TEXT MINING

L10. BIOMEDICAL TEXT MINING

SUZAN VERBERNE 2022

Universiteit Leiden

TODAY'S LECTURE

- Quiz about week 9
- Assignment 2
- Biomedical text mining
  - Motivation
  - Biomedical research questions that can be answered with TM
  - Exercise
  - Text mining modules in the biomedical domain
  - State of the art
- Introduction of the final assignment

QUIZ ABOUT WEEK 9

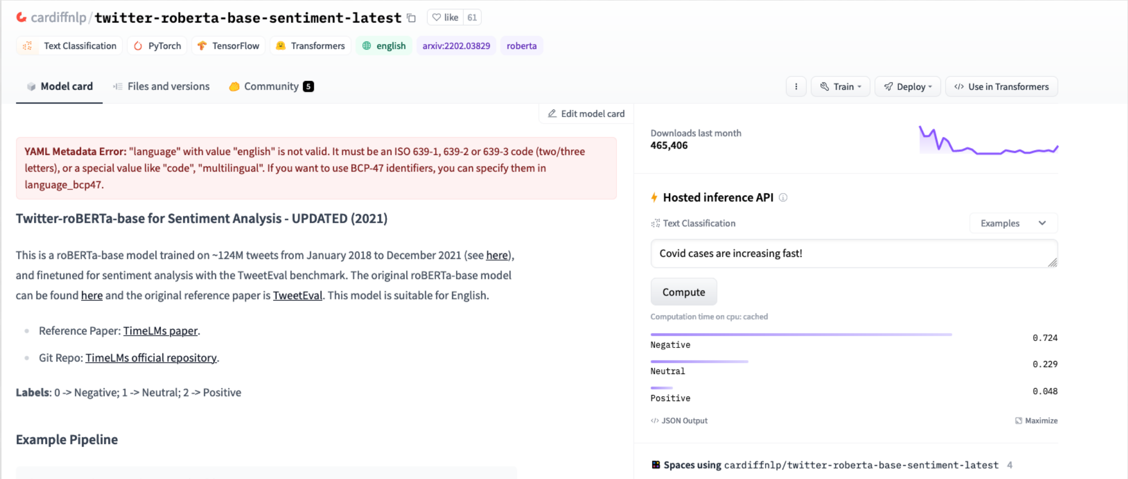
- If we download a sentiment classification model from Huggingface, and we want to use it for classification of customer reviews on a scale of 1-5, what do we need? (multiple answers possible)

  a. Nothing

  b. Labelled customer reviews for finetuning the model

  c. Change the loss function of the model to (ordinal) regression

  d. GPU computing



YAML Metadata Error: "language" with value "english" is not valid. It must be an ISO 639-1, 639-2 or 639-3 code (two/three letters), or a special value like "code", "multilingual". If you want to use BCP-47 identifiers, you can specify them in language_bcp47.

Twitter-roBERTa-base for Sentiment Analysis - UPDATED (2021)

This is a roBERTa-base model trained on ~124M tweets from January 2018 to December 2021 (see here), and finetuned for sentiment analysis with the TweetEval benchmark. The original roBERTa-base model can be found here and the original reference paper is TweetEval. This model is suitable for English.

- Reference Paper: TimeLMs paper.
- Git Repo: TimeLMs official repository.

Labels: 0 -> Negative; 1 -> Neutral; 2 -> Positive

Example Pipeline

Downloads last month
465,406

Hosted inference API

Text Classification

Covid cases are increasing fast!

Computation time on cpu: cached

```
Negative    0.724
Neutral     0.229
Positive    0.048
```

Spaces using `cardiffnlp/twitter-roberta-base-sentiment-latest` 4

https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

QUIZ ABOUT WEEK 9

- If we download a sentiment classification model from Huggingface, and we want to use it for classification of customer reviews on a scale of 1-5, what do we need? (multiple answers possible)

a. Nothing

b. Labelled customer reviews for finetuning the model

c. Change the loss function of the model to (ordinal) regression

d. GPU computing

https://lajavaness.medium.com/regression-with-text-input-using-bert-and-transformers-71c155034b13

QUIZ ABOUT WEEK 9

- Suppose we have binary sentiment classification, positive vs negative. How is the recall for the negative class defined?

a. Of all messages automatically classified as negative, how many are correct

b. Of all messages with true label negative, how many are classified correctly

c. Of all messages, how many are negative

QUIZ ABOUT WEEK 9

- Suppose we have binary sentiment classification, positive vs negative. How is the recall for the negative class defined?

a. Of all messages automatically classified as negative, how many are correct

b. Of all messages with true label negative, how many are classified correctly

c. Of all messages, how many are negative

QUIZ ABOUT WEEK 9

- What are differences between sentiment classification and stance detection? (multiple answers possible)

a. The labels are different (positive/negative vs pro/con)

b. Sentiment classification is ordinal and stance detection not

c. Sentiment classification takes one text as input, stance detection is about the relation between two texts

d. Stance detection is about political issues, sentiment is about customer reviews

QUIZ ABOUT WEEK 9

- What are differences between sentiment classification and stance detection? (multiple answers possible)

  a. The labels are different (positive/negative vs pro/con)

  b. Sentiment classification is ordinal and stance detection not

  c. Sentiment classification takes one text as input, stance detection is about the relation between two texts

  d. Stance detection is about political issues, sentiment is about customer reviews

QUIZ ABOUT WEEK 9

- Why is the standard deviation over runs important in reproducing results?

a. Because there is variation between seeds of the Transformer models

b. Because reproduction is difficult and we never get the exact same result, so we want to be close

c. Because not only the mean but also the standard deviation should be the same

d. Because we want to be sure we compare to the correct baseline

QUIZ ABOUT WEEK 9

- Why is the standard deviation over runs important in reproducing results?

a. Because there is variation between seeds of the Transformer models

b. Because reproduction is difficult and we never get the exact same result, so we want to be close

c. Because not only the mean but also the standard deviation should be the same

d. Because we want to be sure we compare to the correct baseline

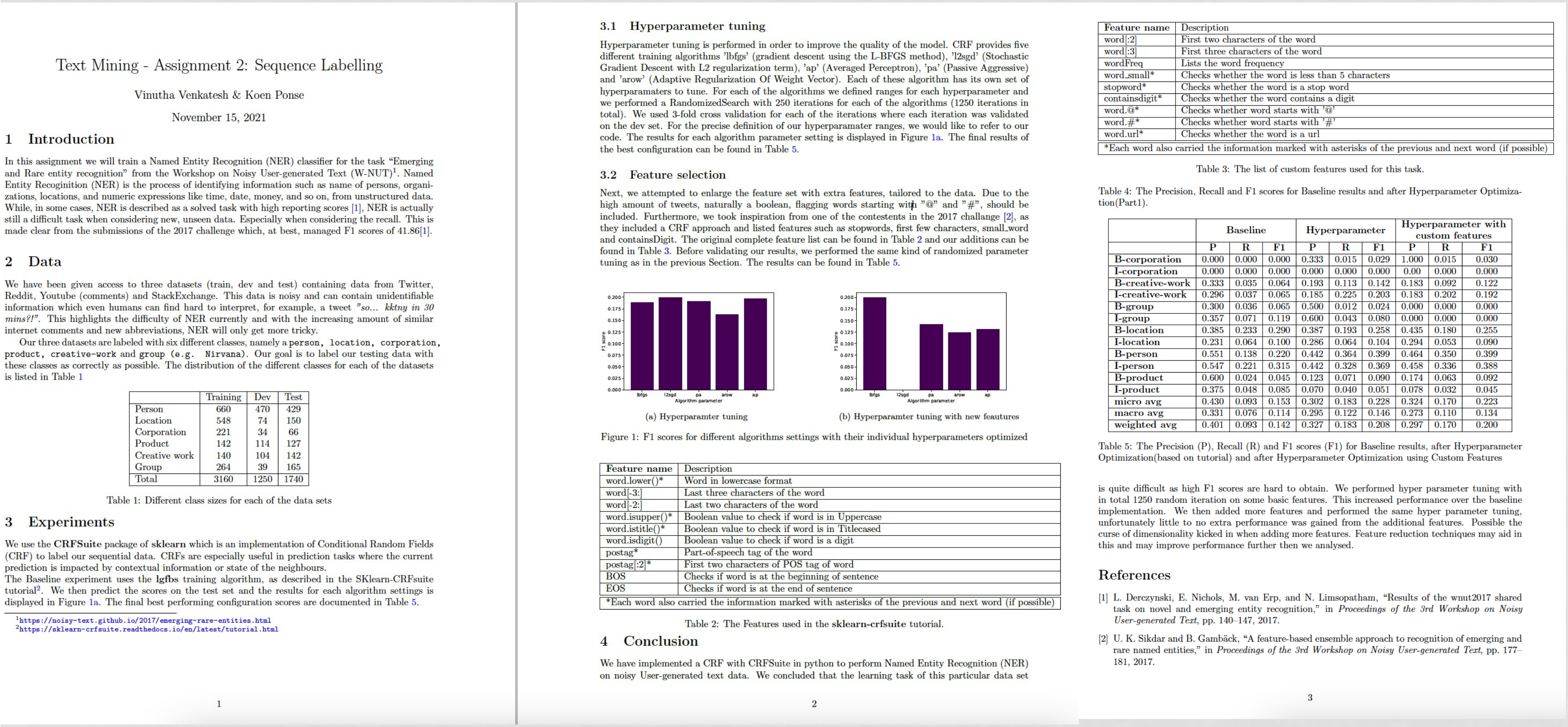
ASSIGNMENT 2

ASSIGNMENT 2 – NER

- Grading :
  - 5 criteria; max 2 points per criterion

1. General: length correct (2-3 pages) and proper writing + formatting
2. Description of the task and the data
3. Description of the adapted features
4. Baseline run with features from tutorial & experimental runs with adapted features (show results in table: Precision, Recall, F-score for the B and I tags)
5. Sensible conclusions

EXAMPLE REPORT



Text Mining - Assignment 2: Sequence Labelling

Vinutha Venkatesh & Koen Ponse

November 15, 2021

1 Introduction

In this assignment we will train a Named Entity Recognition (NER) classifier for the task "Emerging and Rare entity recognition" from the Workshop on Noisy User-generated Text (W-NUT)[1]. Named Entity Recoginition (NER) is the process of identifying information such as name of persons, organizations, locations, and numeric expressions like time, date, money, and so on, from unstructured data. While, in some cases, NER is described as a solved task with high reporting scores [1], NER is actually still a difficult task when considering new, unseen data. Especially when considering the recall. This is made clear from the submissions of the 2017 challenge which, at best, managed F1 scores of 41.86[1].

2 Data

We have been given access to three datasets (train, dev and test) containing data from Twitter, Reddit, Youtube (comments) and StackExchange. This data is noisy and can contain unidentifiable information which even humans can find hard to interpret, for example, a tweet *"so... kktny in 30 mins?!"*. This highlights the difficulty of NER currently and with the increasing amount of similar internet comments and new abbreviations, NER will only get more tricky.

Our three datasets are labeled with six different classes, namely a `person, location, corporation, product, creative-work` and `group (e.g. Nirvana)`. Our goal is to label our testing data with these classes as correctly as possible. The distribution of the different classes for each of the datasets is listed in Table 1

	Training	Dev	Test
Person	660	470	429
Location	548	74	150
Corporation	221	34	66
Product	142	114	127
Creative work	140	104	142
Group	264	39	165
Total	3160	1250	1740

Table 1: Different class sizes for each of the data sets

3 Experiments

We use the **CRFSuite** package of **sklearn** which is an implementation of Conditional Random Fields (CRF) to label our sequential data. CRFs are especially useful in prediction tasks where the current prediction is impacted by contextual information or state of the neighbours.
The Baseline experiment uses the **lgfbs** training algorithm, as described in the SKlearn-CRFsuite tutorial[2]. We then predict the scores on the test set and the results for each algorithm settings is displayed in Figure 1a. The final best performing configuration scores are documented in Table 5.

[1] https://noisy-text.github.io/2017/emerging-rare-entities.html
[2] https://sklearn-crfsuite.readthedocs.io/en/latest/tutorial.html

3.1 Hyperparameter tuning

Hyperparameter tuning is performed in order to improve the quality of the model. CRF provides five different training algorithms 'lbfgs' (gradient descent using the L-BFGS method), 'l2sgd' (Stochastic Gradient Descent with L2 regularization term), 'ap' (Averaged Perceptron), 'pa' (Passive Aggressive) and 'arow' (Adaptive Regularization Of Weight Vector). Each of these algorithm has its own set of hyperparamaters to tune. For each of the algorithms we defined ranges for each hyperparameter and we performed a RandomizedSearch with 250 iterations for each of the algorithms (1250 iterations in total). We used 3-fold cross validation for each of the iterations where each iteration was validated on the dev set. For the precise definition of our hyperparamater ranges, we would like to refer to our code. The results for each algorithm parameter setting is displayed in Figure 1a. The final results of the best configuration can be found in Table 5.

3.2 Feature selection

Next, we attempted to enlarge the feature set with extra features, tailored to the data. Due to the high amount of tweets, naturally a boolean, flagging words starting with "@" and "#", should be included. Furthermore, we took inspiration from one of the contestents in the 2017 challange [2], as they included a CRF approach and listed features such as stopwords, first few characters, small_word and containsDigit. The original complete feature list can be found in Table 2 and our additions can be found in Table 3. Before validating our results, we performed the same kind of randomized parameter tuning as in the previous Section. The results can be found in Table 5.

(a) Hyperparamter tuning

(b) Hyperparamter tuning with new feautures

Figure 1: F1 scores for different algorithms settings with their individual hyperparameters optimized

Feature name	Description
word.lower()*	Word in lowercase format
word[-3:]	Last three characters of the word
word[-2:]	Last two characters of the word
word.isupper()*	Boolean value to check if word is in Uppercase
word.istitle()*	Boolean value to check if word is in Titlecased
word.isdigit()	Boolean value to check if word is a digit
postag*	Part-of-speech tag of the word
postag[:2]*	First two characters of POS tag of word
BOS	Checks if word is at the beginning of sentence
EOS	Checks if word is at the end of sentence
*Each word also carried the information marked with asterisks of the previous and next word (if possible)	

Table 2: The Features used in the **sklearn-crfsuite** tutorial.

4 Conclusion

We have implemented a CRF with CRFSuite in python to perform Named Entity Recognition (NER) on noisy User-generated text data. We concluded that the learning task of this particular data set is quite difficult as high F1 scores are hard to obtain. We performed hyper parameter tuning with in total 1250 random iteration on some basic features. This increased performance over the baseline implementation. We then added more features and performed the same hyper parameter tuning, unfortunately little to no extra performance was gained from the additional features. Possible the curse of dimensionality kicked in when adding more features. Feature reduction techniques may aid in this and may improve performance further then we analysed.

Feature name	Description
word[:2]	First two characters of the word
word[:3]	First three characters of the word
wordFreq	Lists the word frequency
word_small*	Checks whether the word is less than 5 characters
stopword*	Checks whether the word is a stop word
containsdigit*	Checks whether the word contains a digit
word.@*	Checks whether word starts with '@'
word.#*	Checks whether word starts with '#'
word.url*	Checks whether the word is a url
*Each word also carried the information marked with asterisks of the previous and next word (if possible)	

Table 3: The list of custom features used for this task.

Table 4: The Precision, Recall and F1 scores for Baseline results and after Hyperparameter Optimization(Part1).

	Baseline			**Hyperparameter**			**Hyperparameter with custom features**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
B-corporation	0.000	0.000	0.000	0.333	0.015	0.029	1.000	0.015	0.030
I-corporation	0.000	0.000	0.000	0.000	0.000	0.000	0.00	0.000	0.000
B-creative-work	0.333	0.035	0.064	0.193	0.113	0.142	0.183	0.092	0.122
I-creative-work	0.296	0.037	0.065	0.185	0.225	0.203	0.183	0.202	0.192
B-group	0.300	0.036	0.065	0.500	0.012	0.024	0.000	0.000	0.000
I-group	0.357	0.071	0.119	0.600	0.043	0.080	0.000	0.000	0.000
B-location	0.385	0.233	0.290	0.387	0.193	0.258	0.435	0.180	0.255
I-location	0.231	0.064	0.100	0.286	0.064	0.104	0.294	0.053	0.090
B-person	0.551	0.138	0.220	0.442	0.364	0.399	0.464	0.350	0.399
I-person	0.547	0.221	0.315	0.442	0.328	0.369	0.458	0.336	0.388
B-product	0.600	0.024	0.045	0.123	0.071	0.090	0.174	0.063	0.092
I-product	0.375	0.048	0.085	0.070	0.040	0.051	0.078	0.032	0.045
micro avg	0.430	0.093	0.153	0.302	0.183	0.228	0.324	0.170	0.223
macro avg	0.331	0.076	0.114	0.295	0.122	0.146	0.273	0.110	0.134
weighted avg	0.401	0.093	0.142	0.327	0.183	0.208	0.297	0.170	0.200

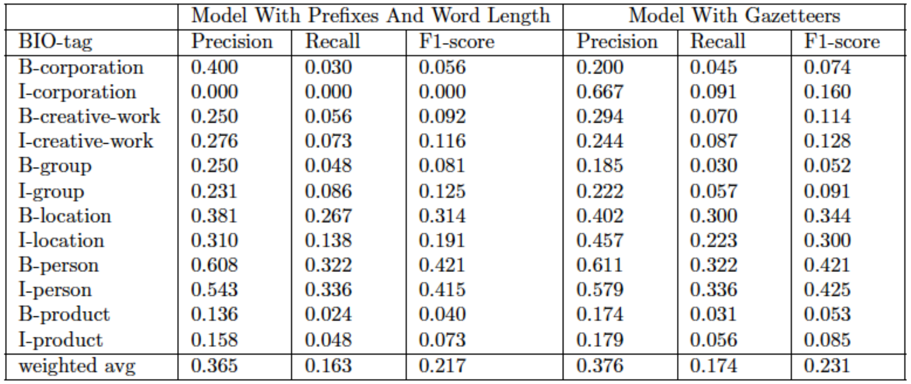
Table 5: The Precision (P), Recall (R) and F1 scores (F1) for Baseline results, after Hyperparameter Optimization(based on tutorial) and after Hyperparameter Optimization using Custom Features

References

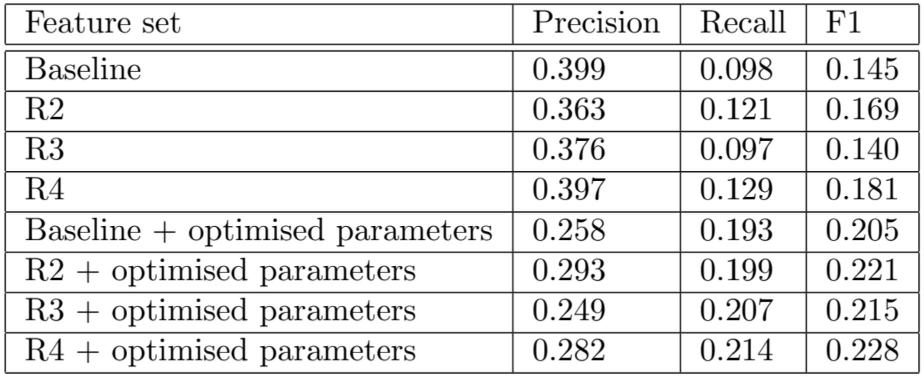
[1] L. Derczynski, E. Nichols, M. van Erp, and N. Limsopatham, "Results of the wnut2017 shared task on novel and emerging entity recognition," in *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 140–147, 2017.

[2] U. K. Sikdar and B. Gambäck, "A feature-based ensemble approach to recognition of emerging and rare named entities," in *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 177–181, 2017.

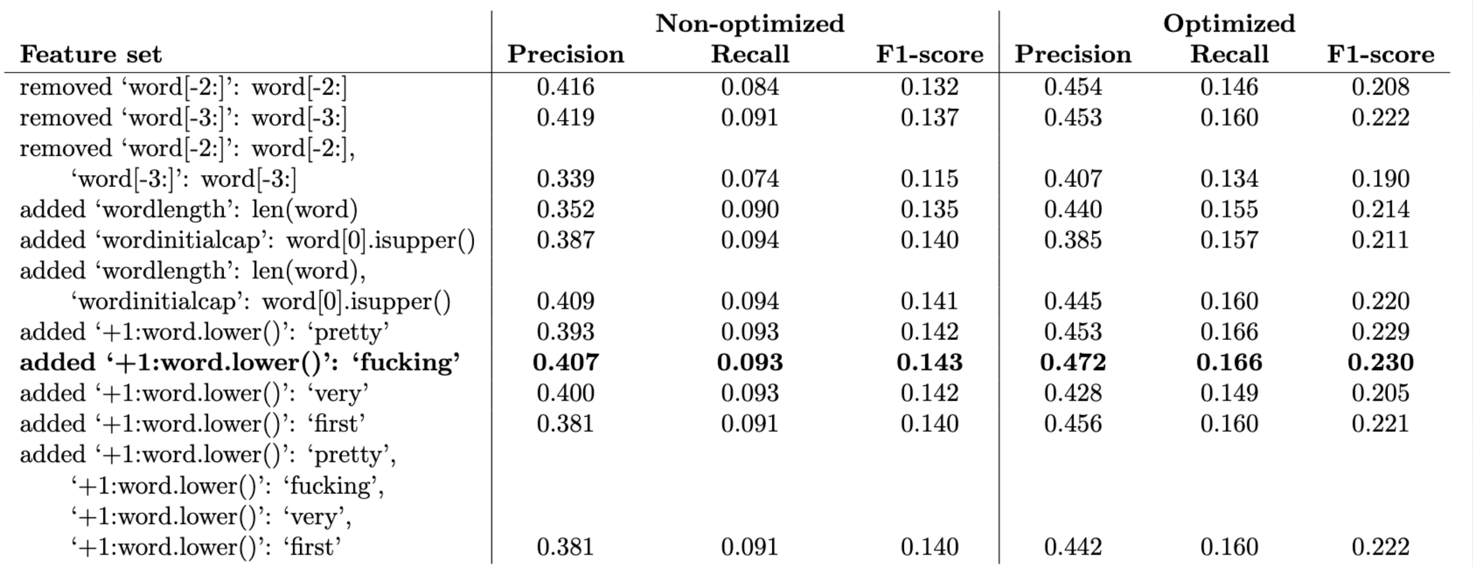
EXAMPLE RESULTS TABLES



	Model With Prefixes And Word Length			Model With Gazetteers		
BIO-tag	Precision	Recall	F1-score	Precision	Recall	F1-score
B-corporation	0.400	0.030	0.056	0.200	0.045	0.074
I-corporation	0.000	0.000	0.000	0.667	0.091	0.160
B-creative-work	0.250	0.056	0.092	0.294	0.070	0.114
I-creative-work	0.276	0.073	0.116	0.244	0.087	0.128
B-group	0.250	0.048	0.081	0.185	0.030	0.052
I-group	0.231	0.086	0.125	0.222	0.057	0.091
B-location	0.381	0.267	0.314	0.402	0.300	0.344
I-location	0.310	0.138	0.191	0.457	0.223	0.300
B-person	0.608	0.322	0.421	0.611	0.322	0.421
I-person	0.543	0.336	0.415	0.579	0.336	0.425
B-product	0.136	0.024	0.040	0.174	0.031	0.053
I-product	0.158	0.048	0.073	0.179	0.056	0.085
weighted avg	0.365	0.163	0.217	0.376	0.174	0.231



Feature set	Precision	Recall	F1
Baseline	0.399	0.098	0.145
R2	0.363	0.121	0.169
R3	0.376	0.097	0.140
R4	0.397	0.129	0.181
Baseline + optimised parameters	0.258	0.193	0.205
R2 + optimised parameters	0.293	0.199	0.221
R3 + optimised parameters	0.249	0.207	0.215
R4 + optimised parameters	0.282	0.214	0.228



		Non-optimized			Optimized	
Feature set	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
removed ‘word[-2:]’: word[-2:]	0.416	0.084	0.132	0.454	0.146	0.208
removed ‘word[-3:]’: word[-3:]	0.419	0.091	0.137	0.453	0.160	0.222
removed ‘word[-2:]’: word[-2:], ‘word[-3:]’: word[-3:]	0.339	0.074	0.115	0.407	0.134	0.190
added ‘wordlength’: len(word)	0.352	0.090	0.135	0.440	0.155	0.214
added ‘wordinitialcap’: word[0].isupper()	0.387	0.094	0.140	0.385	0.157	0.211
added ‘wordlength’: len(word), ‘wordinitialcap’: word[0].isupper()	0.409	0.094	0.141	0.445	0.160	0.220
added ‘+1:word.lower()’: ‘pretty’	0.393	0.093	0.142	0.453	0.166	0.229
added ‘+1:word.lower()’: ‘fucking’	**0.407**	**0.093**	**0.143**	**0.472**	**0.166**	**0.230**
added ‘+1:word.lower()’: ‘very’	0.400	0.093	0.142	0.428	0.149	0.205
added ‘+1:word.lower()’: ‘first’	0.381	0.091	0.140	0.456	0.160	0.221
added ‘+1:word.lower()’: ‘pretty’, ‘+1:word.lower()’: ‘fucking’, ‘+1:word.lower()’: ‘very’, ‘+1:word.lower()’: ‘first’	0.381	0.091	0.140	0.442	0.160	0.222

Table 2: Precision, Recall, F1-score for B and I tags of different feature set combinations based on alterations of default feature set.

Universiteit Leiden

BIOMEDICAL TEXT MINING

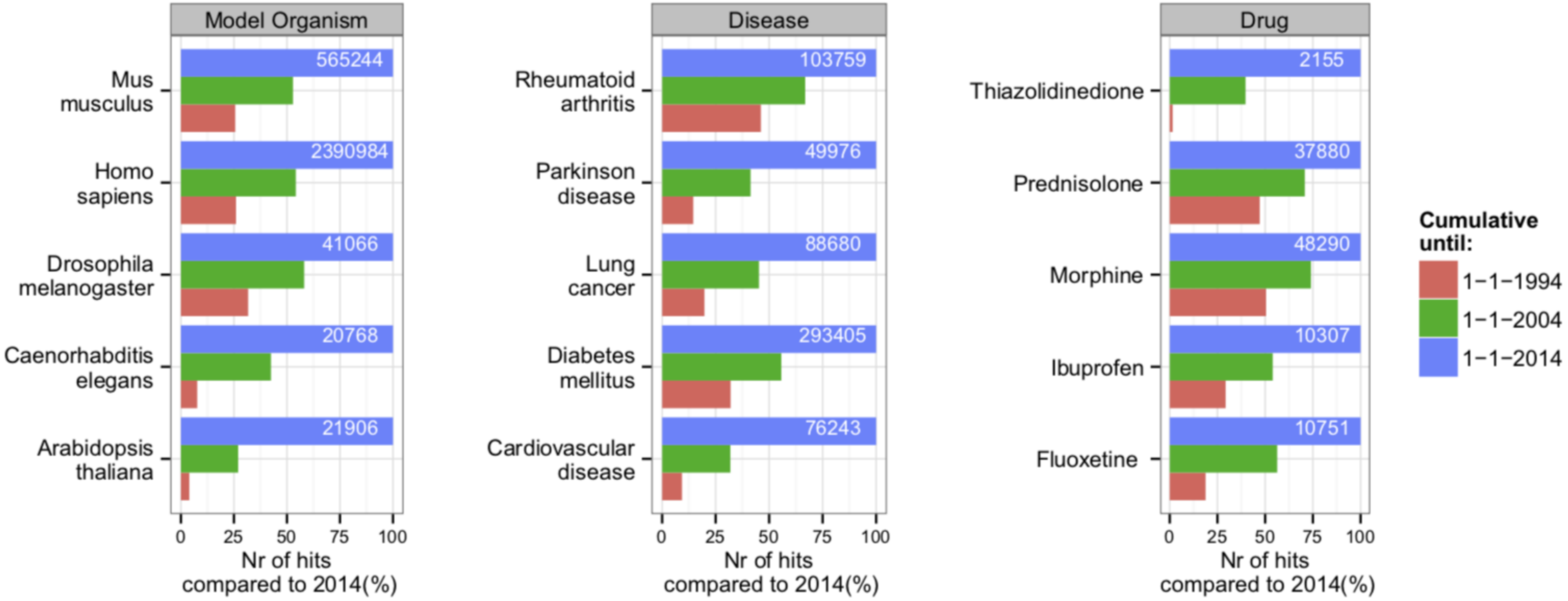
Suzan Verberne 2022

MOTIVATION

- Large amounts of data in the biomedical & health domain:
  - Scientific literature
  - Experimental data
  - Patents
  - Electronic Health Records
  - Patient surveys
  - Health social media (e.g. patient support groups)

PUBMED GROWTH

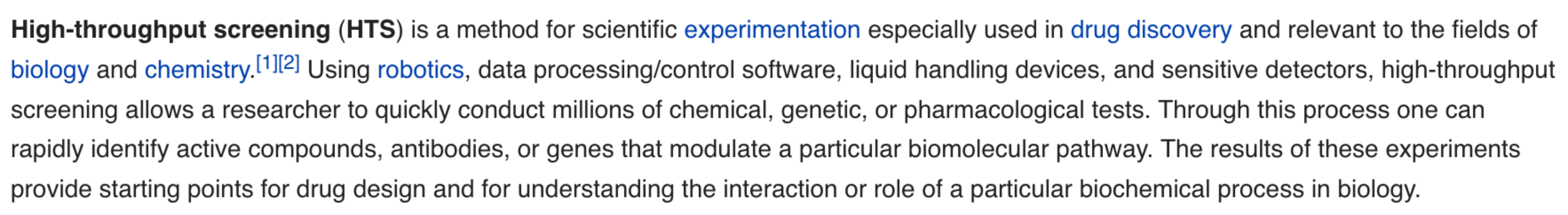
- The number of articles that are added to the literature databases (MEDLINE/PubMed) is growing fast



Fleuren and Alkema (2015). Application of text mining in the biomedical domain

DATA SIZE GROWTH

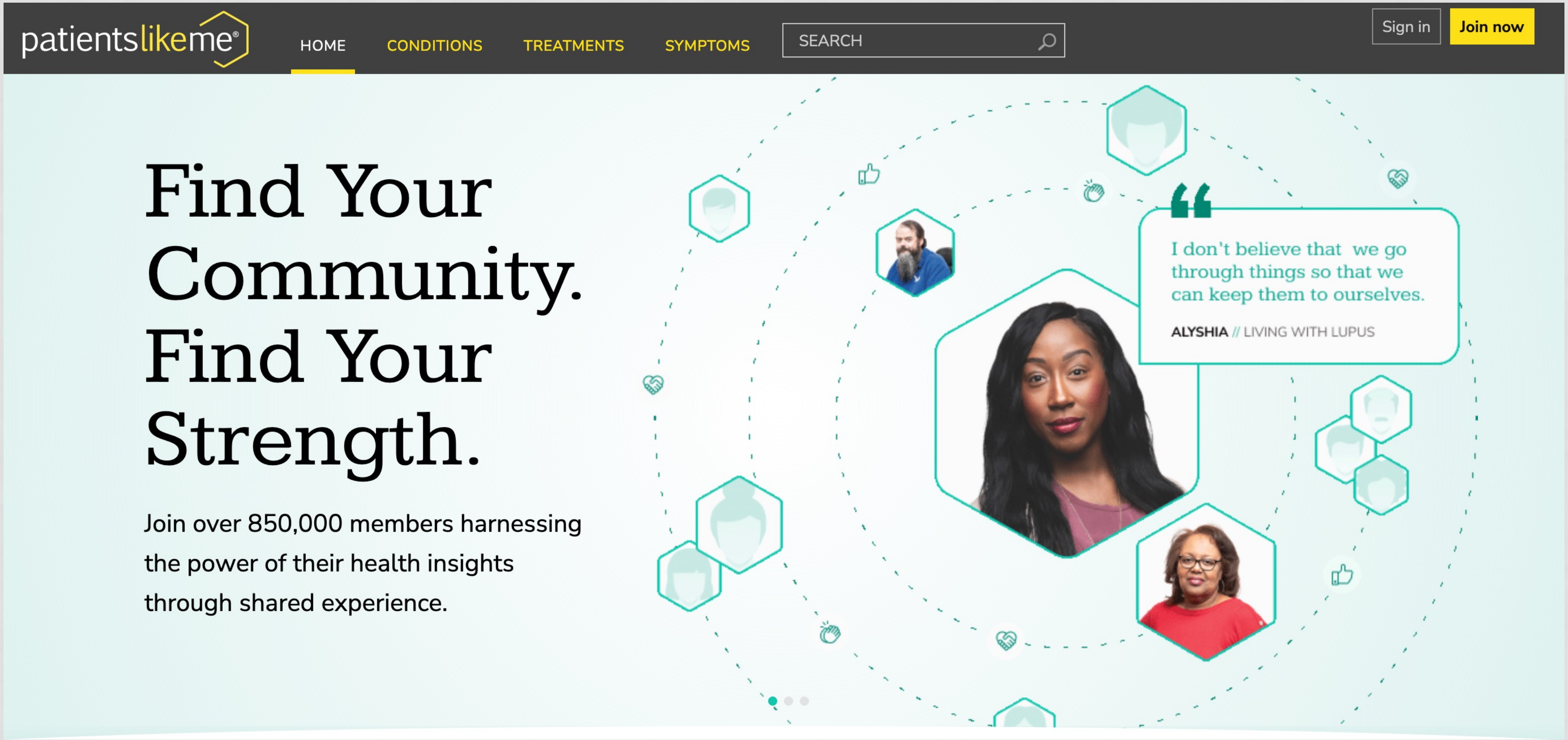
- Not only exponential growth of the scientific literature,
- but also of experimental data
    - E.g. for gene expression profiling or proteomics experiments, regulation of hundreds or thousands of genes and proteins is measured under multiple experimental conditions



High-throughput screening (**HTS**) is a method for scientific experimentation especially used in drug discovery and relevant to the fields of biology and chemistry.[1][2] Using robotics, data processing/control software, liquid handling devices, and sensitive detectors, high-throughput screening allows a researcher to quickly conduct millions of chemical, genetic, or pharmacological tests. Through this process one can rapidly identify active compounds, antibodies, or genes that modulate a particular biomolecular pathway. The results of these experiments provide starting points for drug design and for understanding the interaction or role of a particular biochemical process in biology.

- And of patent data, health social media, electronic health records, patient surveys

ONLINE PATIENT SUPPORT GROUPS



Find Your Community. Find Your Strength.

Join over 850,000 members harnessing the power of their health insights through shared experience.

"I don't believe that we go through things so that we can keep them to ourselves."

ALYSHIA // LIVING WITH LUPUS

GOALS

- Interactive knowledge discovery: assisting the expert in finding the information they need
- Text Mining can assist researchers in
  - finding, evaluating and interpreting the scientific literature and patented biomedical inventions
  - generating new medical hypothesis using information extracted from patient information (health records, social media data)

BIOMEDICAL RESEARCH QUESTIONS

THAT CAN BE ANSWERED WITH TM

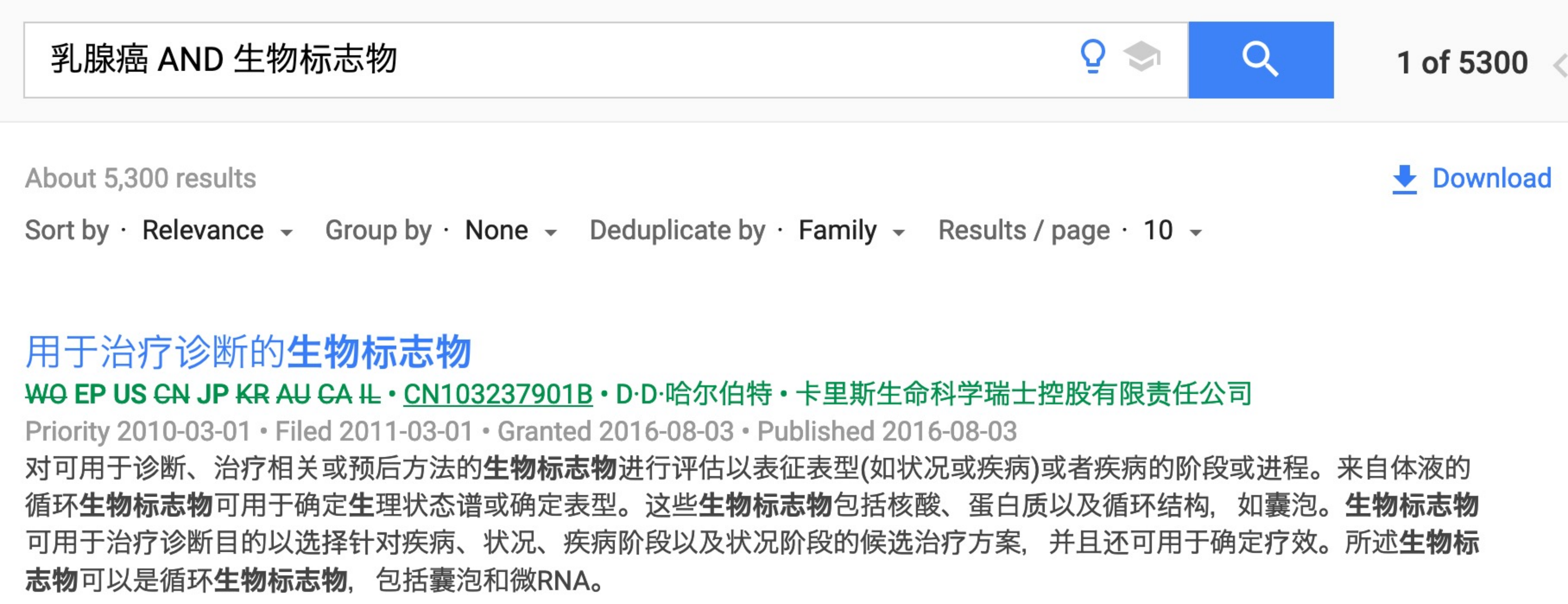
Universiteit Leiden

TEXT MINING FOR BIOMEDICAL RESEARCH

- Systematic reviewing: papers include/exclude
- Gene/protein/disease extraction
- Adverse events (side effects)
- Predictive models for electronic health records
  - Predicting diagnosis codes (discover misclassifications)
  - Predicting time-to-death
  - Predicting hospital discharge
  - Classifying the urgency of medical situations
- Drug interactions

GENE/PROTEIN/DISEASE EXTRACTION

- GNormPlus: a tool for tagging genes, gene families, and protein domains
- How about non-English texts?



乳腺癌 AND 生物标志物

1 of 5300

About 5,300 results

Download

Sort by · Relevance · Group by · None · Deduplicate by · Family · Results / page · 10

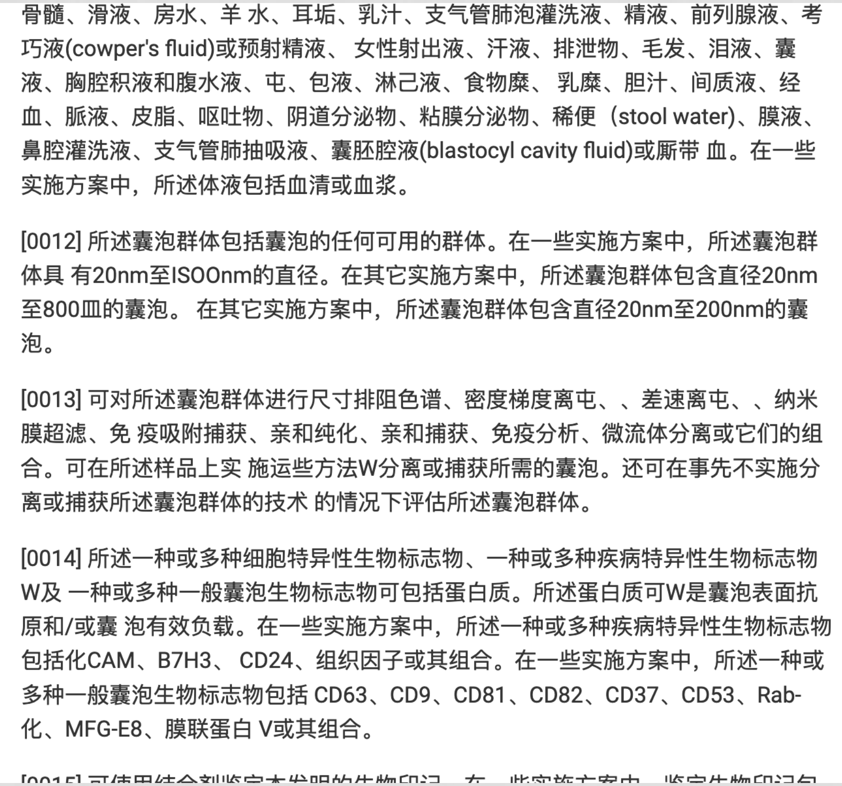
用于治疗诊断的**生物标志物**

~~WO~~ EP US ~~CN~~ JP ~~KR~~ ~~AU~~ ~~CA~~ ~~IL~~ • CN103237901B • D·D·哈尔伯特 • 卡里斯生命科学瑞士控股有限责任公司

Priority 2010-03-01 • Filed 2011-03-01 • Granted 2016-08-03 • Published 2016-08-03

对可用于诊断、治疗相关或预后方法的**生物标志物**进行评估以表征表型(如状况或疾病)或者疾病的阶段或进程。来自体液的循环**生物标志物**可用于确定**生**理状态谱或确定表型。这些**生物标志物**包括核酸、蛋白质以及循环结构，如囊泡。**生物标志物**可用于治疗诊断目的以选择针对疾病、状况、疾病阶段以及状况阶段的候选治疗方案，并且还可用于确定疗效。所述**生物标志物**可以是循环**生物标志物**，包括囊泡和微RNA。

GENE/PROTEIN/DISEASE EXTRACTION



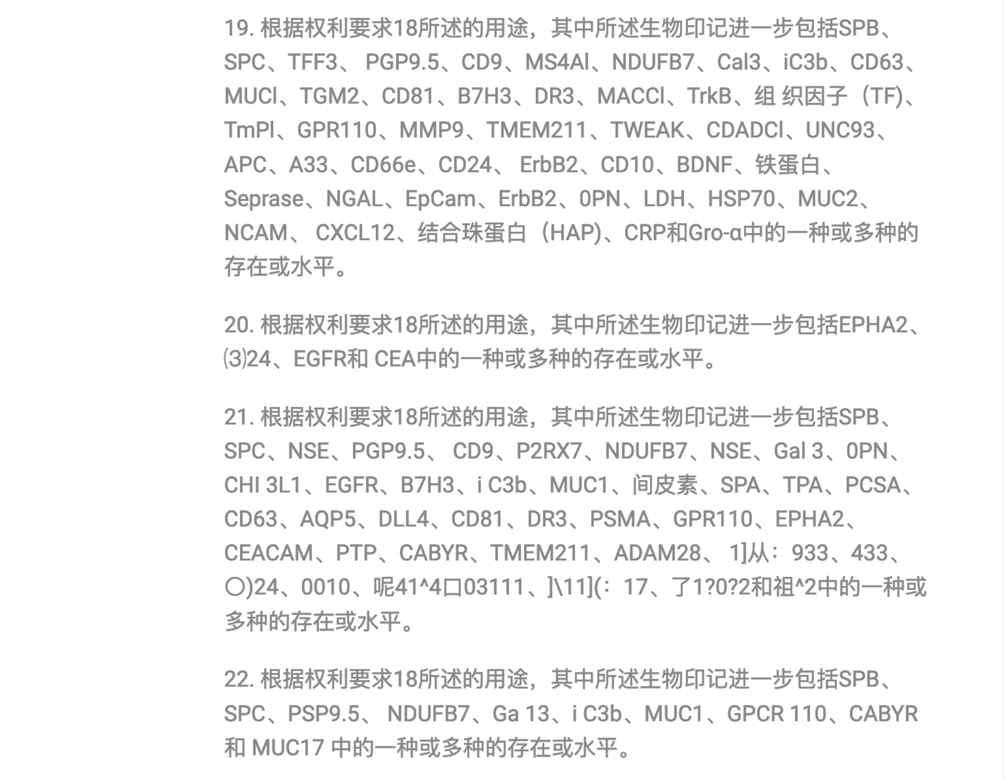
骨髓、滑液、房水、羊 水、耳垢、乳汁、支气管肺泡灌洗液、精液、前列腺液、考巧液(cowper's fluid)或预射精液、 女性射出液、汗液、排泄物、毛发、泪液、囊液、胸腔积液和腹水液、屯、包液、淋己液、食物糜、 乳糜、胆汁、间质液、经血、脈液、皮脂、呕吐物、阴道分泌物、粘膜分泌物、稀便（stool water)、膜液、鼻腔灌洗液、支气管肺抽吸液、囊胚腔液(blastocyl cavity fluid)或厮带 血。在一些实施方案中，所述体液包括血清或血浆。

[0012] 所述囊泡群体包括囊泡的任何可用的群体。在一些实施方案中，所述囊泡群体具 有20nm至ISOOnm的直径。在其它实施方案中，所述囊泡群体包含直径20nm至800皿的囊泡。 在其它实施方案中，所述囊泡群体包含直径20nm至200nm的囊泡。

[0013] 可对所述囊泡群体进行尺寸排阻色谱、密度梯度离屯、、差速离屯、、纳米膜超滤、免 疫吸附捕获、亲和纯化、亲和捕获、免疫分析、微流体分离或它们的组合。可在所述样品上实 施运些方法W分离或捕获所需的囊泡。还可在事先不实施分离或捕获所述囊泡群体的技术 的情况下评估所述囊泡群体。

[0014] 所述一种或多种细胞特异性生物标志物、一种或多种疾病特异性生物标志物W及 一种或多种一般囊泡生物标志物可包括蛋白质。所述蛋白质可W是囊泡表面抗原和/或囊 泡有效负载。在一些实施方案中，所述一种或多种疾病特异性生物标志物包括化CAM、B7H3、 CD24、组织因子或其组合。在一些实施方案中，所述一种或多种一般囊泡生物标志物包括 CD63、CD9、CD81、CD82、CD37、CD53、Rab-化、MFG-E8、膜联蛋白 V或其组合。

[0015] 可使用结合剂鉴定本发明的生物印记。在一些实施方案中，鉴定生物印记包



19. 根据权利要求18所述的用途，其中所述生物印记进一步包括SPB、SPC、TFF3、 PGP9.5、CD9、MS4Al、NDUFB7、Cal3、iC3b、CD63、MUCl、TGM2、CD81、B7H3、DR3、MACCl、TrkB、组 织因子（TF)、TmPl、GPR110、MMP9、TMEM211、TWEAK、CDADCl、UNC93、APC、A33、CD66e、CD24、 ErbB2、CD10、BDNF、铁蛋白、Seprase、NGAL、EpCam、ErbB2、0PN、LDH、HSP70、MUC2、NCAM、 CXCL12、结合珠蛋白（HAP)、CRP和Gro-α中的一种或多种的存在或水平。

20. 根据权利要求18所述的用途，其中所述生物印记进一步包括EPHA2、⑶24、EGFR和 CEA中的一种或多种的存在或水平。

21. 根据权利要求18所述的用途，其中所述生物印记进一步包括SPB、SPC、NSE、PGP9.5、 CD9、P2RX7、NDUFB7、NSE、Gal 3、0PN、CHI 3L1、EGFR、B7H3、i C3b、MUC1、间皮素、SPA、TPA、PCSA、CD63、AQP5、DLL4、CD81、DR3、PSMA、GPR110、EPHA2、CEACAM、PTP、CABYR、TMEM211、ADAM28、 1]从：933、433、〇)24、0010、呢41^4口03111、]\11](：17、了1?0?2和祖^2中的一种或多种的存在或水平。

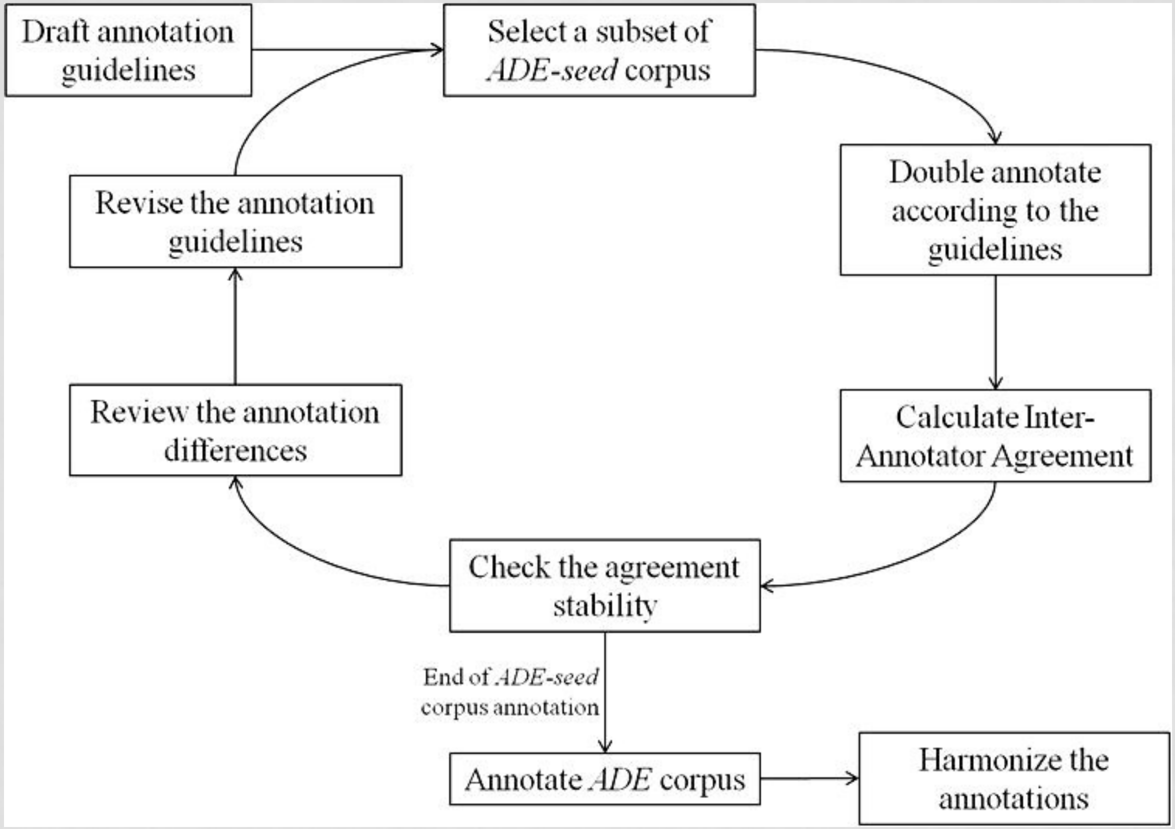
22. 根据权利要求18所述的用途，其中所述生物印记进一步包括SPB、SPC、PSP9.5、 NDUFB7、Ga 13、i C3b、MUC1、GPCR 110、CABYR和 MUC17 中的一种或多种的存在或水平。

Universiteit Leiden

ADVERSE EVENTS

- Mine adverse effects from medical case reports
- Adverse Drug Effect (ADE) benchmark corpus
  - a set of nearly 3000 case reports
  - manually annotated with 5063 drugs and 5776 conditions
  - + ontology of adverse events
- “With these methods, a number of drug label changes for the drugs rituximab, efalizumab, and natalizumab could successfully be predicted”

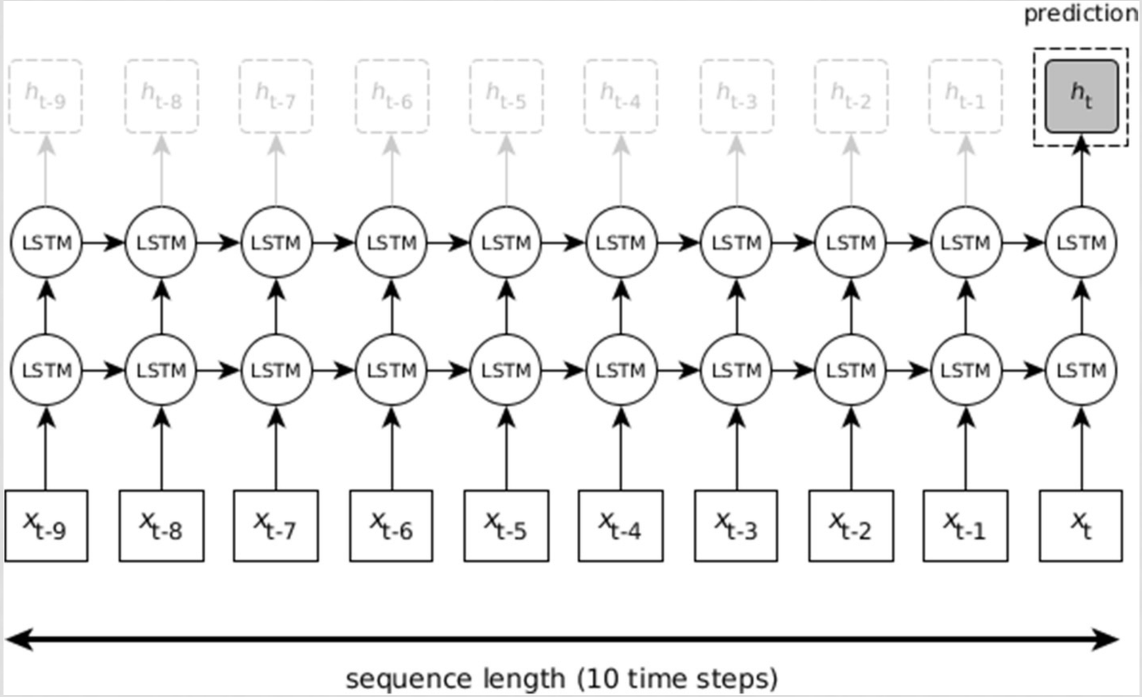
ADE BENCHMARK CORPUS


Draft annotation guidelines
Select a subset of *ADE-seed* corpus
Double annotate according to the guidelines
Calculate Inter-Annotator Agreement
Check the agreement stability
Review the annotation differences
Revise the annotation guidelines
End of *ADE-seed* corpus annotation
Annotate *ADE* corpus
Harmonize the annotations

H. Gurulingappa et al., Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports J. Biomed. Inf. 45 (5) (2012) 885–892.

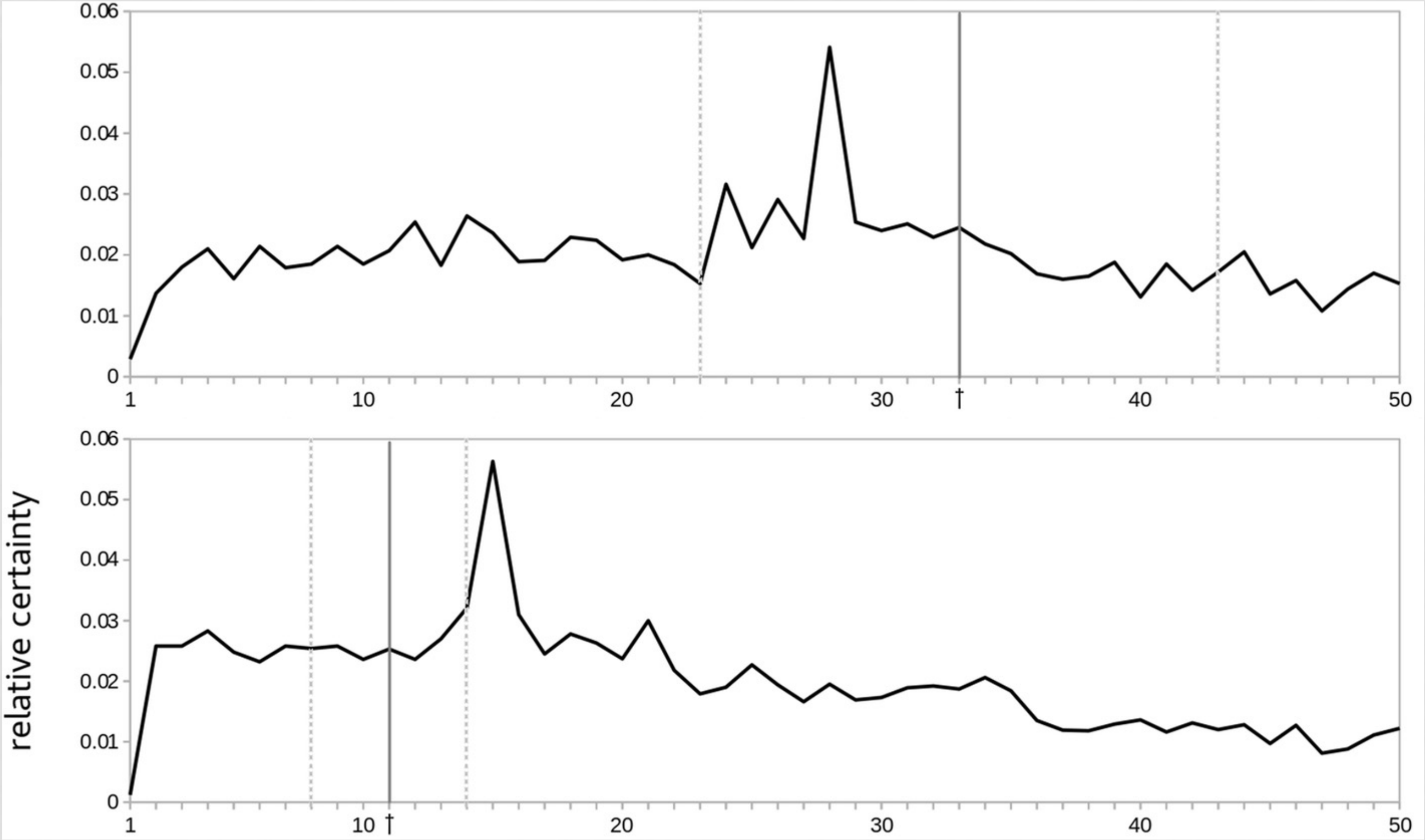
PREDICTIVE MODELS FOR EHRS

- "The model creates a probability distribution by predicting the chance that the end of life will occur during each specific month."



Merijn Beeksma, Suzan Verberne, Antal van den Bosch, Iris Hendrickx, Enny Das, Stef Groenewoud (2019). Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records. BMC Medical Informatics and Decision Making. 19:36. https://doi.org/10.1186/s12911-019-0775-2

PREDICTIVE MODELS ON EHRS



Suzan Verberne 2022

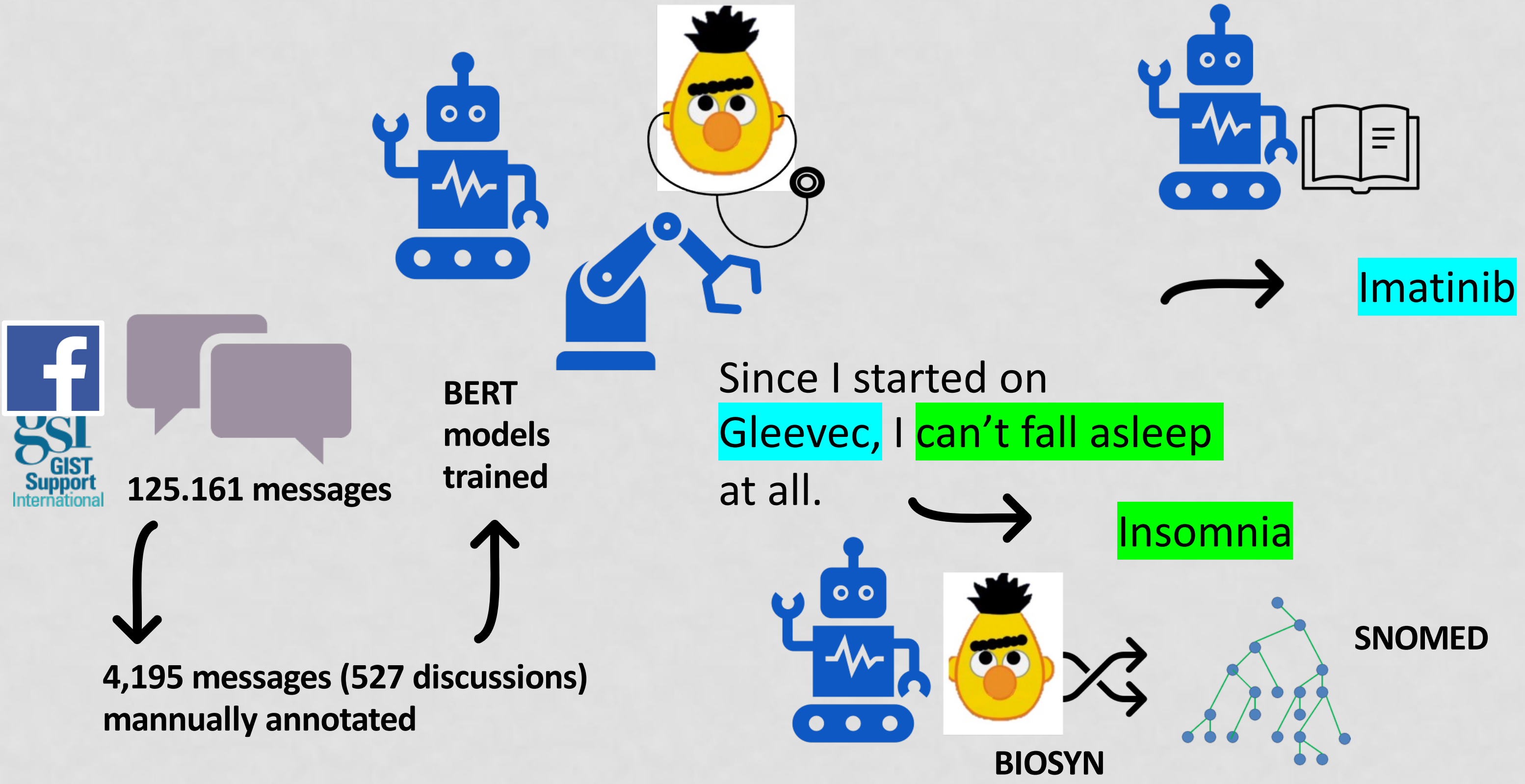
EXERCISE

WHAT IS NEEDED FOR BIO-TM?

- Task: find side effects for medications in online cancer patient forum discussions

- How would you approach this? What do you need?

WHAT IS NEEDED FOR BIO-TM?



125.161 messages

4,195 messages (527 discussions) mannually annotated

BERT models trained

Since I started on Gleevec, I can’t fall asleep at all.

Imatinib

Insomnia

BIOSYN

SNOMED

Universiteit Leiden

Anne Dirkson, Suzan Verberne, Wessel Kraaij, Gerard van Oortmerssen and Hans Gelderblom. Automated gathering of real-world data from online patient forums can complement pharmacovigilance for rare cancers. Nature Scientific Reports 12, 10317 (2022). https://doi.org/10.1038/s41598-022-13894-8

WHAT IS NEEDED FOR BIO-TM?

Steps to take

1. Filter the potentially relevant messages
2. Get/create training data for NER
3. Train an NER model to identify drug names and side effects in the messages
4. Normalize the side effects (map to ontology)
5. Relation extraction: cooccurrences of drug names and side effects in one message
6. (Match the found relations to an existing knowledge base to identify which relations are new)

Needed

- Lists/ontologies of drug names and known side effects (e.g. SNOMED CT)
- Pre-processing
- Pre-trained BERT models for NER and ontology linking
- Labelled data for supervised NER finetuning and evaluation

EVALUATION

Entity extraction (ADR)

Recall	74%
Precision	70%
F1	0.72
Human pairwise F1	0.80

Entity-ontology linking (SNOMED)

Accuracy@1	65%
Accuracy@5	79%

Model applied to the whole GIST forum

Treatment type	Drug	# of ADE found
First-line	Imatinib	13,376
Second-line	Sunitinib	2,335
Third-line	Regorafenib	319
Fourth-line	Ripretinib	319
PDGFRA exon 18 mutations	Avapritinib	297
Off-label	Nilotinib	59
Off-label	Pazopanib	51
Off-label	Sorafenib	47
Off-label	Ponatinib	17
	Unknown	2,948
	Total	**21,051**

Universiteit Leiden

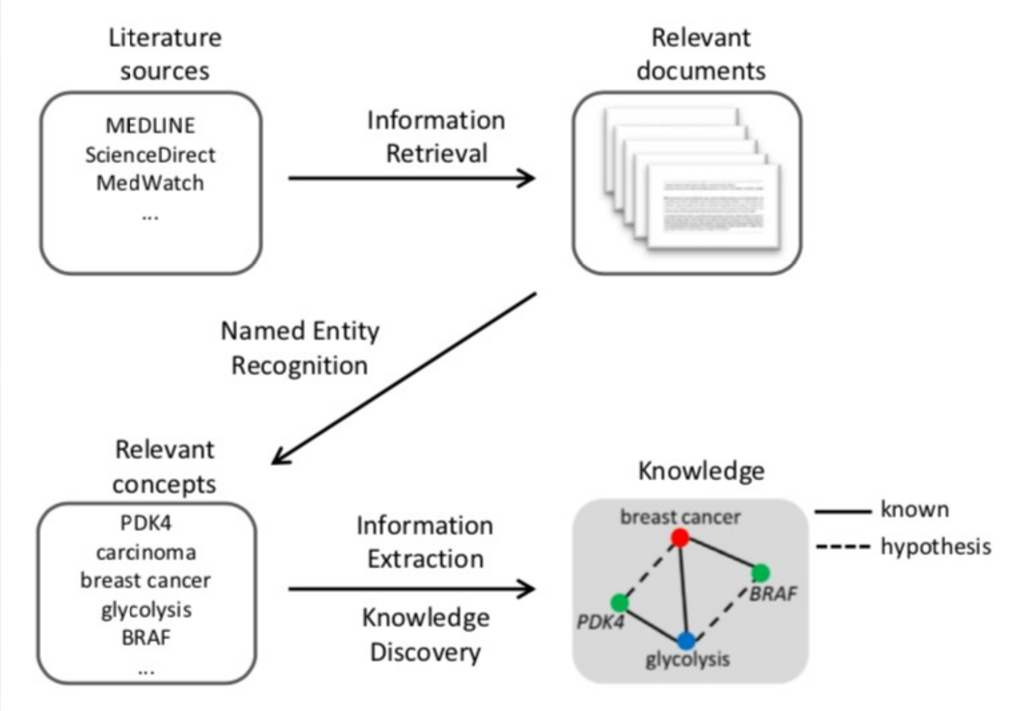
BIO-TM MODULES

Universiteit Leiden

Suzan Verberne 2022

MODULES IN BIO-TM

- Information Retrieval
- Named Entity Recognition
- Ontology linking
- Relation Extraction
- Knowledge Discovery
- Visualization

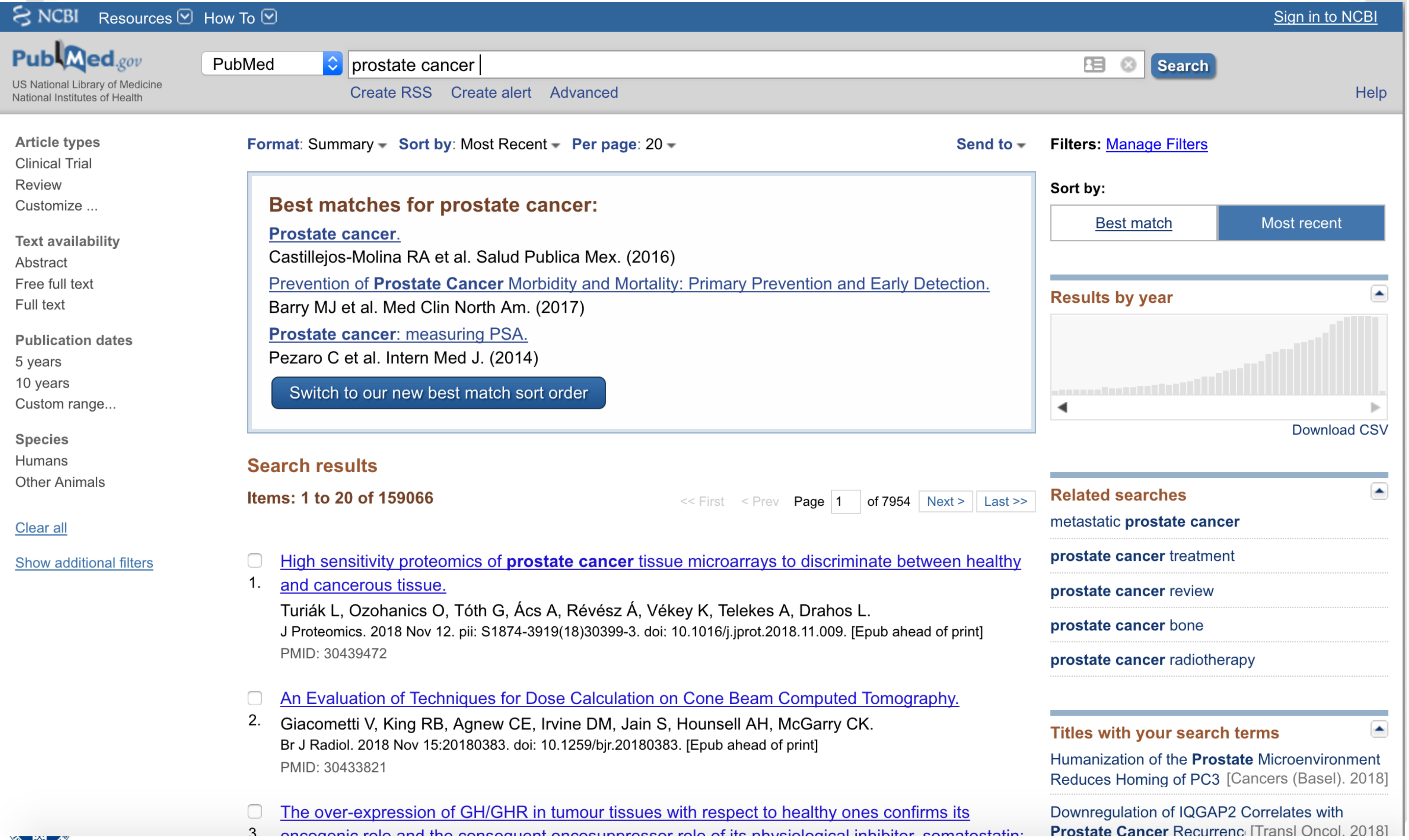


Fleuren and Alkema (2015). Application of text mining in the biomedical domain

INFORMATION RETRIEVAL

- Most used IR system: PubMed
  - Underlying database: MEDLINE
  - MEDLINE has full text papers and annotated abstracts with Medical Subject Heading (MeSH) Terms.
  - Search terms: formulated by expert (query)

PUBMED



NCBI Resources How To Sign in to NCBI

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed | prostate cancer | Search

Create RSS Create alert Advanced Help

Article types
Clinical Trial
Review
Customize ...

Text availability
Abstract
Free full text
Full text

Publication dates
5 years
10 years
Custom range...

Species
Humans
Other Animals

Clear all

Show additional filters

Format: Summary **Sort by**: Most Recent **Per page**: 20 **Send to**

Best matches for prostate cancer:

Prostate cancer.
Castillejos-Molina RA et al. Salud Publica Mex. (2016)

Prevention of **Prostate Cancer** Morbidity and Mortality: Primary Prevention and Early Detection.
Barry MJ et al. Med Clin North Am. (2017)

Prostate cancer: measuring PSA.
Pezaro C et al. Intern Med J. (2014)

Switch to our new best match sort order

Search results

Items: 1 to 20 of 159066

<< First < Prev Page 1 of 7954 Next > Last >>

□ 1. High sensitivity proteomics of **prostate cancer** tissue microarrays to discriminate between healthy and cancerous tissue.
Turiák L, Ozohanics O, Tóth G, Ács A, Révész Á, Vékey K, Telekes A, Drahos L.
J Proteomics. 2018 Nov 12. pii: S1874-3919(18)30399-3. doi: 10.1016/j.jprot.2018.11.009. [Epub ahead of print]
PMID: 30439472

□ 2. An Evaluation of Techniques for Dose Calculation on Cone Beam Computed Tomography.
Giacometti V, King RB, Agnew CE, Irvine DM, Jain S, Hounsell AH, McGarry CK.
Br J Radiol. 2018 Nov 15:20180383. doi: 10.1259/bjr.20180383. [Epub ahead of print]
PMID: 30433821

□ 3. The over-expression of GH/GHR in tumour tissues with respect to healthy ones confirms its oncogenic role and the consequent oncosuppressor role of its physiological inhibitor, somatostatin:

Filters: Manage Filters

Sort by:
Best match | Most recent

Results by year

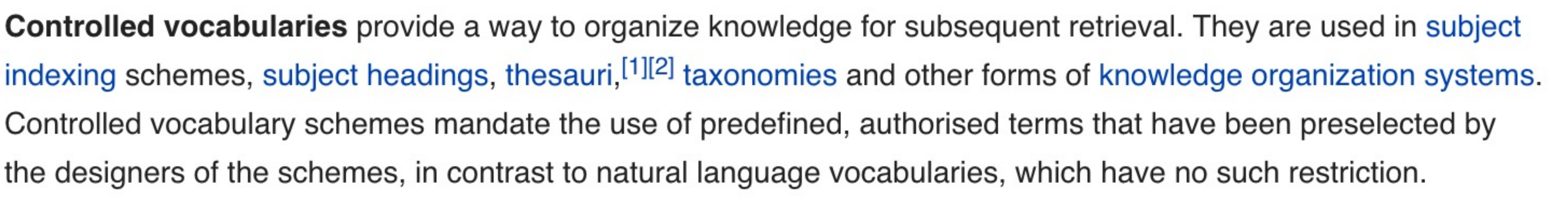
Download CSV

Related searches
metastatic **prostate cancer**
prostate cancer treatment
prostate cancer review
prostate cancer bone
prostate cancer radiotherapy

Titles with your search terms
Humanization of the **Prostate** Microenvironment Reduces Homing of PC3 [Cancers (Basel). 2018]
Downregulation of IQGAP2 Correlates with **Prostate Cancer** Recurrenc [Transl Oncol. 2018]

INFORMATION RETRIEVAL

- TM systems sometimes have advanced query options:
  - ‘concepts’: organizing similar keywords such as synonyms and alternative names into one concept based on a controlled vocabulary and subsequently incorporating all keywords of the same concept into the query.



Controlled vocabularies provide a way to organize knowledge for subsequent retrieval. They are used in subject indexing schemes, subject headings, thesauri,[1][2] taxonomies and other forms of knowledge organization systems. Controlled vocabulary schemes mandate the use of predefined, authorised terms that have been preselected by the designers of the schemes, in contrast to natural language vocabularies, which have no such restriction.

NAMED ENTITY RECOGNITION

Universiteit Leiden

Suzan Verberne 2022

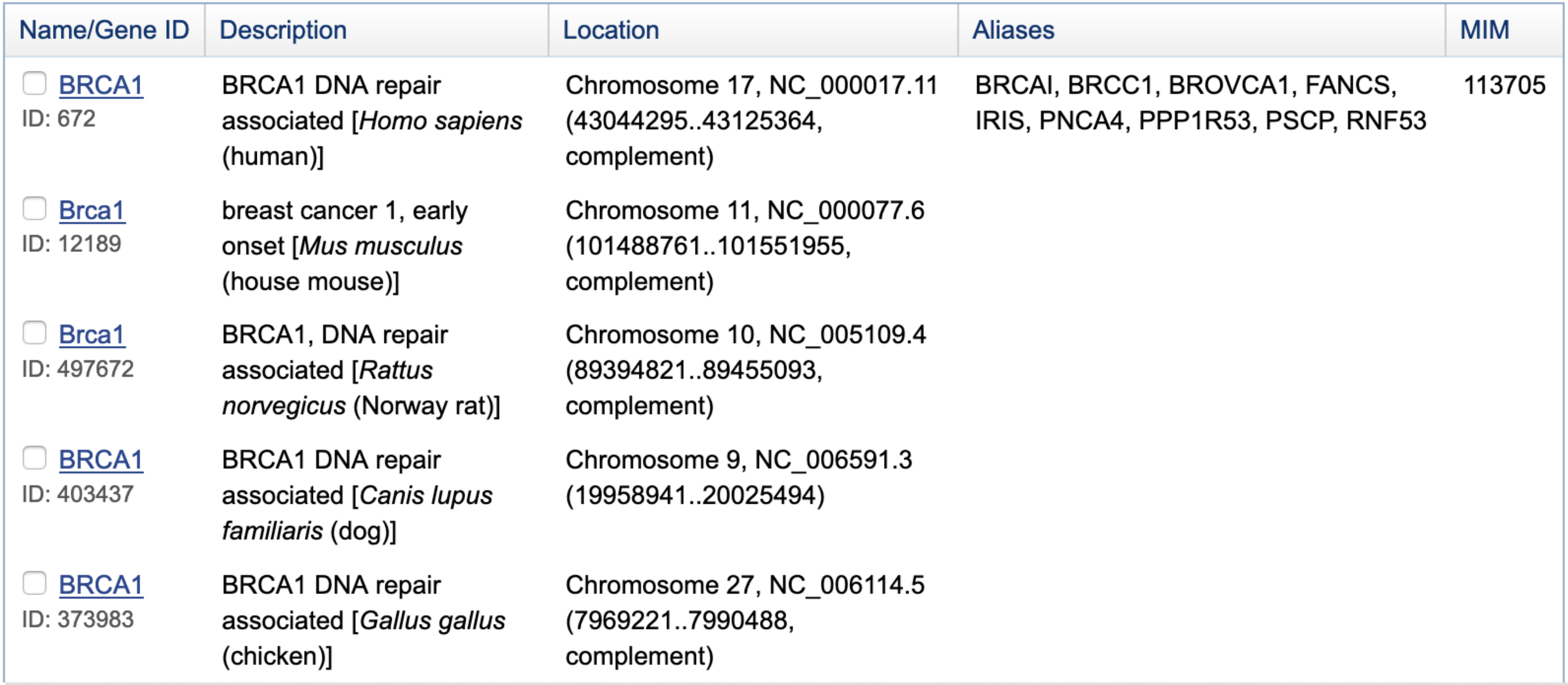
NAMED ENTITY RECOGNITION

- Identifying biomedical entities in retrieved documents
- Mentions of entities are highlighted and linked to the specific concept in a controlled vocabulary (thesaurus or ontology)
  - Unified Medical Language System (UMLS)
  - “The UMLS integrates and distributes key terminology, classification and coding standards, and associated resources to promote creation of more effective and interoperable biomedical information systems and services, including electronic health records.”
- Ambiguity and variation are challenges

BIO-NER

- One of the biggest challenges of bio-NER is the recognition of genes and protein names in scientific text
  - These are often described using different names and symbols and multiple genes share symbols and names
  - “Results from the gene normalization task of the BioCreative II contest underline this challenge, since none of the participating systems was able to correctly extract all human genes from a set of expert-curated MEDLINE abstracts”
    - (experts agreed in 90–95% of the cases)

GENE NAMES



Name/Gene ID	Description	Location	Aliases	MIM
□ BRCA1 ID: 672	BRCA1 DNA repair associated [*Homo sapiens* (human)]	Chromosome 17, NC_000017.11 (43044295..43125364, complement)	BRCAI, BRCC1, BROVCA1, FANCS, IRIS, PNCA4, PPP1R53, PSCP, RNF53	113705
□ Brca1 ID: 12189	breast cancer 1, early onset [*Mus musculus* (house mouse)]	Chromosome 11, NC_000077.6 (101488761..101551955, complement)		
□ Brca1 ID: 497672	BRCA1, DNA repair associated [*Rattus norvegicus* (Norway rat)]	Chromosome 10, NC_005109.4 (89394821..89455093, complement)		
□ BRCA1 ID: 403437	BRCA1 DNA repair associated [*Canis lupus familiaris* (dog)]	Chromosome 9, NC_006591.3 (19958941..20025494)		
□ BRCA1 ID: 373983	BRCA1 DNA repair associated [*Gallus gallus* (chicken)]	Chromosome 27, NC_006114.5 (7969221..7990488, complement)		

https://www.ncbi.nlm.nih.gov/gene

BIO-NER

- Benchmark data for biomedical NER

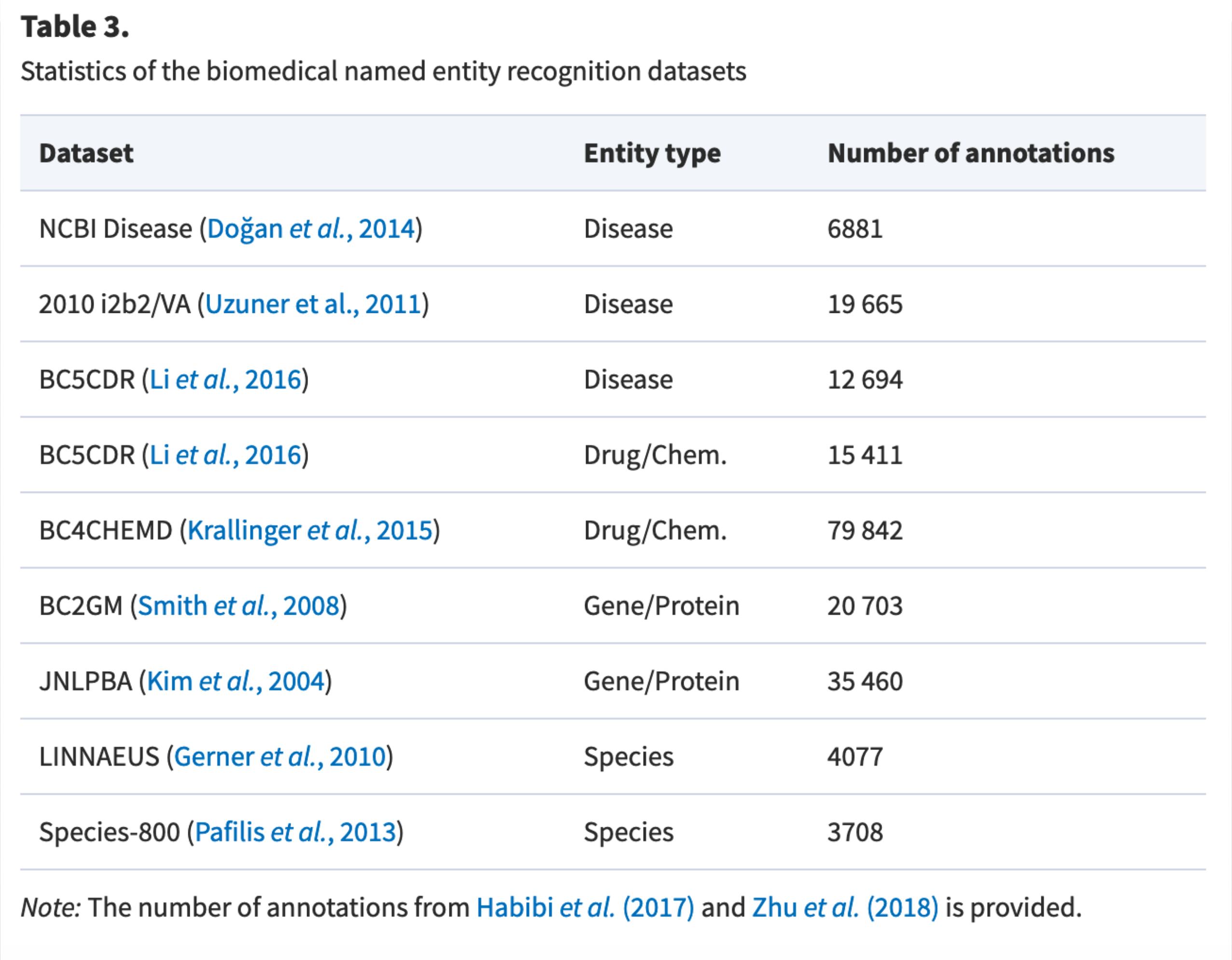


Table 3.

Statistics of the biomedical named entity recognition datasets

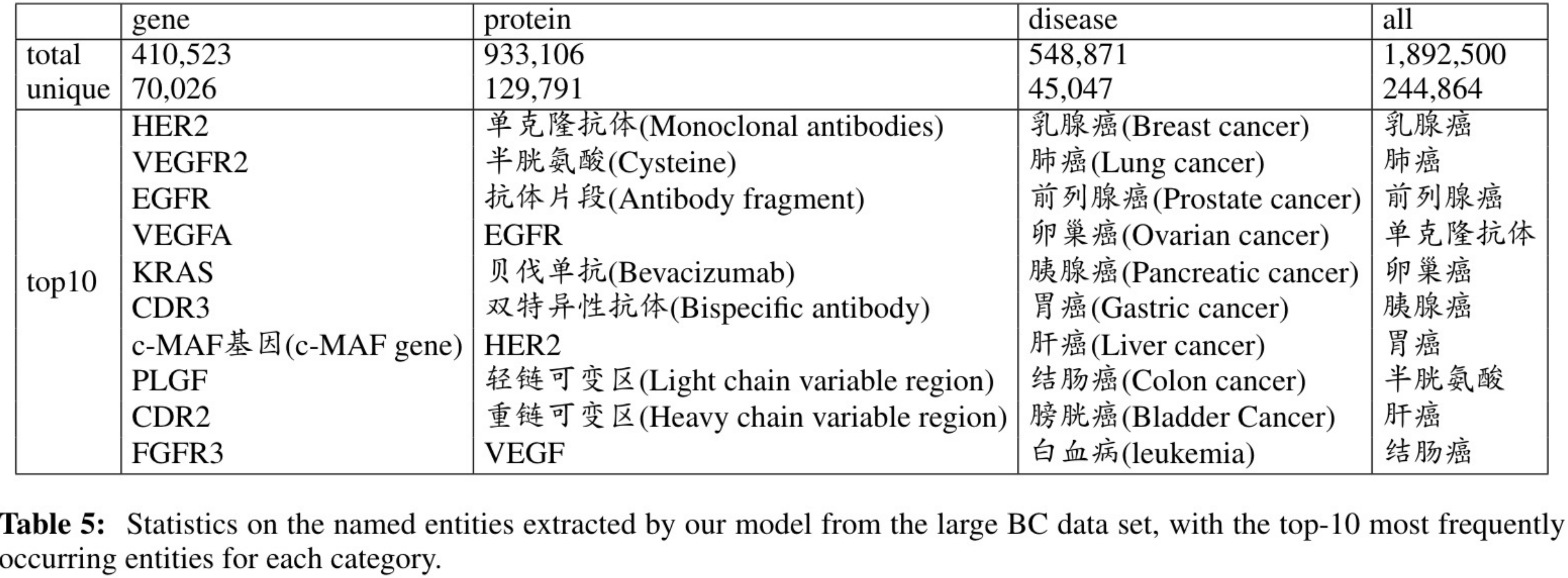
Dataset	Entity type	Number of annotations
NCBI Disease (Doğan *et al.*, 2014)	Disease	6881
2010 i2b2/VA (Uzuner et al., 2011)	Disease	19 665
BC5CDR (Li *et al.*, 2016)	Disease	12 694
BC5CDR (Li *et al.*, 2016)	Drug/Chem.	15 411
BC4CHEMD (Krallinger *et al.*, 2015)	Drug/Chem.	79 842
BC2GM (Smith *et al.*, 2008)	Gene/Protein	20 703
JNLPBA (Kim *et al.*, 2004)	Gene/Protein	35 460
LINNAEUS (Gerner *et al.*, 2010)	Species	4077
Species-800 (Pafilis *et al.*, 2013)	Species	3708

Note: The number of annotations from Habibi *et al.* (2017) and Zhu *et al.* (2018) is provided.

Lee et al (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. https://academic.oup.com/bioinformatics/article/36/4/1234/5566506

OTHER LANGUAGES THAN ENGLISH

Biomedical entities extracted from a Chinese patent dataset related to Breast cancer



	gene	protein	disease	all
total	410,523	933,106	548,871	1,892,500
unique	70,026	129,791	45,047	244,864
top10	HER2	单克隆抗体(Monoclonal antibodies)	乳腺癌(Breast cancer)	乳腺癌
	VEGFR2	半胱氨酸(Cysteine)	肺癌(Lung cancer)	肺癌
	EGFR	抗体片段(Antibody fragment)	前列腺癌(Prostate cancer)	前列腺癌
	VEGFA	EGFR	卵巢癌(Ovarian cancer)	单克隆抗体
	KRAS	贝伐单抗(Bevacizumab)	胰腺癌(Pancreatic cancer)	卵巢癌
	CDR3	双特异性抗体(Bispecific antibody)	胃癌(Gastric cancer)	胰腺癌
	c-MAF基因(c-MAF gene)	HER2	肝癌(Liver cancer)	胃癌
	PLGF	轻链可变区(Light chain variable region)	结肠癌(Colon cancer)	半胱氨酸
	CDR2	重链可变区(Heavy chain variable region)	膀胱癌(Bladder Cancer)	肝癌
	FGFR3	VEGF	白血病(leukemia)	结肠癌

Table 5: Statistics on the named entities extracted by our model from the large BC data set, with the top-10 most frequently occurring entities for each category.

Universiteit Leiden

Yuting Hu and Suzan Verberne (2020). Named Entity Recognition for Chinese biomedical patents. In the Proceedings of the 28th International Conference on Computational Linguistics (COLING)

RELATION EXTRACTION

RELATION EXTRACTION

- Co-occurrence-based methods
- NLP-based methods

RELATION EXTRACTION

- Co-occurrence based methods assume that two concepts that often occur together in the same text are related
  - E.g. the co-occurrence of retinol-binding protein 4 (RBP4) and insulin resistance in MEDLINE abstracts suggests a functional relationship between gene and disease

RELATION EXTRACTION

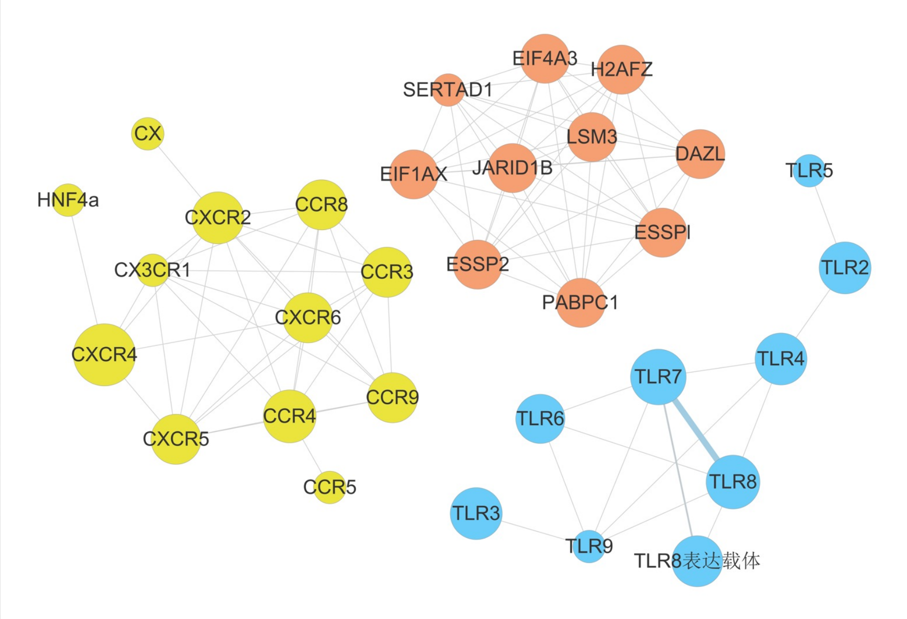
- Co-occurrence based methods assume that two concepts that often occur together in the same text are related
  - E.g. the co-occurrence of retinol-binding protein 4 (RBP4) and insulin resistance in MEDLINE abstracts suggests a functional relationship between gene and disease
- Statistics for co-occurrence frequencies:
  - actual number of cooccurrences
  - expected number of cooccurrences based on the frequencies of both entities
  - a statistical test to decide if the cooccurrence is statistically significant (e.g. Chi-square. Null hypothesis: they are independent)

RELATION EXTRACTION

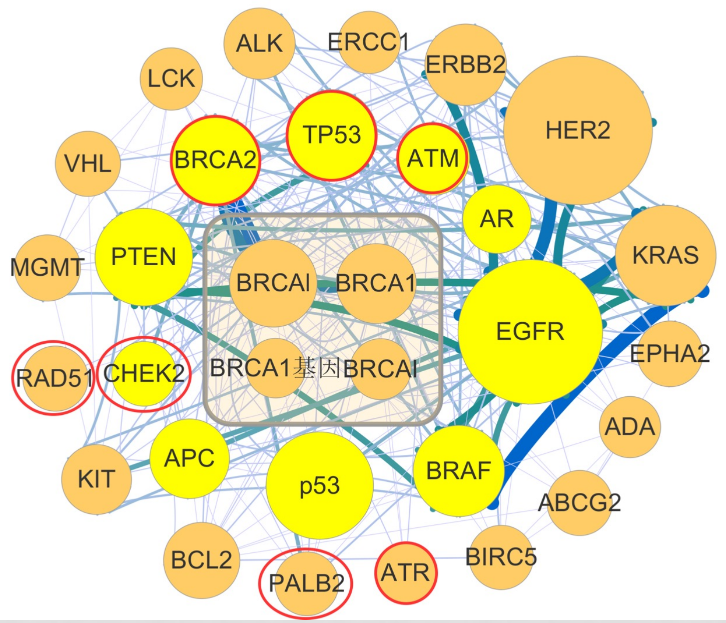
- Structure-based methods are phrase based and are able to detect triples in text e.g. gene A *inhibits* gene B or gene C *is involved in* disease G
  - Provides information about the type of relationship between two concepts
  - Structure-based methods often have a higher precision than co-occurrence based methods but lower recall (limited set of relations)

VISUALISATION

- Networks of entities and relations (in this case genes)



Part of the gene–gene connection network for the human gene dataset.

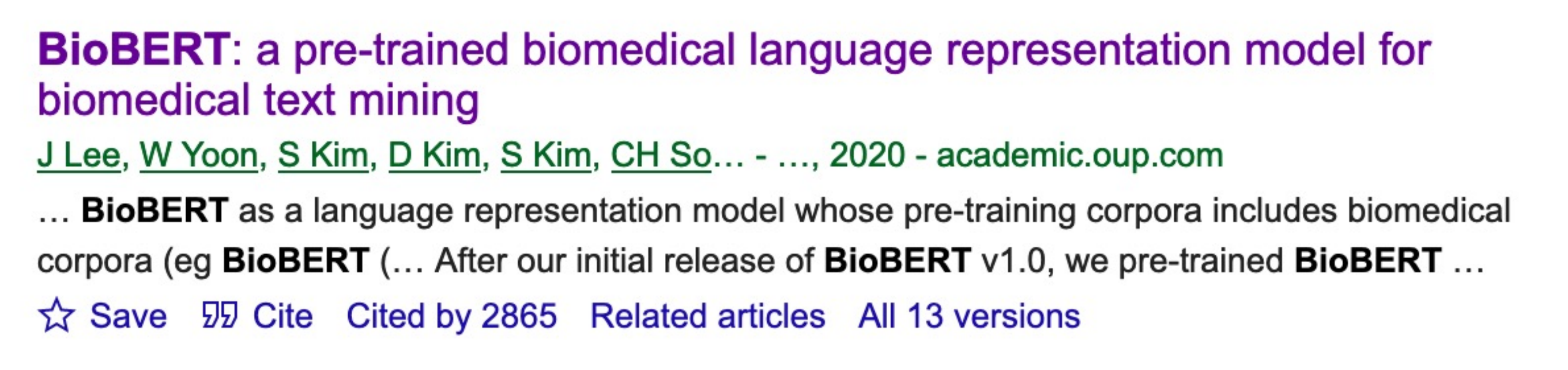


The BRCA1 gene network generated from our breast cancer dataset. Nodes in red circles are the nodes that are also related to BRCA1 in the STRING database.

Yuting Hu and Suzan Verberne (2020). Named Entity Recognition for Chinese biomedical patents. In the Proceedings of the 28th International Conference on Computational Linguistics (COLING)

STATE OF THE ART

DOMAIN-SPECIFIC MODELS



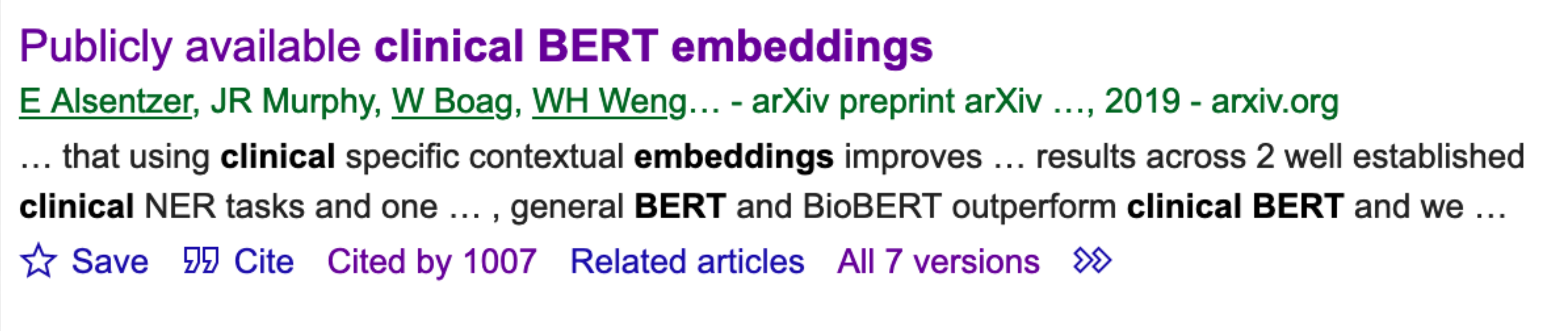
BioBERT: a pre-trained biomedical language representation model for biomedical text mining

J Lee, W Yoon, S Kim, D Kim, S Kim, CH So… - …, 2020 - academic.oup.com

… **BioBERT** as a language representation model whose pre-training corpora includes biomedical corpora (eg **BioBERT** (… After our initial release of **BioBERT** v1.0, we pre-trained **BioBERT** …

☆ Save ❞ Cite Cited by 2865 Related articles All 13 versions

- https://academic.oup.com/bioinformatics/article/36/4/1234/5566506



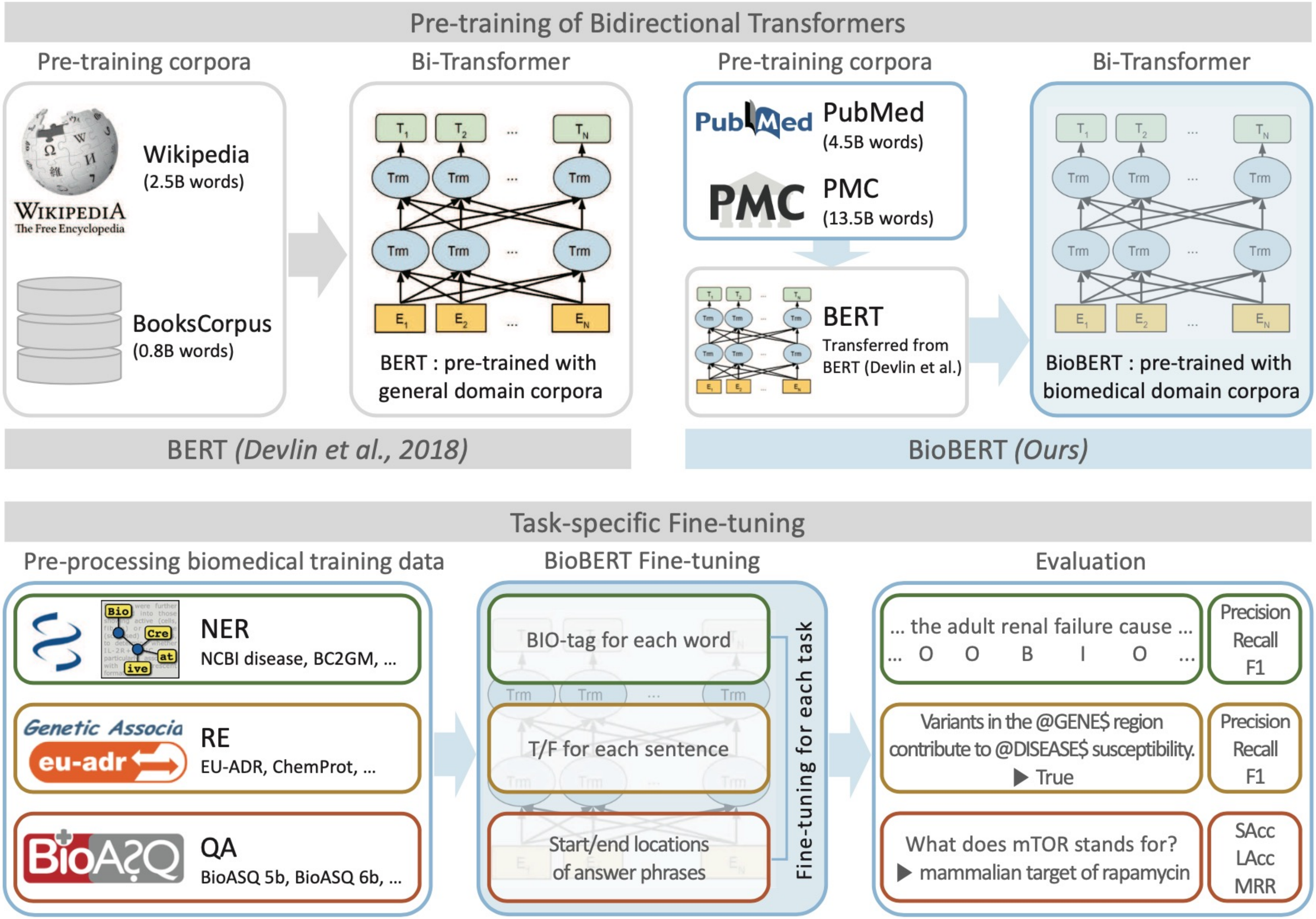
Publicly available **clinical BERT embeddings**

E Alsentzer, JR Murphy, W Boag, WH Weng… - arXiv preprint arXiv …, 2019 - arxiv.org

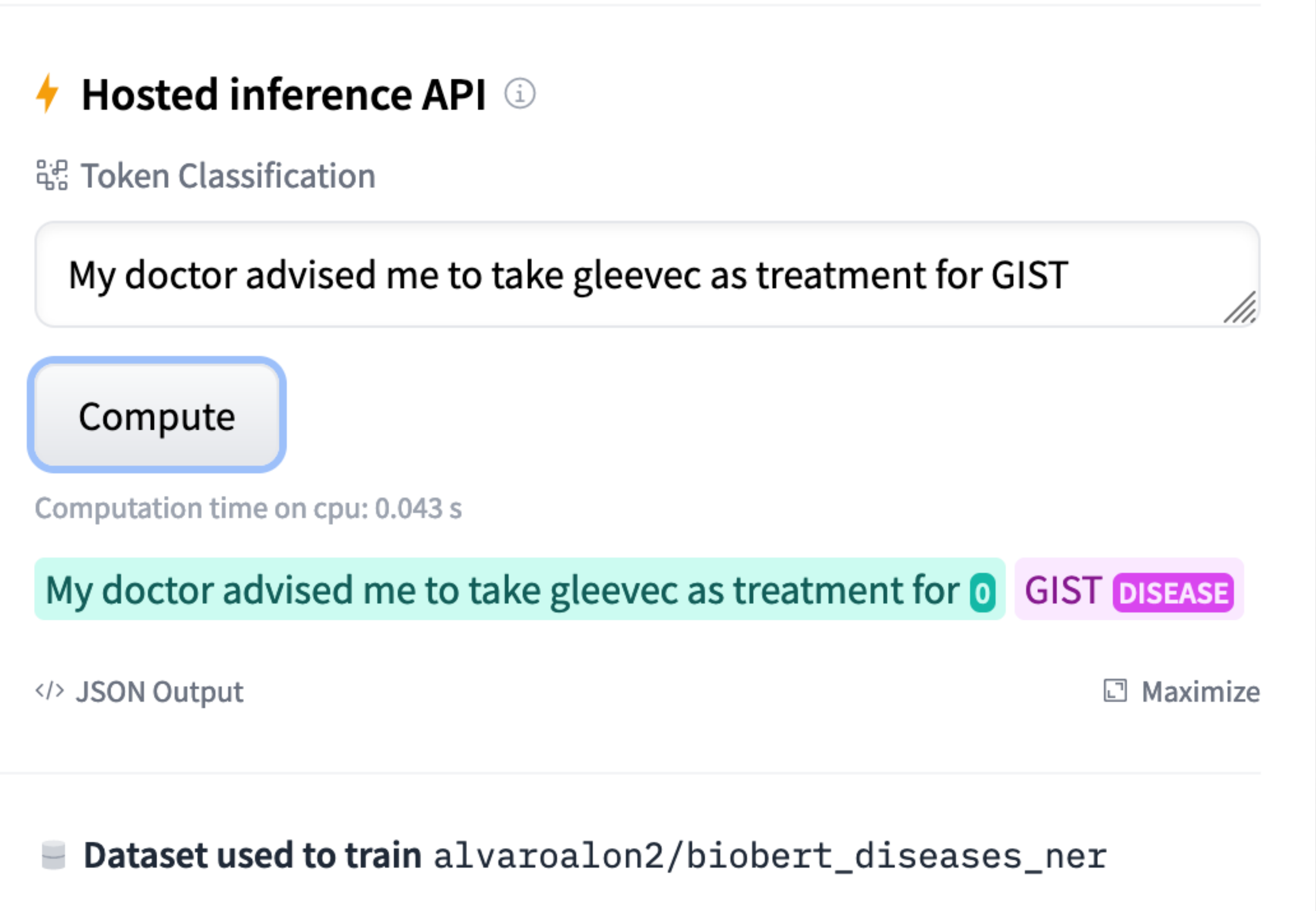
… that using **clinical** specific contextual **embeddings** improves … results across 2 well established **clinical** NER tasks and one … , general **BERT** and BioBERT outperform **clinical BERT** and we …

☆ Save ❞ Cite Cited by 1007 Related articles All 7 versions »

- https://arxiv.org/abs/1904.03323


Pre-training of Bidirectional Transformers
Pre-training corpora
Bi-Transformer
Pre-training corpora
Bi-Transformer
Wikipedia
(2.5B words)
WIKIPEDIA
The Free Encyclopedia
BooksCorpus
(0.8B words)
T1
T2
TN
Trm
E1
E2
EN
BERT : pre-trained with general domain corpora
PubMed
PubMed
(4.5B words)
PMC
PMC
(13.5B words)
BERT
Transferred from BERT (Devlin et al.)
BioBERT : pre-trained with biomedical domain corpora
BERT (Devlin et al., 2018)
BioBERT (Ours)
Task-specific Fine-tuning
Pre-processing biomedical training data
BioBERT Fine-tuning
Evaluation
NER
NCBI disease, BC2GM, ...
Bio
Cre
at
ive
RE
Genetic Associa
eu-adr
EU-ADR, ChemProt, ...
QA
BioASQ
BioASQ 5b, BioASQ 6b, ...
BIO-tag for each word
T/F for each sentence
Start/end locations of answer phrases
Fine-tuning for each task
... the adult renal failure cause ...
... O O B I O ...
Precision
Recall
F1
Variants in the @GENE$ region contribute to @DISEASE$ susceptibility.
▶ True
Precision
Recall
F1
What does mTOR stands for?
▶ mammalian target of rapamycin
SAcc
LAcc
MRR

EXAMPLE ON HUGGINGFACE



Hosted inference API

Token Classification

My doctor advised me to take gleevec as treatment for GIST

Compute

Computation time on cpu: 0.043 s

My doctor advised me to take gleevec as treatment for 0 GIST DISEASE

</> JSON Output

Maximize

Dataset used to train `alvaroalon2/biobert_diseases_ner`

DOMAIN-SPECIFIC MODELS

Differences in pre-training of domain-specific models

- Further pre-training vs pre-training from scratch
  - The collection has to be huge for pre-training from scratch
  - Therefore, domain-specific models are often further trained
  - This has an effect on the vocabulary of the model. Why?

  - WordPiece vocabulary is optimized for the pre-training corpus
    - BioBERT uses the BERT$_{BASE}$ vocabulary
    - Unknown terms are split in subwords:
      Immunoglobulin => I ##mm ##uno ##g ##lo ##bul ##in

FINAL ASSIGNMENT

ABOUT THE FINAL ASSIGNMENT

- Topics to choose from:
  - Text Classification
  - Information Extraction
  - Sentiment Analysis
- Data: provided by us
- Experiments: choose your own method. You can build on tutorials
- Report: 8 pages + max 2 pages for references and appendix (research paper). Advice: use LaTeX on Overleaf
  - Template suggestion: https://www.overleaf.com/latex/templates/springer-conference-proceedings-template-updated-2022-01-12/wcvbtmwtykqj

RESEARCH PAPER STRUCTURE

1. Introduction
2. Background/related work
3. Data
4. Methods
5. Results
6. Discussion
7. Conclusion
8. Contributions of the team members

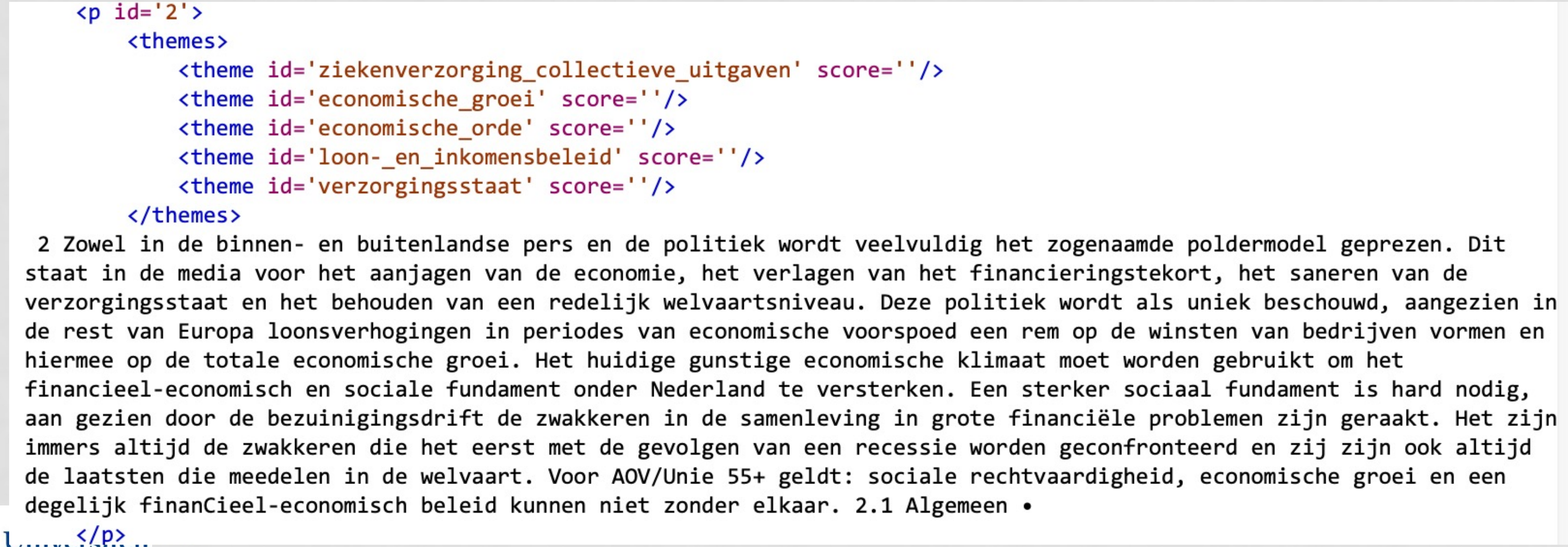
Grading criteria can be found on Brightspace: Assignments -> Criteria for final assignment

Don’t copy text from external sources! This is considered plagiarism and will be reported to the board of examiners

1. TEXT CLASSIFICATION

Multi-label classification of Dutch election political manifestos

- Data: Political_election_manifestos.zip (Brightspace)
- Paper: Verberne_2014_Automatic thematic classification of election manifestos.pdf (Brightspace)

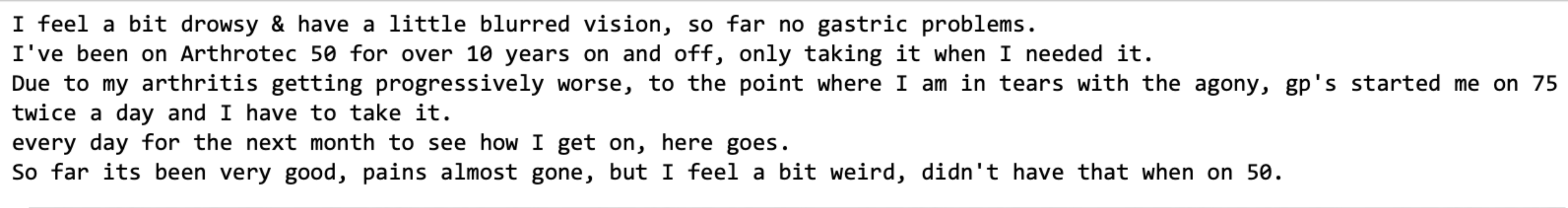


```
    <p id='2'>
        <themes>
            <theme id='ziekenverzorging_collectieve_uitgaven' score=''/>
            <theme id='economische_groei' score=''/>
            <theme id='economische_orde' score=''/>
            <theme id='loon-_en_inkomensbeleid' score=''/>
            <theme id='verzorgingsstaat' score=''/>
        </themes>
 2 Zowel in de binnen- en buitenlandse pers en de politiek wordt veelvuldig het zogenaamde poldermodel geprezen. Dit
staat in de media voor het aanjagen van de economie, het verlagen van het financieringstekort, het saneren van de
verzorgingsstaat en het behouden van een redelijk welvaartsniveau. Deze politiek wordt als uniek beschouwd, aangezien in
de rest van Europa loonsverhogingen in periodes van economische voorspoed een rem op de winsten van bedrijven vormen en
hiermee op de totale economische groei. Het huidige gunstige economische klimaat moet worden gebruikt om het
financieel-economisch en sociale fundament onder Nederland te versterken. Een sterker sociaal fundament is hard nodig,
aan gezien door de bezuinigingsdrift de zwakkeren in de samenleving in grote financiële problemen zijn geraakt. Het zijn
immers altijd de zwakkeren die het eerst met de gevolgen van een recessie worden geconfronteerd en zij zijn ook altijd
de laatsten die meedelen in de welvaart. Voor AOV/Unie 55+ geldt: sociale rechtvaardigheid, economische groei en een
degelijk finanCieel-economisch beleid kunnen niet zonder elkaar. 2.1 Algemeen •
    </p>
```

2. INFORMATION EXTRACTION

CSIRO Adverse Drug Event Corpus (Cadec)

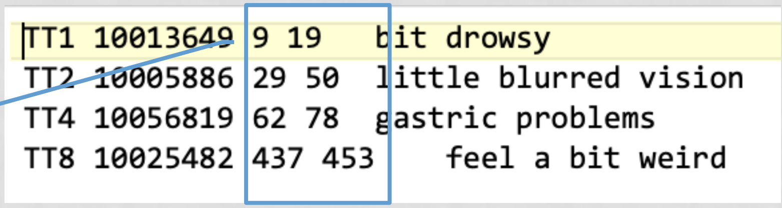
- Data: CADEC.v2.zip (Brightspace)
- Paper: Karimi_2015_Cadec- A corpus of adverse drug event annotations.pdf (Brightspace)



```
I feel a bit drowsy & have a little blurred vision, so far no gastric problems.
I've been on Arthrotec 50 for over 10 years on and off, only taking it when I needed it.
Due to my arthritis getting progressively worse, to the point where I am in tears with the agony, gp's started me on 75
twice a day and I have to take it.
every day for the next month to see how I get on, here goes.
So far its been very good, pains almost gone, but I feel a bit weird, didn't have that when on 50.
```

cadec/text/ARTHROTEC.1.txt

Transformation to IOB needed (using start and end character positions)



```
TT1 10013649 9 19    bit drowsy
TT2 10005886 29 50   little blurred vision
TT4 10056819 62 78   gastric problems
TT8 10025482 437 453     feel a bit weird
```

cadec/meddra/ARTHROTEC.1.ann

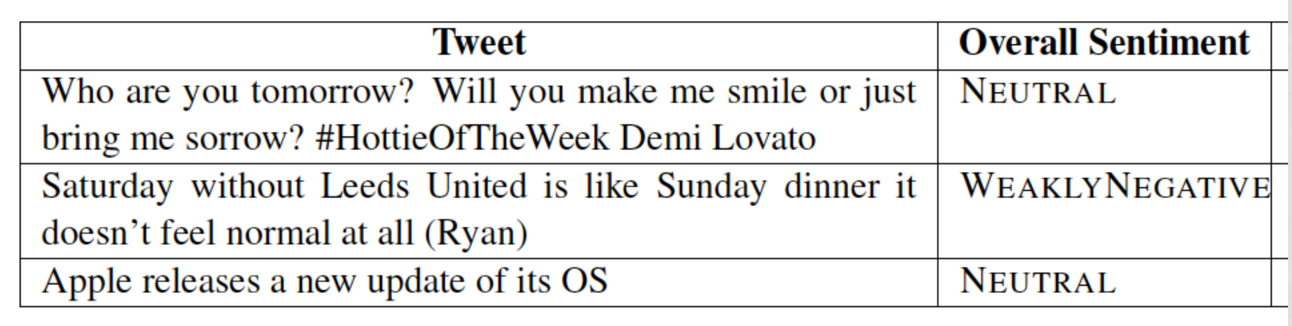
https://data.csiro.au/collections/collection/CIcsiro:10948/SQcadec

3. SENTIMENT ANALYSIS

SemEval 2017 Task 4: Sentiment Analysis in Twitter (subtask A)

- Data: semeval-2017-tweets_Subtask-A.zip (attached)
- Paper: Rosenthal_2017_SemEval-2017 Task 4- Sentiment Analysis in Twitter.pdf

- Subtask A: sentiment classification on a fivepoint scale



Tweet	Overall Sentiment
Who are you tomorrow? Will you make me smile or just bring me sorrow? #HottieOfTheWeek Demi Lovato	NEUTRAL
Saturday without Leeds United is like Sunday dinner it doesn't feel normal at all (Ryan)	WEAKLYNEGATIVE
Apple releases a new update of its OS	NEUTRAL

https://alt.qcri.org/semeval2017/task4/

OWN TOPIC

- You are allowed to propose your own topic. In that case, submit a proposal on or before December 1st:
  - One paragraph describing the task (with one or two references to papers)
  - Your research question(s)
  - A reference to the data and a table summarizing the data set size
- We will provide feedback on this

DEADLINES

- November 28: select a topic (choose one of three)
  - You don’t have to send this to us
- (December 1: proposal for your own topic; if you want to)
- December 7: online lab session for practical help with data/code
- December 13: submit a draft of your introduction, data, and method sections.
  - You receive 1 point out of the 10 for the final assignment by completing this step.
- January 8: submit the full paper
  - If your submission is late, then it will be counted as re-sit (maximum grade: 6). The re-sit deadline is February 8.
  - (Because of the grading deadline, I cannot push the deadline further away from the exam date)
- Everything can be submitted to the Brightspace item ‘Final assignment’

GENERAL GUIDELINES

1. Your introduction needs to contain a description of the task and your research questions. What is the problem and how will you solve it?
2. In your background section you describe a few relevant papers.
3. In your data section you provide a description and some statistics of the data. What are the labels and how are they distributed?
4. In your methods section you describe what you did and how
5. In your results section you provide clear tables with the results, and a description. Don’t forget a baseline comparison
6. Add relevant points of discussion (limitations, implications)
7. In the conclusion section you answer your research questions

CONCLUSIONS

SUZAN VERBERNE 2022

Universiteit Leiden

HOMEWORK

- Read:
  - Lee et al. (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining
- Final assignment:
  - Do you stay in the same team?
  - Choose a topic before November 28 or submit your own proposal before December 1.
- Next week: guest lecture by Vincent Slot from TextKernel about text mining in a commercial context

AFTER THIS LECTURE...

- You give three examples of biomedical research questions that can be answered with the help of text mining
- You can describe the text mining components that are needed for biomedical knowledge discovery
- You can explain the value of ontologies in the biomedical domain
- You can define supervised methods for biomedical entity recognition
- You can explain relation extraction using co-occurrences and natural language processing
- You can describe the training procedures for domain-specific BERT models
- You know what to expect from the final assignment