TREC 2017 Precision Medicine

NIST, Gaithersburg, MD - Nov 17, 2017

Medical University of Graz

TEAM: imi_mug

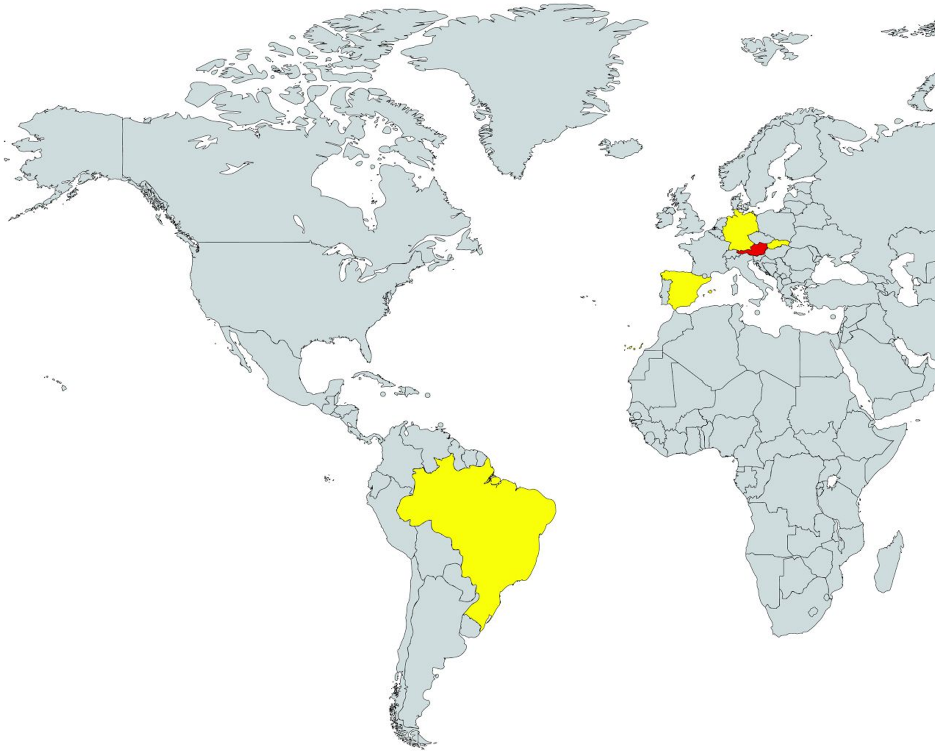
Institute for Medical Informatics, Statistics and Documentation

Pablo López-García

Michel Oleynik

Zdenko Kasáč

Stefan Schulz

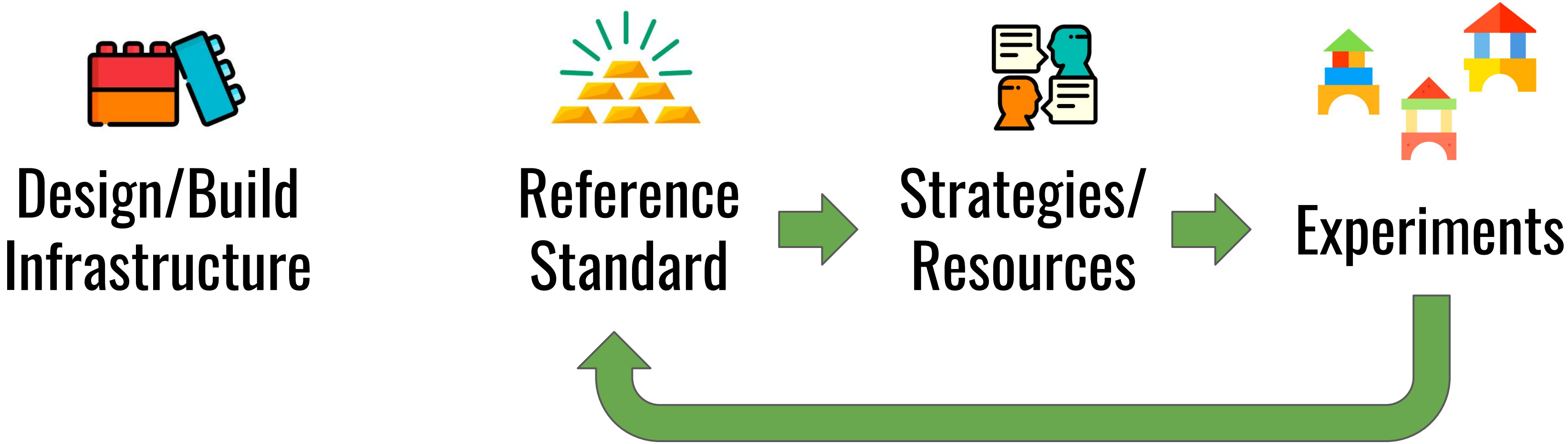


Premises

Fully automatic, Reusable, Open Source, on GitHub

Biomedical articles

Organization



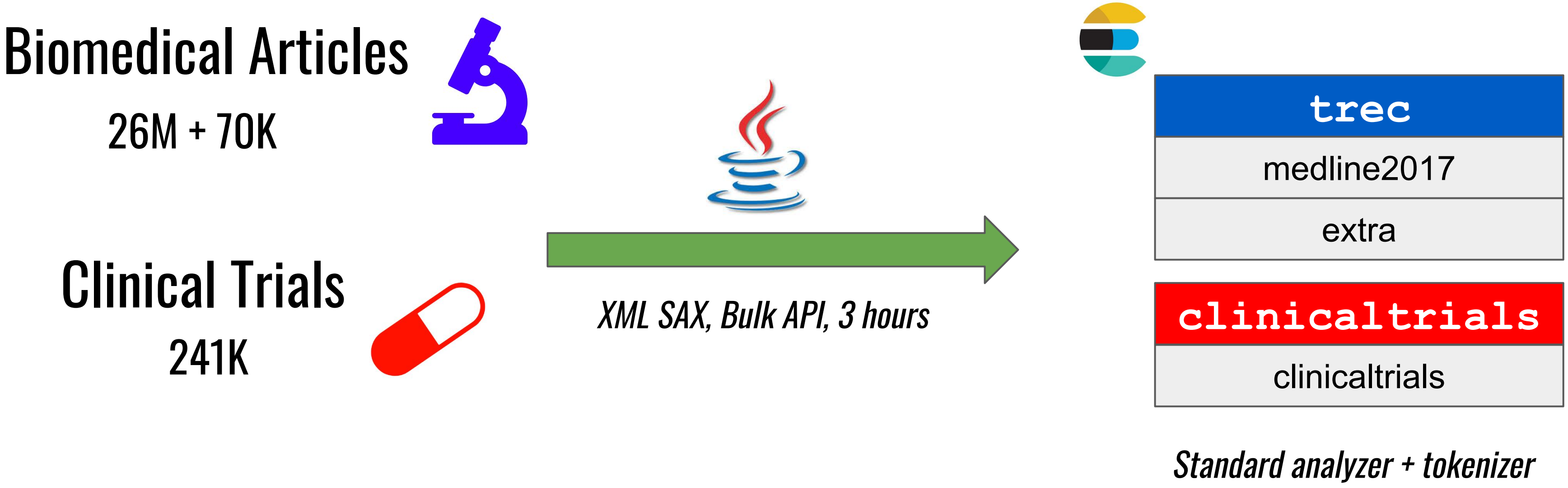
Design/Build Infrastructure

Reference Standard

Strategies/ Resources

Experiments

Infrastructure: Documents



Biomedical Articles

26M + 70K

Clinical Trials

241K

XML SAX, Bulk API, 3 hours

trec

medline2017

extra

clinicaltrials

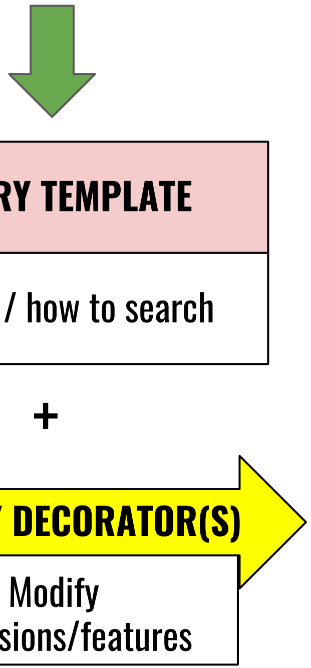
clinicaltrials

Standard analyzer + tokenizer

Queries?

Infrastructure: Flexible Query Generation

Colon cancer	RB1, TP53, KRAS	57-year-old female	None



QUERY TEMPLATE

Where / how to search

+

QUERY DECORATOR(S)

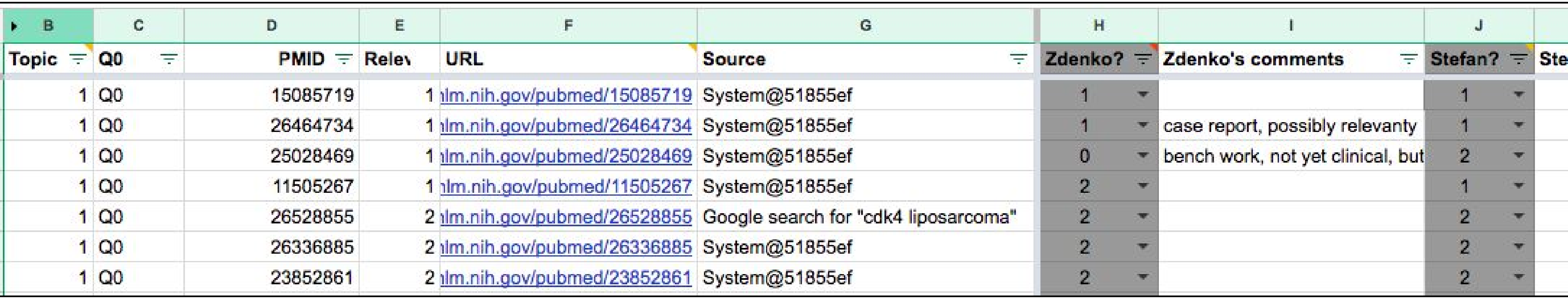
Modify dimensions/features

=

ELASTICSEARCH QUERY

```
"bool": {
  "must": [
    {
       "multi_match": {
         "query": "colon cancer
                  cancer of colon
                  neoplasm of colon",
         "fields": [
           "title^2",
           "abstract",
           "keyword",
           "meshTags"   ],
       "tie_breaker": 0.3,
       "type": "best_fields",
       "boost": 1
     },
...
```

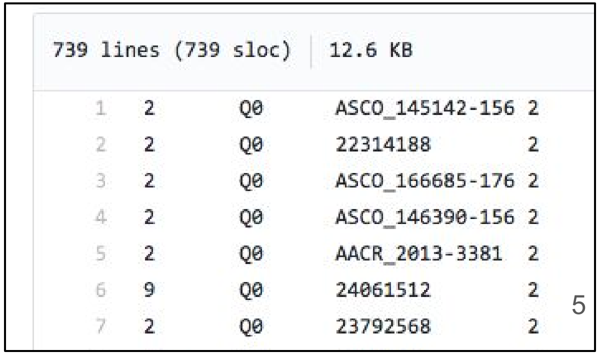
Reference Standard



B	C	D	E	F	G	H	I	J	
Topic	Q0	PMID	Relev	URL	Source	Zdenko?	Zdenko's comments	Stefan?	Ste
1	Q0	15085719	1	lm.nih.gov/pubmed/15085719	System@51855ef	1		1	
1	Q0	26464734	1	lm.nih.gov/pubmed/26464734	System@51855ef	1	case report, possibly relevanty	1	
1	Q0	25028469	1	lm.nih.gov/pubmed/25028469	System@51855ef	0	bench work, not yet clinical, but	2	
1	Q0	11505267	1	lm.nih.gov/pubmed/11505267	System@51855ef	2		1	
1	Q0	26528855	2	lm.nih.gov/pubmed/26528855	Google search for "cdk4 liposarcoma"	2		2	
1	Q0	26336885	2	lm.nih.gov/pubmed/26336885	System@51855ef	2		2	
1	Q0	23852861	2	lm.nih.gov/pubmed/23852861	System@51855ef	2		2	

739 documents / 1 to 3 assessors

Publicly available on GitHub in TREC qrels format



```
739 lines (739 sloc) | 12.6 KB

2    Q0    ASCO_145142-156 2
2    Q0    22314188        2
2    Q0    ASCO_166685-176 2
2    Q0    ASCO_146390-156 2
2    Q0    AACR_2013-3381  2
9    Q0    24061512        2
2    Q0    23792568        2
```

Strategies / Resources

Elasticsearch

must/shold, topic dimensions, doc fields, boosting

```
"bool": {
  "must": [
    {
       "multi_match": {
         "query": {{disease}} + {{gene}},
         "fields": ["title^2",
                    "abstract",
                    "keyword",
                    "meshTags"],
       "tie_breaker": 0.3,
       "type": "best_fields",
     }
  "should": [

     {

        "match": { "_type": "extra" }

     }, (...)
```

Keywords boosting

TREC extra topics, gold standard, med. knowledge, chemo. suffixes

cancer
carcinoma
tumor

gene
genotype
DNA
base

malignancy
outcome
patient
prognosis
prognostic
recurrence
resistance
study
surgery
survival
targets
therapeutical
therapy
treatment

detection
embedded
fish
formalin
paraffin
probes
screening
tissue
transcript
tumorigenesis

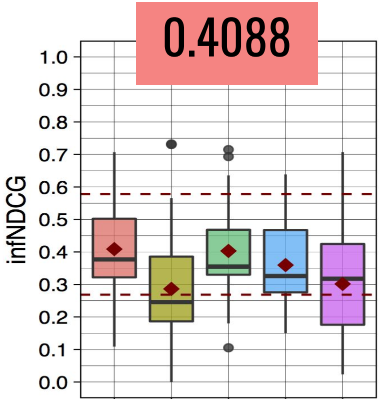
-ate
-cin
-lin
-one
-mab
-mus
-nib

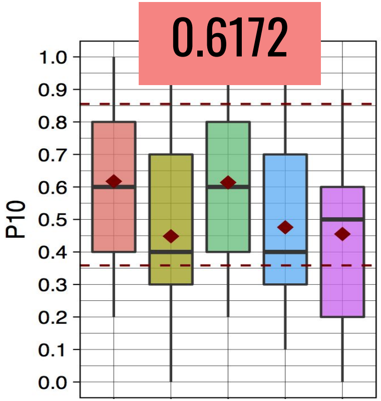
Query Expansion

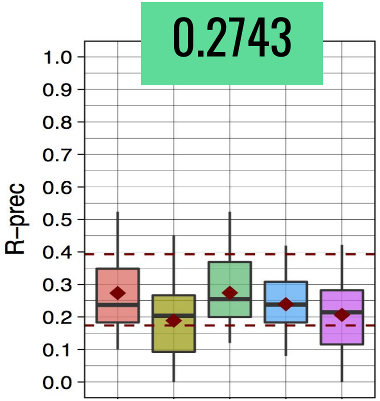
Diseases (Lexigram/SNOMED), Genes (NCBI Homo Sapiens)

Biomedical Articles

	mugpubboost	mugpubshould	mugpubbase	mugpubdiseas	mugpubgene
Match disease & gene	must	should	must	must	must
Boost+ keywords	✔	✔	✔	✔	✔
Boost- keywords	✔	✔		✔	✔
Boost+ chemo. suffixes	✔	✔		✔	✔
Disease expansion				✔	
Gene expansion					✔

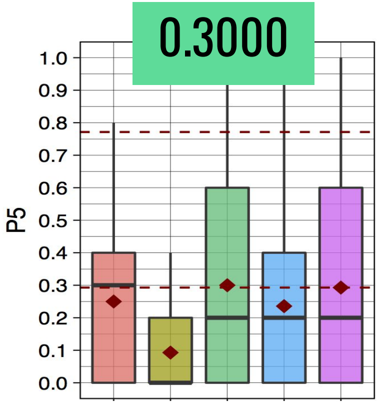

0.4088

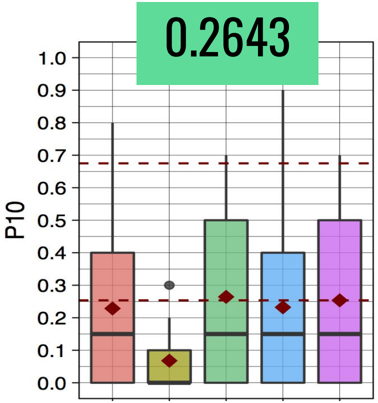

0.6172

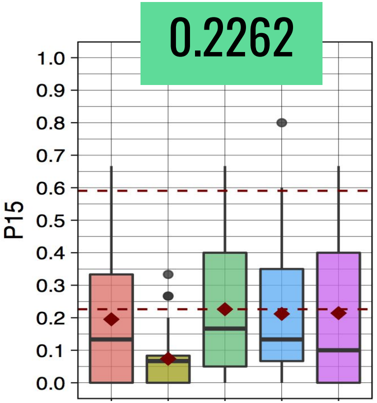

0.2743

Clinical Trials

	mugctboost	mugctdisease	mugctbase	mugctgene	mugctmust
Match age range & sex	must	must	must	must	should
Match disease & gene	must	must	should	must	should
Comorbidities NOT excl.	✔	✔		✔	✔
Boost+ keywords	✔	✔		✔	
Disease expansion		✔			
Gene expansion				✔	







Limitations, Discussion, Future Work

1. Judgements coherent between our reference standard / TREC
2. Baseline strategy was best / nearly best
3. Results: often < 1,000
4. Explore other strategies:
   a. All dimensions
   b. Other ES analyzers
5. Framework: more flexible/orthogonal, sub-templates, command-line
6. Future participation

Contact, More Info

Pablo López-García: pablo.lopez@medunigraz.at - Michel Oleynik: michel.oleynik@stud.medunigraz.at

GitHub: https://github.com/bst-mug/trec2017

IMI/MUG: https://www.medunigraz.at/imi/en/

BST/Stefan Schulz: http://user.medunigraz.at/stefan.schulz/

Acknowledgments

Lexigram API

Brazilian National Research Council - CNPq (Project No. 206892/2014-4)