Topic Discovery and Evolution Through Social Media

Zachary Douglas Vink

Division of Science and Mathematics
University of Minnesota, Morris
Morris, Minnesota, USA

5 December 2015

Outline

Outline

Purpose

Problem: Increasing number of articles and papers that need categories

Solution: Discover how to categorize these documents by their topics

Outline

Key Terms

Document: A collection of text conveying at least one idea

Corpus: A collection of documents

Stemming: Reducing a word to its base form (Ex. *Exercising* -> *Exercise*)

Vocabulary: Each stemmed word in the corpus

Key Terms (Cont.)

Distribution: Assigns a probability to each item in a set

Model: A tool used to discover topics of documents within a corpus

Bag-of-Words Model: In this model, a document is represented as a set of words ignoring all grammar and order

Key Terms (Cont.)

Solution: Discover how to categorize documents by their topics

Topic discovery: The process of identifying a set of topics describing the documents in a corpus

Labeling a document: Assigning one or more topics from the discovered set to a document

Key Terms (Cont.)

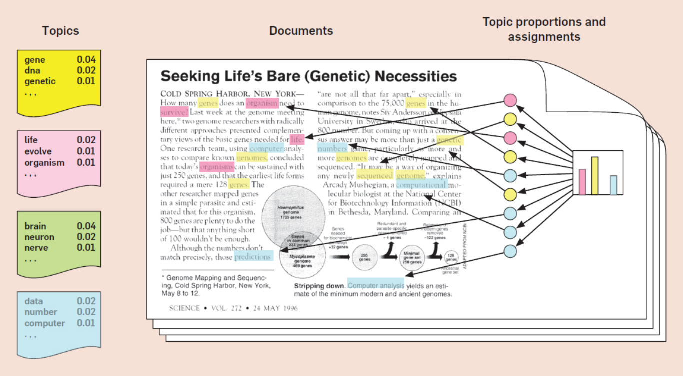
Topic evolution: The description of changes within a set of features showing how those features describe topics differently or similarly over a *period of time*

Iterative Approaches

1. Gather a corpus
2. Set number of topics
3. Generate model
4. Check for accuracy
5. Repeat until satisfactory results

Outline

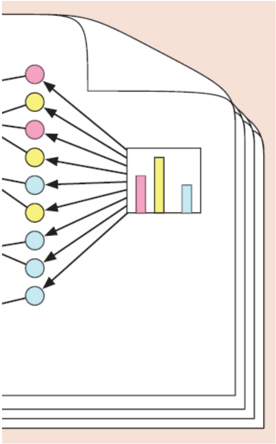
Latent Dirichlet Allocation (LDA)



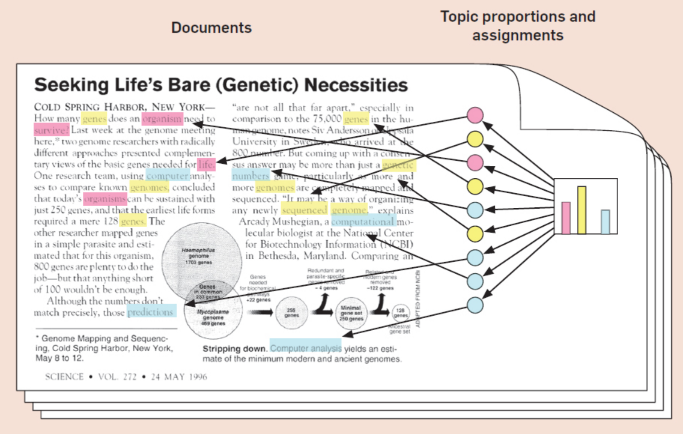
- LDA assumes documents are created at random

LDA Assumptions

- Each document is created with a set distribution of topics
- We assume each document is created one word at a time
- Each word is based off the topic chosen
- The topic is chosen with a probability based on the documents topic distribution



Topic Assumption Description



- Words are chosen based on the topic chosen

LDA Assumptions

- Each word in the *vocabulary* has a chance of coming from any topic
- Topics are named based on the words with the highest probability of being chosen

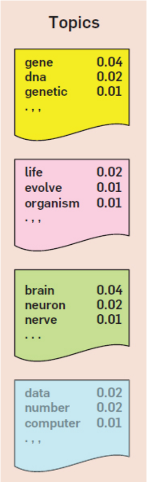
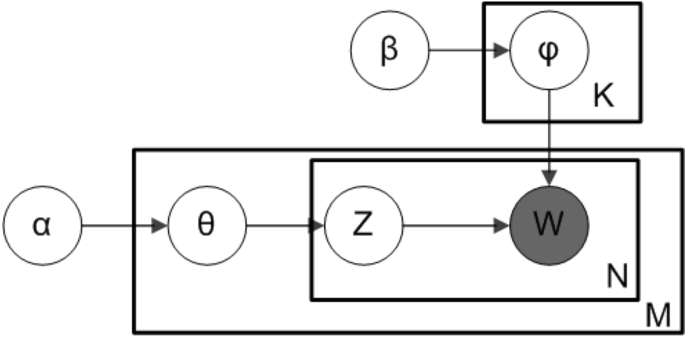
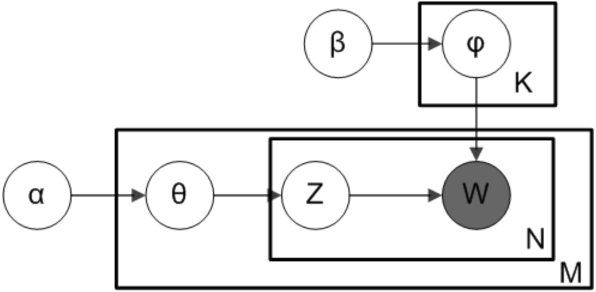


Plate Description



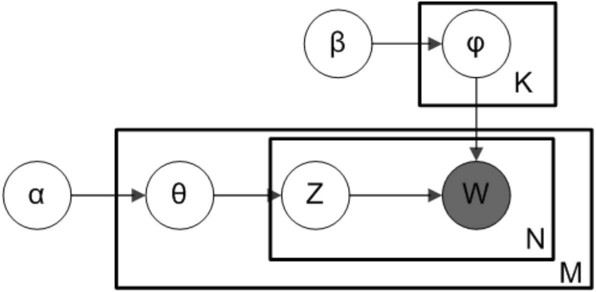
- Blank plates are the variables we are trying to create
- The solid plate is the observable data

Variables in LDA



- α and β starting parameters
- Topic Word Distributions: $\phi_{1:K}$
- W is observed word
- Words in Document: **N**
- Topic Distributions: θ_d
- Chosen topic: z
- Documents in Corpus: **M**
- Total Topics: **K**

Discovering the Topics



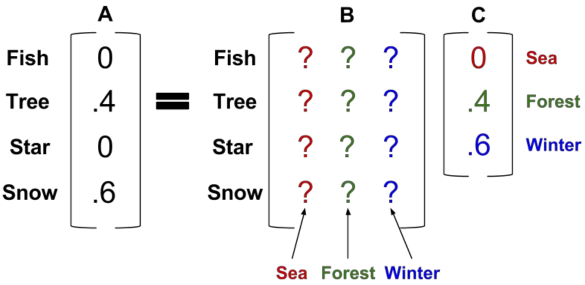
- Check the probability of obtaining the observed corpus
- Modify parameters to increase probability
- Repeat until the probability of producing the observed corpus is maximized

Outline

Dictionaries

- A dictionary is a matrix that encodes associations between two features.
- Topic Discover and Evolution use them to encode textual features and their relationship with topics.

Training Dictionaries



- **A** represents a document by word proportions
- **B** represents a dictionary
- **C** represents a document by topic proportions

Dictionary Complexity

$$\begin{array}{r}\text{Fish}\\ \text{Tree}\\ \text{Star}\\ \text{Snow}\end{array}\overset{\textbf{A}}{\begin{bmatrix}0\\ .4\\ 0\\ .6\end{bmatrix}} = \begin{array}{r}\text{Fish}\\ \text{Tree}\\ \text{Star}\\ \text{Snow}\end{array}\underset{\begin{array}{ccc}\text{Sea} & \text{Forest} & \text{Winter}\end{array}}{\overset{\textbf{B}}{\begin{bmatrix}0 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 0\\ 0 & 0 & 1\end{bmatrix}}}\overset{\textbf{C}}{\begin{bmatrix}0\\ .4\\ .6\end{bmatrix}}\begin{array}{l}\text{Sea}\\ \text{Forest}\\ \text{Winter}\end{array}$$

- **Fish** should be related in some way to *Sea*
- **Tree** should not *only* by related to *Forest*
- The vector **C** needs to be assigned topics by hand

Training Dictionaries

$$
\begin{array}{cl}
 & \begin{array}{c} \text{Fish} \\ \text{Tree} \\ \text{Star} \\ \text{Snow} \end{array} \\
\textbf{X} & \begin{bmatrix} .2 \\ .2 \\ .3 \\ .3 \end{bmatrix} \\
\textbf{Y} & \begin{bmatrix} .1 \\ .1 \\ .8 \\ 0 \end{bmatrix} \\
\textbf{Z} & \begin{bmatrix} .2 \\ .3 \\ .1 \\ .4 \end{bmatrix} \\
\textbf{W} & \begin{bmatrix} .6 \\ .3 \\ 0 \\ .1 \end{bmatrix}
\end{array}
$$

$$
\textbf{B} = \begin{array}{c} \text{Fish} \\ \text{Tree} \\ \text{Star} \\ \text{Snow} \end{array}
\begin{bmatrix} ? & ? & ? \\ ? & ? & ? \\ ? & ? & ? \\ ? & ? & ? \end{bmatrix}
$$

Sea Forest Winter

$$
\begin{array}{lcl}
\begin{array}{c} \text{Sea} \\ \text{Forest} \\ \text{Winter} \end{array} & & \\
\begin{bmatrix} 0 \\ .4 \\ .6 \end{bmatrix} & & \textbf{X} \\
\begin{bmatrix} .2 \\ .2 \\ .6 \end{bmatrix} & & \textbf{Y} \\
\begin{bmatrix} .8 \\ .1 \\ .1 \end{bmatrix} & & \textbf{Z} \\
\begin{bmatrix} .4 \\ .3 \\ .3 \end{bmatrix} & & \textbf{W}
\end{array}
$$

Non-Negative Matrix Factorization (NMF)

- A loss function returns the total error
- NMF is a *process* that generates the dictionary with the least amount of error

The l_2-norm

- The l_2-norm measures the distance of a vector
- We create an error vector for each row
- By taking the l_2-norm of the error vector for each row in a matrix, we derive the amount of error for the matrix

The l_2-norm

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \sqrt{\sum_{i=1}^{n} |x_i|^2}$$

The formula on the right shows how to obtain the l_2-norm for the vector on the left.

Non-Negative Matrix Factorization

$$\min_{B \geq 0} \sum_{i=1}^{N} l(\mathbf{y}_i, \mathbf{B}\theta_i)$$

- Set N equal to the number of rows in the dictionary $\mathbf{B}$
- Set $\mathbf{y}_i$ to the document i by word proportions
- Set θ_i to the document i by topic proportions
- The function $l(\mathbf{y}_i, \mathbf{B}\theta_i)$ returns the l_2-norm of row i in $\mathbf{B}$

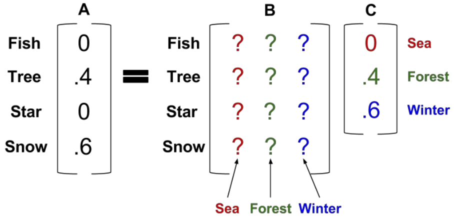
Outline

The Badge Model

- Utilizes user descriptions to assign topics to documents
- Automates the topic assignment of documents before training a dictionary

The Badge Model Foundation Equation

$$\mathbf{y}_i \approx \mathbf{B}\theta_i$$



- The formula can be described by our earlier example
- *Badges* are words users describe themselves with
- The topics are created from user *badges*

Setting Topics

- User 1: Liberal, Minnesotan
- User 2: Liberal, Athlete
- User 3: Conservative, Athlete

$$\begin{bmatrix} 1/3 & Liberal \\ 1/3 & Athlete \\ 1/6 & Conservative \\ 1/6 & Minnesotan \end{bmatrix}$$

The Loss Objective

$$\min_{B \geq 0} \sum_{i=1}^{N} l(\mathbf{y}_i, \mathbf{B}\theta_i) + \lambda_B \sum_{j=1}^{V} \sum_{k=1}^{K} |\mathbf{B}_{jk}|$$

- The gray part of the loss objective is the l_2-norm as discussed earlier
- The new addition is a penalty for non-sparse matrices
- To increase the sparsity of the resulting dictionary, increase λ

Discovering Topics

$$\min_{\theta \geq 0} \sum_{i=1}^{N} l(\mathbf{y}_i, \mathbf{B}\theta_i) + \lambda_\theta \sum_{j=1}^{V} \sum_{k=1}^{K} |\theta_{jk}|$$

- The formula above discovers the minimum θ
- $\mathbf{B}$ is known from the previous step
- Document i does not have assigned topics

Outline

LTECS

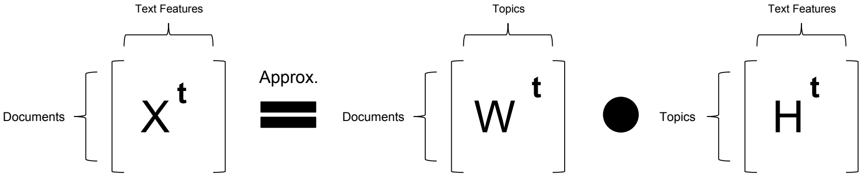
- Topic Evolution Model
- Uses topic discovery methods (such as the *badge model*) to form a foundation
- Defines topics based on users and content
- Introduces an expansion on NMF called collective factorization

LTECS

- How do dictionaries *evolve* over time?
- Do they change based on content or users?
- How do we map the changes?

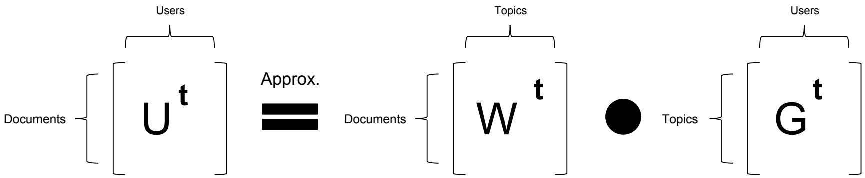
The Content Formula

Symbol	Description
t	an arbitrary time
d	a document in the corpus for training
f	a textual feature in a document, typically a non-stop word
k	the number of topics describing all the documents in the training corpus
N_d^t	the number of documents in the corpus associated with time t
N_f	the number of textual features in the corpus associated with time t
$\mathbf{W}^{\mathrm{t}}$	An N_d^t x k matrix
$\mathbf{H}^{\mathrm{t}}$	An k x N_f matrix
$\mathbf{X}^{\mathrm{t}}$	An N_d^t x N_f matrix



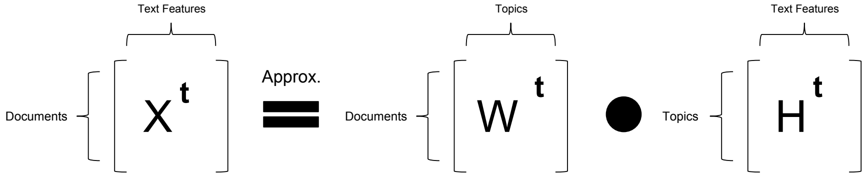
The Content Formula

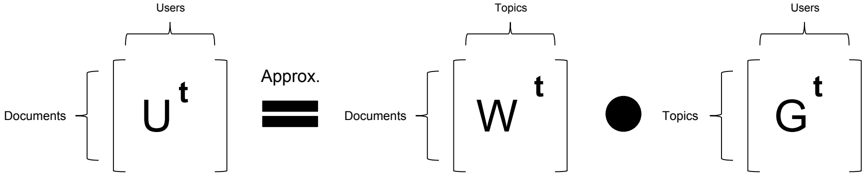
Symbol	Description
t	an arbitrary time
d	a document in the corpus for training
k	the number of topics describing all the documents in the training corpus
N_d^t	the number of documents in the corpus associated with time t
N_u	the number of users who shared a document at time t
$\mathbf{W}^{\mathrm{t}}$	An N_d^t x k matrix
$\mathbf{G}^{\mathrm{t}}$	An k x N_u matrix
$\mathbf{U}^{\mathrm{t}}$	An N_d^t x N_u matrix



Discovering the Dictionary

Solve for $\mathbf{W}^t$





Topic Evolution Matrix

$$\mathbf{X}^t \approx \mathbf{W}^t \mathbf{M}_T^t \mathbf{H}^{t-1}$$
$$\mathbf{U}^t \approx \mathbf{W}^t \mathbf{M}_C^t \mathbf{G}^{t-1}$$

The matrices $\mathbf{M}_T^t$ and $\mathbf{M}_C^t$ are *topic evolution matrices*. Multiplying them by $\mathbf{H}^{t-1}$ or $\mathbf{G}^{t-1}$ is approximately equal to $\mathbf{H}^t$ or $\mathbf{G}^t$ respectively

The Loss Objective

$$L = \mu L_T + (1 - \mu) L_C + R$$

$$\begin{aligned}
L_T &= ||\mathbf{X}^t - \mathbf{W}^t\mathbf{H}^t||_F^2 + ||\mathbf{X}^t - \mathbf{W}^t\mathbf{M}_T^t\mathbf{H}^{t-1}||_F^2, \\
L_C &= ||\mathbf{U}^t - \mathbf{W}^t\mathbf{G}^t||_F^2 + ||\mathbf{U}^t - \mathbf{W}^t\mathbf{M}_C^t\mathbf{G}^{t-1}||_F^2, \\
\mathrm{R} &= \alpha(||\mathbf{W}^{\mathbf{t}}||_1 + ||\mathbf{H}^t||_1 + ||\mathbf{G}^t||_1 + ||\mathbf{M}_T^t||_1 \\
&\qquad + ||\mathbf{M}_C^t||_1) + \lambda(||\mathbf{M}_T^t - I||_F^2 + ||\mathbf{M}_C^t - I||_F^2).
\end{aligned}$$

The Loss Objective

$$L = \mu L_T + (1 - \mu) L_C + R$$

$$L_T = ||\mathbf{X}^t - \mathbf{W}^t \mathbf{H}^t||_F^2 + ||\mathbf{X}^t - \mathbf{W}^t \mathbf{M}_T^t \mathbf{H}^{t-1}||_F^2$$

The Loss Objective

$$L = \mu L_T + (1 - \mu) L_C + R$$

$$L_C = ||\mathbf{U}^t - \mathbf{W}^t \mathbf{G}^t||_F^2 + ||\mathbf{U}^t - \mathbf{W}^t \mathbf{M}_C^t \mathbf{G}^{t-1}||_F^2$$

The Loss Objective

$$\boldsymbol{L} = \mu L_T + (1 - \mu) L_C + \boldsymbol{R}$$

$$\begin{aligned}\mathrm{R} = \alpha(||\mathbf{W^t}||_1 + ||\mathbf{H}^t||_1 + ||\mathbf{G}^t||_1 + ||\mathbf{M}_T^t||_1 \\ + ||\mathbf{M}_C^t||_1) + \lambda(||\mathbf{M}_T^t - I||_F^2 + ||\mathbf{M}_C^t - I||_F^2).\end{aligned}$$

Outline

Summary

- *LDA*
  - useful with little document processing beforehand
  - can fail to capture intent in an article
- The *badge model*
  - captures the intent of articles
  - can return opposite of expected results
- *LTECS*
  - returned data must be closely analyzed
  - requires a large amount of processing power
  - less error in results

Questions and Thanks

Questions?

References

See the UMM Vink '15 paper for all references on figures and data