Best Practices in Building Topic Models with LDA for Mining Regulatory Textual Documents

NCTR CTP Working Group
Speaker: Weizhong Zhao, Ph.D.
Supervisor: James Chen Ph.D. and Wen Zou Ph.D.

Motivation

Topic Modeling Goal Is To Facilitate Information Retrieval

Facilitating finding needles in haystack by shrinking the haystack

Inherent Limitations to Bear in Mind:

•Shortcoming of Topic Modeling, as well as <u>All</u> text mining of unstructured corpora: Model validation is mainly subjective

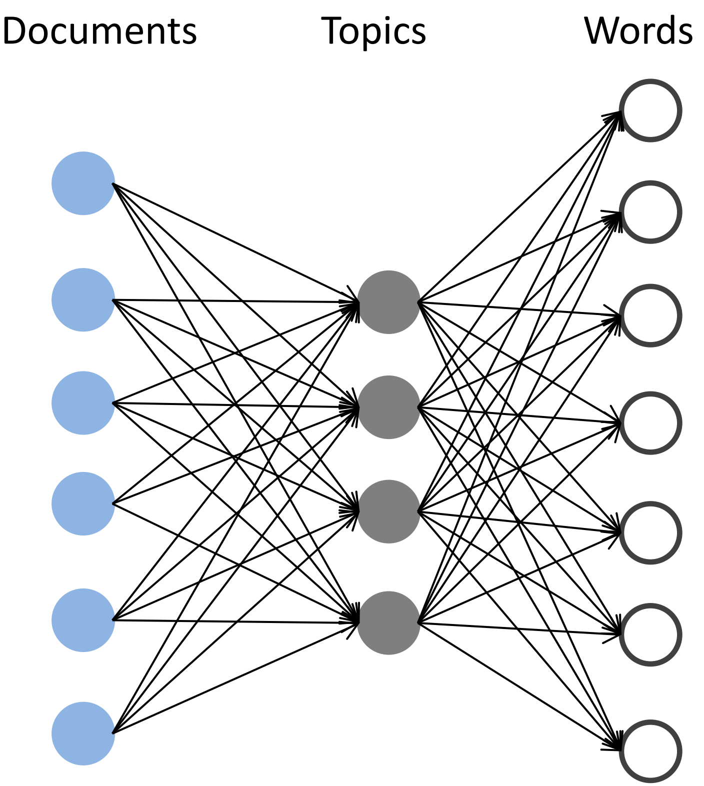
•No quantitative means to measure if truth has been found, when truth is not known *a priori*

•Limited quantitative means to measure fit to data or prediction accuracy

•Topic modeling is data-driven, unsupervised learning

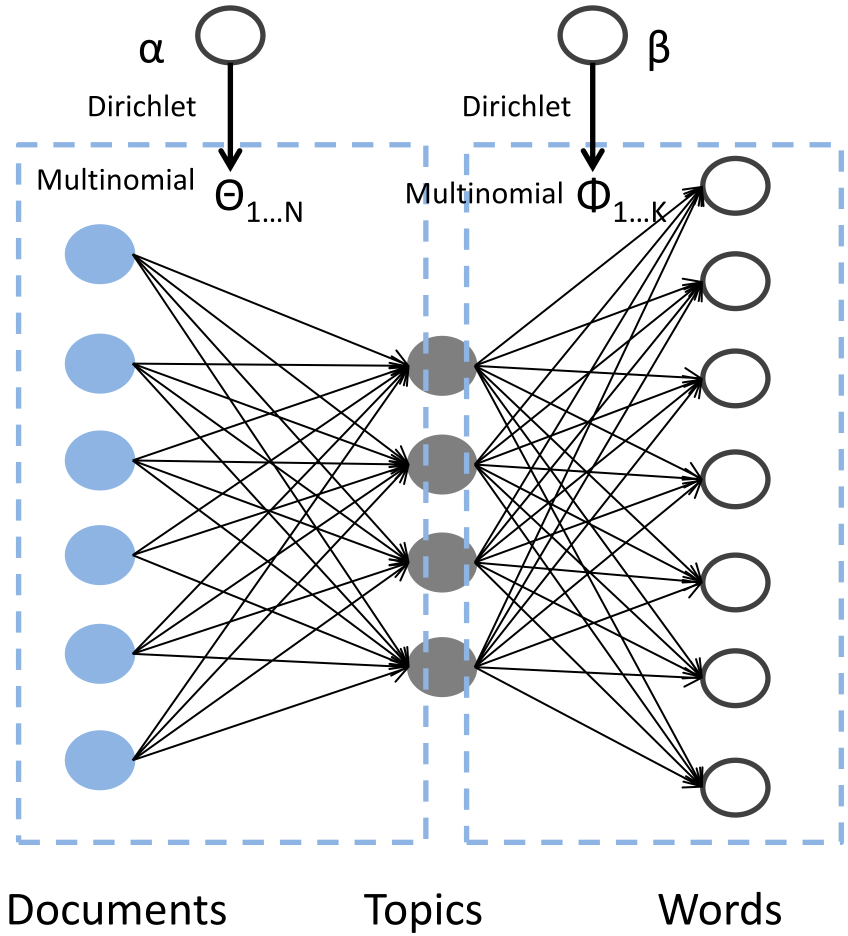
Topic modeling

- Topic models are algorithms for discovering the main themes that pervade a large collection of documents.
- Definitions
  - Word: an item from a vocabulary indexed by $\{1,\ldots,V\}$.
  - Document: sequence of M words denoted by $d = \{w_1, w_2, \ldots, w_M\}$, where w_i is the ith word in the sequence.
  - Corpus is a collection of N documents, denoted by $D = \{d_1, d_2, \ldots, d_N\}$



Latent Dirichlet Allocation

- Latent Dirichlet Allocation (LDA), which is the most popular topic modeling approach, has proved to be an effective tool in text mining field.



Illustrative Workflow

Using a ground truth corpora

Develop ground truth data set

Remove duplicates

Topic modeling (LDA)

Compare topic word distributions with themes of PubMed searches (visualize with word clouds)

Search PubMed abstracts using pertinent MeSH[1] terms

e.g.,

Search: ("Tobacco Use"[MeSH] OR "Smoking"[MeSH]) AND (“Lung Neoplasms”[MeSH]) to retrieve a “**theme”** of abstracts related to “smoking ” and “lung cancer”;

We had **41** themes in all, with many intentionally overlapping

[1]MeSH (Medical Subject Headings) is the NLM controlled vocabulary thesaurus used for indexing PubMed articles

Use MeSH Terms to Search PubMed for Themes

("Tobacco Use"[MeSH] OR "Smoking"[MeSH]) AND " themes below":

We have 41 themes

Smoking related diseases;
number of abstracts; **18 themes**

- Cardiovascular diseases; 15418
- Brain diseases; 3033
- Heart diseases; 7091
- Stroke; 1003
- Cerebrovascular disorders; 2216
- Pulmonary diseases; 3033
- Asthma; 2009
- Trachea; 141
- Pharynx; 89
- Nervous system diseases; 5210
- Parkinson disease; 228
- Immune system diseases; 4634
- Diabetes Mellitus; 5079
- HIV; 87
- Periodontal diseases; 1790
- Thromboangiitis Obliterans; 141
- Kidney diseases; 1062
- Osteoporosis; 411

Smoking related other health issues; number of abstracts; **7 themes**

- Cessation; 15429
- Pregnancy; 6951
- Aging; 1558
- Gender difference; 765
- Addiction; 638
- Dependency; 12
- Child malnutrition; 10

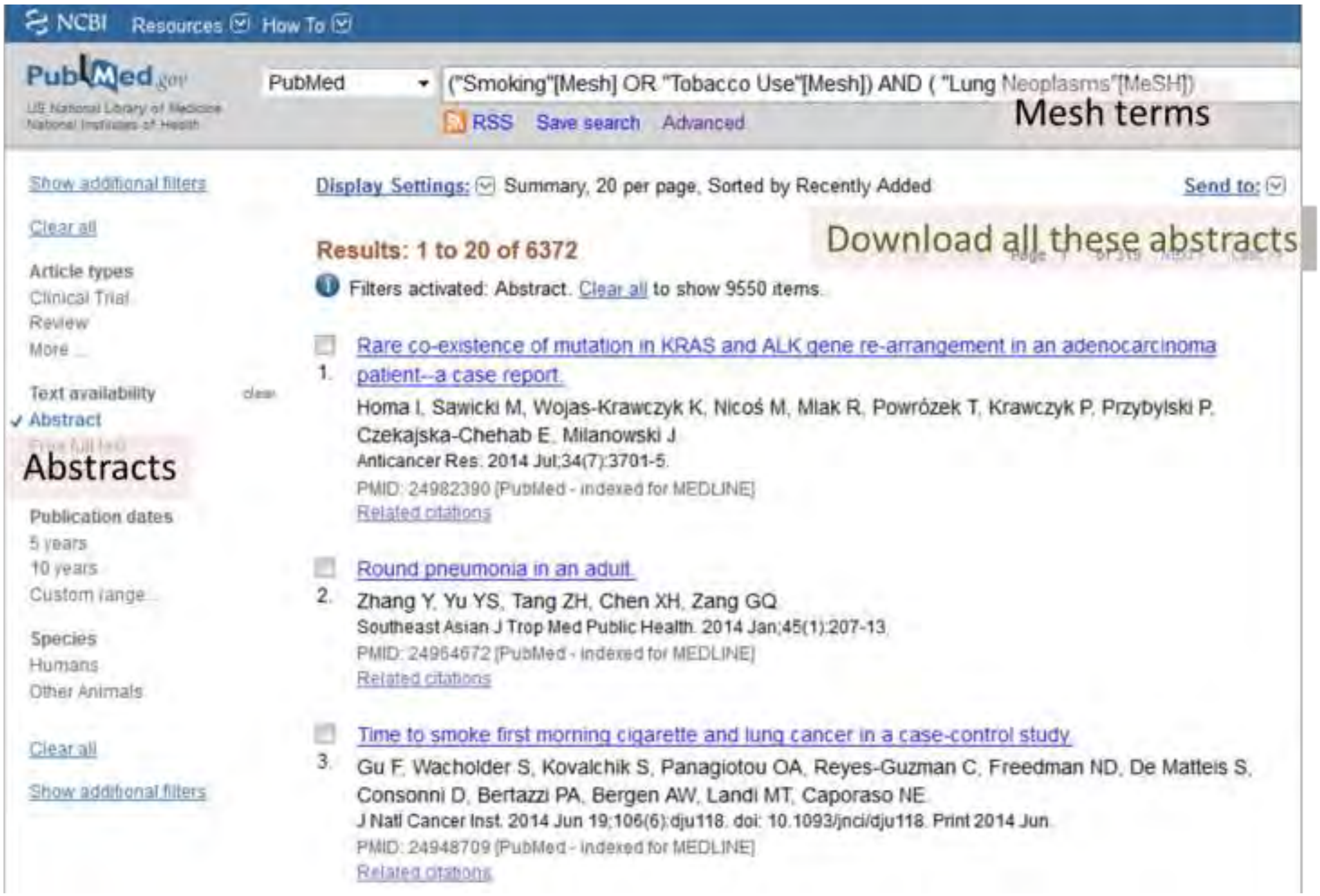
Smoking related cancers;
number of abstracts; **14 themes**

- Lung cancer; 6345
- Laryngeal cancer; 617
- Kidney cancer; 203
- Mouth cancer; 1551
- Stomach cancer; 526
- Nose cancer; 77
- Urinary bladder cancer; 838
- Uterine cervical cancer; 488
- Ureteral cancer; 26
- Breast cancer; 812
- Pancreatic cancer; 411
- Hematologic cancer; 16
- Esophageal cancer; 661
- Liver cancer; 276

Negative controls having
no association with smoking;
2 themes

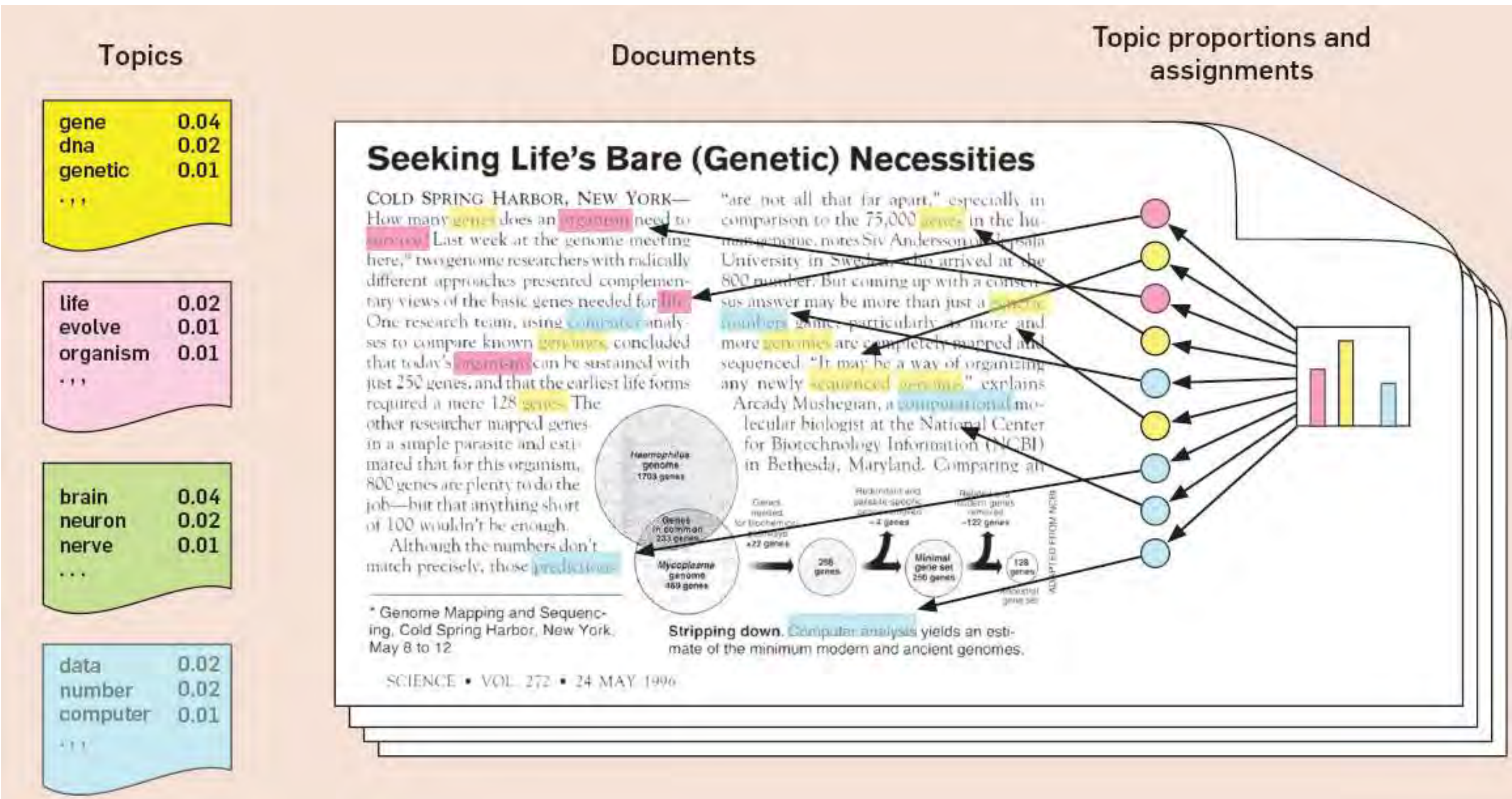
- Foot injury; 2201
- Lupus Vulgaris; 466

PubMed Document Retrieval



Topic Modeling

Topic modeling(LDA)

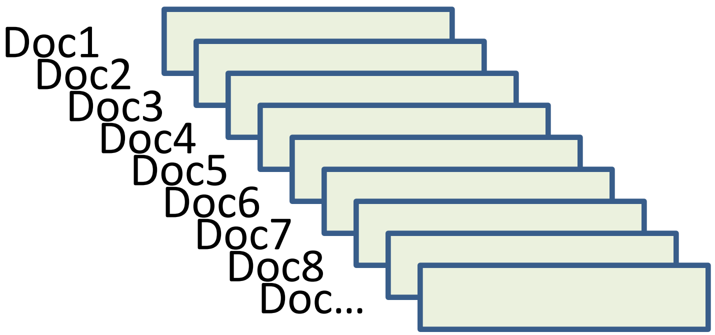


Identifying the relevant topics for themes

Topics with higher normalized mean probability values in each theme are the relevant topics

Normalization

Docs in theme 1



Per-doc topic distribution

Topic distribution of doc1

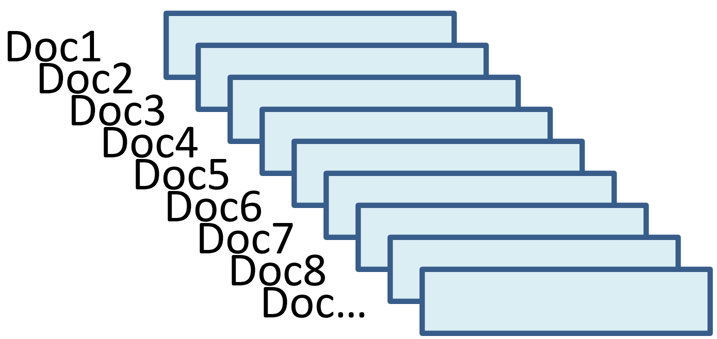
Topic distribution of doc2

Topic distribution of doc3

……

Calculate the averaged topic distribution for all docs in theme 1

Docs in theme 2



Topic distribution of doc1

Topic distribution of doc2

Topic distribution of doc3

……

Calculate the averaged topic distribution for all docs in theme 2

.
.
.

Visualizing Topic-Word Multinomial Distributions

Topic 34: heart diseases

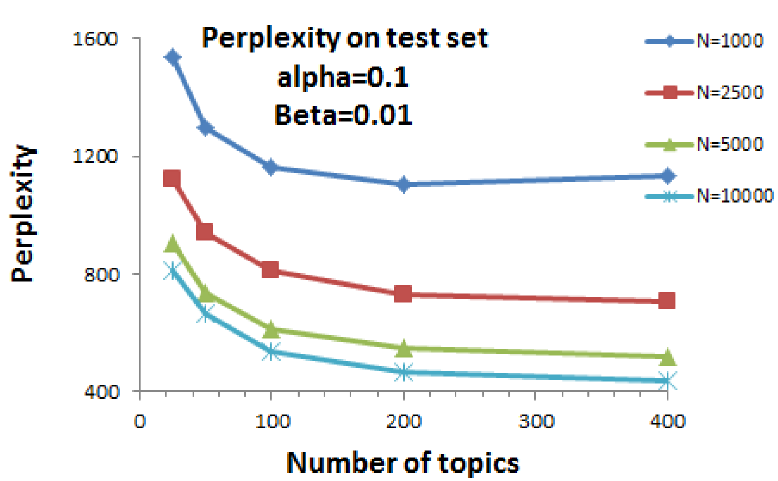
I: Sensitivity studies*: determine modeling parameters for topic modeling

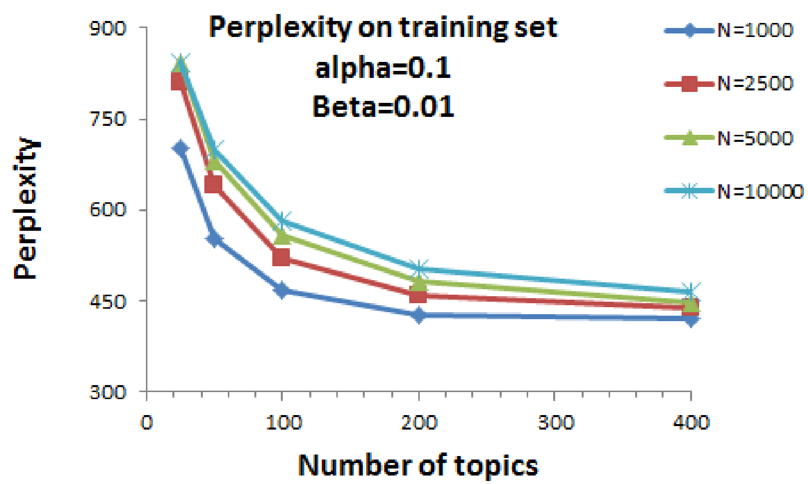
- Parameters:
  - Topic number, T
    - How good to characterize the dataset
  - Alpha
    - Control document topic matrix
  - Beta
    - Control topic word matrix
- Perplexity and 4-fold cross validation

* LDA will usually quickly yield good and usable models just using default code parameters, but sensitivity studies are warranted for obtaining best models

I: Sensitivity studies: determine modeling parameters for topic modeling

How number of topics affects perplexity



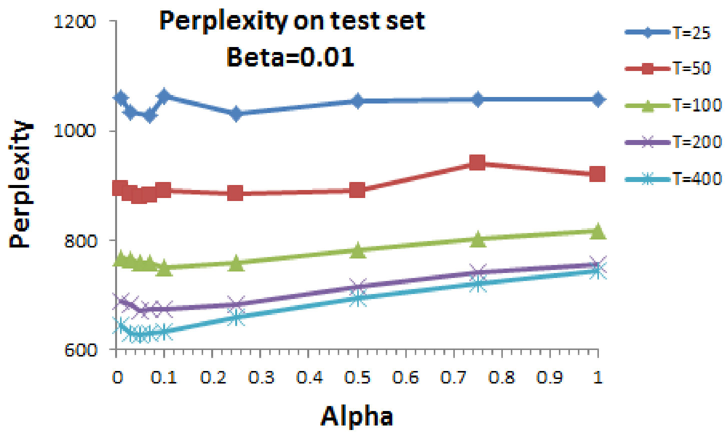


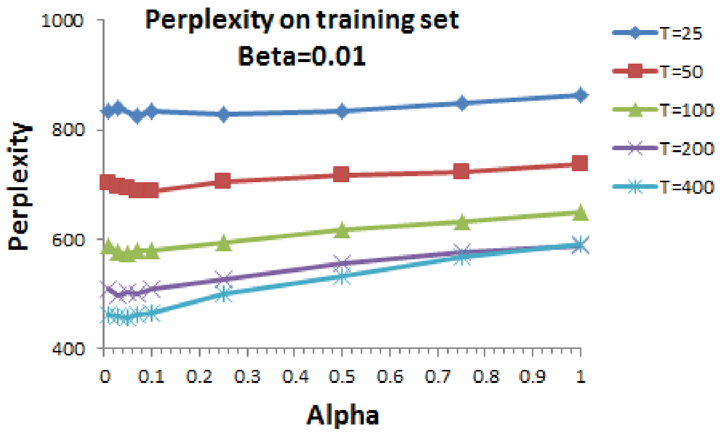
- Beta: 0.01; - Alpha: 0.1; - T: 25, 50, 100, 200, 400; - Size of training dataset: N=1000, 2500, 5000, 10000
- Test set (the remaining 25% of the whole data); - #iteration=200; - Model evaluation (Perplexity)
- LDA implementation: Mallet LDA

With statistical perplexity the surrogate for model quality, a good number of topics is 100~200

I: Sensitivity studies: determine modeling parameters for topic modeling

Dirichlet hyperparameter α affects perplexity



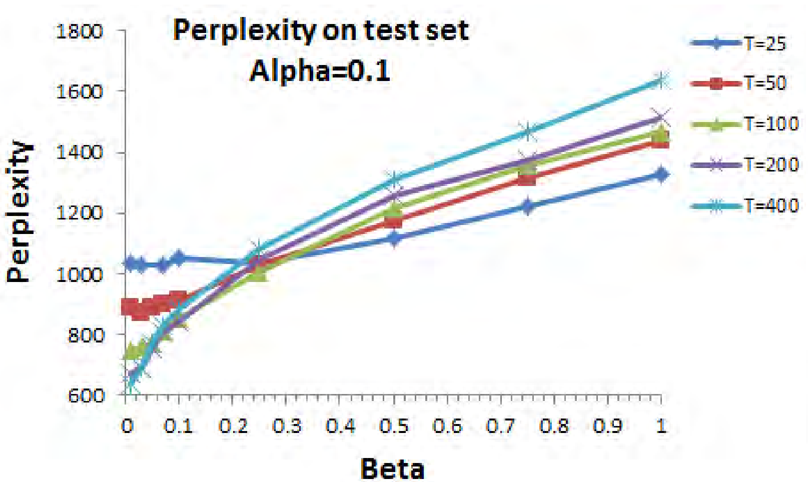


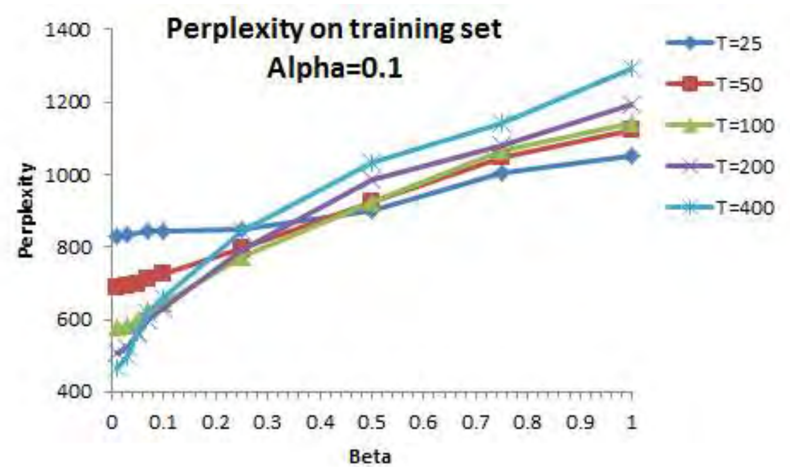
- Alpha: 0.01-1.0; - beta: 0.01; - T: 25, 50, 100, 200, 400; - Training set (75% of the combined data);
- Test set (the remaining 25% of the whole data); - #iteration=200; - Model evaluation (Perplexity)
- LDA implementation: Mallet LDA

The α "sweet spot" is [0.01, 0.1]
Over fitting not yet apparent even for T = 400

I: Sensitivity studies: determine modeling parameters for topic modeling

Dirichlet hyperparameter β affects perplexity





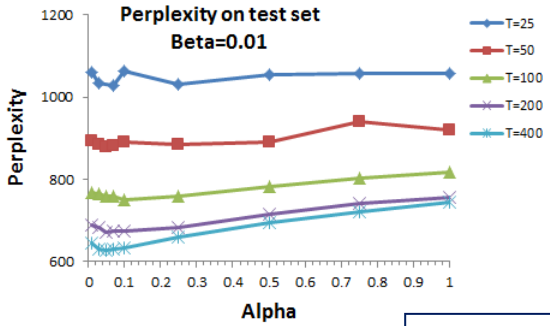
- Beta: 0.01-1; - Alpha: 0.1; - T: 25, 50, 100, 200, 400; - Training set (75% of the combined data);
- Test set (the remaining 25% of the whole data); - #iteration=200; - Model evaluation (Perplexity)
- LDA implementation: Mallet LDA

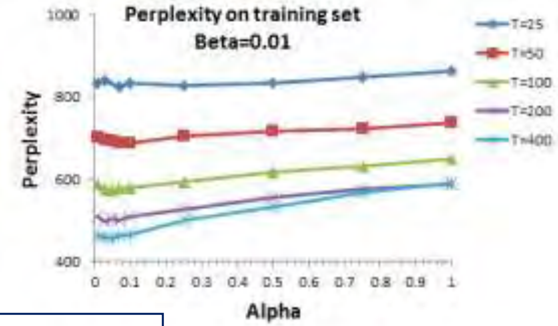
The Beta value 0.01 usually derives the best topic model for the dataset

I: Sensitivity studies: determine modeling parameters for topic modeling

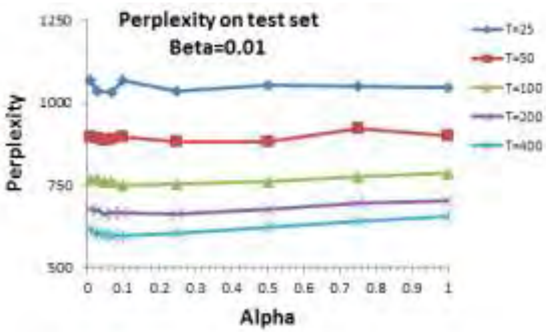
Symmetric Alpha Vs. Asymmetric Alpha

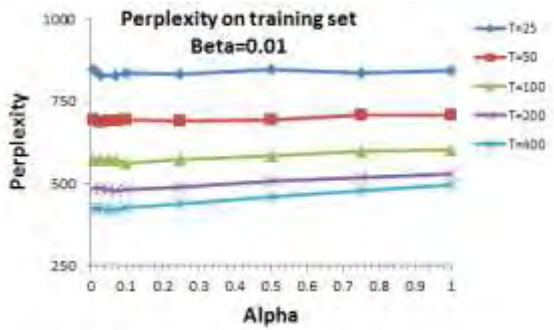
Symmetric alpha





Asymmetric alpha





Perplexity from asymmetric alpha is more stable than symmetric alpha in range of 0.01-1.0

II: Validation: find the ground truths embedded in the documents

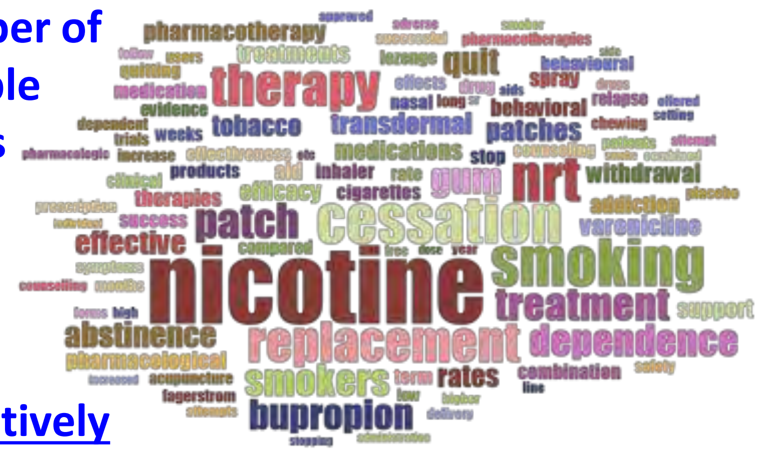
Q1: Can topic modeling find ground truths?

Topics most relevant to ground truth: smoking and cessation (26% of total abstracts)

group=max,tota	40	58	91	63	27	82	46	37	99	20	Topic ID
c[0]=40,15429	0.273349	0.261702	0.238897	0.212858	0.203085	0.164646	0.14902	0.131505	0.123813	0.11775	Normalized prob.

Topic 40: cessation programs

Topic 58: cessation therapy / treatment



Themes with large number of abstracts have multiple relevant subthemes

Topic concept is subjectively defined by the prevalence of words in topics

Topic 91: studies of intervention for cessation

Topic 63: training and education for cessation

Validation: find the ground truths embedded in the documents

Question: Can topic modeling delineate intentionally overlapped ground truths?

Highly overlapped ground truths

C[1]-Ground truth: smoking and cardiovascular diseases; 26% of total abstracts

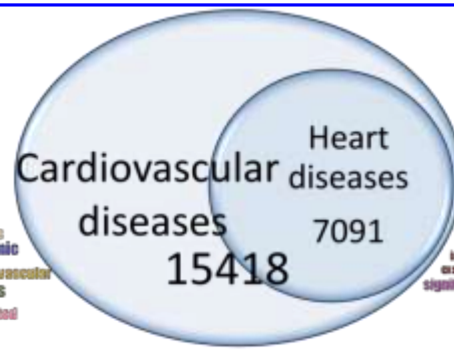
C[2]-Ground truth: Smoking and heart diseases; 12% of total abstracts

group=max,tota	21	36	29	34	6	18	31	45	23	3
c[1]=21,15418	0.14336	0.125199	0.125079	0.121004	0.104548	0.092294	0.089697	0.085339	0.079403	0.066964
group=max,tota	21	34	29	6	45	36	31	18	93	23
c[2]=21,7091	0.29152	0.225163	0.142722	0.121275	0.109409	0.077454	0.075152	0.071983	0.071689	0.05739

For first-10 topics relevant to these 2 themes, 90% are overlapped

First-2 topics relevant to these 2 themes differentiate overlapped truths

Topic 21: cardiovascular diseases



Topic 34: heart diseases

Topic 36: Hypertension

Less overlapped ground truths

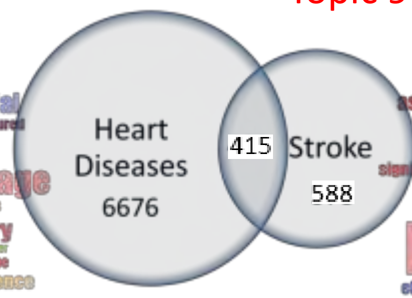
C[2]-Ground truth: Smoking and heart diseases; 12% of total abstracts

group=max,total	21	34	29	6	45	36	31	18	93	23
c[2]=21,7091	0.29152	0.225163	0.142722	0.121275	0.109409	0.077454	0.075152	0.071983	0.071689	0.05739

C[16]-Ground truth: Smoking and stroke; 1.7% of total abstracts

group=max,tota	88	97	6	34	31	36	93	62	18	23
c[16]=88,1003	0.302609	0.092546	0.090957	0.079409	0.07697	0.076393	0.072904	0.069036	0.065412	0.060535

Topic 88: stroke



Topic 97: mortality of cardiovascular diseases

The overlapped themes are observed

Topic 6: risk factor of cardiovascular diseases

Topic 34: heart diseases

II: Validation: find the ground truths embedded in the documents

Q3: How sensitive are topic models in detecting themes with fewer documents?

Truth sets with fewer abstracts

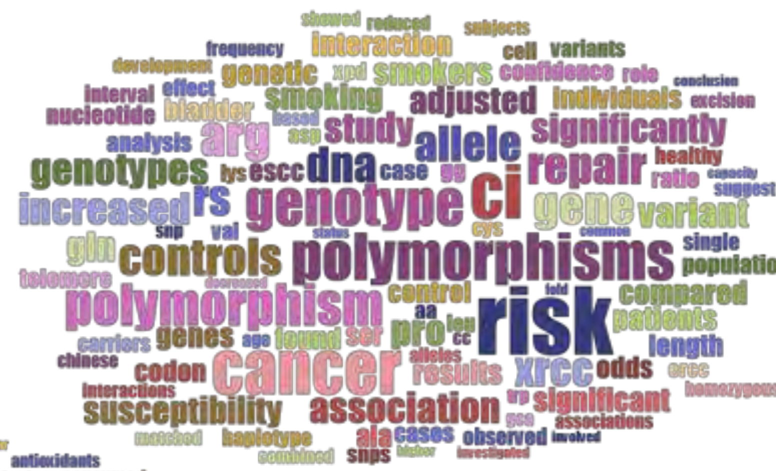
C[23]-Ground truth: Smoking and stomach cancer; 0.9% of total abstracts

group=max,tota	35	7	0	83	14	95	48	57	4	10
c[23]=35,526	0.154949	0.151049	0.136513	0.076421	0.074841	0.063053	0.049868	0.049189	0.045205	0.044938

With 0.9% of total docs, the relevant topics are associated with the corresponding theme

Topic 35: gastric and bladder cancer

Topic 7: gene polymorphisms



Topic 0: nutrition

Associations between genetic polymorphisms and gastric cancer

-pubmed/19375306

Note: Nutrition is an important *stomach cancer Treatment*

-pubmed/8850434

Truth sets with fewest abstracts

C[37]-Ground truth: Smoking and child malnutrition; 0.017% of total abstracts

group=max,tota	52	65	78	37	46	13	33	96	89	10	Topic ID
c[37]=52,10	0.483682	0.263421	0.21407	0.199289	0.191036	0.148235	0.123547	0.108536	0.098856	0.098538	Relative prob.

EVEN **with minuscule 0.017% of total docs (10/59000), topic is well differentiated**

Topic 52: children's exposure of smoking

Topic 65: physical examination

II: Validation: find the ground truths embedded in the documents

Q4: Can topic modeling identify the intruding documents, i.e., negative controls?

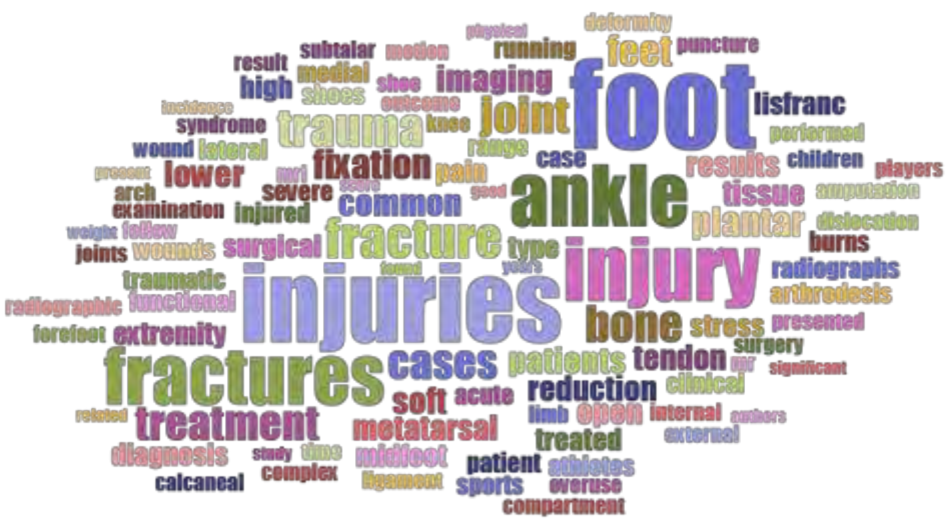
Negative control truth set

C[39]-Ground truth: Foot injury; 3.7% of total abstracts

group=max,tota	66	24	92	71	45	84	5	80	9	2
c[39]=66,2201	0.885649	0.62826	0.12692	0.080118	0.06674	0.061733	0.043651	0.036649	0.026148	0.025881

Obtuse negative control themes topic differentiated by distinct subthemes

Topic 66: foot injuries



Topic 24: foot reconstruction

Conclusions

- Topic modeling easily distinguishes ground truths in quality documents across ***<u>many</u>*** themes, and even reveals numerous subthemes.
- Topic modeling also differentiates overlapped ground truths.
- Themes with minimal documents (e.g., <.1% of total documents) can be detected by topic modeling.
- Topic modeling can recognize the intruding themes (i.e., negative controls).
- Topic modeling appears to find the truth, if it's there to be found.

Acknowledgement

- Supervisors: Wen Zou and James Chen
- CTP Project team members:

Roger Perkins, Yijun Ding, Ke Yu, Shiheng Wang, and Joe Meehan

- Weigong Ge

Thanks to ORISE and CTP for supporting these studies.