

Text Mining & Bioinformatics

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



Swiss Institute of
Bioinformatics

Hes·SO

Haute Ecole Spécialisée
de Suisse occidentale

University of Applied Sciences
Western Switzerland

SIB Text Mining Group

- SIB Literature services

- <https://sibils.org/>

Biodiversity PMC

BiotXplorer

Variomes

SynVar

Celltrriage

SPARQL endpoint

INTRODUCTION TO



- 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 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 !**

- 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

- Precision
- Recall
- Other metrics...

- 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
- $\text{Recall} = 3/5 = 0.60$ 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 ?

[Example of “ground truth” or “benchmark collections; in such collections, the number of relevant papers is estimated a priori, see <https://trec.nist.gov/>]

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>

Variant P53 (R213L)

MEDLINE (2 documents)

PubMed Central (37 documents)

Clinical Trials (1 document)

Supplementary data (60 documents)

Sort and highlight

+

 Sort

+

 Highlight

Filters

-

 Date

2013

2023

-

 Sections

37 documents

1

The spectrum of subclonal TP53 mutations in chronic lymphocytic leukemia: A next generation sequencing retrospective study

PMID:35961859. PMC10086786. Giuseppa De Luca, Giannamaria Cerruti, Sonia Lastraioli, Romana Conte, Adalberto Ibatici, et al. 2022. Hematological Oncology. research-article. **MEDLINE PubMedCentral EuropePMC SIBiLS**

gene

full_text

spliceite_5 (ex.5) c.376-2A>T Aberrant splicing 2.43 10,433 254 6 c.570_573delTCCT p.(Pro191fs) Frameshift deletion 6.3 11,723 740 6 c.638G>T p.(Arg213Leu) Missense 1.65 5260 87 7 c.742C>T p.(Arg248Trp

full_text

.(Ile232Phe) Missense 29.1 14,904 513 8 c.826G>A p.(Ala276Thr) Missense 47.4 4247 2013 8 c.842 A>G p.(Asp281Gly) Missense 2.2 8235 177 6 c.638G>T p.(Arg213Leu) Missense 31.4 1241 389 7 c.716A>C p.(Asn239Thr

score

1.00

17

Synthetic metrics

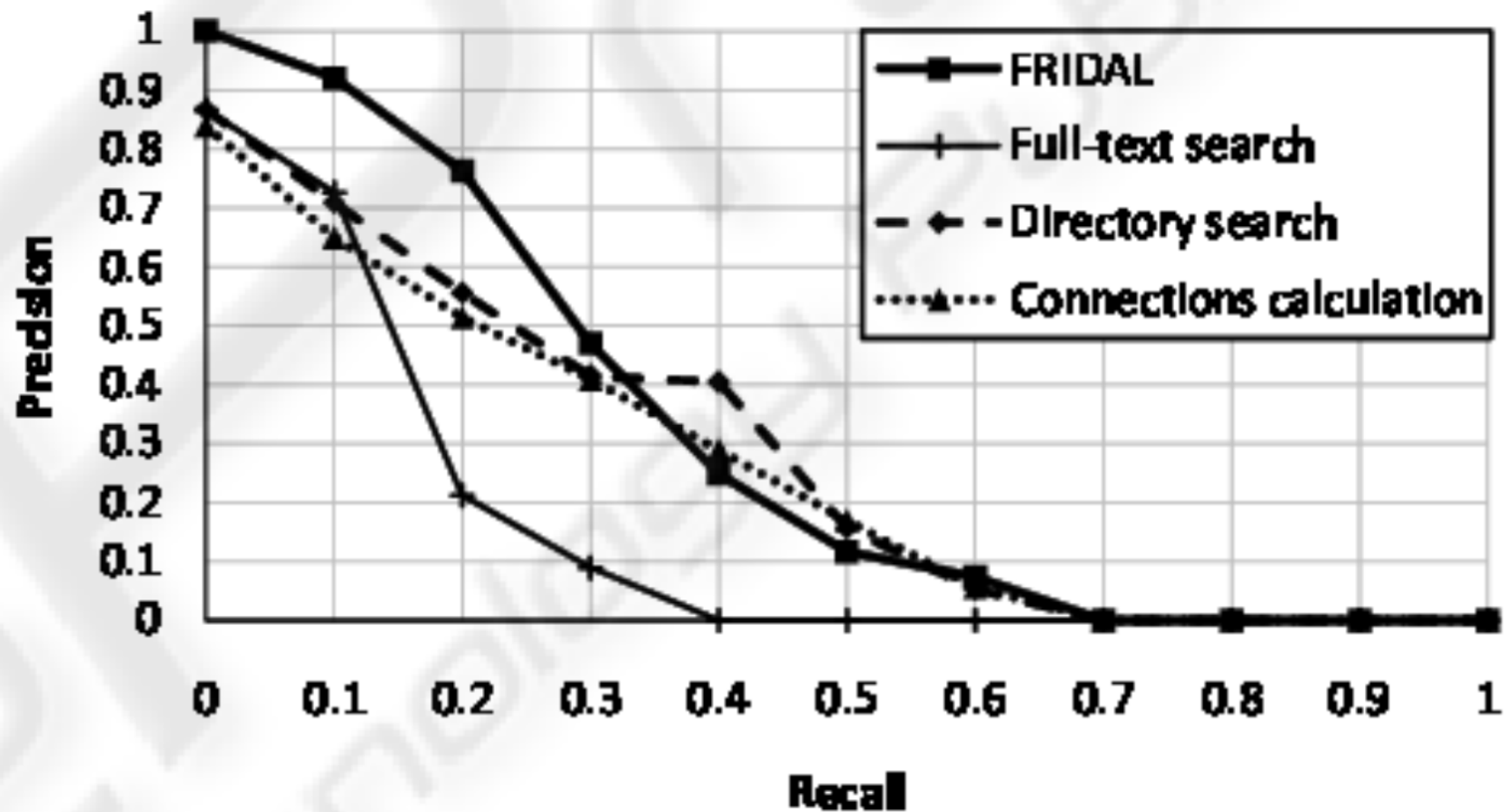
■ Rank

- 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., $0.9 \times \text{Recall} + 0.1 \times \text{Precision}$

Example



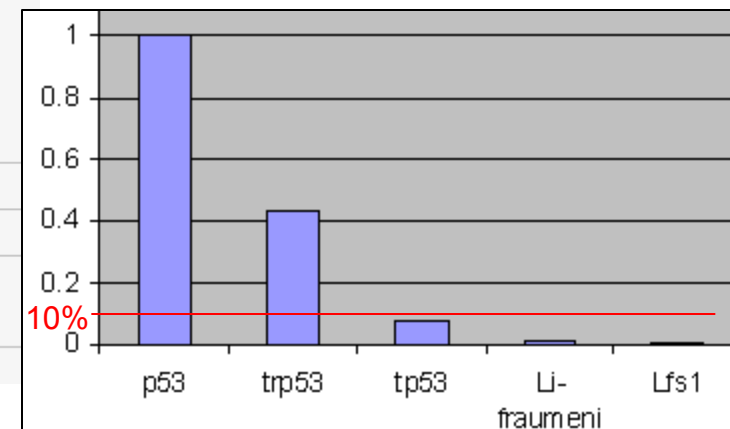
Feature normalization

- 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 !

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

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



- i, ii, iii → 1, 2, 3 (e.g. *histone deacetylase iii*)
- Greek letters (e.g. α -*tubulin*)
- Hyphenation «-»: {alphan-tubulin, 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

Sequence to encode

"This is a superduper complicatted sequence,
but this sequence can be encoded."

Dictionary state at iteration_i

Encoded sequence at iteration_i

Iteration₀

[<unk>, <s>, </s>, e, ., c, s, u, d, i, n, t, a, p, b, h, o, q, r, ., ., T, l, m]

[., T, h, i, s, ., i, s, ., a, ., s, u, p, e, r, d, u, p, e, r, ., c, o, m, p, l, i, c, a, t, t, e, d, ., s, e, q, u, e, n, c, e, ., ., b, u, t, ., t, h, i, s, ., s, e, q, u, e, n, c, e, ., c, a, n, ., b, e, ., e, n, c, o, d, e, d, .]

Iteration₁

[<unk>, <s>, </s>, en, e, ., c, s, u, d, i, n, t, a, p, b, h, o, q, r, ., ., T, l, m]

[., T, h, i, s, ., i, s, ., a, ., s, u, p, e, r, d, u, p, e, r, ., c, o, m, p, l, i, c, a, t, t, e, d, ., s, e, q, u, en, c, e, ., ., b, u, t, ., t, h, i, s, ., s, e, q, u, en, c, e, ., c, a, n, ., b, e, ., en, c, o, d, e, d, .]

...

Iteration₄

[<unk>, <s>, </s>, en, is, .s, enc, e, ., c, s, u, d, i, n, t, a, p, b, h, o, q, r, ., ., T, l, m]

[., T, h, i, s, ., is, ., a, .s, u, p, e, r, d, u, p, e, r, ., c, o, m, p, l, i, c, a, t, t, e, d, .s, e, q, u, enc, e, ., ., b, u, t, ., t, h, i, s, .s, e, q, u, enc, e, ., c, a, n, ., b, e, ., enc, o, d, e, d, .]

...

Iteration₃₀

[<unk>, <s>, </s>, en, is, .s, enc, ca, ed, eq, er, up, .b, equ, his, ence, uper, .sequ, .sequence, co, li, mp, od, tt, ut, .T, .a, .t, can, .be, .co, .is, catt, e, ., c, s, u, d, i, n, t, a, p, b, h, o, q, r, ., ., T, l, m]

[.T, his, .is, .a, .s, uper, d, uper, .co, mp, li, catt, ed, .sequence, ., .b, ut, .t, his, .sequence, ., can, .be, ., enc, od, ed, .]

...

Iteration₄₄

[<unk>, <s>, </s>, en, is, .s, enc, ca, ed, eq, er, up, .b, equ, his, ence, uper, .sequ, .sequence, co, li, mp, od, tt, ut, .T, .a, .t, can, .be, .co, .is, catt, mpli, oded, .but, .can, .enc, duper, .This, .this, catted, .super, .compli, .encoded, .superduper, .complicatted, e, ., c, s, u, d, i, n, t, a, p, b, h, o, q, r, ., ., T, l, m]

[.This, .is, .a, .superduper, .complicatted, .sequence, ., .but, .this, .sequence, .can, .be, .encoded, .]

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 !

[HERE]

Thank you for your attention !