# Advanced Parser for Biomedical Texts


## Anton Karazeev[1)], Maxim Holmatov[2)]

[1)] Moscow Institute of Physics and Technology, [2)] Saint-Petersburg State Pediatric Medical University

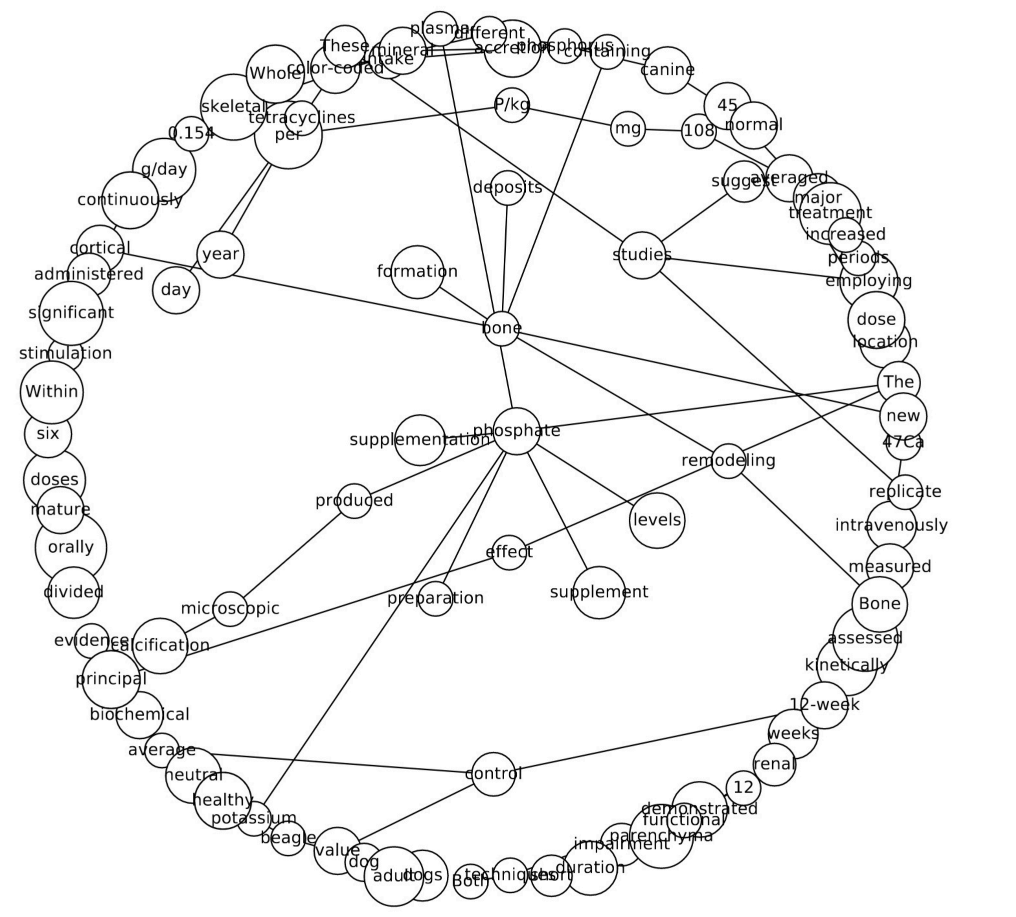
Laboratory of Functional Analysis of the Genome


## 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



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.

## 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.

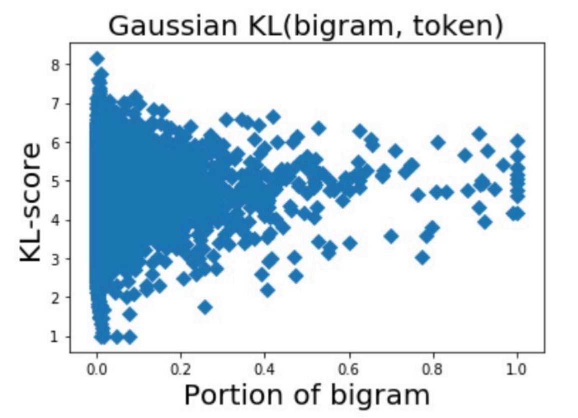
## Kullback-Leibler Divergence

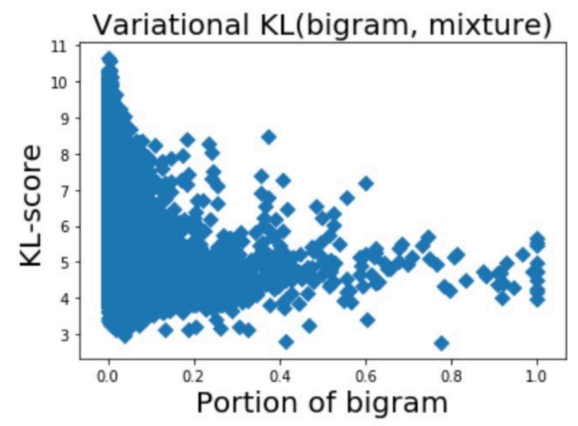
In the context of machine learning, $D_{KL}(P \| Q)$ is often called the information gain achieved if P is used instead of Q.

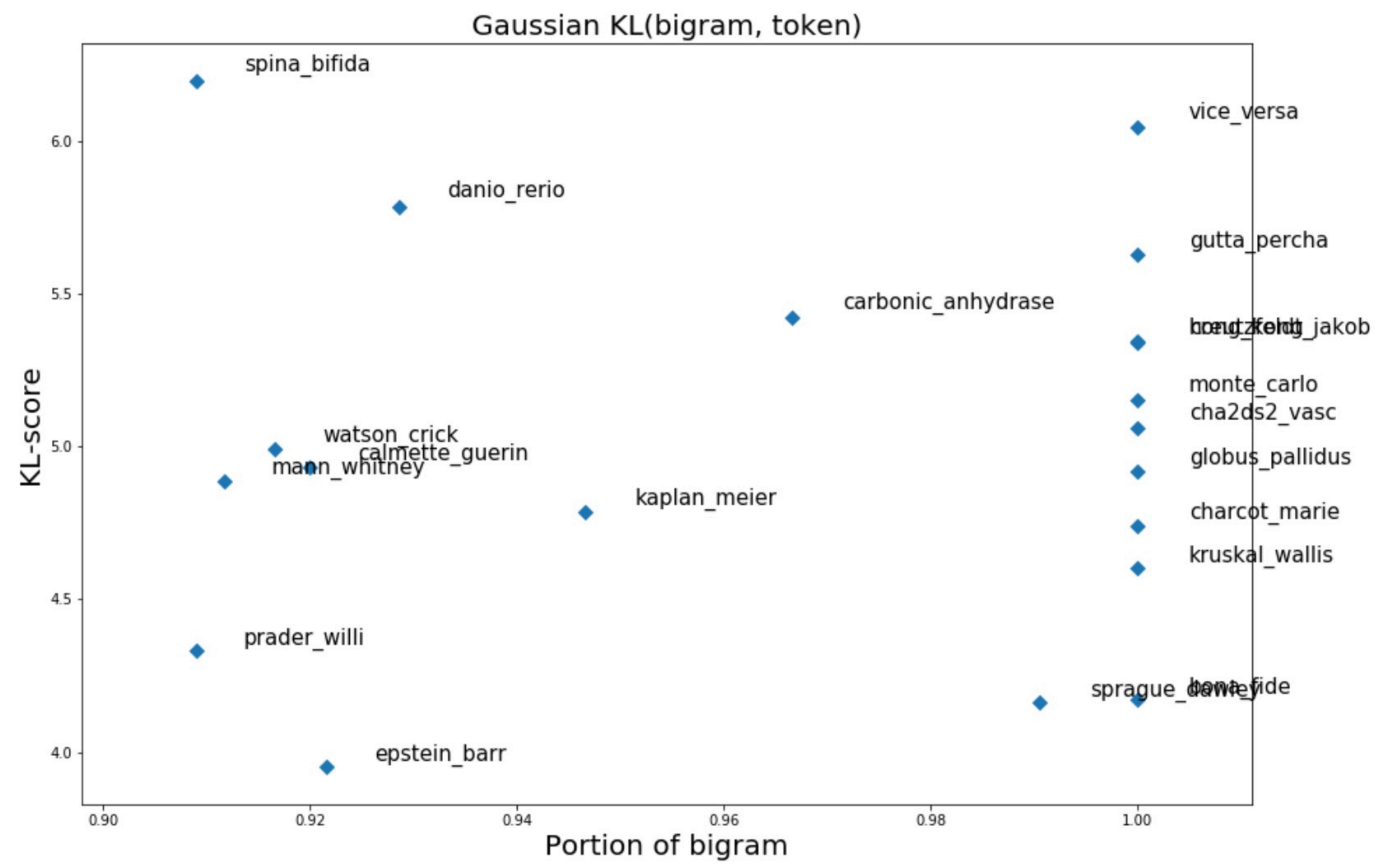
$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}.$$

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

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



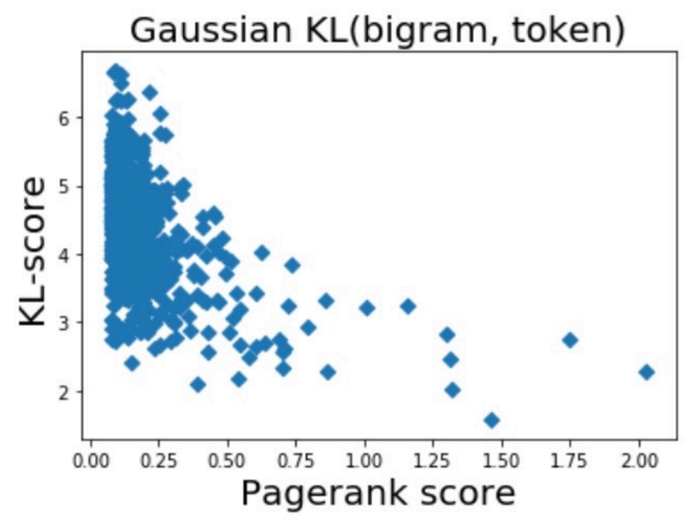



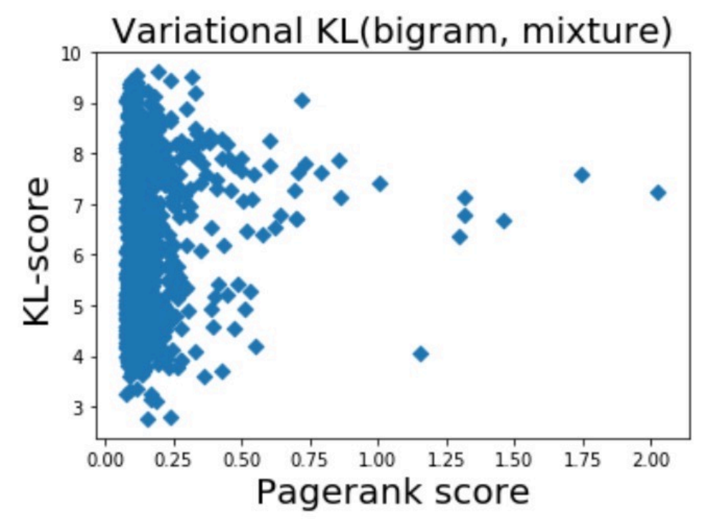



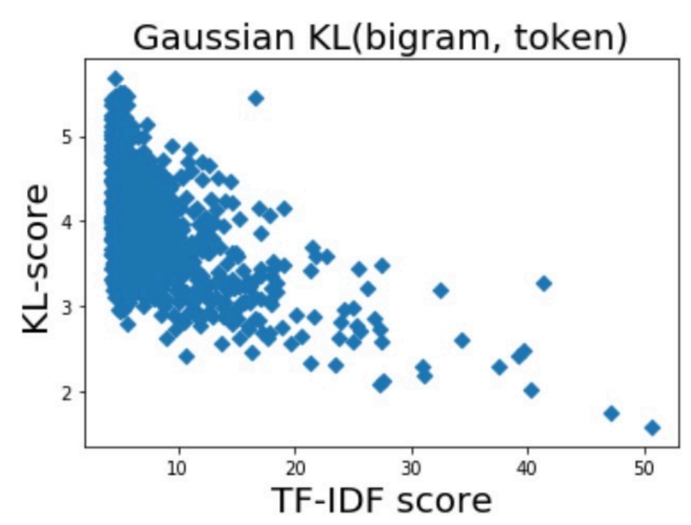

## Results

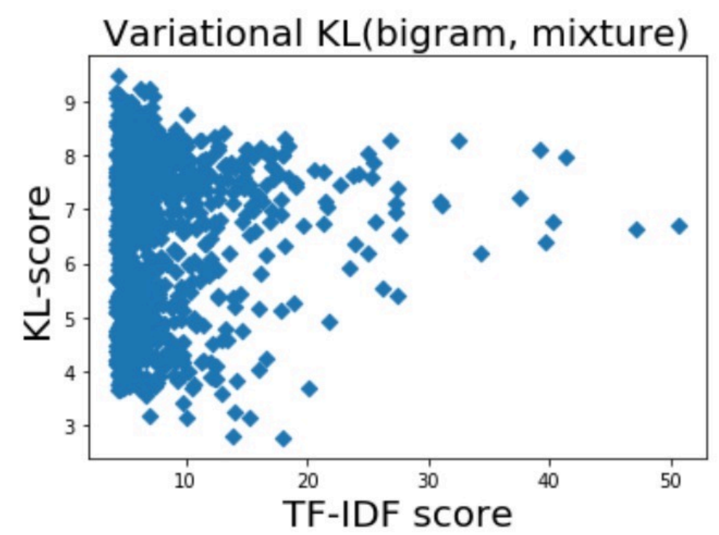
| PageRank | Gaussian KL(bigram, token) | Gaussian KL(token, bigram) | Variational KL(bigram, mixture) | Variational KL(mixture, bigram) |
|---|---|---|---|---|
| breast_cancer | ang_iii | citron_kinase | coli_isolates | early_disease |
| cancer_cells | citron_kinase | biliary_complications | liver_cancer | 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.