Practical Text Analytics: Latent Semantic Analysis

Fan Dai

Iowa State University

October 15, 2019

Recall: Text analytics process

Text analytics process

- Planning the text analytics projects
- Preparing and preprocessing text
- Analyzing data
  1. *Latent Semantic Analysis (LSA)*

     (*Chapter 6, Anandarajan et al.(2019)*)
- Interpreting results

LSA: Motivating Example

Source: **https://kids.nationalgeographic.com/**

- What a computer can learn from text describing "cheetah" by looking at word frequency, proximity. . . ?
  - "cheetah" and "cat" are semantically related.
  - "cheetah" and "fastest"are more closely related than "cat" and "fastest".
- The computer makes the connection that *cheetah is the fastest cat*.

LSA: Definition

- The underlying idea

  Extract and reveal information that conveyed from how words co-occur with each other across documents

  1. Reflect a shared latent concept, e.g., {"fastest", "cat"}$\Rightarrow$"cheetah".
  2. Classify the documents, e.g., {"dog", "cat", "apple", "blueberry", "orange"}$\Rightarrow$ group documents into "animal" and "fruit"

- The analysis object: tokens

  1. Meaning of words
  2. Relationships with other tokens

LSA: singular value decomposition (SVD)

How to do when two words are related through a third word only?

e.g., Two words from Iowa's winery blogs: "Ackerman", "Tassel"

- Rarely together within the same document.
- Related through their frequent shared co-occurrence with other terms like “producer” or "price".

Solution:

Singular value decomposition (SVD) on the weighted term-document matrix (TDM)

TDM

- Rows: words/terms. Columns: documents
- Entries: the number of times that the ith term appears in the jth documents.

LSA: singular value decomposition (SVD)

Local weight for word *i* in document *j*		Global weight for word *i*	
Raw	$tf_{i,j}$	None	1
Binary	$\begin{cases} tf_{i,j} \geq 1: 1 \\ tf_{i,j} = 0: 0 \end{cases}$	IDF	$1 + \log_2\left(\frac{n}{df(i)}\right)$
Log	$\log_2(tf_{i,j} + 1)$	Entropy	$1 + \sum_{j=document}^{corpus} \frac{\frac{tf_{i,j}}{gf_i} \times \log_2\left(\frac{tf_{i,j}}{gf_i}\right)}{\log_2(n)}$
$tf_{i,j}$	Term frequency: number of times word i appears in the document j		
df_i	Document frequency: Number of documents word i appears in at least once in the corpus		
gf_i	Global frequency: Number of times word i appears across the entire corpus		
n	Number of documents in the corpus		
$Final\ weight = local\ weight \times global\ weight$			

Figure: Weighting for TDM

Source: **David Gefen**; **et al.** **(2017)**

$tfidf$-weighted TDM is prefered in LSA

LSA: singular value decomposition (SVD)

How does SVD work?

- identify underlying factors in the weighted TDM by transforming it into three matrices that represent terms, documents, and a matrix multiplier for reconstruction, respectively.

$$M = U \times \Sigma \times V^\top$$

  - M is the weighted TDM
  - U contains the left singular vectors of terms
  - Σ is a matrix with weight values on the diagonal
  - V contains the right singular vectors of documents

LSA: singular value decomposition (SVD)

- in LSA, we apply a truncated SVD

$$M \approx A_k = U_k \times \Sigma_k \times V_k^\top$$

A_k represents the LSA space out of the rank r matrix M

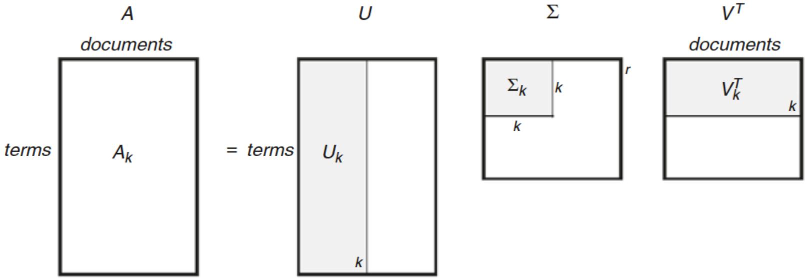


Figure: Truncated SVD process

Source: **Martin and Berry (2007)**

LSA: singular value decomposition (SVD)

- the k remaning singular vectors in U and V correspond to k "hidden concepts" where the terms and documents participate.
- k is too small: vectors that are related conceptually are not combined
- k is too large: redundant information is included (with those singular values that are "too small" and thus "negligible").
- determine k: e.g., scree plot showing variance explained by number of singular vectors.

- SVD v.s. Principal component analysis (PCA): *reference link*

What can we do with the SVD in LSA?

Example: 10 respondents and descriptions of their dog (Anandarajan et al.(2019)):

- *Document 1: My Favorite Dog Is Fluffy and Tan.*
  Document 2: the dog is brown and cat is brown.
  Document 3: My favorite coat is brown and hat is pink
  Document 4: My dog has a hat and leash. ♡

  $\vdots$

  Document 10: MY fluffy dog has a white coat and hat!

LSA: singular value decomposition (SVD)

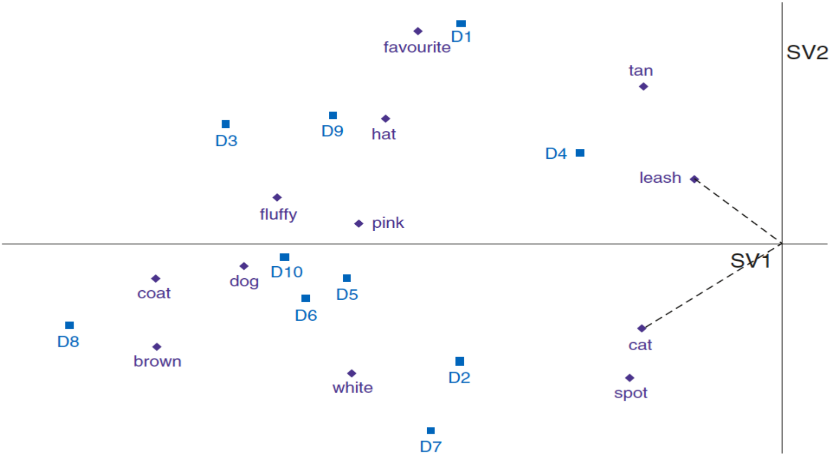
$tfidf$-weighted TDM

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Brown	0.0	3.0	1.5	0.0	1.5	3.0	1.5	1.5	1.5	0.0
Cat	0.0	4.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Coat	0.0	0.0	2.0	0.0	4.0	2.0	0.0	2.0	0.0	2.0
Dog	1.3	1.3	0.0	1.3	0.0	1.3	1.3	2.6	1.3	1.3
Favorite	2.7	0.0	2.7	0.0	0.0	0.0	0.0	0.0	2.7	0.0
Fluffy	1.7	0.0	0.0	0.0	1.7	1.7	0.0	1.7	1.7	1.7
Hat	0.0	0.0	2.3	2.3	0.0	0.0	0.0	0.0	2.3	2.3
Leash	0.0	0.0	0.0	4.3	0.0	0.0	0.0	0.0	0.0	0.0
Pink	0.0	0.0	3.3	0.0	0.0	0.0	0.0	3.3	0.0	0.0
Spot	0.0	0.0	0.0	0.0	0.0	0.0	4.3	0.0	0.0	0.0
Tan	4.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
White	0.0	0.0	0.0	0.0	0.0	0.0	2.7	2.7	0.00	2.7

Source: **Practical text analytics: maximizing the value of text data**

LSA: singular value decomposition (SVD)

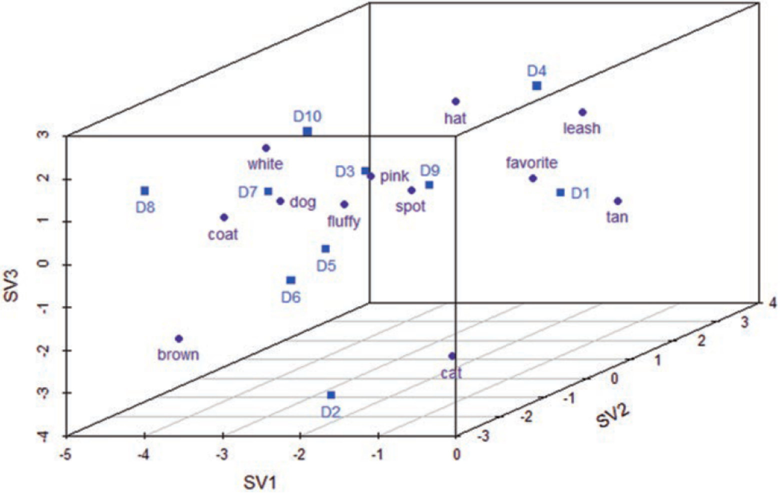
Apply truncated SVD with $k = 2$ and plot each terms and documents according to the row vectors in $U_2\Sigma_2$ and $V_2\Sigma_2$, respectively.



Source: **Practical text analytics: maximizing the value of text data**

LSA: singular value decomposition (SVD)

Apply truncated SVD with $k = 3$.



Source: **Practical text analytics: maximizing the value of text data**

LSA: cosine similarity

Measure the similarity between two vectors in the LSA space.

- Cosine similarity: $\cos(\mathbf{v_1}, \mathbf{v_2}) = \frac{\mathbf{v_1}^\top \mathbf{v_2}}{||\mathbf{v_1}||||\mathbf{v_2}||}$, which can be applied to terms, documents (or both), and queries.
- Query: a scaled, weighted sum of the component term vectors.

$$\text{query} = q^\top U_k \Sigma_k^{-1}$$

- LSA uses the cosine measures to find documents that are similar to words that designated as query terms

Example: Cosine values between the query (brown, pink, tan) and documents.

brown	*cat*	*coat*	*dog*	*favorite*	*fluffy*	*hat*	*leash*	*pink*	*spot*	*tan*	*white*
1	0	0	0	0	0	0	0	1	0	1	0

Document	Cosine
6	0.81
5	0.78
9	0.73
2	0.71
1	0.69
3	0.66
8	0.24
10	−0.08
7	−0.30
4	−0.30

Source: **Practical text analytics: maximizing the value of text data**

LSA: summaries

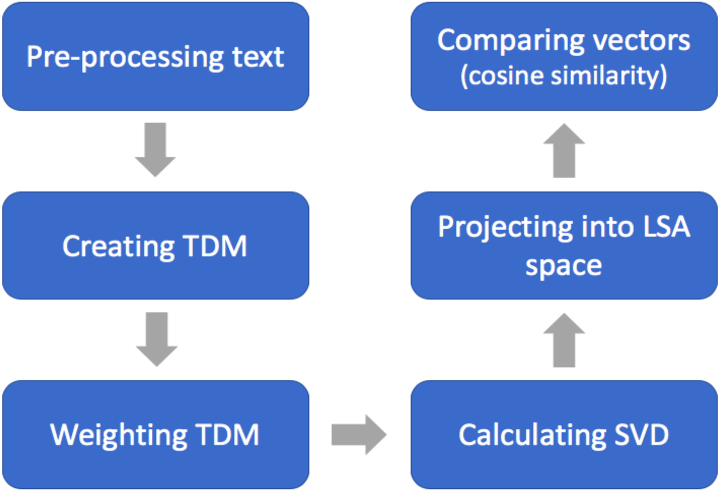


Figure: Main Steps for latent semantic analysis