Bayer META (Message Exchange Text Analytics)

Mid-term Presentation

Maggie Lu, Siddharth Menon, Yiling Zhong, Olivia Zheng

October 26, 2018

Agenda

- Project Overview
- NLP (Natural Language Processing) Algorithm
  - Data collection and cleaning
  - Algorithms
  - Topic/Author modeling
  - Dynamic Modeling
- Application Pipeline
- UI Mockup
- Q & A

Project Overview

- Problem Statement

The problem of	Siloed communication among different teams and departments
Affects	All who work on innovative projects
The impact of which is	Lack of collaboration resulting in unachieved potential in productivity
A successful solution would be	A tool that can synergize information and model topics from email correspondence, connect relevant parties with useful information and provide the opportunities to collaborate.

Project Overview

- Product Position Statement

For	Bayer scientists, technicians, as well as business professionals
Who	Work on multiple projects across multiple functional departments
Our solution	Would connect previously isolated parties with keywords from their email correspondence, intranet knowledge base, and academic publications as well as provide relevant internet resources
That	Promote effective and efficient collaboration

Needs and Features

- Top Priority Needs
  - Keyword Identification
  - User Collaboration Identification
  - Dynamic Learning
- Secondary Needs
  - Resource Discovery
  - User Notification
  - Tiered Access

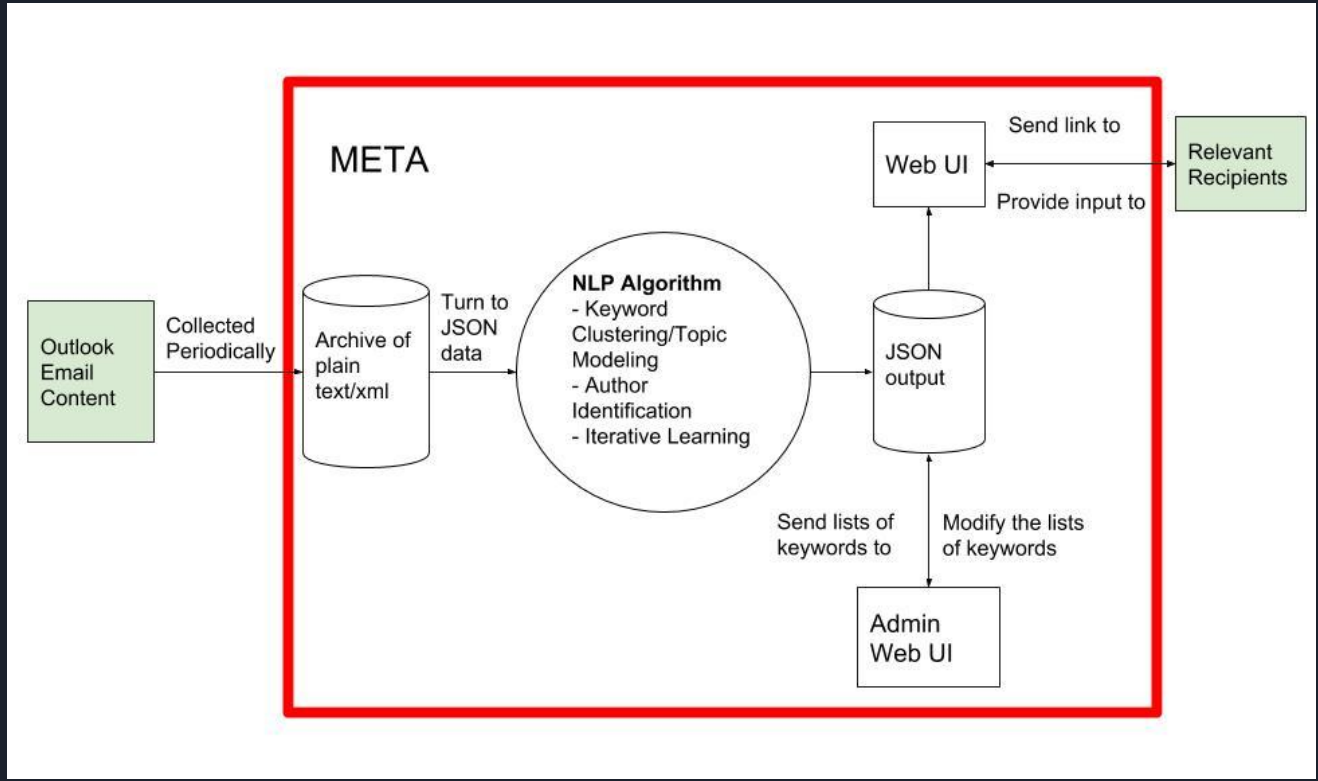
Project Timeline and Scope

- POC(Proof of Concept): late-August 2018 to mid-October 2018
  - Define data sources
  - Research Natural Language Processing Algorithms
  - Research and demo the application pipeline
- Development and Implementation: late-October 2018 to mid-December 2018
  - Develop and construct the application pipeline
  - Implement working NLP algorithms
  - Address medium priority needs and features
  - Configure and debug code
- Deployment: late-December 2018 and onward
  - Deploy the system within Bayer's internal network

POC Deliverables

- Design Document that details:
  - System Architecture
  - Analytical Methodologies
  - Integration Requirements
- Web Application Demo
  - Hosted on publicly accessible PaaS (Platform as a Service)
  - Does not include Bayer email server connectivity

High-Level System Diagram



Analytical Methodology: Natural Language Processing (NLP)

- ❏ Data Collection and Cleaning
- ❏ NLP Algorithm Deep-Dive
- ❏ Author-Topic Modeling
- ❏ Dynamic Modeling

Data Collection and Cleaning

From email topic grouping to knowledge discovery and knowledge sharing

Data Collection and Cleaning

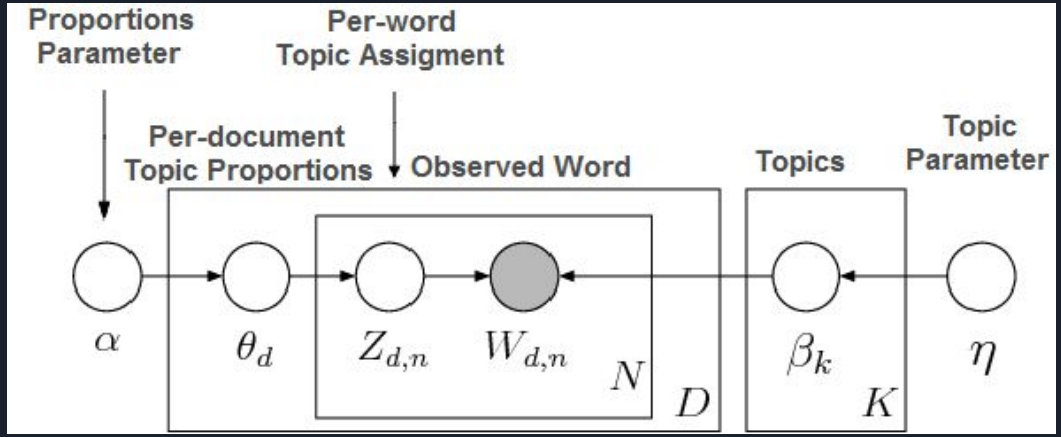
Data collected from Scholarly Journals and Converted to Email Format

- ➔ Sender : Author 1
- ➔ Receiver: Author 2
- ➔ Title: Title of Journal Article
- ➔ Content: Abstract Text

The properly formatted data were then stored in json files

NLP Deep Dive-**LDA**

Latent Dirichlet Allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.



NLP Deep Dive-**TFIDF**

TFIDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

Benefit of TF IDF: helps us to filter out common but not important adjective or verb in the documents, and therefore gives the model a clearer result.

$$\mathrm{TF-IDF} = \mathrm{tf}(t_i, q) \times IDF(t_i)$$

$$TF(t, d) = freq(t, d)$$

$$IDF(t, D) = log_2 \frac{|D|}{|\{d \in D | t \in d\}|}$$

NLP Deep Dive-**Mallet**

Developed by University of Massachusetts, MALLET topic model package includes:

- An extremely fast and highly scalable implementation of Gibbs sampling, efficient methods for document-topic hyperparameter optimization, and tools for inferring topics for new documents given trained models.
- This option turns on hyperparameter optimization, which allows the model to better fit the data by allowing some topics to be more prominent than others. Optimization every 10 iterations is reasonable.

Source: http://mallet.cs.umass.edu/

NLP Deep Dive-**Models Comparison**

The state-of-the-art in terms of topic coherence are the intrinsic measure UMass and the extrinsic measure UCI, both based on the same high level idea. Both measure compute the sum:

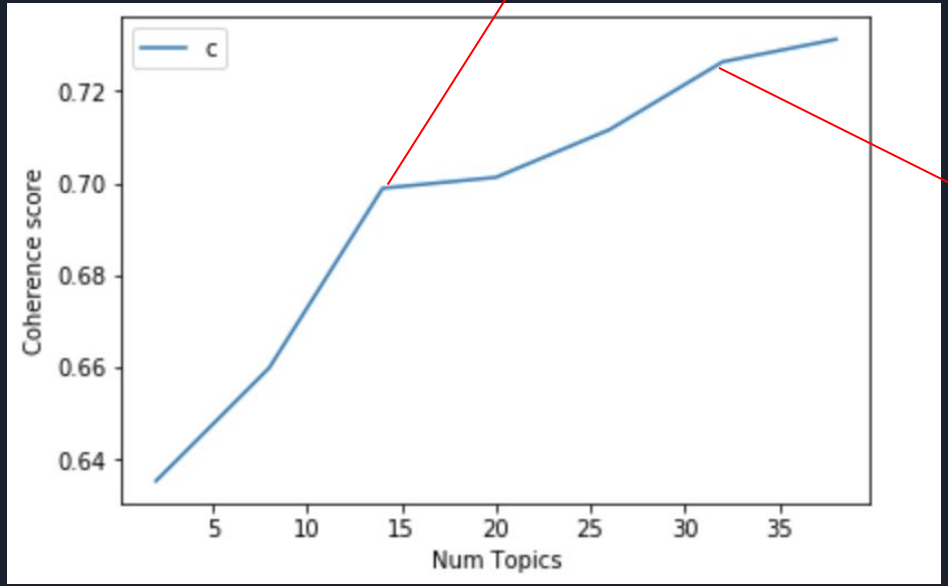
$$\text{Coherence}=\sum_{i<j}\text{score}(w_i,w_j)$$

of pairwise scores on the words w1, ..., wn used to describe the topic, usually the top n words by frequency p(w|k).

Result : LDA_Mallet > LDA >LDA_tfidf

NLP Deep Dive-**Choose K topics**

Local optimal n = 15



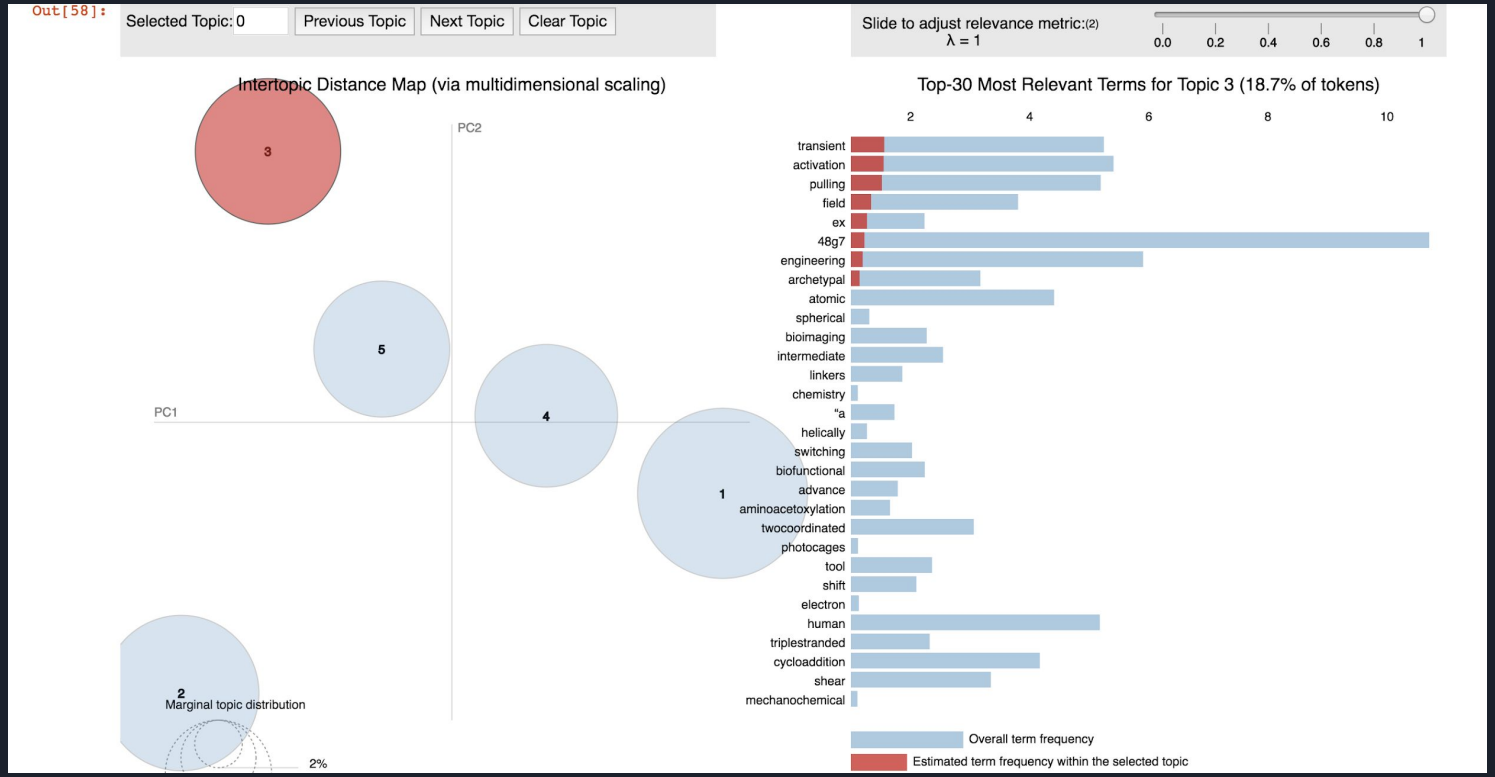
Local optimal n = 32

NLP Deep Dive-Initial Outcome

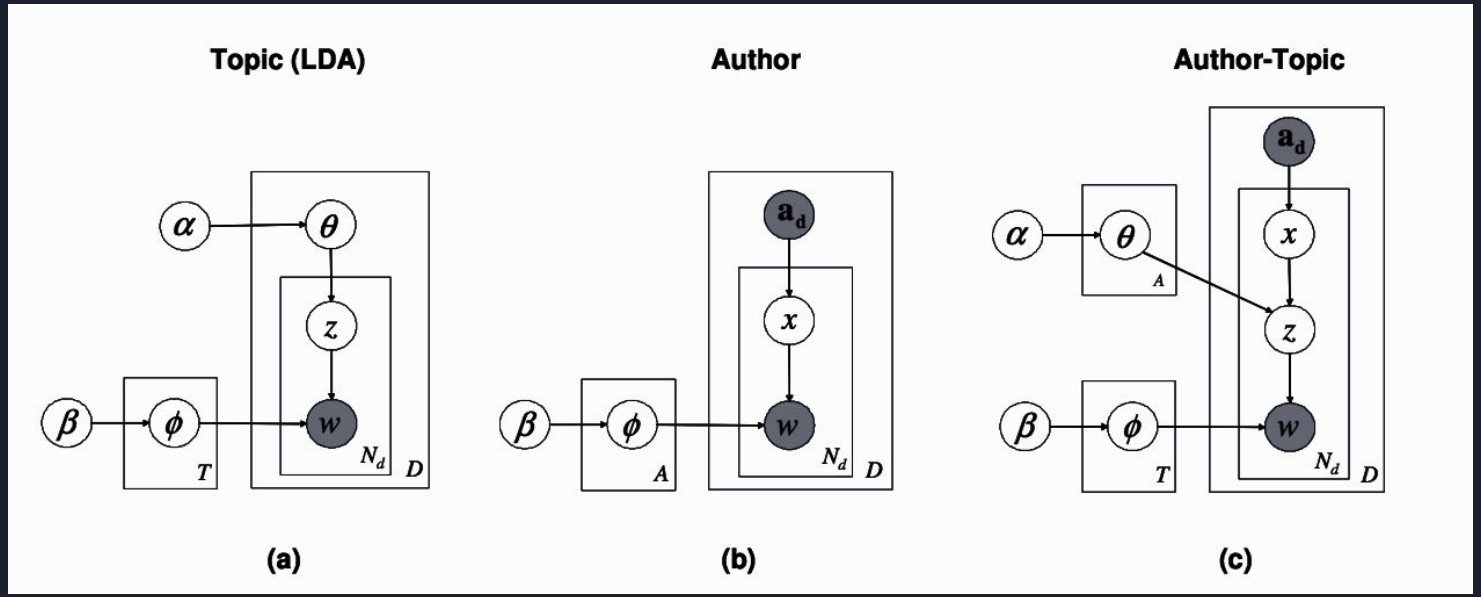
```
for idx, topic in lda_model_tfidf.print_topics(-1):
    print('Topic: {} \nWords: {}'.format(idx+1, topic))
```

```
Topic: 1
Words: 0.015*"48g7" + 0.007*"pulling" + 0.007*"atomic" + 0.006*"activation" + 0.005*"enhancement" + 0.004*"nb5" + 0.004*"relaxation" + 0.004*"
ilica" + 0.004*"metal–organic"
Topic: 2
Words: 0.005*"transient" + 0.005*"activation" + 0.005*"pulling" + 0.005*"field" + 0.004*"ex" + 0.004*"48g7" + 0.004*"engineering" + 0.004*"arc
*"atomic" + 0.003*"spherical"
Topic: 3
Words: 0.005*"strategy" + 0.005*"48g7" + 0.004*"eu" + 0.004*"twocoordinated" + 0.004*"large" + 0.003*"relaxation" + 0.003*"derived" + 0.003*"c
*"dramatic" + 0.003*"fefe"
Topic: 4
Words: 0.012*"engineering" + 0.005*"human" + 0.004*"transient" + 0.004*"48g7" + 0.004*"nanowire" + 0.004*"physiological" + 0.004*"group" + 0.0
0.003*"cycloaddition" + 0.003*"activation"
Topic: 5
Words: 0.009*"automated" + 0.007*"peroxisomal" + 0.007*"silica" + 0.006*"48g7" + 0.005*"human" + 0.005*"eu" + 0.005*"cycloaddition" + 0.005*"p
*"ice" + 0.004*"assembly"
```

NLP Deep Dive-**Visualization**



Author-Topic Modeling



Source: Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (n.d.). *The Author-Topic Model for Authors and Documents,* retrievable at: https://mimno.infosci.cornell.edu/info6150/readings/398.pdf

Author-Topic Modeling

- Compare the Performance of Three Models
  - LDA
  - LDA-tfidf
  - LDA Mallet

- Explore and Draw Insight using
  - Coherence Scores for Model Performance Comparison
  - Hellinger Distance to measure closeness of authors
  - t-SNE flattening for Visualization

Author-Topic Modeling

```
Topic 9:
worldwide reference provide_insight
provide_insight reference worldwide offer science critical functional cognitive frequency surgery

Topic 10:
extract march reveal
reveal march extract apply conclude help depend dose china document

Topic 11:
use previously disorder
disorder previously use variable acid easily record sectional fix breast_cancer

Topic 12:
therefore cancer intrinsic
intrinsic cancer therefore exposure site since demand laboratory energy simulate

Topic 13:
vitro strongly nuclear
nuclear strongly vitro worldwide coefficient duration functional theoretical cochrane yield

Topic 14:
tumor chemical magnetic_resonance
magnetic_resonance chemical tumor document mode investigation range assembly learning child
```

Author-Topic Modeling

```
Ying Zhang
Docs: [1667, 3
Topics:
[('expression
 ('difference
 ('plant speci
```

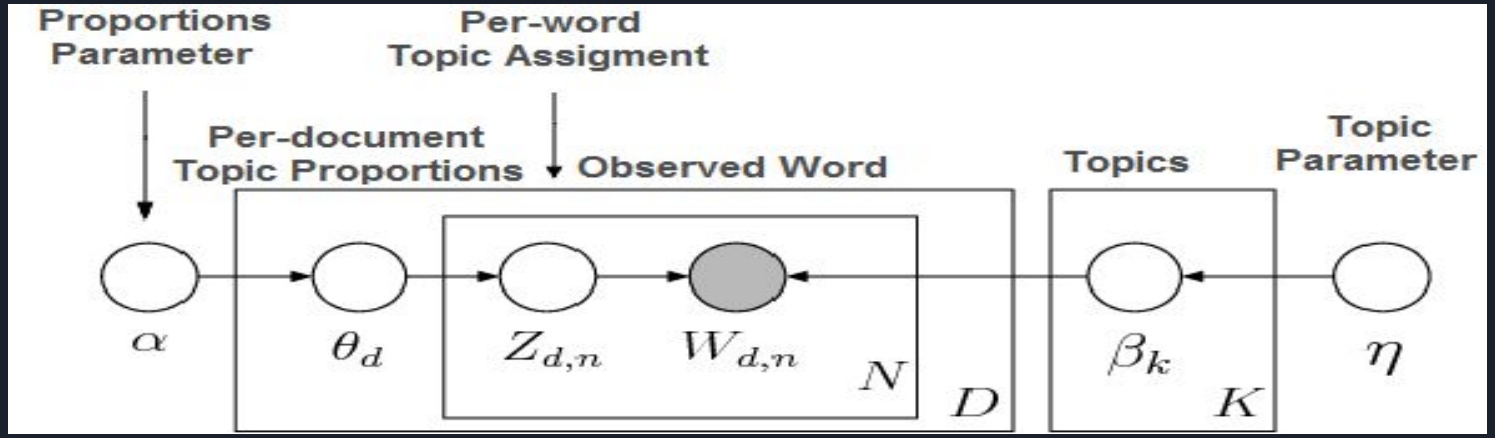
```
36709742),
12291338181),
9109882595)]
```

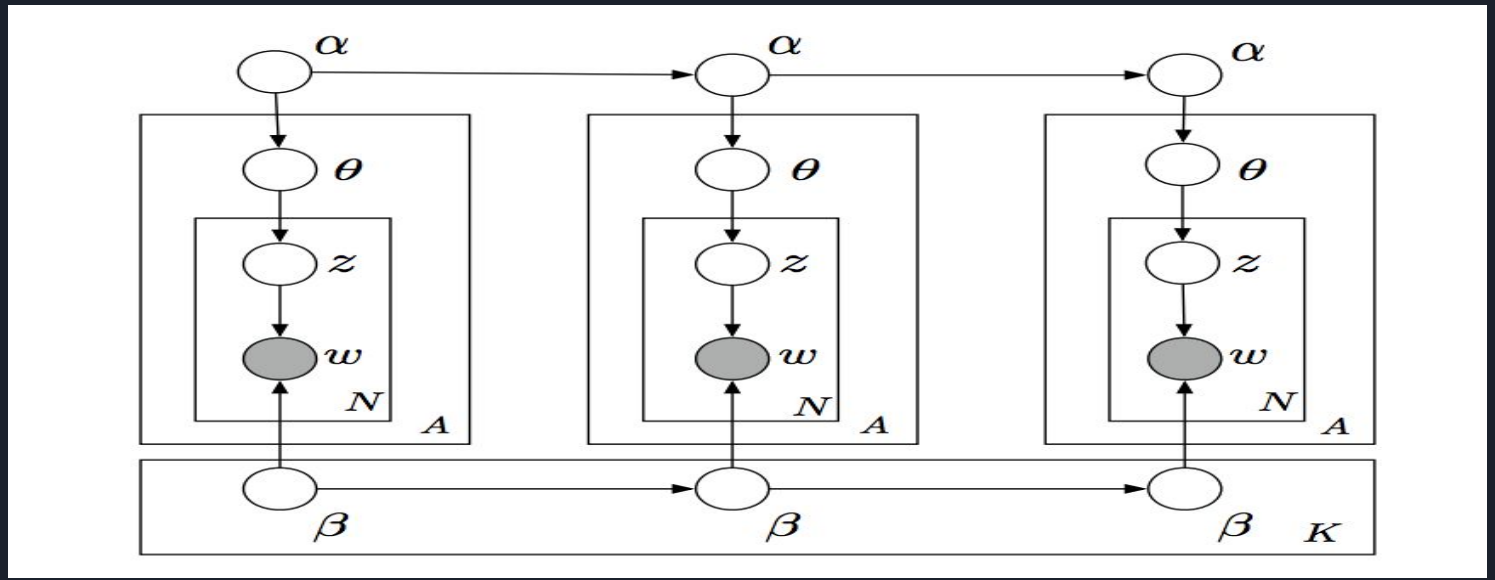
	Author	Score	Size
124	Ying Zhang	1.000000	3
119	Yan Chen	0.695471	4
131	Yuanyuan Li	0.682265	3
125	Yong Liu	0.673982	3
41	Jian Xu	0.665912	4
28	Hui Li	0.664160	3
123	Ying Wang	0.647400	4
71	Nuria K Koteyeva	0.642060	3
133	Yun Li	0.641408	3
127	Yu Zhang	0.615359	3

Dynamic Modeling

In LDA, one underlying assumption is the order of documents does not matter

- ➔ Dynamic modeling captures language and topic changes over time
- ➔ Uses a series of LDA-like topic models that are tied together to model documents in which order is important


Proportions Parameter
Per-word Topic Assigment
Per-document Topic Proportions
Observed Word
Topics
Topic Parameter
α
θ_d
$Z_{d,n}$
$W_{d,n}$
N
D
β_k
K
η

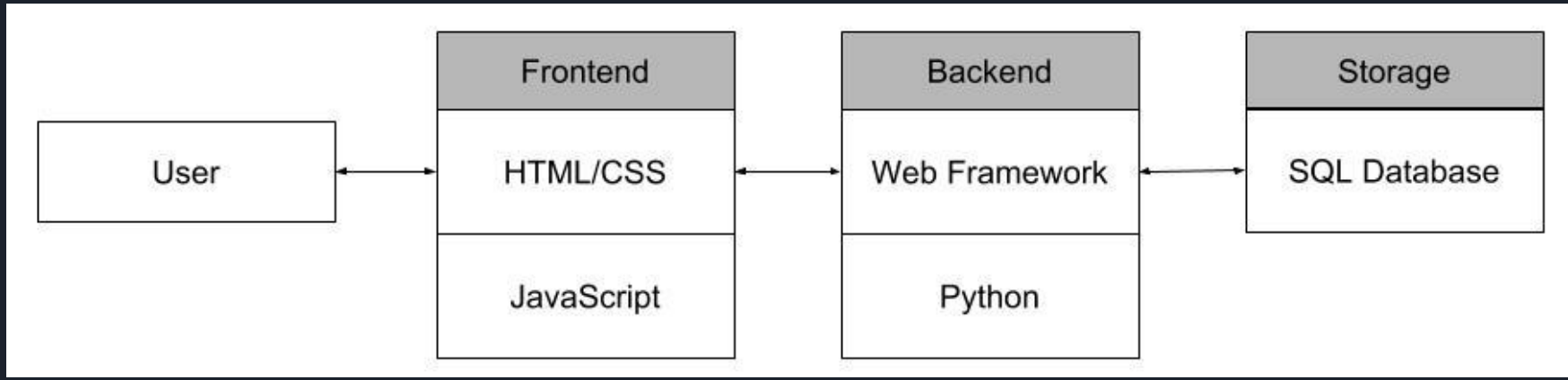

α
θ
z
w
N
A
β
K

Dynamic Modeling

Models the corpora as a collection of topics with keywords evolving over time

- ➔ Take the Harry Potter books as an example
  - ◆ Seven documents, each in its own time group
  - ◆ For the *Voldemort* topic, each document returns different keyword
    - Book 1: Voldemort, Stone, Evil, Quirrell …
    - Book 2: Voldemort, Chamber, Petrify, Basilisk …
    - Book 3: Voldemort, Pettigrew, Hogsmeade, Azkaban ...

Application Pipeline



References

- Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (n.d.). *The Author-Topic Model for Authors and Documents*, retrievable at: https://mimno.infosci.cornell.edu/info6150/readings/398.pdf
- Blei, DM., Lafferty, JD. *Dynamic Topic Models*, retrievable at: https://mimno.infosci.cornell.edu/info6150/readings/dynamic_topic_models.pdf
- http://mallet.cs.umass.edu/
- https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

Thank You!