UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Yevhen Tyshchenko

Thesis Title

Master's Thesis (30 ECTS)

Supervisor: Kairit Sirts, PhD

Tartu 2018

Type Inference for Fourth Order Logic Formulae

Abstract:

Abstract

Keywords:

List of keywords

CERCS:

CERCS code and name: https://www.etis.ee/Portal/Classifiers/Details/d3717f7b-bec8-4cd9-8ea4-c89cd56ca46e

Tüübituletus neljandat järku loogikavalemitele

Lühikokkuvõte:

One or two sentences providing a basic introduction to the field, comprehensible to a scientist in any discipline.

Two to three sentences of more detailed background, comprehensible to scientists in related disciplines.

One sentence clearly stating the general problem being addressed by this particular study.

One sentence summarising the main result (with the words “here we show´´ or their equivalent).

Two or three sentences explaining what the main result reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.

One or two sentences to put the results into a more general context.

Two or three sentences to provide a broader perspective, readily comprehensible to a scientist in any discipline, may be included in the first paragraph if the editor considers that the accessibility of the paper is significantly enhanced by their inclusion.

Võtmesõnad:

List of keywords

CERCS:

CERCS kood ja nimetus: https://www.etis.ee/Portal/Classifiers/Details/d3717f7b-bec8-4cd9-8ea4-c89cd56ca46e

Contents

1 Introduction **7**

2 Previous Work **8**

3 Technical background **8**
3.1 Web data retrieval 8
3.2 Text representation 9
3.2.1 Bag-of-words 9
3.2.2 Term Frequency-Inverse Document Frequency 9
3.2.3 10
3.3 Topic modeling 10
3.3.1 Latent Dirichlet Allocation 10
3.4 Classification 12
3.4.1 Support Vector Machines 12
3.4.2 Convolutional Neural Networks 13

4 Method **14**
4.1 Data scraping 14
4.2 Data preprocessing 14
4.3 Topic modeling 15
4.3.1 Model evaluation and parameter search 16
4.4 Classification setup 16

5 Results and analysis **16**
5.1 Topic model 16

6 Definitions **22**

7 Conclusion **22**
II. Licence 23

Unsolved issues

Abstract 2
List of keywords 2
CERCS code and name: https://www.etis.ee/Portal/Classifiers/Details/d3717f7b-bec8-4cd9-8ea4-c89cd56ca46e 2
One or two sentences providing a basic introduction to the field, comprehensible to a scientist in any discipline. 2
Two to three sentences of more detailed background, comprehensible to scientists in related disciplines. 2
One sentence clearly stating the general problem being addressed by this particular study. 2
One sentence summarising the main result (with the words "here we show´´ or their equivalent). 2
Two or three sentences explaining what the main result reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge. 2
One or two sentences to put the results into a more general context. 2
Two or three sentences to provide a broader perspective, readily comprehensible to a scientist in any discipline, may be included in the first paragraph if the editor considers that the accessibility of the paper is significantly enhanced by their inclusion. 2
List of keywords 2
CERCS kood ja nimetus: https://www.etis.ee/Portal/Classifiers/Details/d3717f7b-bec8-4cd9-8ea4-c89cd56ca46e 2
What is it in simple terms (title)? 7
Why should anyone care? 7
What was my contribution? 7
What you are doing in each section (a sentence or two per section) 7
Describe previous work 8
Web data extraction, applications and techniques: A survey 8
A Brief Survey of Web Data Extraction Tools 8
Original paper G. Salton and M. McGill, editors. Introduction to Modern Information Retrieval. McGraw-Hill, 1983. 9
Original LDA paper http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf . . 10
describe batch variational Bayes: Online Learning for Latent Dirichlet Allocation article 11
add formulas $https://www.cl.cam.ac.uk/teaching/1213/L101/clark_lectures/lect7.pdf$ 11
http://www.ece.umd.edu/ smiran/LDA.pdf 11
Original paper https://rdcu.be/MlPW 12

link to image: https://www.quora.com/What-does-support-vector-machine-SVM-mean-in-laymans-terms . . . 12
what is CNN? . . . 13
CNN architecture . . . 13
CNN layers . . . 13
Scraping procedure flowchart . . . 14
Yoon Kim paper $https://arxiv.org/pdf/1408.5882.pdf$ . . . 14
Yoon Kim code $https://github.com/yoonkim/$ . . . 14
Topic modelling and methods http://www.cs.columbia.edu/ blei/papers/BleiLafferty2009.pdf . . . 15
Papers with topic model evaluation approaches. . . . 16
What methods will be covered? . . . 16
Why they were used? . . . 16
topic number vs svm accuracy plot . . . 16
The Chi-square test of independence Mary L. McHugh . . . 17
Should I provide a strict mathematical definitions here? . . . 22
Multinomial distribution . . . 22
what did you do? . . . 22
What are the results? . . . 22
future work? . . . 22

1 Introduction

What is it in simple terms (title)?

Why should anyone care?

What was my contribution?

What you are doing in each section (a sentence or two per section)

Usually the data about person's mental state is confidential and only doctors have an access for it. Fortunately, people tend to share their thoughts, stories and quite personal feelings on the web resources such as social network insights, blog platforms, forums etc. In particular, some research are based on Twitter posts data while others collect post entries from mental health support forums that are then labeled by metal health specialists. These resources are publicly available and often can be quired and filtered using keywords.

2 Previous Work

Describe previous work

3 Technical background

This section incorporates technical description of methods and approaches applied for this natural language processing task including web data retrieval, numerical text representation, probabilistic topic modeling, and text classification methods. In particular, web data extraction approaches are described in 3.1, text representation methods are covered in 3.2...

3.1 Web data retrieval

Web data extraction, applications and techniques: A survey

A Brief Survey of Web Data Extraction Tools

Web data information retrieval field has developed considerably since the rapid growth of Word Wide Web pages and websites. The existing approaches could be grouped into two categories by their underlying work principle: *tree-based approaches*, *web wrappers*.

The first category is *tree-based approaches* that consider the DOM web page representation which is basically a labeled rooted tree hierarchy over a mixture of text and HTML tags. This particular web page representation motivates the usage of tree-based algorithms and mechanisms of addressing specific page tree nodes containing desired data. This is usually performed by XPath queries to a single page element or a group of similar page elements enclosed between HTML tags.

The second category is *web wrappers* that in terms of web data extraction is often described as a process aimed to find, extract, and transform an unstructured target data for further analysis by computer program in automated or semi-automated way. This approach could be decomposed on the following three steps:

1. Initialization — wrapper is created;
2. Execution — wrapper runs and collects the data;
3. Maintenance — if the data source structure changes then the respective wrapper should be also tuned to handle it.

3.2 Text representation

3.2.1 Bag-of-words

Most of the natural language processing tasks require input text numerical representation rather than raw words meaning that each word must be encoded in order to be processed by machines. One of the most popular and straightforward approaches to encode text is to use a bag-of-words (BoW) text representation. This model encodes words with their frequencies and outputs a matrix where rows are documents and columns represent unique words. The word importance in this model is just its frequency which is the main weakness of this representation. In other words, frequently used words such as "the" that are not discriminative across the collection of the documents will get higher weights thus hiding the less frequent but more informative key words.

3.2.2 Term Frequency-Inverse Document Frequency

//

Original paper G. Salton and M. McGill, editors. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

Therefore, term frequency-inverse document frequency (TFIDF) is applied to resolve this issue. TFIDF is basically a numeric statistic that describes the importance of a word with respect to each document in a corpora. The term frequency (TF) cab be calculated as follows:

$$tf(w,d) = \frac{n_w}{\sum_k n_k},$$

where n_w is the number of times the word appeared in a document d and $\sum_k n_k$ is the total number of words in a document.

The inverse document frequency (IDF) diminishes the weights of frequent words and strengthen rare words if they don't appear often in other documents except the current one, and can be calculated as follows:

$$idf(w,C) = log\frac{|C|}{|\{c \in C : w \in c\}|},$$

where $|C|$ is the corpus size and $|\{c \in C : t \in c\}|$ is the number of documents where the word w appears. Finally, TFIDF itself is a product of the two aforementioned frequencies:

$$tfidf(w,d,C) = tf(w,d) \cdot idf(w,C).$$

The word will be ranked with greater TFIDF value if it is present in a particular document and absent in other documents.

3.2.3

3.3 Topic modeling

Nowadays, the amounts of written text information often do not fit the computational capacity and make scientists apply text mining methods to discover useful hidden text structures or similarities. One important task in natural language processing field is detecting the between document similarities by ideas and topics mentioned there. Topic modeling is an unsupervised text mining approach that discovers these similarities in a text or corpora. Topic model itself refer to statistical algorithms that explore the topics over the collection of documents and helps to annotate these documents according to detected topics.

3.3.1 Latent Dirichlet Allocation

//

Original LDA paper http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Latent Dirichlet Allocation (LDA) is a generative probabilistic model that models a text corpus and is helpful in discovering its underlying topic structures. The main idea behind LDA is that each document is represented as a probability distribution over latent topics. Every topic in this model is treated as a distribution over words existing in the collection of documents. The LDA model could be decomposed into two steps: *model generation* and *model inference*.

Model generation: On model generation stage document-to-topic and topic-to-word distributions are drawn from two Dirichlet distributions with α and β parameters. Then, for each word position in document choose topic and, finally, given the topic distribution pick the word for this position. LDA model is often represented using plate diagram (Figure 1) with the following notations:

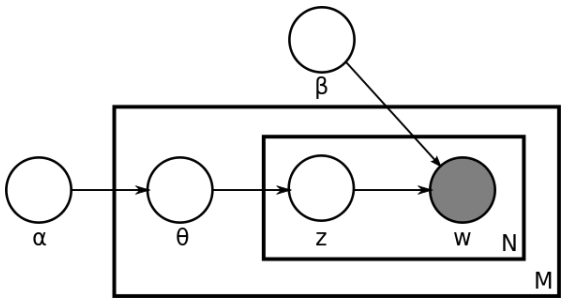

Figure 1. Plate notation of the LDA model.

- The number of topics K is fixed and defined.
- The number of all documents in a corpus D is M.
- ϕ_k is a k-th topic word distribution over a fixed vocabulary where $1 \leq k \leq K$.
- Θ_m is a m-th document topic distribution where $1 \leq m \leq M$.
- z_m is the topic assignment for m-th document, and $z_{m,n}$ is the topic assignment for n-th word in m-th document.
- $w_{m,n}$ is actually the n-th word in m-th document.
- α, β are hyperparameters of a Dirichlet distribution.

Then, the generative process could be described in following steps:

1. $\forall$ document $d_i \in D$ draw $\Theta_{d_i} \sim Dirichlet(\alpha)$
2. $\forall$ topic $k \in K$ draw $\phi_k \sim Dirichlet(\beta)$
3. $\forall i, j : 1 \leq i \leq M, 1 \leq j \leq |d_i|$ where $|d_i|$ is the number of selected words in document d_i draw:
   - topic $z_{i,j} \sim Multinomial(\Theta_i)$
   - word $w_{i,j} \sim Multinomial(\phi_{z_{i,j}})$

Model inference: Now the task is to infer word-to-topic assignments $z_{i,j}$, document-to-topic distributions Θ_i and corpus topic distributions ϕ_k. The original paper [Blei] suggests a *variational inference* approach to estimate the posterior distribution of the hidden variable w — gray circle on figure 1.

describe batch variational Bayes: Online Learning for Latent Dirichlet Allocation article

add formulas $https://www.cl.cam.ac.uk/teaching/1213/L101/clark_lectures/lect7.pdf$

http://www.ece.umd.edu/ smiran/LDA.pdf

Finally, apart from the vector space corpus representation LDA outputs document-to-topic and word-to-topic distribution matrices which are useful for further topic analysis, and, for instance, can be applied as features for text classification tasks.

3.4 Classification

3.4.1 Support Vector Machines

//

Original paper https://rdcu.be/MlPW

Support Vector Machine (SVM) is a supervised machine learning method that can be used for both classification and regression tasks. The main idea behind it is to divide a linearly separable data into two classes with maximum between-class distance. If we consider two-dimensional example illustrated on Figure? SVM finds an optimal line or set of lines that lie as far from the nearest class data points as possible. The dashed lines are called support vectors.

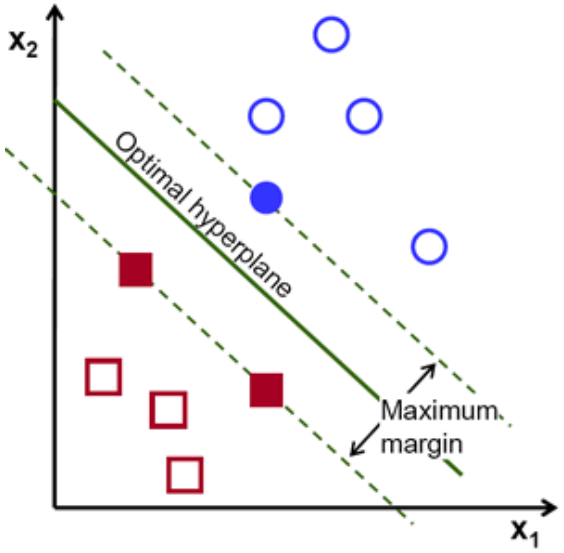

Figure 2. SVM illustration for two-dimensional case.

link to image: https://www.quora.com/What-does-support-vector-machine-SVM-mean-in-laymans-terms

In general, SVM finds an optimal hyperplane or a set of hyperplanes in high-dimensional space that separate the input data with the maximum margin between this hyperplane and nearest training data points of any class. If the input data is not linearly separable a kernel function is applied to map it into higher-dimensional space to make it linearly separable. The most popular non-linear kernels are: polynomial kernel, Gaussian radial basis function, and sigmoid kernel.

3.4.2 Convolutional Neural Networks

//

what is CNN?

CNN architecture

CNN layers

4 Method

4.1 Data scraping

Scraping procedure flowchart

Web data retrieval technique applied in this work falls into the category of web wrappers although the XPath queries were used to get specific DOM elements. The goal of this step is to extract two corpus of blog posts: control and clinical, such that they will be similar in terms of the discussed topics.

The data for this work was extracted from the Blogger platform in an semi-automated way using Python script and Selenium package for browser automation. Initially, the idea was to collect the control and clinical blogs by keyword interest "*depression*" using the platform built-in filter, then manually verify scraped blog URLs, and finally get desired blog posts by these URLs using scrapper. The drawback of the built-in filter is that it brought mostly control group containing psychiatrists, volunteers, psychologists and blogs about religion. Therefore, another query was constructed with the help of Google Advanced Search tool and can be interpreted as follows:

- Find pages such that contain all these words: *my, life, clinical, depression, anxiety*
- Any of the words: *anxiety, depression*
- Within the website: *blogspot.com*.

my life clinical depression anxiety anxiety OR depression OR disorder site:blogspot.com

This detailed search query brought much more blogs where people mention their clinical or self-stated diagnosis.

The scraping method introduced more than 100 blog links that were then manually evaluated. This brought in 47 clinical and 36 control subject blogs which were then scraped entirely. All in all, there are 10799 and 6176 control and clinical blog posts respectively, and later restricted to the same size to avoid data imbalance.

4.2 Data preprocessing

Yoon Kim paper $https://arxiv.org/pdf/1408.5882.pdf$

Yoon Kim code $https://github.com/yoonkim/$

This section describes data preprocessing - a vital step in any natural language processing task. Firstly, all input data was cleaned using regular expressions which are actually the sequences of characters that define the search string in a text. In particular,

the initial string cleaning pipeline were taken from the Yoon Kim's preprocessing code, modified and extended with additional regular expressions matching and replacing the URL addresses, contractions and redundant whitespaces with a single whitespace. Additionally, brackets, colons, dashes, punctuation marks, contractions and all newline symbols were removed. Finally, the posts were lowercased and saved as separate text files in order to perform the preprocessing step only once thus considerably reducing the overall execution time.

In order to prevent an undesirable impact on final topic model the most frequent and rare words were removed. In particular, the word was filtered out if it occurs in less than 3 documents or more than 80% of all documents. However, further experiments on non-filtered data showed better classification results so this cleaning step was rejected.

The *lemmatization* were chosen because unlike *stemming* it aims to remove word endings based on vocabulary and morphological word analysis. This ensures that the word endings will not be roughly dropped often resulting in senseless word pieces (*stemming*) but rather transformed into word lemmas conformant with dictionary. For example, the lemmatization output for words *"good", "better", "best"* is just the common lemma *"good"* while stemming will result in *"good", "bet", "best"*. Furthermore, all words were enriched with *part-of-speech tags* (POS-tags) to represent their word classes and improve lemmatization results.

Then, the data was split into three parts statically with 70%, 15% and 15% partitions sizes for training, development and test set respectively. To ensure that each author belongs to only one partition splitting was done based on blog sizes measured in number of posts per author. Moreover, it was attempted to keep the distribution of blog sizes similar across these parts so as to prevent data imbalance.

The bag of words text representation was constructed using the scikit-learn CountVectorizer with 100000 features i.e. word tokens. It produced a sparse matrix representation of a corpus $n \times m$ where n stands for the number of documents in corpus and m is the number of features.

4.3 Topic modeling

//

Topic modelling and methods http://www.cs.columbia.edu/ blei/papers/BleiLafferty2009.pdf

Topic modeling experiments were performed using the *scikit-learn* Python LDA implementation. Out-of-the-box method also allows to choose the learning method for the model either *"online"* or *"batch"*. The main difference between these methods is that *"online"* parameter refers to Online variational Bayes method which uses a mini-batch of data on each iteration to update the resulting document-to-topic matrix while *"batch"* refers to Batch variational Bayes method and uses all data to update this matrix on each

step. The latter method was chosen because the final model will be trained only once and the amount of input data is fixed and small enough to treat it as a single batch.

Papers with topic model evaluation approaches.

4.3.1 Model evaluation and parameter search

The optimal number of n topics for the model was chosen based on the classification performance of previously tuned SVM classifier using the grid search parameter optimization procedure.

4.4 Classification setup

What methods will be covered?

Why they were used?

This section provides the optimal configurations applied for the classification methods and describes their technical details. Additionally, it also discusses the evaluation measures used in this work.

LDA+SVM The initial approach is similar to [Blei] and consists of SVM classifier on top of LDA features. SVM implementation was taken from the *scikit-learn* library. The optimal SVM parameters were tuned with the built-in parameter search procedure and fixed: regularization term $C = xx$ and $gamma = yy$.

5 Results and analysis

This section covers the topic modeling results in section 5.1, and provides the performance tables for the applied classification methods 6.2.

5.1 Topic model

The resulting LDA model requires further analysis to reach better understanding of the scraped data. The topic model parameter search procedure described in 4.3.1 resulted in the best classification score of $blabla$ with $blaaa$ topics. It was also observed that SVM's accuracy fluctuated during the parameter tuning as illustrated on Figure $100500topfigure$.

topic number vs svm accuracy plot

Particularly, the highest classification accuracies were faced with n and m topic numbers that unexpectedly differ quite a lot. Thus, the topics were manually investigated

and labeled according to their twenty most representative words. To evaluate the scraped data in terms of its topic model two document-to-topic probability distribution matrices for control and clinical text corpus were visualized on 3. They outlined the shift towards $sometopic$ in clinical group comparing to controls. Potentially, this imbalance could make classifier predict based on topic assignments so the aim was to statistically estimate how dependent are these two categorical variables: *control* and *clinical*.

The Chi-square test of independence Mary L. McHugh

This work covers the results of two statistical tests: *Chi-square test of independence* and *Fisher's exact test*. They both require a contingency table as input so the one was constructed with 15 columns (LDA topics) and two rows — *control* and *clinical* that contain document counts (Figure 100).

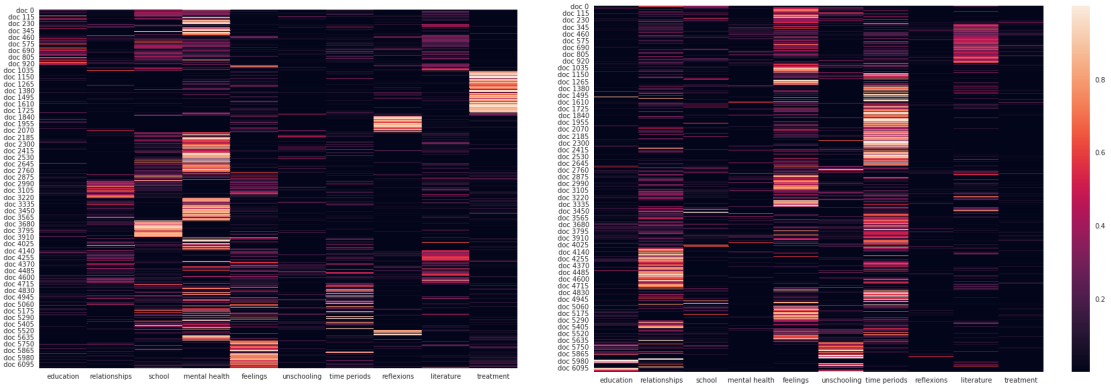

Figure 3. Heatmap visualizations of document-to-topic probabilities for control (*left*) and clinical (*right*) text corpuses.

The resulting document-to-topic distribution analysis showed that the assumptions about the data collection were confirmed.

Table 1. My caption

	school student counselor college high resource peer counseling program teen parent teacher information suzyblue great training youth bullying class idea
	help yourself others change friend important social find something goal person self positive health often going situation doing myself someone
	self such rather thought u say experience may mind sense world understanding something doe question between idea belief itself patient

	autism nate nathan kid vaccine antibiotic issue right study really child boy doctor week supplement infection first yeast u pediatrician
	brain pain system memory right left level body feeling therapy consciousness mean patient cortex u awareness human connection doe down
	god u poem jesus through poetry love may word lord man though must faith depression christ where christian kathleen suffering
	study trial patient treatment placebo data drug effect antidepressant result clinical depression author group journal analysis research article fda published
	exercise pain muscle disease body help symptom doctor blood leg hand physical breathing kennedy several chronic while cold problem every
	physician care amp medical company health business interest managed patient medicine organization industry conflict pharmaceutical manager practice personality professional cost
	child parent family jack kid home school boy son say autism adult mom u nate really mother daughter holiday said
	anxiety panic attack fear help symptom mental disorder feeling going amy anxious health situation share thought something may control really
	blog post week new month facebook comment write page writing read email reader blogging link might thank last please share
	blog site post twitter comment blogger web read follower article internet tweet doe opinion person lot follow say issue list
	study diet brain level fat risk high acid low vitamin paper omega disease increase increased blood diabetes serotonin woman depression
	relationship felt mother problem couple often help emotional might feeling family partner told person each psychotherapist friend mary spouse article

	health dr clinical psychologist mental university social marijuana drug service use trust national psychology nh state medical trainee foundation abuse
	food eating eat weight body diet sugar sleep fat healthy meal disorder chocolate calorie health bad recipe fruit vegetable use
	woman sexual love father friend sex men mother bob young both family u old man memory swing never mike sister
	little back really got around love going first down night said went last look home new before house room friend
	dsm burner triple diagnosis sleep point diagnostic criterion apnea burnout qi el iv back cpap fluid machine lower yuan between
	alcohol image drinking drink strain alcoholic anger pixel int java alcoholism awt double sober public return value hd w drunk
	week today run mile back last running going hour still again got long feeling really morning home better right little
	mental patient health under hospital act amhp sec assessment police service case social treatment care mha person detained local decision
	medication drug effect patient side use taking problem mg pain addiction prescription dose doctor prescribed should treatment disorder sleep may
	therapy therapist client psychotherapist help trauma feeling psychotherapy often emotional might problem article emdr experience nyc self experiencing part somatic
	psychiatrist problem patient psychiatry psychiatric case treatment care medical point addiction based approach state physician md number diagnosis may pubmed
	test border exam u sat study bank patrol state derrick ben illegal may deductive arizona through security new inductive immigration

	thought depression brain thinking book b dear mind curtis fear negative exercise feeling pattern read thank use pain always chemical
	drug health company zyprexa marketing lilly said should cancer document industry such evidence u mental article new case state research
	disease kennedy cell research article neuron researcher muscle sbma ar treatment motor patient kda muscular study spinal mouse gene new
	mother stress baby brain early birth system level trauma child may later hormone pregnancy change during experience pain those womb
	com u money new world million system book american market company myth doe google job business technology country link say
	qi heat heart yin kidney shen chinese yang blood ren deficiency channel point tongue fire jing woman red liver phlegm
	care hospital patient health state mental psychiatric treatment system service managed problem inpatient unit medical facility staff illness cost bed
	book unschooling learning read really school writing child reading story learn lot unschoolers something education world those might different writer
	patient psychiatrist treatment doe paul talk dr doctor say may should someone tell session medication psychiatric therapy psychotherapy issue therapist
	gun war law veteran firearm soldier weapon bill bear those military u sue country nation american lesbian abraham where violent
	gene dna protein organism genome virus codon c base cell g human genetic sequence bacteria strand mouse acid enzyme two
	code use page using line user data script text app web file should java javascript doe document software content design
	product greece european debt germany herbal greek herb plan ronald vegetarian german energy payment chair wheelchair directive c bank loan

	disorder bipolar mental illness symptom treatment diagnosis mood personality schizophrenia patient psychiatric medication adhd may depression behavior condition episode often
	dog family animal pet support center alzheimer cisco injury self member owner care brian sun emotional harm service dead behavior
	friend painting new place dream u world love where moment beautiful art each through paint experience around never those back
	depression illness depressed mental symptom depressive ect health help treatment brain cause those mood person doctor major feeling better physical
	feeling pain patient u therapy brain those never level primal deep where system imprint love first early idea down mean
	call phone say question doe job insurance pay ask should someone company something number said really tell never money told
	http com www x html org blogspot n b function h video size noise var pseudogenes w new canvas yahoo
	myself feeling thought never going really something help love say back still through better fear friend those bad felt again
	suicide mental health week u book death law prevention violence president article mass jesse new france story act york post
	shrink psychiatrist roy patient psychiatry rap post podcast dinah clink blog new u doctor comment clinkshrink talk doc medical duck

6 Definitions

Should I provide a strict mathematical definitions here?

The Dirichlet distribution is a multivariate probability distribution that represents the probabilities x_i of $K > 2$ distinct categories such that

$$\sum_{n=1}^{K} x_i = 1, x_i \in (0, 1)$$

Multinomial distribution

The Multinomial distribution is often referred as *categorical distribution* in natural language processing context. It is a discrete probability distribution that represents the outcomes of a random variable that can take one of N categories.

7 Conclusion

what did you do?

What are the results?

future work?

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Alice Cooper**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

   1.1 reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

   1.2 make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

   of my thesis

   Type Inference for Fourth Order Logic Formulae

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, dd.mm.yyyy