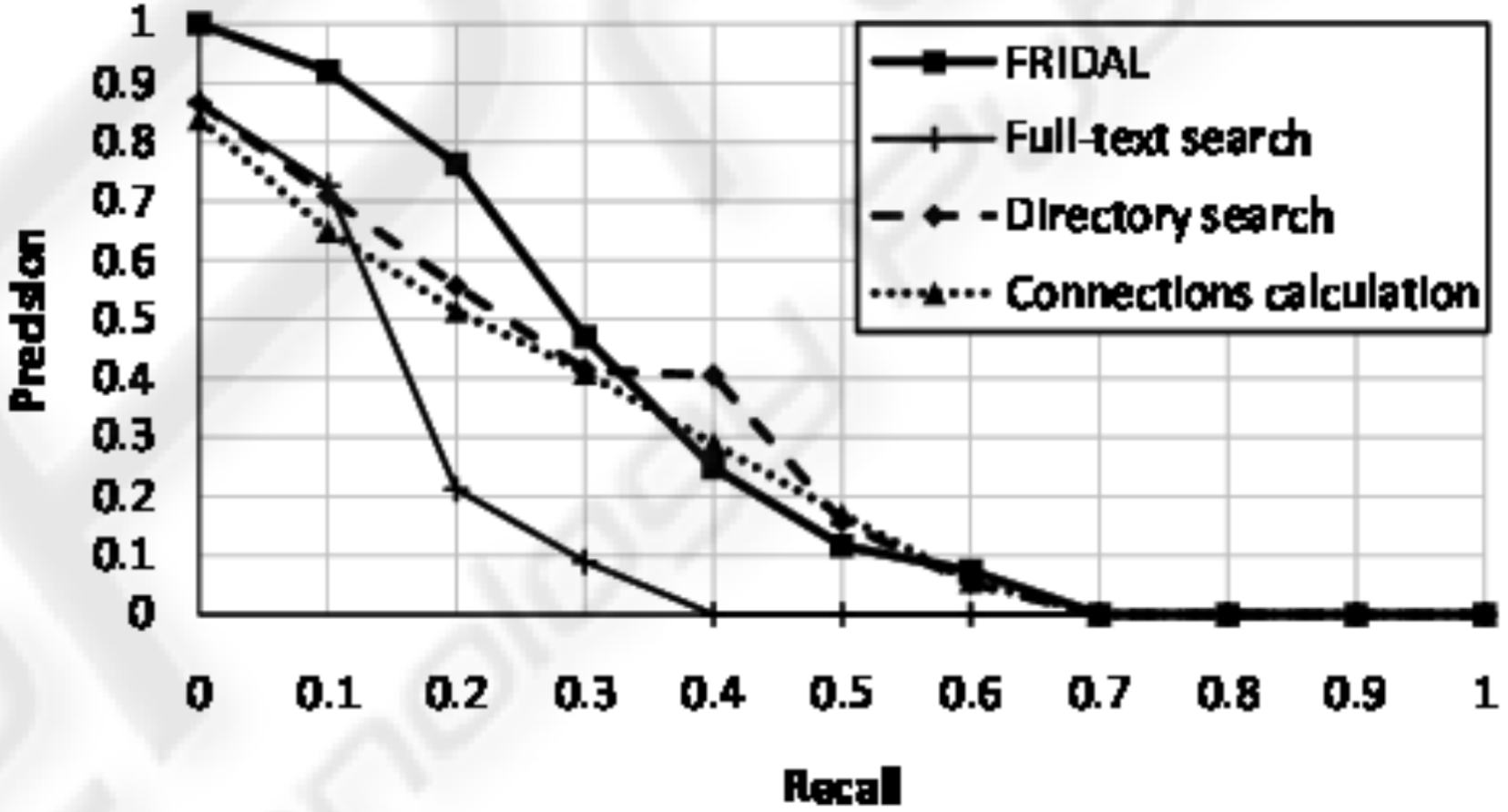
https://goldorak.hesge.ch/synvar



Synthetic metrics

- Rank
  - R^{th-1} is more important than R^{th}
  - So, we compute average precision at different rank values (10, 20, … 30%, …)
  - Mean average precision
- F1 and related metrics
  - Harmonic or geometric mean
  - Utility metrics
    - E.g., 0.9 x Recall + 0.1 x Precision

Example


Precision
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Recall
FRIDAL
Full-text search
Directory search
Connections calculation

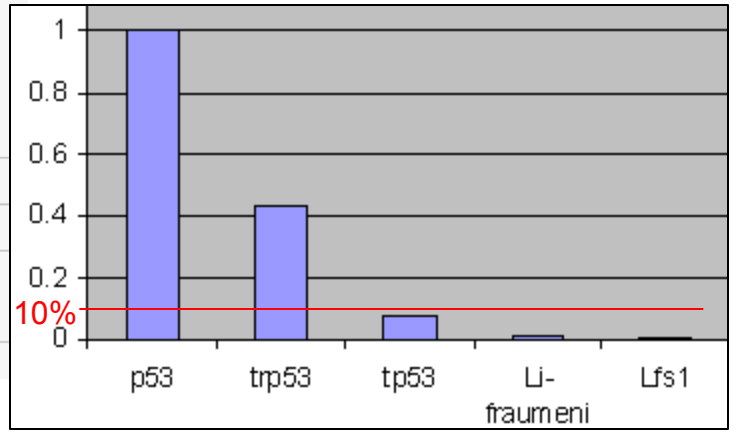
Feature normalization

- Words
- Subwords (character N-grams)
- Stems
- Word N-grams
- Syntactic entities (noun phrases, verb phrases, …),
- Semantic entities (gene names, chem. compounds, diseases, …)

Term normalization: database & ontology vs. reality !

Antigen NY-CO-13	Protein	SwissProt:P04637
Cellular tumor antigen p53	Protein [preferred]	SwissProt:P04637
FLJ92943	Gene	EntrezGene:7157
LFS1	Gene	EntrezGene:7157
		HGNC:11998
Li-Fraumeni syndrome	Gene	HGNC:11998
p53	Gene	EntrezGene:7157
		HGNC:11998
P53	Gene	OMIM:191170
		SwissProt:P04637
p53 antigen	Gene	EntrezGene:7157
p53 transformation suppressor	Gene	EntrezGene:7157
p53 tumor suppressor	Gene	EntrezGene:7157
phosphoprotein p53	Gene	EntrezGene:7157
Phosphoprotein p53	Protein	SwissProt:P04637
TP53	Gene [preferred]	HGNC:11998
		SwissProt:P04637
	Gene	EntrezGene:7157
		OMIM:191170
transformation-related protein 53	Gene	EntrezGene:7157
TRANSFORMATION-RELATED PROTEIN 53	Gene	OMIM:191170
TRP53	Gene	EntrezGene:7157
		OMIM:191170
tumor protein p53	Gene [preferred]	HGNC:11998

Synonyms	#
p53	53362
trp53	23364
tp53	4156
li-fraumeni	775
lfs1	431



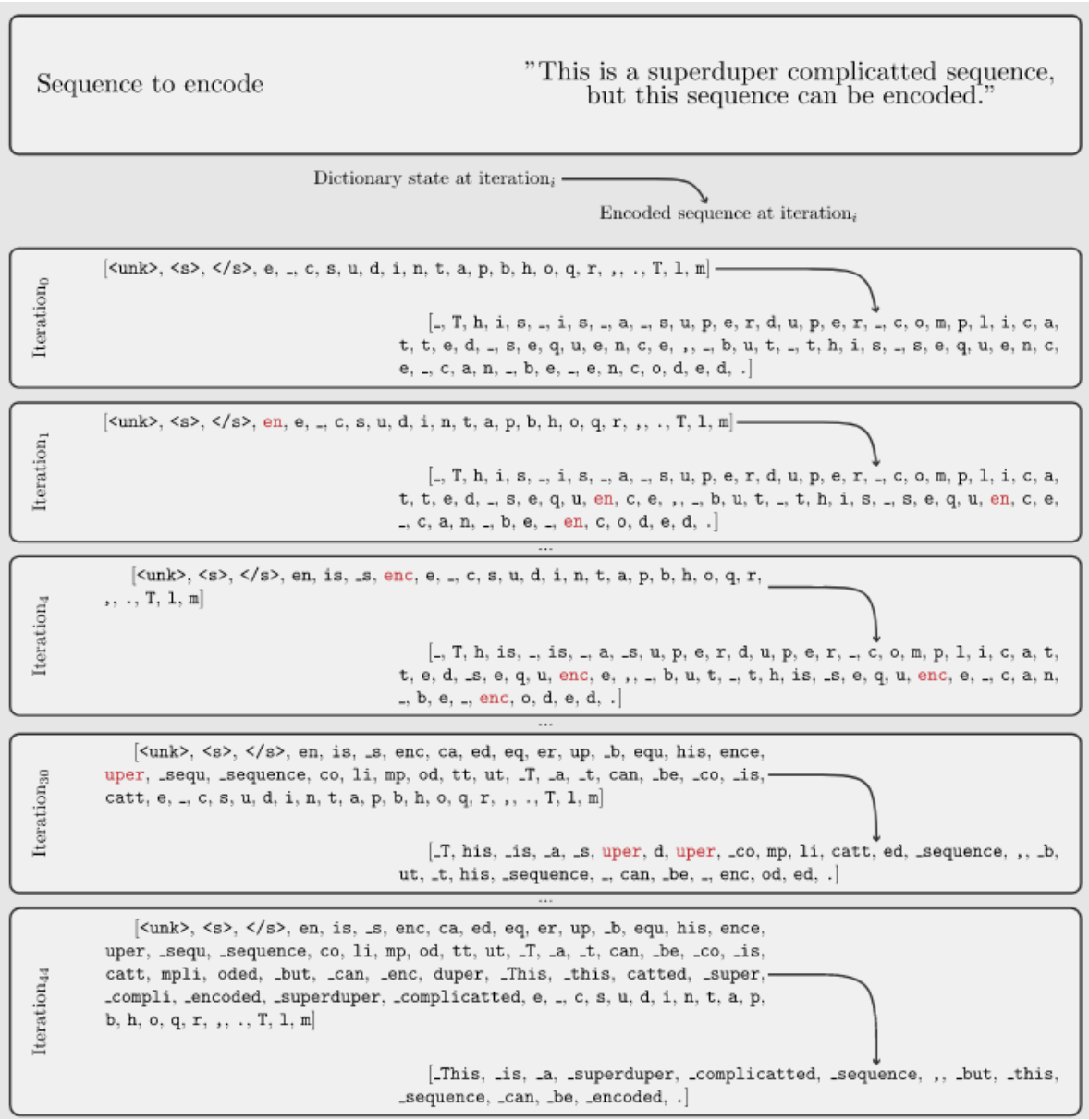
Transformations (ETL, reg. expressions, …)

- i, ii, iii → 1, 2, 3 (e.g. *histone deacetylase iii*)
- Greek letters (e.g *α-tubulin*)
- Hyphenation «-»: {alphatubulin, alpha, tubulin)
- Chemistry
  - Inchi
  - SMILES
  - PubChem, chEBI, DrugBank…

Stemming vs. Lemmatization (needs syntactic analysis)

Original	Stemming	Lemmatization
New	New	New
York	York	York
is	is	be
the	the	the
most	most	most
densely	dens	densely
populated	popul	populated
city	citi	city
in	in	in
the	the	the
United	Unite	United
States	State	States

Byte Pair Encoding (BPE) → Embeddings



Impact of normalization and expansion ?

- Recall

- Precision

Normalization and expansion impact

- Recall

Normalization/expansion improves recall

- Precision

Normalization/expansion degrades precision

→ No free lunch and fine-tuning are needed !

Back end tasks – intelligent stuff that computers can do ☺

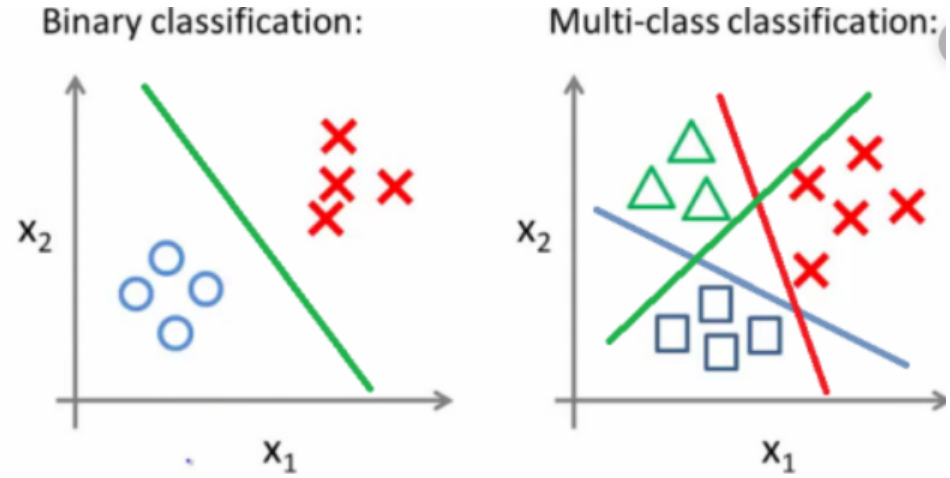
- Ranker
- Classifier
- … and both are the same: a classifier is a ranker with a threshold !
- Regression (ok for structured data)

Ranking and text mining features

- Given an objective function, rank a collection of text !
- Objective function = distance, precision, recall, …

Classification: one-class, binary, ...

- With n binary classifiers, we can design n-class categorizers



Information retrieval – so called « ad hoc »

- PubMed
  - Boolean & Ante-chronological
  - « Best match »
    - Vector-space
    - Learning to rank
- EuropePMC
  - Vector-space (Lucene)
- SIBiLS
  - Vector-space (Lucene)
  - Combination of weighting schema (Terrier)

Boolean vs. Vector-space systems

- Boolean: would return results satisfying the query using AND, OR, NOT operators

- Vector-space is like Boolean but the ranking is based on the differential weighting of:
  - TF: Term frequency

The more frequent a term in a document the stronger the association with that document

  - IDF: Inverse document frequency

The less frequent a term in a document collection the stronger the association with a document it appears in

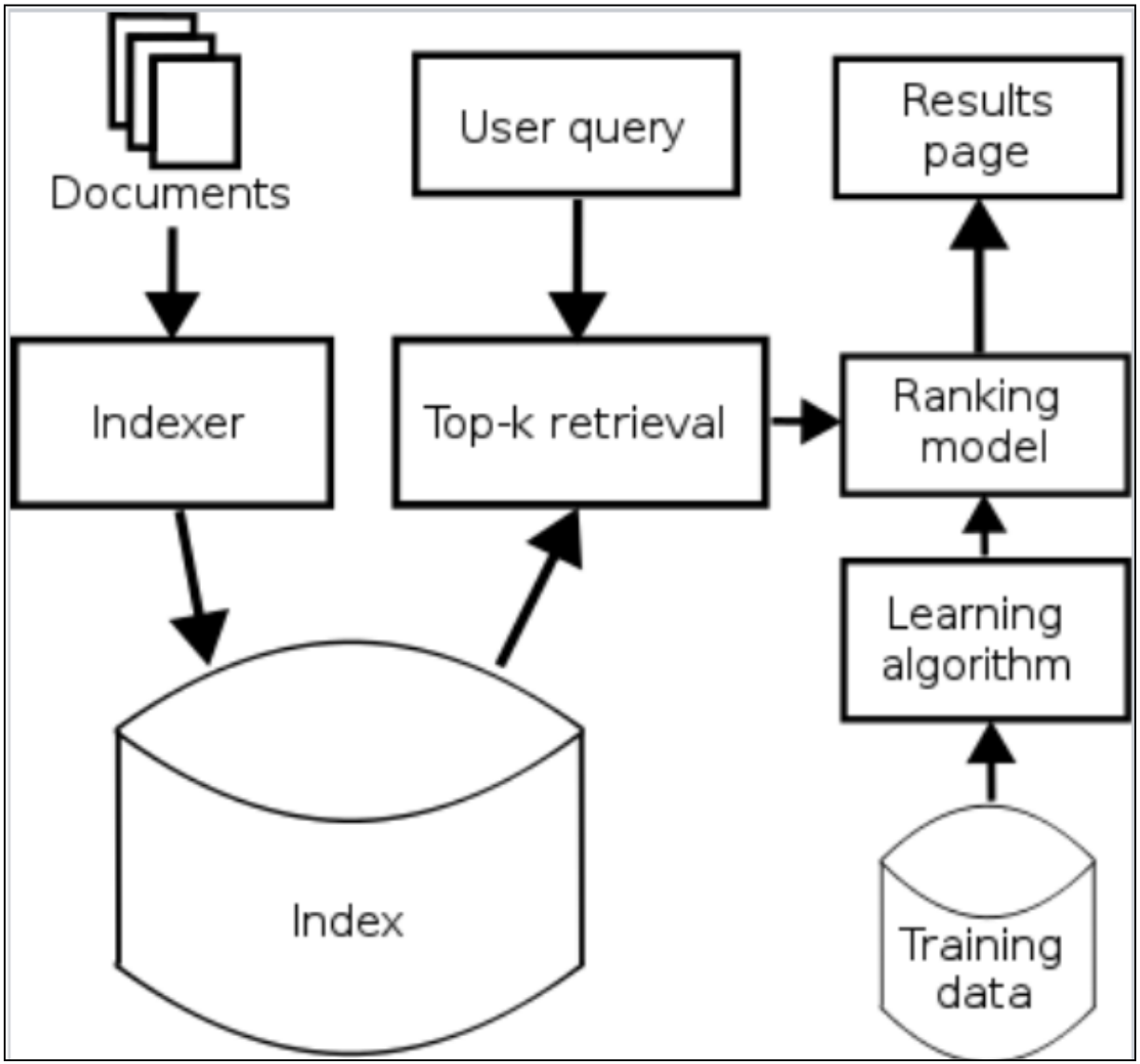
  - Document normalization factor

Longer documents tend to have more words irrespective on the relevance of those words

Weighting schema

Term frequency		Document frequency		Normalization	
n (natural)	$\mathrm{tf}_{t,d}$	n (no)	1	n (none)	1
l (logarithm)	$1 + \log(\mathrm{tf}_{t,d})$	t (idf)	$\log \frac{N}{\mathrm{df}_t}$	c (cosine)	$\frac{1}{\sqrt{w_1^2+w_2^2+\ldots+w_M^2}}$
a (augmented)	$0.5 + \frac{0.5 \times \mathrm{tf}_{t,d}}{\max_t(\mathrm{tf}_{t,d})}$	p (prob idf)	$\max\{0, \log \frac{N-\mathrm{df}_t}{\mathrm{df}_t}\}$	u (pivoted unique)	$1/u$ (Section 6.4.4)
b (boolean)	$\begin{cases} 1 & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$			b (byte size)	$1/CharLength^{\alpha}, \alpha < 1$
L (log ave)	$\frac{1+\log(\mathrm{tf}_{t,d})}{1+\log(\mathrm{ave}_{t \in d}(\mathrm{tf}_{t,d}))}$				

Machine learning for retrieval ? Yes, learning to rank

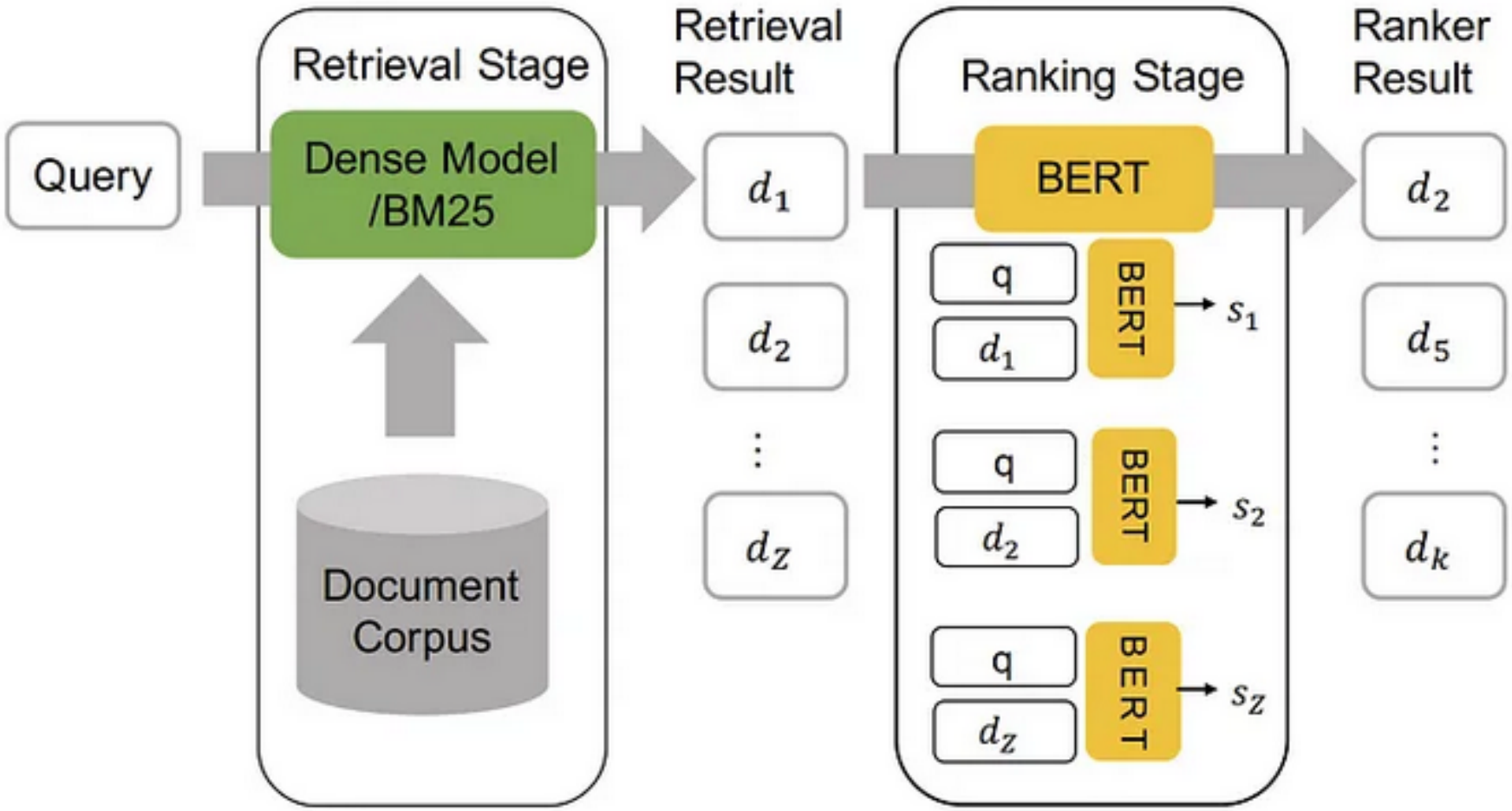


Zero-shot learning does not exist...

Language models did improved the ability to learn with little data

Typically in IR, you need 2-3 iterations or min. 20-30 queries to start learning something useful !

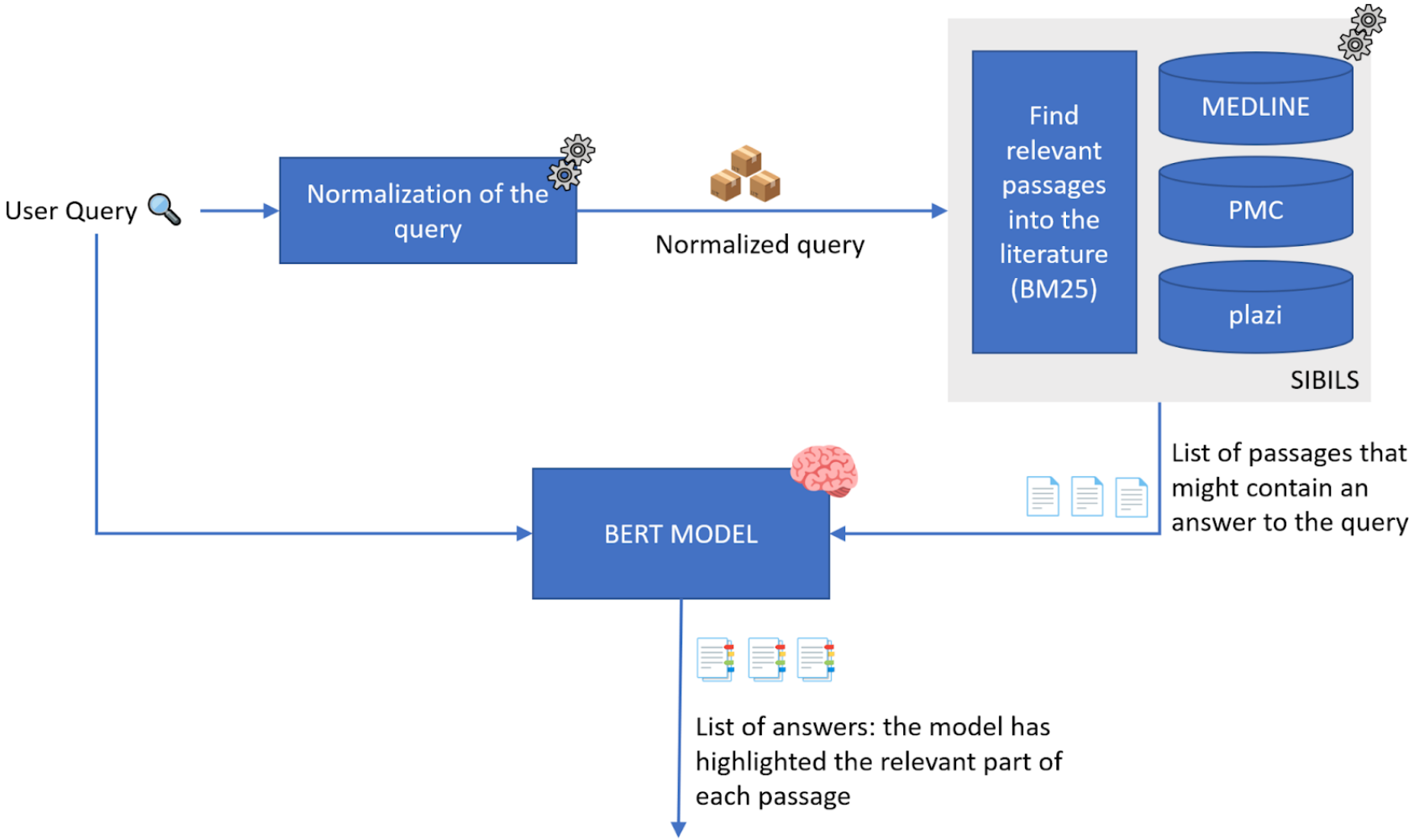
Search + pre-trained language models such as BERT



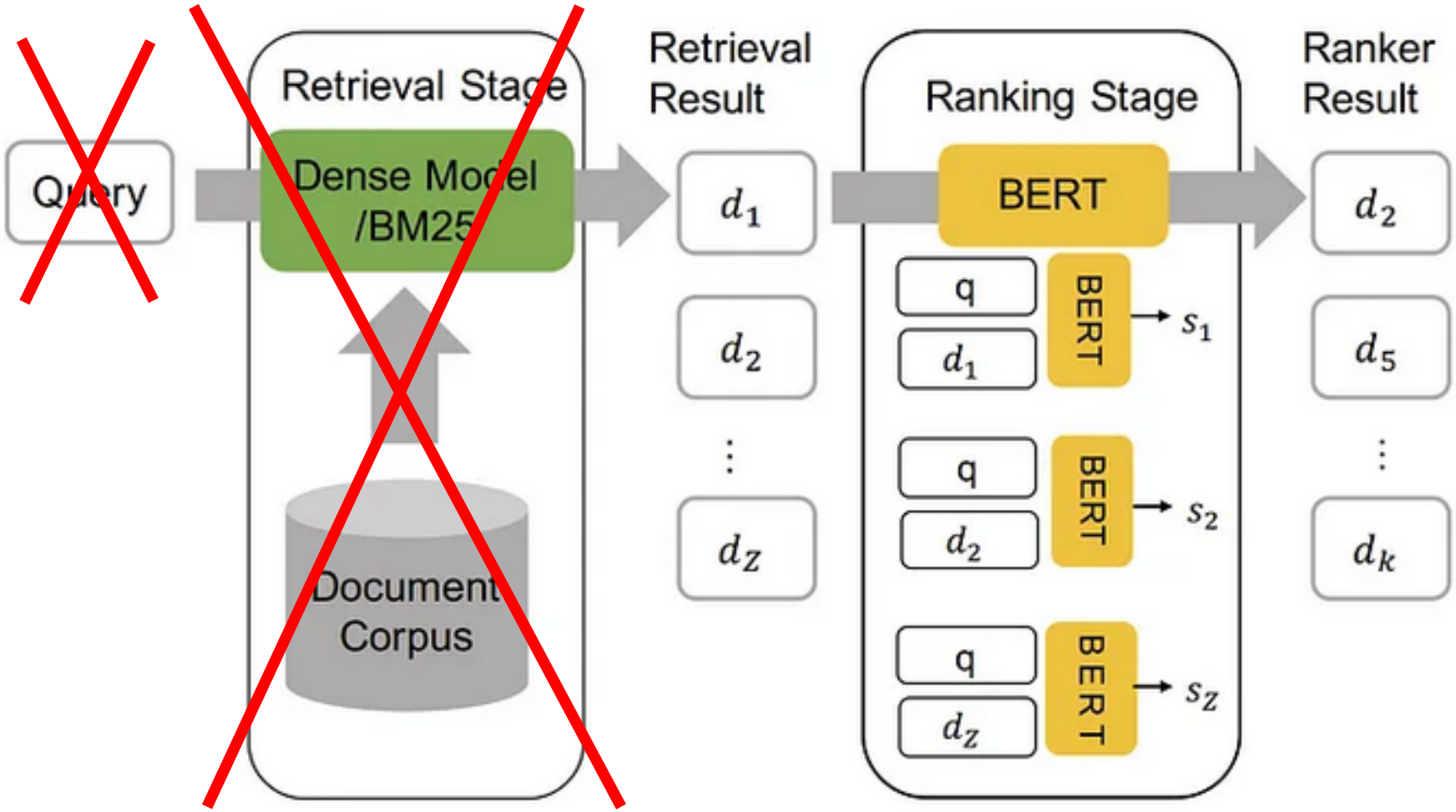
Sparse retrievers → Large matrix, mostly sparse
Dense retrievers → Embeddings

Question Answering

- https://sibils.text-analytics.ch/search/

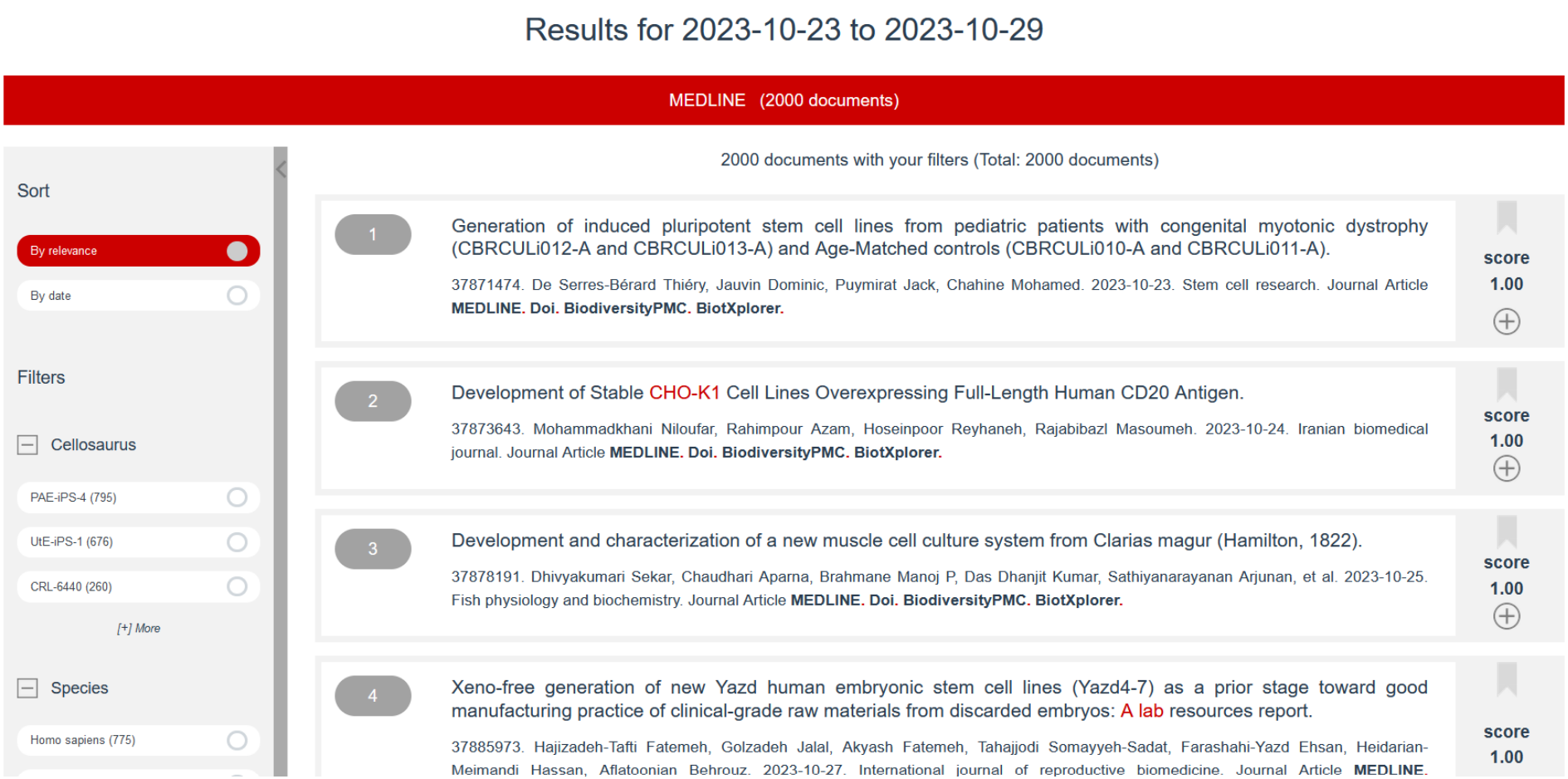


Zero shot also can occur → hallucinations ?



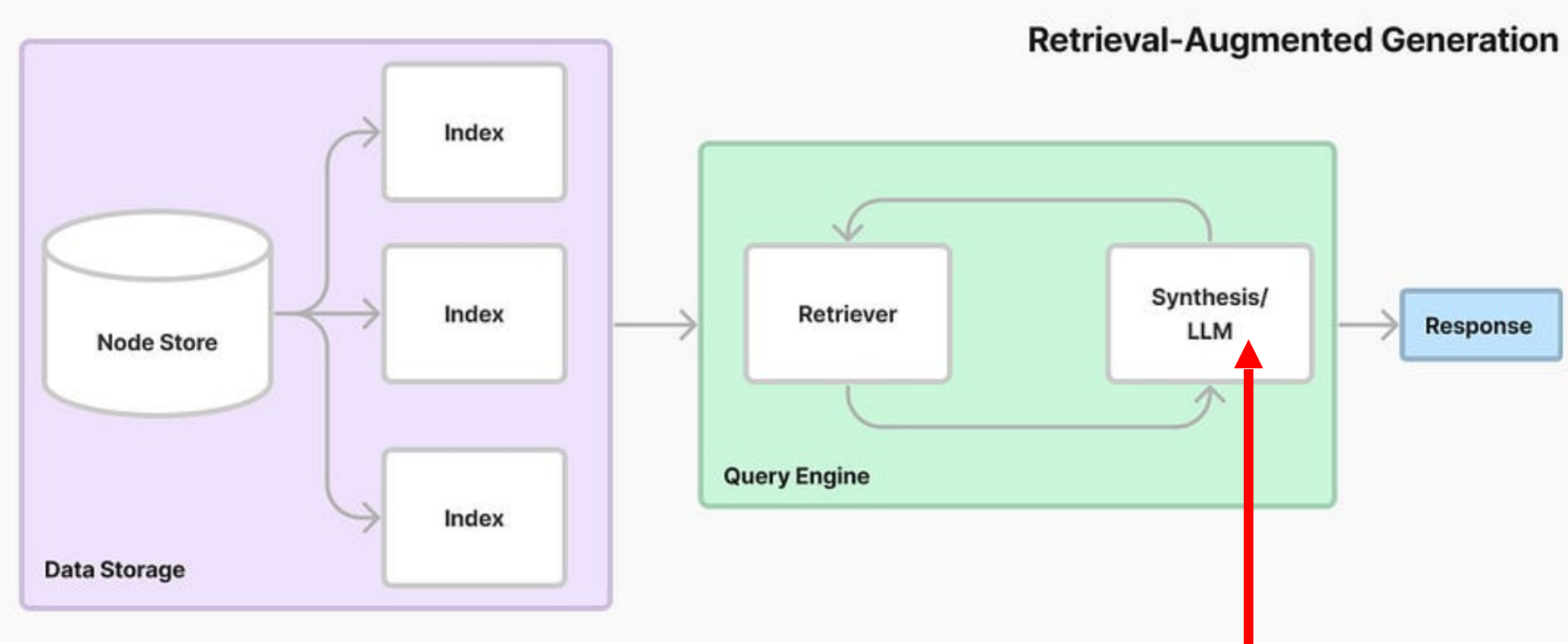
Sparse retrievers → Large matrix, mostly sparse
Dense retrievers → Embeddings

CellTriage: query independent model-based curation support



https://sibils.text-analytics.ch/search/celltriage

Search + generative language models



Drawbacks: compute costs++, hallucinations

Thank you for your attention !