Advanced Parser for Biomedical Texts

Mikhail Skoblov,
Laboratory of Functional analysis of the Genome

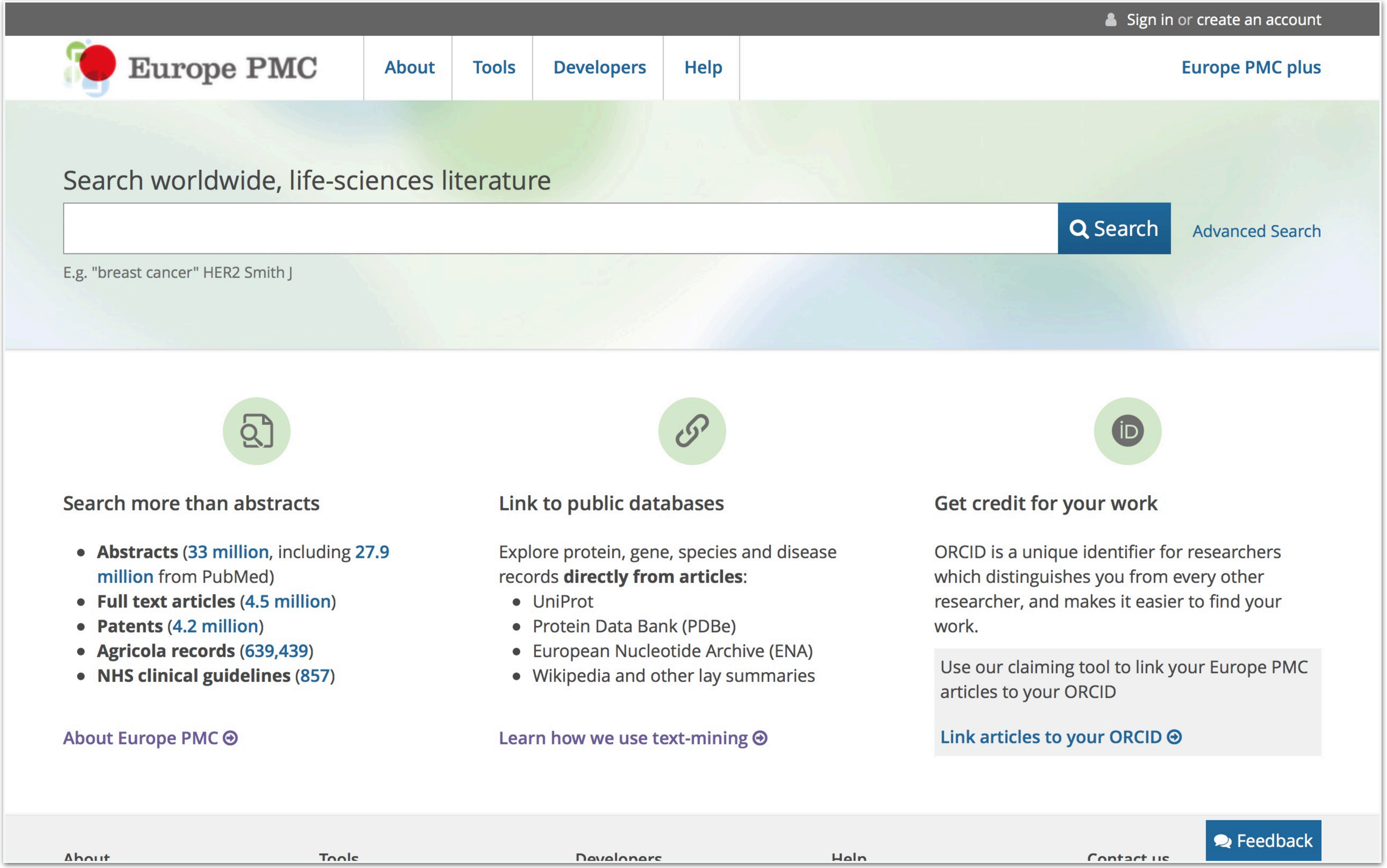
Anton Karazeev,
Department of Innovation and High Technologies

2017

Outline

- Task in general
- Existent methods
- Intermediate result: Information Gain method
- Further improvements
- References

Task in general



Search worldwide, life-sciences literature

E.g. "breast cancer" HER2 Smith J

Search more than abstracts

- **Abstracts** (33 million, including 27.9 million from PubMed)
- **Full text articles** (4.5 million)
- **Patents** (4.2 million)
- **Agricola records** (639,439)
- **NHS clinical guidelines** (857)

About Europe PMC

Link to public databases

Explore protein, gene, species and disease records **directly from articles**:

- UniProt
- Protein Data Bank (PDBe)
- European Nucleotide Archive (ENA)
- Wikipedia and other lay summaries

Learn how we use text-mining

Get credit for your work

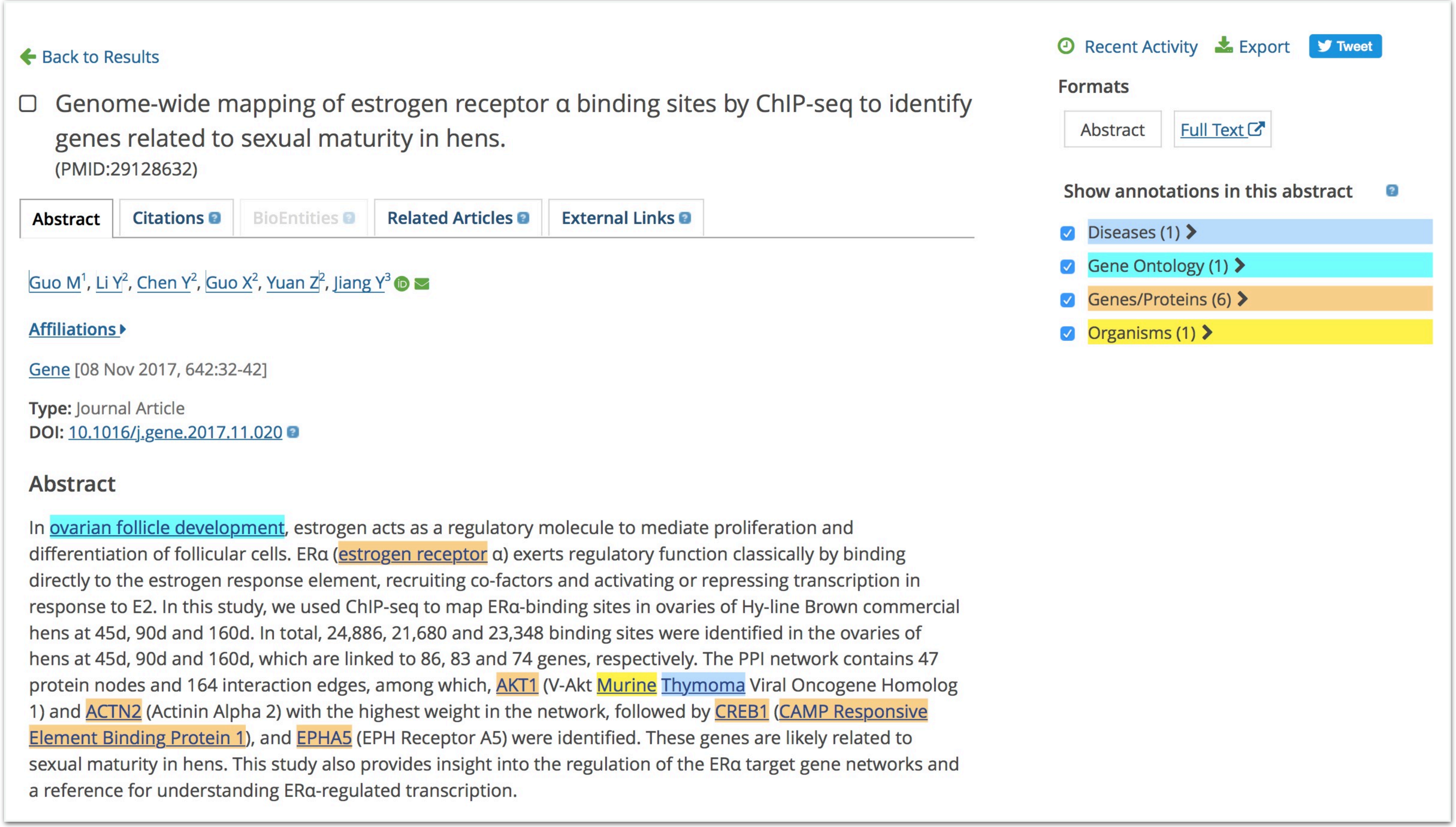
ORCID is a unique identifier for researchers which distinguishes you from every other researcher, and makes it easier to find your work.

Use our claiming tool to link your Europe PMC articles to your ORCID

Link articles to your ORCID

Europe PubMed Central - https://europepmc.org

Task in general



☐ Genome-wide mapping of estrogen receptor α binding sites by ChIP-seq to identify genes related to sexual maturity in hens.
(PMID:29128632)

Guo M[1], Li Y[2], Chen Y[2], Guo X[2], Yuan Z[2], Jiang Y[3]

Affiliations

Gene [08 Nov 2017, 642:32-42]

Type: Journal Article
DOI: 10.1016/j.gene.2017.11.020

Show annotations in this abstract

- ☑ Diseases (1)
- ☑ Gene Ontology (1)
- ☑ Genes/Proteins (6)
- ☑ Organisms (1)

Abstract

In ovarian follicle development, estrogen acts as a regulatory molecule to mediate proliferation and differentiation of follicular cells. ERα (estrogen receptor α) exerts regulatory function classically by binding directly to the estrogen response element, recruiting co-factors and activating or repressing transcription in response to E2. In this study, we used ChIP-seq to map ERα-binding sites in ovaries of Hy-line Brown commercial hens at 45d, 90d and 160d. In total, 24,886, 21,680 and 23,348 binding sites were identified in the ovaries of hens at 45d, 90d and 160d, which are linked to 86, 83 and 74 genes, respectively. The PPI network contains 47 protein nodes and 164 interaction edges, among which, AKT1 (V-Akt Murine Thymoma Viral Oncogene Homolog 1) and ACTN2 (Actinin Alpha 2) with the highest weight in the network, followed by CREB1 (CAMP Responsive Element Binding Protein 1), and EPHA5 (EPH Receptor A5) were identified. These genes are likely related to sexual maturity in hens. This study also provides insight into the regulation of the ERα target gene networks and a reference for understanding ERα-regulated transcription.

Europe PubMed Central - https://europepmc.org

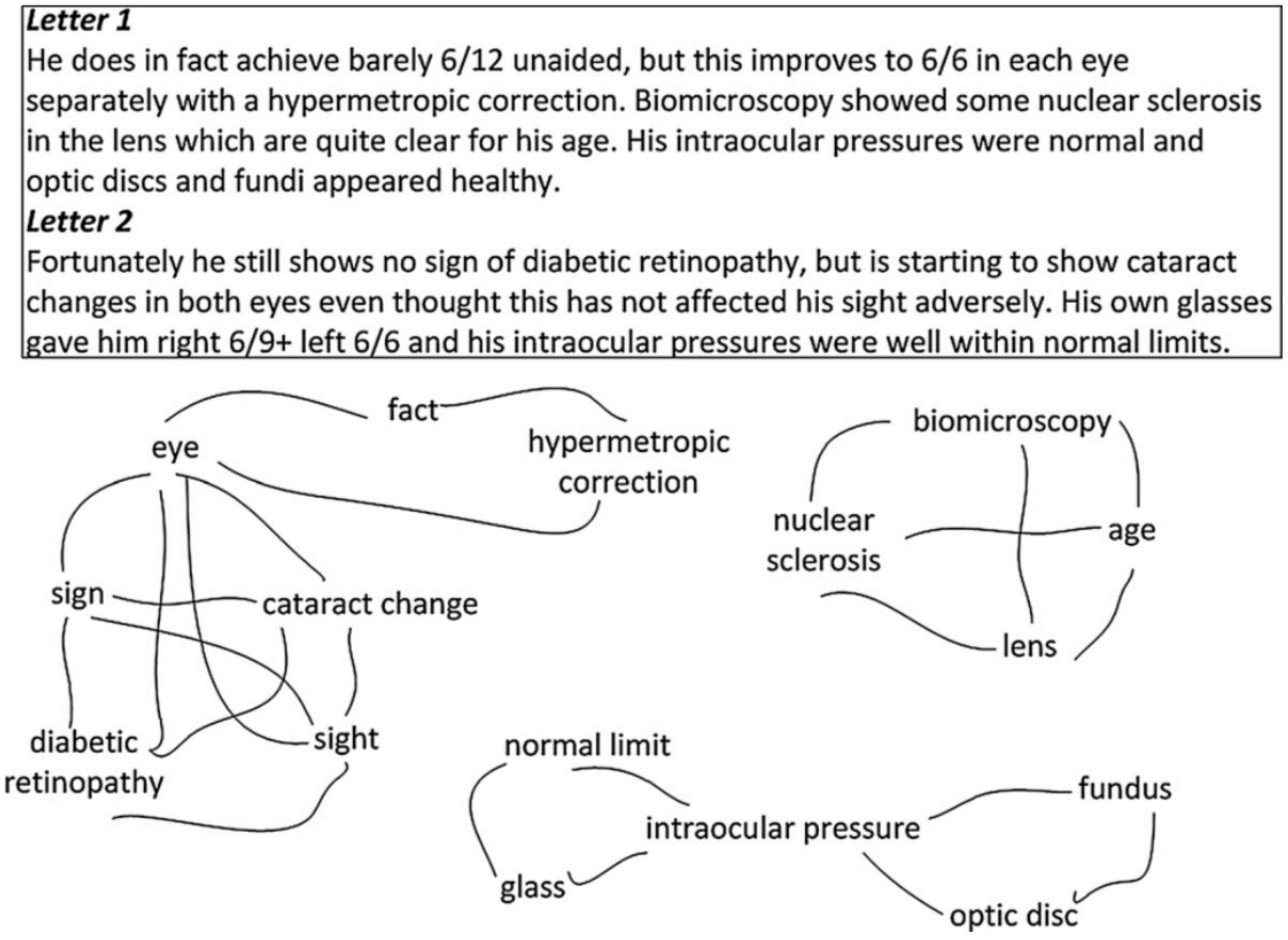
Existent methods

- TF-IDF
- TextRank* (PageRank)

* - R. Mihalcea and P. Tarau, *TextRank: Bringing Order into Texts,* 2004.

Existent methods: TextRank

$$S(v_i) = (1-d) + d \times \sum_{j \in in(v_i)} \frac{1}{|out(v_j)|} S(v_j)$$



Letter 1

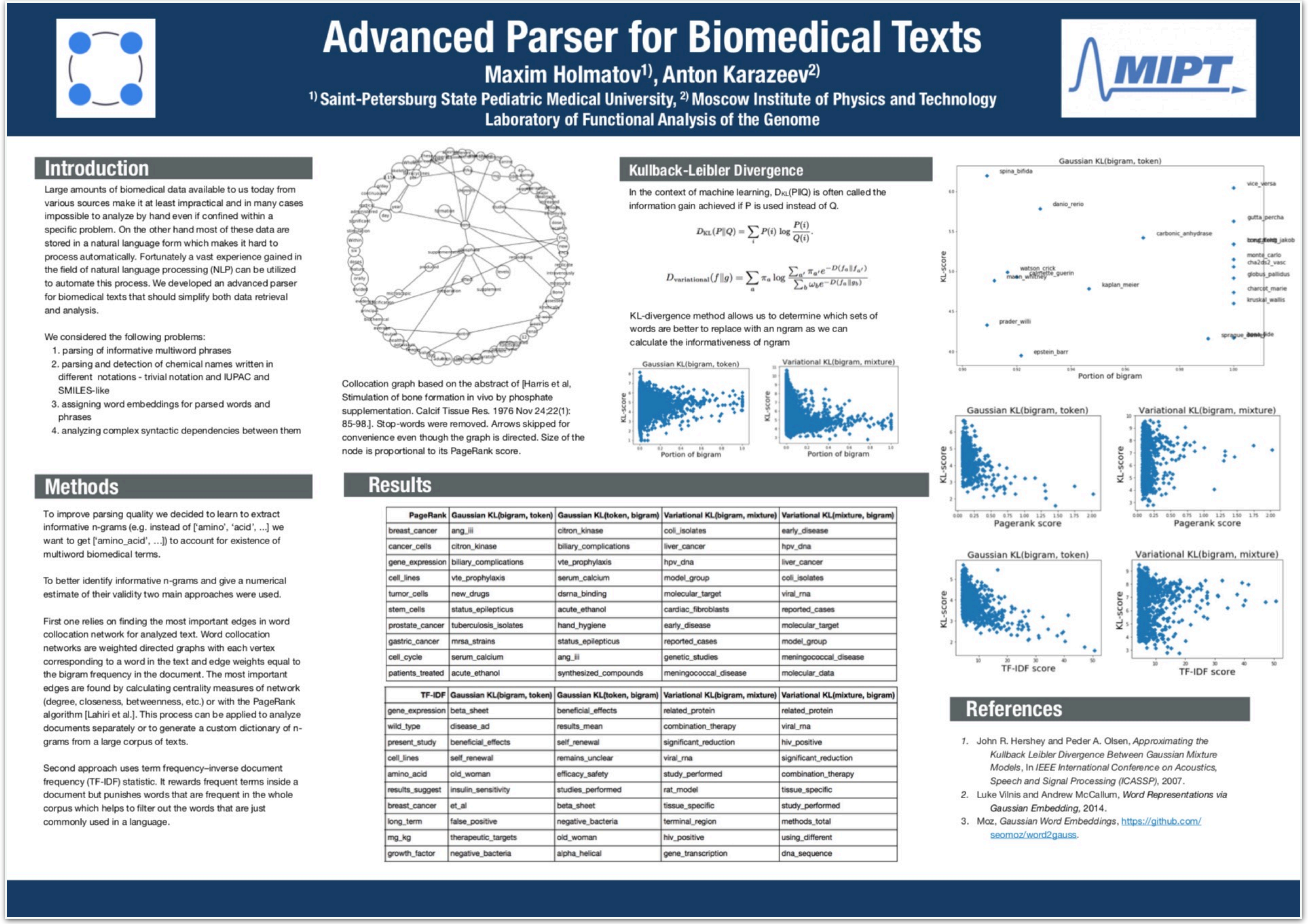
He does in fact achieve barely 6/12 unaided, but this improves to 6/6 in each eye separately with a hypermetropic correction. Biomicroscopy showed some nuclear sclerosis in the lens which are quite clear for his age. His intraocular pressures were normal and optic discs and fundi appeared healthy.

Letter 2

Fortunately he still shows no sign of diabetic retinopathy, but is starting to show cataract changes in both eyes even thought this has not affected his sight adversely. His own glasses gave him right 6/9+ left 6/6 and his intraocular pressures were well within normal limits.

W. Liu et. al., *A genetic algorithm enabled ensemble for unsupervised medical term extraction from clinical letters*, 2015.

Intermediate result: Information Gain method



Advanced Parser for Biomedical Texts

Maxim Holmatov[1], Anton Karazeev[2]
[1] Saint-Petersburg State Pediatric Medical University, [2] Moscow Institute of Physics and Technology
Laboratory of Functional Analysis of the Genome

Introduction

Large amounts of biomedical data available to us today from various sources make it at least impractical and in many cases impossible to analyze by hand even if confined within a specific problem. On the other hand most of these data are stored in a natural language form which makes it hard to process automatically. Fortunately a vast experience gained in the field of natural language processing (NLP) can be utilized to automate this process. We developed an advanced parser for biomedical texts that should simplify both data retrieval and analysis.

We considered the following problems:

1. parsing of informative multiword phrases
2. parsing and detection of chemical names written in different notations - trivial notation and IUPAC and SMILES-like
3. assigning word embeddings for parsed words and phrases
4. analyzing complex syntactic dependencies between them

Methods

To improve parsing quality we decided to learn to extract informative n-grams (e.g. instead of ['amino', 'acid', ...] we want to get ['amino_acid', ...]) to account for existence of multiword biomedical terms.

To better identify informative n-grams and give a numerical estimate of their validity two main approaches were used.

First one relies on finding the most important edges in word collocation network for analyzed text. Word collocation networks are weighted directed graphs with each vertex corresponding to a word in the text and edge weights equal to the bigram frequency in the document. The most important edges are found by calculating centrality measures of network (degree, closeness, betweenness, etc.) or with the PageRank algorithm [Lahiri et al.]. This process can be applied to analyze documents separately or to generate a custom dictionary of n-grams from a large corpus of texts.

Second approach uses term frequency–inverse document frequency (TF-IDF) statistic. It rewards frequent terms inside a document but punishes words that are frequent in the whole corpus which helps to filter out the words that are just commonly used in a language.

Collocation graph based on the abstract of [Harris et al, Stimulation of bone formation in vivo by phosphate supplementation. Calcif Tissue Res. 1976 Nov 24;22(1): 85-98.]. Stop-words were removed. Arrows skipped for convenience even though the graph is directed. Size of the node is proportional to its PageRank score.

Kullback-Leibler Divergence

In the context of machine learning, $D_{KL}(P\|Q)$ is often called the information gain achieved if P is used instead of Q.

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}.$$

$$D_{\mathrm{variational}}(f\|g) = \sum_a \pi_a \log \frac{\sum_{a'} \pi_{a'} e^{-D(f_a\|f_{a'})}}{\sum_b \omega_b e^{-D(f_a\|g_b)}}$$

KL-divergence method allows us to determine which sets of words are better to replace with an ngram as we can calculate the informativeness of ngram

Results

PageRank	Gaussian KL(bigram, token)	Gaussian KL(token, bigram)	Variational KL(bigram, mixture)	Variational KL(mixture, bigram)
breast_cancer	ang_iii	citron_kinase	coli_isolates	early_disease
cancer_cells	citron_kinase	biliary_complications	liver_cancer	hpv_dna
gene_expression	biliary_complications	vte_prophylaxis	hpv_dna	liver_cancer
cell_lines	vte_prophylaxis	serum_calcium	model_group	coli_isolates
tumor_cells	new_drugs	dsrna_binding	molecular_target	viral_rna
stem_cells	status_epilepticus	acute_ethanol	cardiac_fibroblasts	reported_cases
prostate_cancer	tuberculosis_isolates	hand_hygiene	early_disease	molecular_target
gastric_cancer	mrsa_strains	status_epilepticus	reported_cases	model_group
cell_cycle	serum_calcium	ang_iii	genetic_studies	meningococcal_disease
patients_treated	acute_ethanol	synthesized_compounds	meningococcal_disease	molecular_data

TF-IDF	Gaussian KL(bigram, token)	Gaussian KL(token, bigram)	Variational KL(bigram, mixture)	Variational KL(mixture, bigram)
gene_expression	beta_sheet	beneficial_effects	related_protein	related_protein
wild_type	disease_ad	results_mean	combination_therapy	viral_rna
present_study	beneficial_effects	self_renewal	significant_reduction	hiv_positive
cell_lines	self_renewal	remains_unclear	viral_rna	significant_reduction
amino_acid	old_woman	efficacy_safety	study_performed	combination_therapy
results_suggest	insulin_sensitivity	studies_performed	rat_model	tissue_specific
breast_cancer	et_al	beta_sheet	tissue_specific	study_performed
long_term	false_positive	negative_bacteria	terminal_region	methods_total
mg_kg	therapeutic_targets	old_woman	hiv_positive	using_different
growth_factor	negative_bacteria	alpha_helical	gene_transcription	dna_sequence

References

1. John R. Hershey and Peder A. Olsen, *Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models*, In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.
2. Luke Vilnis and Andrew McCallum, *Word Representations via Gaussian Embedding*, 2014.
3. Moz, *Gaussian Word Embeddings*, https://github.com/seomoz/word2gauss.

* Poster at MCCMB'17

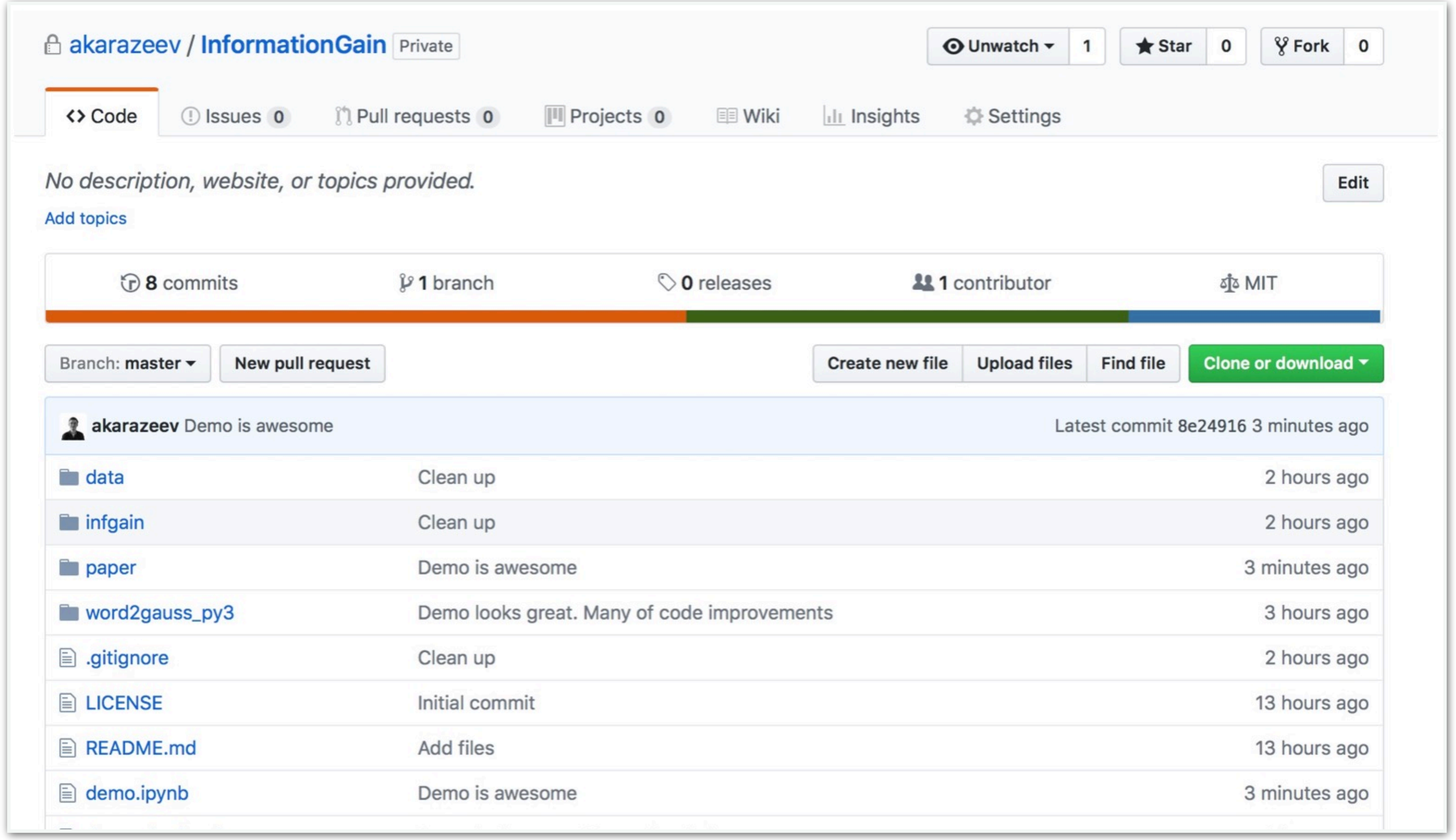
Intermediate result: Information Gain method

- Meaningful keyphrases extraction using Kullback-Leibler divergence-based method

- We called it Information Gain method:

Information Gain: new method of keyphrase extraction

Anton Karazeev[1], Ilia Korvigo[1], Mikhail Skoblov[1]



akarazeev / InformationGain Private

No description, website, or topics provided.

akarazeev Demo is awesome — Latest commit 8e24916 3 minutes ago

data	Clean up	2 hours ago
infgain	Clean up	2 hours ago
paper	Demo is awesome	3 minutes ago
word2gauss_py3	Demo looks great. Many of code improvements	3 hours ago
.gitignore	Clean up	2 hours ago
LICENSE	Initial commit	13 hours ago
README.md	Add files	13 hours ago
demo.ipynb	Demo is awesome	3 minutes ago

https://github.com/akarazeev/informationgain

Before tokens removing

	Closed KL(ngram, tilda)	Closed KL(tilda, ngram)	TF-IDF	TextRank	Variational KL(ngram, tilda)	Variational KL(tilda, ngram)
1	voltage gated	p nnuumm	uunnkk uunnkk	uunnkk effects	tumorigenic clones	db db
2	resonance energy	case series	uunnkk nnuumm	uunnkk plasma	glun1 glun2b	w w
3	direct suppression	nnuumm p	nnuumm uunnkk	nnuumm patients	ophthalmology journals	substantia nigra
4	tumor samples	therapy exposure	patients uunnkk	uunnkk show	egfp nachralpha3	tumorigenic clones
5	prenatal diagnosis	prenatal diagnosis	cells uunnkk	fetal cells	numbered tag	intestinal tract
6	serine threonine	mid log	uunnkk patients	induced transition	substantia nigra	numbered tag
7	n acetylcysteine	tumorigenic clones	uunnkk study	uunnkk studies	volar oblique	alfalfa peroxidase
8	antioxidant potential	l lactis	expression uunnkk	gene expression	elastin matrix	qbeta replicase
9	negative affectivity	resonance energy	study uunnkk	uunnkk cells	nursing home	hydroxymethylglutaryl coenzyme
10	therapy exposure	agarose gel	uunnkk cell	uunnkk rivaroxaban	regenerative medicine	transdermal buprenorphine
11	ovx zol	test whether	results uunnkk	suggested guidelines	hydroxymethylglutaryl coenzyme	lysinibacillus sp
12	light chain	tumor samples	uunnkk results	uunnkk use	lysinibacillus sp	oral contraceptives
13	pedot go	williams syndrome	used uunnkk	complete follow	hypocotyl elongation	tick borne
14	rs10468017 variant	foot protein	associated uunnkk	coagulation function	intestinal tract	distant metastases
15	tumorigenic clones	isolates type	uunnkk used	comparable uunnkk	carbonic anhydrase	lxxll motif
16	mirror neuron	neuron areas	activity uunnkk	uunnkk tumor	mirror neuron	glun1 glun2b
17	neuron areas	mosaic virus	compared uunnkk	related infections	charnley classification	bm examination
18	elevated umbilical	eec syndrome	levels uunnkk	clinical studies	voluntary interruptions	xl cgd
19	proteomic analysis	voltage gated	effects uunnkk	first data	moogoo udder	clinicaltrials gov
20	cultured fibroblasts	nci n78	role uunnkk	uunnkk results	try ends	multivariate logistic

After tokens removing

	Closed KL(ngram, tilda)	Closed KL(tilda, ngram)	TF-IDF	TextRank	Variational KL(ngram, tilda)	Variational KL(tilda, ngram)
1	voltage gated	case series	present study	fetal cells	tumorigenic clones	db db
2	resonance energy	therapy exposure	breast cancer	induced transition	glun1 glun2b	w w
3	direct suppression	prenatal diagnosis	wild type	gene expression	ophthalmology journals	substantia nigra
4	tumor samples	mid log	results suggest	suggested guidelines	egfp nachralpha3	tumorigenic clones
5	prenatal diagnosis	tumorigenic clones	gene expression	complete follow	numbered tag	intestinal tract
6	serine threonine	l lactis	cell lines	coagulation function	substantia nigra	numbered tag
7	n acetylcysteine	resonance energy	long term	related infections	volar oblique	alfalfa peroxidase
8	antioxidant potential	agarose gel	cell cycle	clinical studies	elastin matrix	qbeta replicase
9	negative affectivity	test whether	risk factors	first data	nursing home	hydroxymethylglutaryl coenzyme
10	therapy exposure	tumor samples	polymerase chain	pathway conduction	regenerative medicine	transdermal buprenorphine
11	ovx zol	williams syndrome	growth factor	treated seeds	hydroxymethylglutaryl coenzyme	lysinibacillus sp
12	light chain	foot protein	chain reaction	gene increase	lysinibacillus sp	oral contraceptives
13	pedot go	isolates type	significant difference	similar trends	hypocotyl elongation	tick borne
14	rs10468017 variant	neuron areas	cell death	higher levels	intestinal tract	distant metastases
15	tumorigenic clones	mosaic virus	cancer cells	investigated lesions	carbonic anhydrase	lxxll motif
16	mirror neuron	eec syndrome	mg kg	bone marrow	mirror neuron	glun1 glun2b
17	neuron areas	voltage gated	cell line	cu b	charnley classification	bm examination
18	elevated umbilical	nci n78	significantly higher	molecular computing	voluntary interruptions	xl cgd
19	proteomic analysis	distant metastasis	important role	zokor species	moogoo udder	clinicaltrials gov
20	cultured fibroblasts	clinical importance	cell surface	affected organs	try ends	multivariate logistic

After tokens removing

	Closed KL(ngram, tilda)	Closed KL(tilda, ngram)	TF-IDF	TextRank	Variational KL(ngram, tilda)	Variational KL(tilda, ngram)
1	voltage gated	case series	present study	fetal cells	tumorigenic clones	db db
2	resonance energy	therapy exposure	breast cancer	induced transition	glun1 glun2b	w w
3	direct suppression	prenatal diagnosis	wild type	gene expression	ophthalmology journals	substantia nigra
4	tumor samples	mid log	results suggest	suggested guidelines	egfp nachralpha3	tumorigenic clones
5	prenatal diagnosis	tumorigenic clones	gene expression	complete follow	numbered tag	intestinal tract
6	serine threonine	l lactis	cell lines	coagulation function	substantia nigra	numbered tag
7	n acetylcysteine	resonance energy			volar oblique	alfalfa peroxidase
8	antioxidant potential	agarose gel			elastin matrix	qbeta replicase
9	negative affectivity	test whether	risk factors	first data	nursing home	hydroxymethylglutaryl coenzyme
10	therapy exposure	tumor samples	polymerase chain	pathway conduction	regenerative medicine	transdermal buprenorphine
11	ovx zol	williams syndrome	growth factor	treated seeds	hydroxymethylglutaryl coenzyme	lysinibacillus sp
12	light chain	foot protein	chain reaction	gene increase	lysinibacillus sp	oral contraceptives
13	pedot go	isolates type	significant difference	similar trends	hypocotyl elongation	tick borne
14	rs10468017 variant	neuron areas	cell death	higher levels	intestinal tract	distant metastases
15	tumorigenic clones	mosaic virus	cancer cells	investigated lesions	carbonic anhydrase	lxxll motif
16	mirror neuron	eec syndrome	mg kg	bone marrow	mirror neuron	glun1 glun2b
17	neuron areas	voltage gated	cell line	cu b	charnley classification	bm examination
18	elevated umbilical	nci n78	significantly higher	molecular computing	voluntary interruptions	xl cgd
19	proteomic analysis	distant metastasis	important role	zokor species	moogoo udder	clinicaltrials gov
20	cultured fibroblasts	clinical importance	cell surface	affected organs	try ends	multivariate logistic

Existent methods

After tokens removing

	Closed KL(ngram, tilda)	Closed KL(tilda, ngram)	TF-IDF	TextRank	Variational KL(ngram, tilda)	Variational KL(tilda, ngram)
1	voltage gated	case series	present study	fetal cells	tumorigenic clones	db db
2	resonance energy	therapy exposure	breast cancer	induced transition	glun1 glun2b	w w
3	direct suppression	prenatal diagnosis	wild type	gene expression	ophthalmology journals	substantia nigra
4	tumor samples	mid log	results suggest	suggested guidelines	egfp nachralpha3	tumorigenic clones
5	prenatal diagnosis	tumorigenic clones	gene expression	complete follow	numbered tag	intestinal tract
6	serine threonine	l lactis	cell lines	coagulation function	substantia nigra	numbered tag
7	[illegible]	[illegible]	[illegible]	[illegible]	volar oblique	alfalfa peroxidase
8	[illegible]	[illegible]	[illegible]	[illegible]	elastin matrix	qbeta replicase
9	[illegible]	[illegible]	risk factors	first data	nursing home	hydroxymethylglutaryl coenzyme
10	therapy exposure	tumor samples	polymerase chain	pathway conduction	regenerative medicine	transdermal buprenorphine
11	ovx zol	williams syndrome	growth factor	treated seeds	hydroxymethylglutaryl coenzyme	lysinibacillus sp
12	light chain	foot protein	chain reaction	gene increase	lysinibacillus sp	oral contraceptives
13	pedot go	isolates type	significant difference	similar trends	hypocotyl elongation	tick borne
14	rs10468017 variant	neuron areas	cell death	higher levels	intestinal tract	distant metastases
15	tumorigenic clones	mosaic virus	cancer cells	investigated lesions	carbonic anhydrase	lxxll motif
16	mirror neuron	eec syndrome	mg kg	bone marrow	mirror neuron	glun1 glun2b
17	neuron areas	voltage gated	cell line	cu b	charnley classification	bm examination
18	elevated umbilical	nci n78	significantly higher	molecular computing	voluntary interruptions	xl cgd
19	proteomic analysis	distant metastasis	important role	zokor species	moogoo udder	clinicaltrials gov
20	cultured fibroblasts	clinical importance	cell surface	affected organs	try ends	multivariate logistic

Proposed method (A)

Existent methods

After tokens removing

	Closed KL(ngram, tilda)	Closed KL(tilda, ngram)	TF-IDF	TextRank	Variational KL(ngram, tilda)	Variational KL(tilda, ngram)
1	voltage gated	case series	present study	fetal cells	tumorigenic clones	db db
2	resonance energy	therapy exposure	prenatal diagnosis	induced transition	glun1 glun2b	w w
3	direct suppression	prenatal diagnosis	wild type	gene expression	ophthalmology journals	substantia nigra
4	tumor samples	mid log	results suggest	suggested guidelines	egfp nachralpha3	tumorigenic clones
5	prenatal diagnosis	tumorigenic clones	gene expression	complete follow	numbered tag	intestinal tract
6	serine threonine	l lactis	cell lines	coagulation function	substantia nigra	numbered tag
7	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
8	a	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
9	[illegible]	[illegible]	risk factors	first data	[illegible]	nzyme
10	therapy exposure	tumor samples	polymerase chain	pathway conduction	regenerative medicine	transdermal buprenorphine
11	ovx zol	williams syndrome	growth factor	treated seeds	hydroxymethylglutaryl coenzyme	lysinibacillus sp
12	light chain	foot protein	chain reaction	gene increase	lysinibacillus sp	oral contraceptives
13	pedot go	isolates type	significant difference	similar trends	hypocotyl elongation	tick borne
14	rs10468017 variant	neuron areas	cell death	higher levels	intestinal tract	distant metastases
15	tumorigenic clones	mosaic virus	cancer cells	investigated lesions	carbonic anhydrase	lxxll motif
16	mirror neuron	eec syndrome	mg kg	bone marrow	mirror neuron	glun1 glun2b
17	neuron areas	voltage gated	cell line	cu b	charnley classification	bm examination
18	elevated umbilical	nci n78	significantly higher	molecular computing	voluntary interruptions	xl cgd
19	proteomic analysis	distant metastasis	important role	zokor species	moogoo udder	clinicaltrials gov
20	cultured fibroblasts	clinical importance	cell surface	affected organs	try ends	multivariate logistic

Proposed method (A)

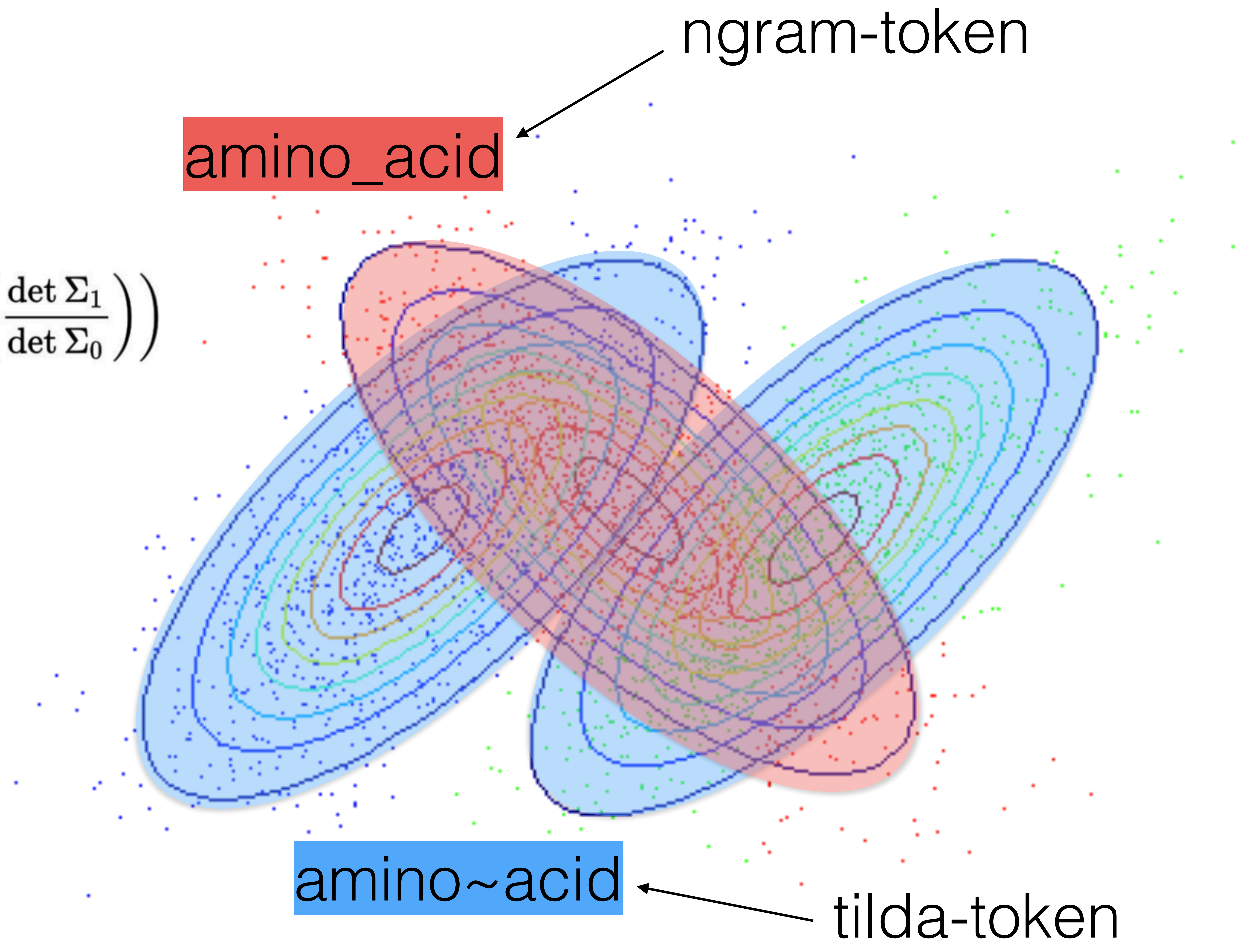
Existent methods

Proposed method (B)

Proposed method (A)

Closed KL divergence formula*

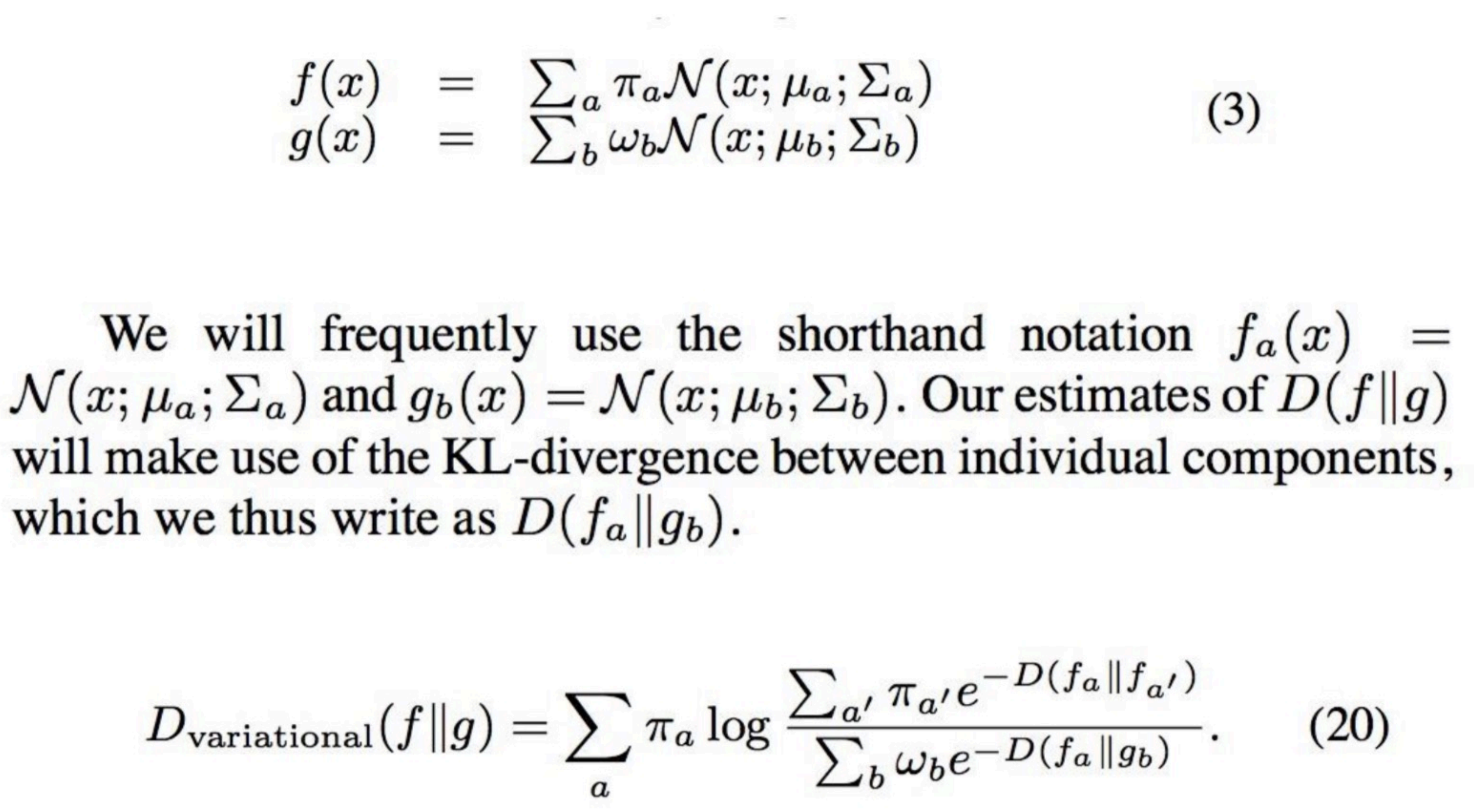
$$D_{\mathrm{KL}}(\mathcal{N}_0 \| \mathcal{N}_1) = \frac{1}{2}\left(\operatorname{tr}\left(\Sigma_1^{-1}\Sigma_0\right) + (\mu_1 - \mu_0)^{\top}\Sigma_1^{-1}(\mu_1 - \mu_0) - k + \ln\left(\frac{\det \Sigma_1}{\det \Sigma_0}\right)\right)$$



* - https://en.wikipedia.org/wiki/Kullback–Leibler_divergence

Proposed method (B)

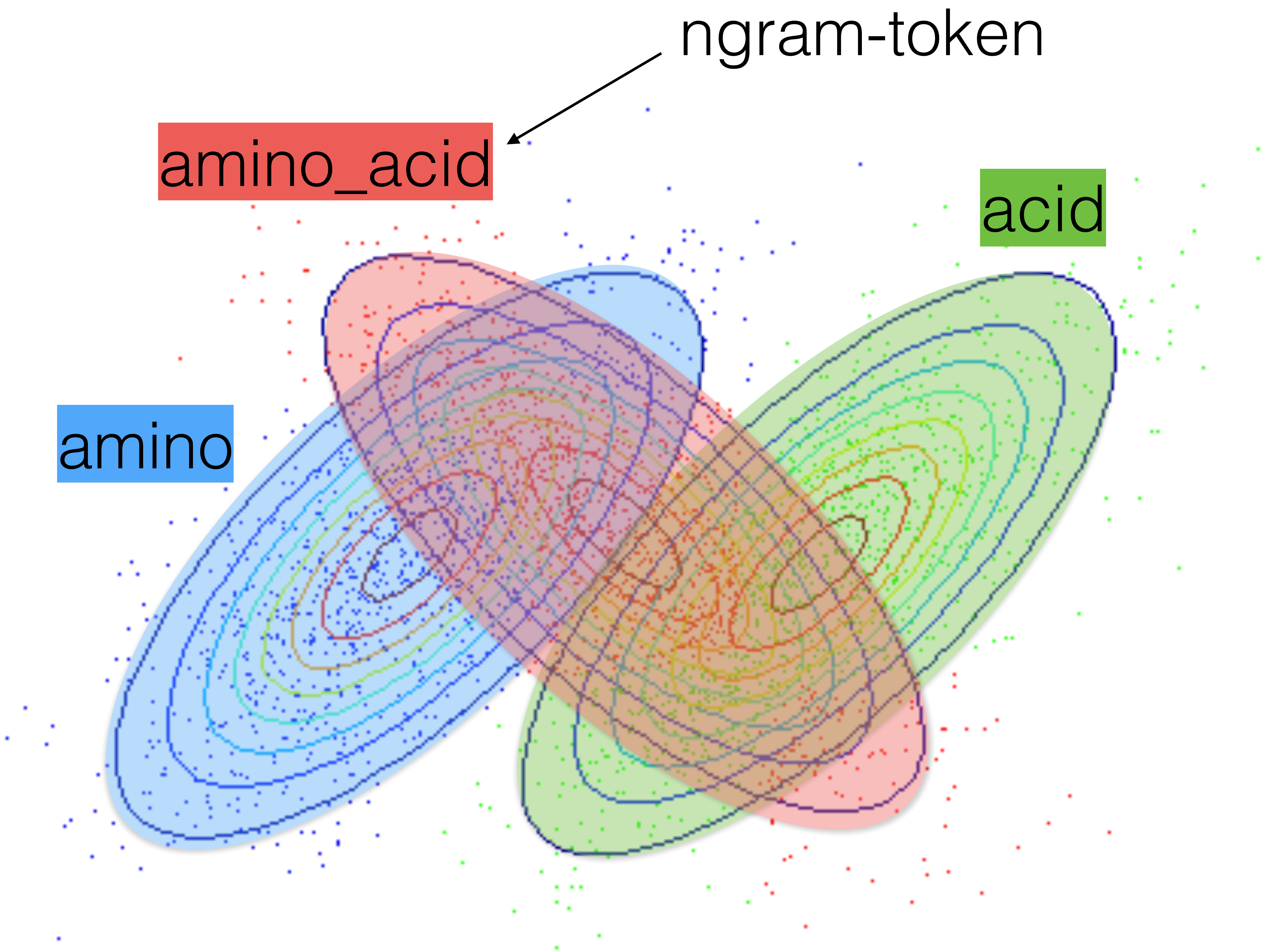
Variational KL divergence formula*



$$\begin{aligned} f(x) &= \textstyle\sum_a \pi_a \mathcal{N}(x;\mu_a;\Sigma_a) \\ g(x) &= \textstyle\sum_b \omega_b \mathcal{N}(x;\mu_b;\Sigma_b) \end{aligned} \qquad (3)$$

We will frequently use the shorthand notation $f_a(x) = \mathcal{N}(x;\mu_a;\Sigma_a)$ and $g_b(x) = \mathcal{N}(x;\mu_b;\Sigma_b)$. Our estimates of $D(f\|g)$ will make use of the KL-divergence between individual components, which we thus write as $D(f_a\|g_b)$.

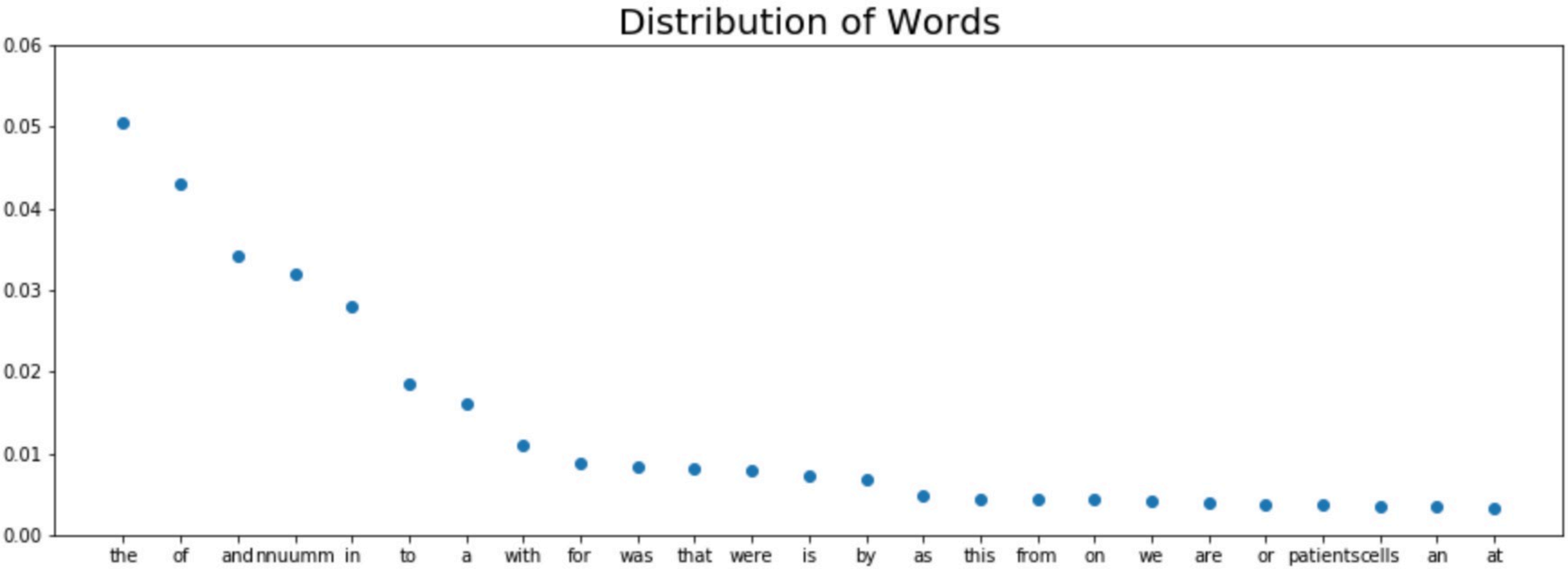
$$D_{\text{variational}}(f\|g) = \sum_a \pi_a \log \frac{\sum_{a'} \pi_{a'} e^{-D(f_a\|f_{a'})}}{\sum_b \omega_b e^{-D(f_a\|g_b)}}. \qquad (20)$$



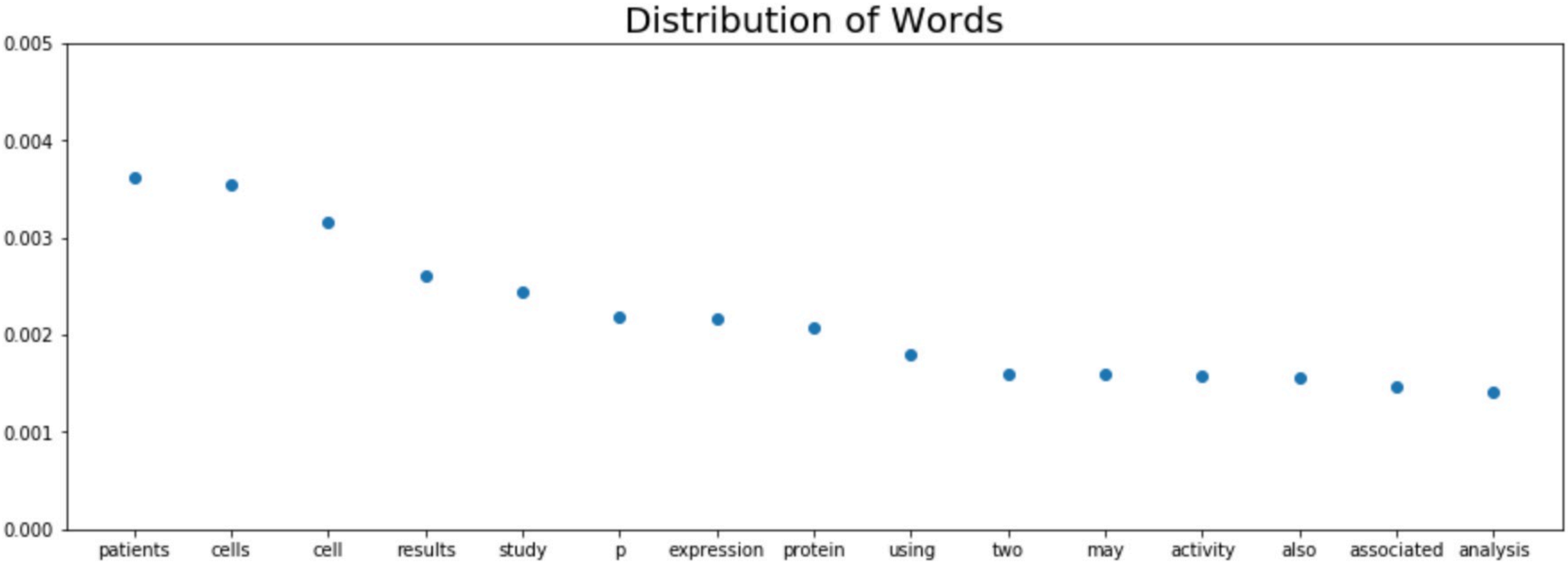
---

* - J. R. Hershey and P. A. Olsen, Approximating the Kullback Leib
Divergence Between Gaussian Mixture Models, 200

Random text

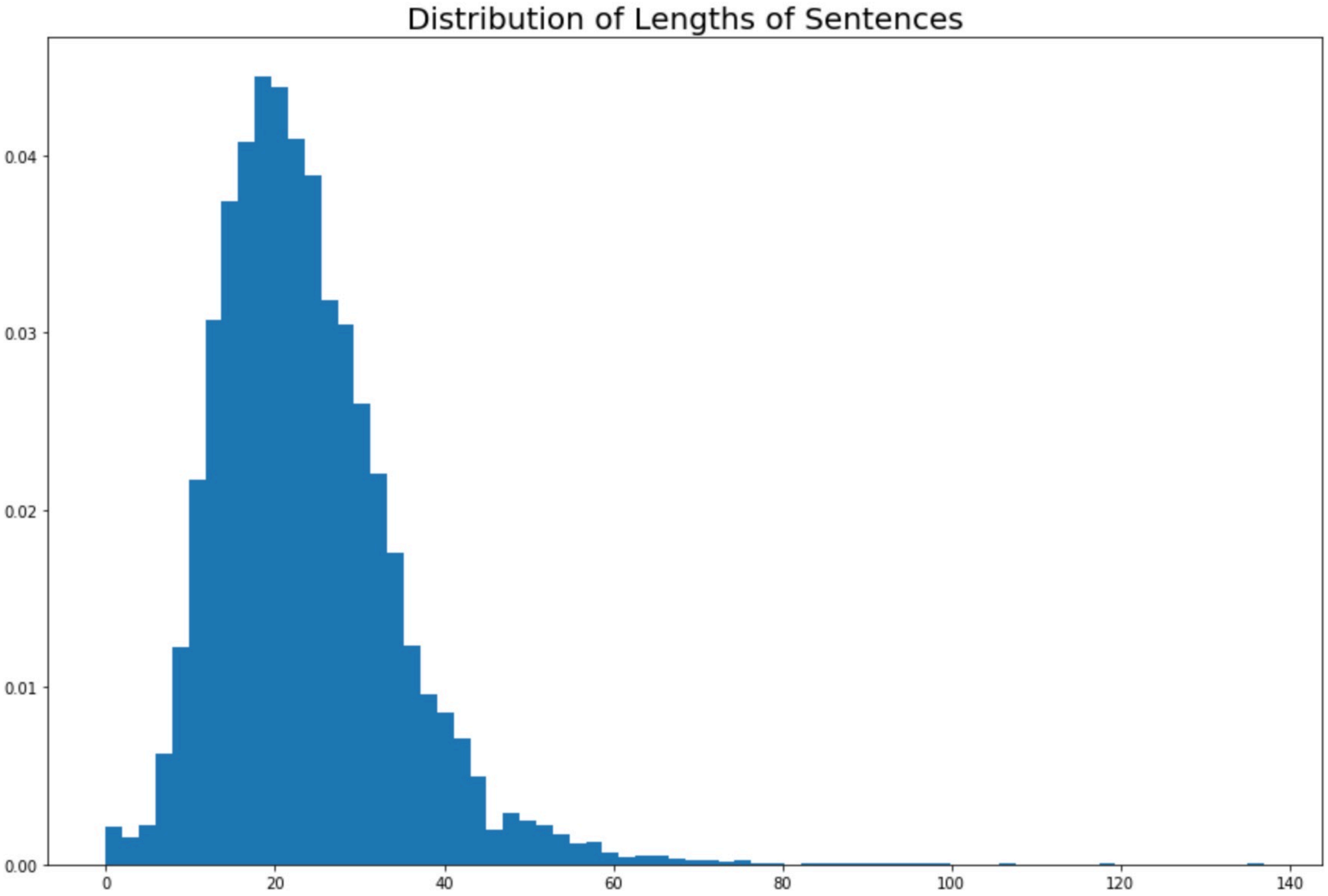

Distribution of Words
0.06
0.05
0.04
0.03
0.02
0.01
0.00
the
of
and
nnuumm
in
to
a
with
for
was
that
were
is
by
as
this
from
on
we
are
or
patients
cells
an
at

Random text



* stopwords like ['to', 'a', 'of', 'the', …] are removed

Random text


Distribution of Lengths of Sentences
0.04
0.03
0.02
0.01
0.00
0
20
40
60
80
100
120
140

Random text: before

	Closed KL(ngram, tilda)	Closed KL(tilda, ngram)	TF-IDF	TextRank	Variational KL(ngram, tilda)	Variational KL(tilda, ngram)
1	p nnuumm	p nnuumm	uunnkk uunnkk	unsuitable nnuumm	nnuumm wt	nnuumm wt
2	nnuumm p	nnuumm p	uunnkk nnuumm	uunnkk potential	nnuumm relevant	nnuumm beyond
3	nnuumm antiviral	patients patients	nnuumm uunnkk	uunnkk use	nnuumm beyond	food nnuumm
4	response using	response using	patients uunnkk	uunnkk protein	randomly nnuumm	nnuumm antiviral
5	artificial nnuumm	nnuumm defined	uunnkk patients	approach cells	food nnuumm	randomly nnuumm
6	high used	nnuumm antiviral	uunnkk cells	p total	nnuumm antiviral	nnuumm mucosa
7	nnuumm example	results two	cells uunnkk	apoptosis analyses	putative nnuumm	nnuumm relevant
8	cells based	useful patients	cell uunnkk	uunnkk differentiation	artificial nnuumm	artificial nnuumm
9	nnuumm injected	pathway cell	uunnkk cell	use enlargement	nnuumm mucosa	nnuumm reaction
10	useful patients	randomly nnuumm	uunnkk results	uunnkk factor	possible nnuumm	nnuumm ras
11	patients patients	nnuumm cytomegalovirus	nnuumm nnuumm	therapeutic suggests	nnuumm reverse	median nnuumm
12	nnuumm cytomegalovirus	cells based	results uunnkk	sumo uunnkk	nnuumm controlled	putative nnuumm
13	nnuumm crucial	nnuumm series	study uunnkk	uunnkk results	sequencing nnuumm	nnuumm renal
14	alcohol nnuumm	also associated	uunnkk study	disc nnuumm	contribute nnuumm	nnuumm injected
15	related cell	nnuumm able	uunnkk expression	method role	nnuumm tumors	possible nnuumm
16	plus nnuumm	nnuumm example	expression uunnkk	number nnuumm	surgical nnuumm	antibodies nnuumm
17	nnuumm impaired	nnuumm prior	protein uunnkk	tissue segments	antibodies nnuumm	detection nnuumm
18	understand nnuumm	beta1 nnuumm	uunnkk protein	distinct conclusion	nnuumm ras	nnuumm problem
19	randomly nnuumm	water nnuumm	using uunnkk	specific uunnkk	nnuumm reaction	nnuumm efficient
20	nnuumm events	nnuumm tandem	uunnkk using	uunnkk case	nnuumm multiple	nnuumm tumors

Random text: after

	Closed KL(ngram, tilda)	Closed KL(tilda, ngram)	TF-IDF	TextRank	Variational KL(ngram, tilda)	Variational KL(tilda, ngram)
1	response using	patients patients	results cell	approach cells	high used	useful patients
2	high used	response using	cells patients	p total	studies alpha	pathway cell
3	cells based	results two	cell cells	apoptosis analyses	useful patients	significant activity
4	useful patients	useful patients	cells cells	use enlargement	p methods	infection cells
5	patients patients	pathway cell	gene specific	therapeutic suggests	cancer using	response using
6	related cell	cells based	patients cells	method role	significant activity	type patients
7	also associated	also associated	mean cells	tissue segments	infection cells	studies alpha
8	pathway cell	n results	high cell	distinct conclusion	pathway cell	high used
9	conclusion expression	related cell	showed expression	indicate dose	response using	addition high
10	demonstrated cell	demonstrated cell	present cell	cardiomyocytes years	type patients	clinical patients
11	studies alpha	showed expression	tumor cells	observed discrepancies	production cells	protein cells
12	n results	proteins cell	methods study	sequences time	addition high	cell cases
13	expression clinical	response patients	patients study	used days	response patients	associated results
14	cell patients	study data	factors cells	work growth	expression clinical	cancer using
15	data cells	conclusion expression	cell effects	mechanisms dependency	well cells	mean cells
16	p methods	data cells	cells results	protein classification	patients one	p methods
17	response patients	cell treatment	data cells	structure patient	conclusion expression	increased increased
18	cell treatment	cell cases	demonstrated cell	higher weight	observed cells	patients one
19	patients associated	high used	patients time	positive values	protein cells	results compared
20	cells cells	patients associated	genetic patients	elevated cells	gene specific	patients time

Further improvements

- Finish description of the method
- Complete module for Python-language
- [next semester] Automatic classification of extracted medical terms

References

- J. R. Hershey and P. A. Olsen, *Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models, In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.
- R. Mihalcea and P. Tarau, *TextRank: Bringing Order into Texts,* 2004.
- L. Vilnis and A. McCallum, *Word Representations via Gaussian Embedding*, 2014.
- W. Liu et. al., *A genetic algorithm enabled ensemble for unsupervised medical term extraction from clinical letters*, 2015.
- Moz, *Gaussian Word Embeddings*, https://github.com/seomoz/word2gauss.