BC COMS 2710: Computational Text Analysis

Lecture 19
Dimensionality Reduction

Announcements – Assignments

- Readings 05:
  - No more this semester – congrats!
- HW02
  - Mostly done grading
  - You all did very well!!!
- HW04
  - Due Thursday

Announcements

- P/D/F deadline is today
- Course Evaluations
  - Due 06/14

Final Project – Deliverables

- Project ideation – Friday May 28st
  - 5 points
- Project proposal – Sunday June 6th
  - 9 points
- Project presentations – Monday June 14th
  - 6 points
- Project submissions – Friday June 18th
  - 15 points
- http://coms2710.barnard.edu/final_project

Project Presentations – Monday June 14th

5 minute presentations by each group

Format:

- Research Question
- Motivation
  - Why should we care?
- Data Collected
  - Where did the data come from? How did you collect it?
  - What filtering was done?
  - Resulting corpus:
    - How many documents? Average size of documents? Vocabulary size?
- Results (preliminary)
  - figures

Goal:

- Publicizing your research is important
- You all to see what everyone else is working on

Evaluation Metrics – Incorrect Classification

"*Congratulations, offer letter from Google*" Spam

"*Congratulations, you won the lottery*" Not Spam

Which mistake is worse?

Evaluation Metrics

Recall

- When we do not want false negatives

Precision

- When we do not want false positives

Logistic Regression

Summary of Logistic Regression

- Optimizes $\mathrm{P}(Y\,|X)$ directly
- Define the features
- Learn a vector of weights for each label $y \epsilon Y$
  - Gradient descent, update weights based on error
- Multiple feature vector by weight vector
- Output is $\mathrm{P}(Y = y\,|X)$ after normalizing
- Choose the most probable Y

Slide from Nate Chambers

Summary of Logistic Regression

- Optimizes $\mathrm{P}(Y\ |X)$ directly
- Define the features
- Learn a vector of weights for each label $y \epsilon Y$
  - Gradient descent, update weights based on error
- **Multiple feature vector by weight vector**
- Output is $\mathrm{P}(Y = y\ |X)$ after normalizing
- Choose the most probable Y

Slide from Nate Chambers

Scoring one document - dot product

$$[f_1, f_2, f_3] \cdot [w_1, w_2, w_3] = (f_1 \times w_1) + (f_2 \times w_2) + (f_3 \times w_3)$$

$$= \sum (f_i \times w_i)$$

Score two documents

$$[[f_{1,1}, f_{1,2}, f_{1,3}] \\ [f_{2,1}, f_{2,2}, f_{2,3}]] \cdot [w_1, \\ w_2, \\ w_3]$$

$$= [(f_{1,1} \times w_1) + (f_{1,2} \times w_2) + (f_{1,3} \times w_3), \\ (f_{2,1} \times w_1) + (f_{2,2} \times w_2) + (f_{2,3} \times w_3)]$$

Matrix Multiplication

We can multiply two matrices A and B if ….

number of columns in A = number of rows in B

The size of the resulting matrix is ….

number of rows in A & the number of columns in B

[Khan Academy]

Matrix Multiplication

$$\begin{bmatrix} 1 & 7 \\ 2 & 4 \end{bmatrix} \cdot \begin{bmatrix} 3 & 3 \\ 5 & 2 \end{bmatrix}$$

A B

Khan Academy

Matrix Multiplication

$$\vec{b}_1 \quad \vec{b}_2$$

$$\begin{matrix} \vec{a}_1 \rightarrow \\ \vec{a}_2 \rightarrow \end{matrix} \begin{bmatrix} 1 & 7 \\ 2 & 4 \end{bmatrix} \cdot \begin{bmatrix} 3 & 3 \\ 5 & 2 \end{bmatrix}$$

$$A \qquad\qquad B$$

[Khan Academy]

Matrix Multiplication

$$\begin{matrix} & \vec{b_1} & \vec{b_2} \\ & \downarrow & \downarrow \end{matrix}$$

$$\begin{matrix}\vec{a_1}\rightarrow \\ \vec{a_2}\rightarrow\end{matrix} \underbrace{\begin{bmatrix} 1 & 7 \\ 2 & 4 \end{bmatrix}}_{A} \cdot \underbrace{\begin{bmatrix} 3 & 3 \\ 5 & 2 \end{bmatrix}}_{B} = \underbrace{\begin{bmatrix} \vec{a_1} \cdot \vec{b_1} & \vec{a_1} \cdot \vec{b_2} \\ \vec{a_2} \cdot \vec{b_1} & \vec{a_2} \cdot \vec{b_2} \end{bmatrix}}_{C}$$

Khan Academy

Matrix Multiplication Example

$$A \qquad\qquad B \qquad\qquad C$$

$$[[1,2,3] \quad 4,5,6]] \cdot [[7,8, \quad 9,10, \quad 11,12]] =$$

Question: What are the dimension of C?

2 rows x 2 columns

Document–Term Matrix

Document-Term Matrix

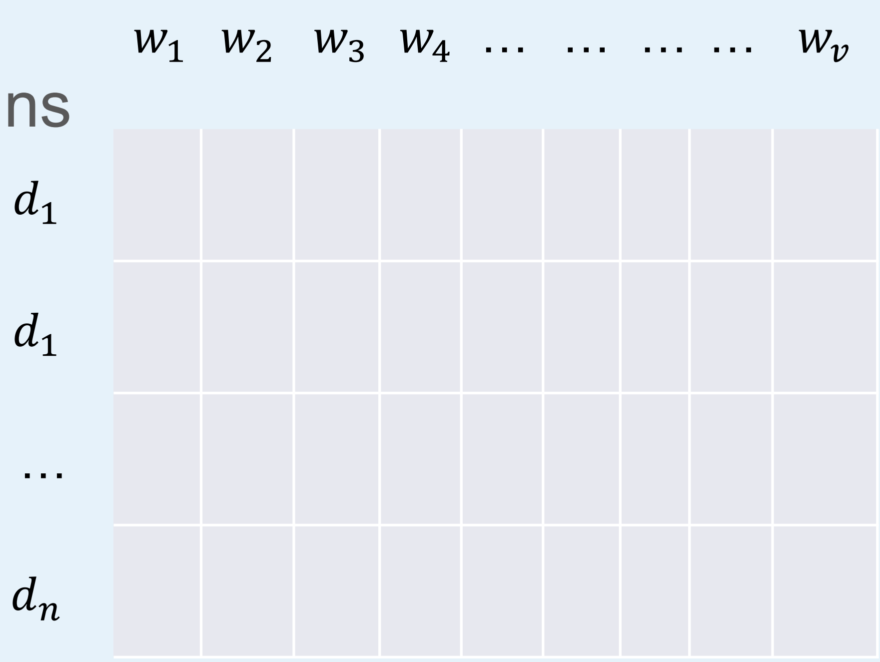
DMT:

- Rows represent a document
- Columns represent a word
- Values represent some feature of word w_i in document d_j

	w_1	w_2	w_3	w_4	$\ldots$	$\ldots$	$\ldots$	$\ldots$	w_v
d_1									
d_1									
$\ldots$									
d_n									10

Properties of Document-Term Matrix

- Sparse matrix
  - Most values are 0
- Very large
  - Many, many, many columns
- Noisy



Goal

- Make values in each cell more meaningful
- Reduce the size of the matrix
  - Dimensionality reduction
- Remove noise

Matrix Factorization/ Dimensionality Reduction

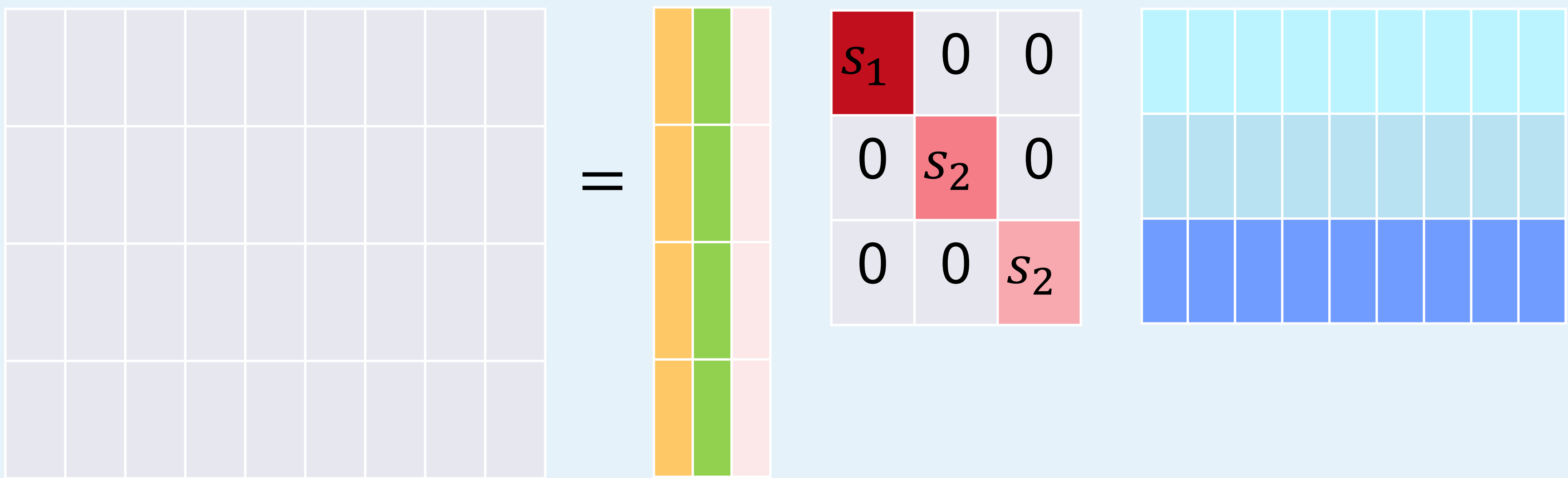
Abstract thought

https://www.youtube.com/watch?v=dROx9Djr7mk

Singular Value Decomposition

$$\underset{n \times v}{M} = \underset{n \times k}{U} \; \underset{k \times k}{S} \; \underset{k \times v}{V^T}$$

Singular Value Decomposition



M = U S V

n x v = n x k, k x k, k x v

Latent Semantic Analysis

- Applying SVD to the DTM is called Latent Semantic Analysis

- The name of LDA is based on this

Week 6

- Monday 06/07 - Matrix Factorization
- Tuesday 06/08 – Word Embeddings
- Wednesday 06/09 – Guest Lecture
  - Attendance required
- Thursday 06/10 – ngrams & phrases