Topic Modeling
With Non-Negative Matrix Factorization

Dan Becker
galvanize

What Is Topic Modeling

- *Unsupervised learning*
  - *Identify underlying (latent) topics in corpus*
  - *Characterize topics in document*

Sample Topics

Sports	Business	Politics
Cute Human Interest	Local	International

What Topic Modeling is Not

Document Classification

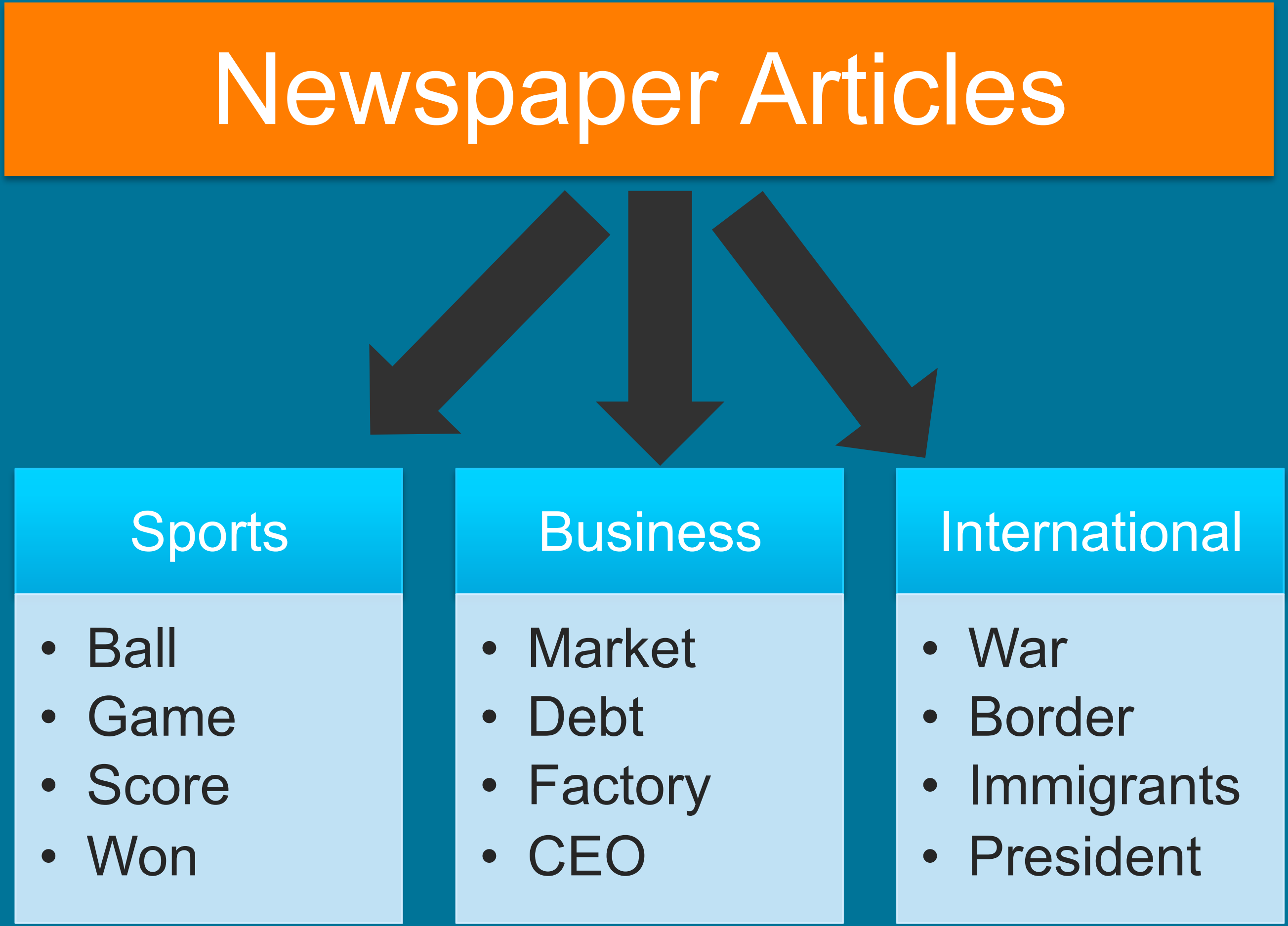
- Supervised
- Predefined categories
- Each document in single category

Why Topic Modeling

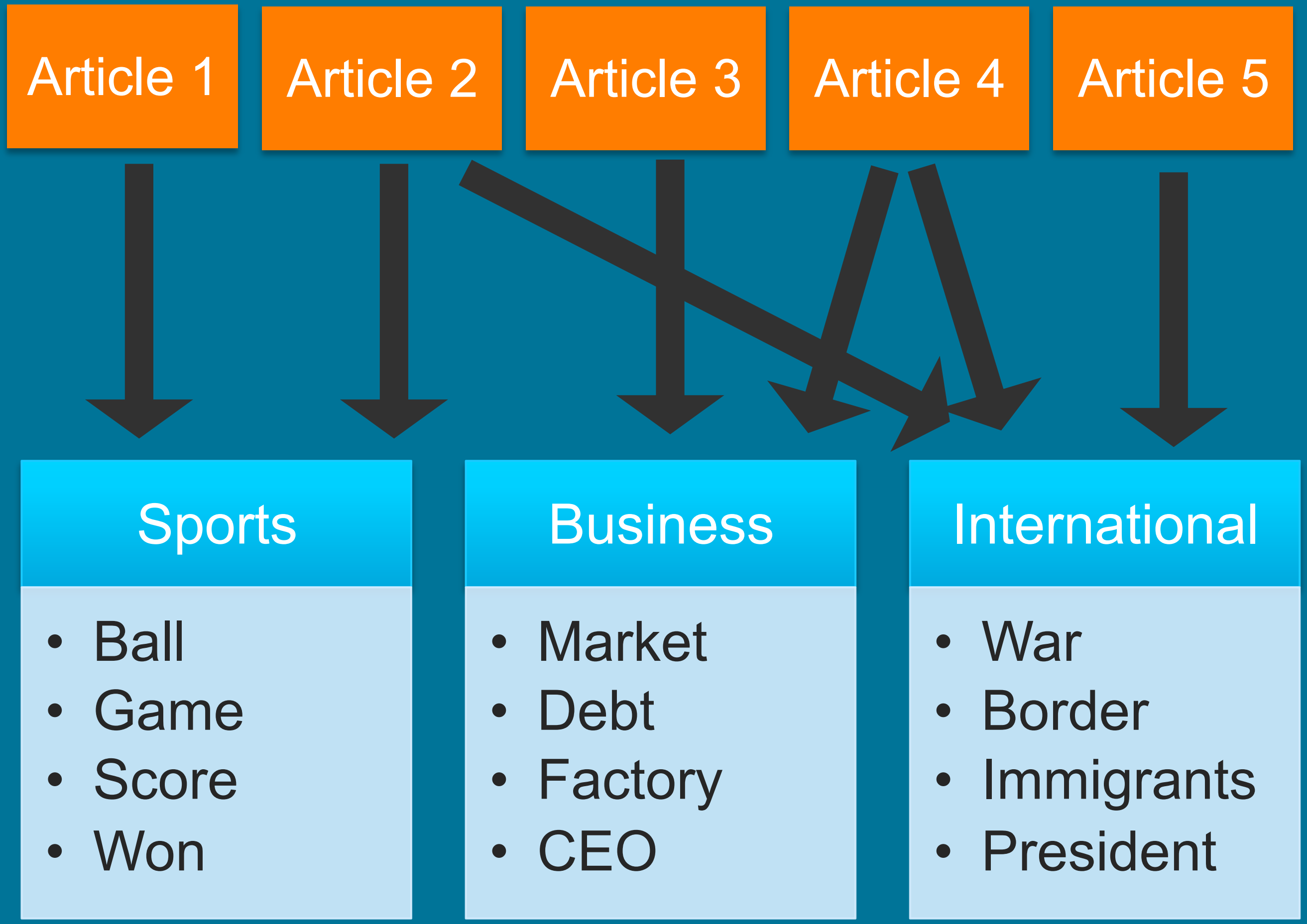
Applications

- Group similar documents
  - Document recommendations
  - Send to correct editor
- Search
- Learn about corpus (find themes)
- Practical alternative for classification

Example Topics



Example Topics and Articles



Review: Document-Term Matrix

Word

Document

	W1	W2	W3	W4	W5	W6
D1	0	0	1	0	1	0
D2	1	1	0	2	0	1
D3	0	3	1	0	2	1
D4	0	0	0	1	0	1
D5	1	0	0	0	1	0

The Ideal Output

Topic

Document	T1	T2	T3	T4
D1	0.2	0	0.8	0
D2	0	1	0	0
D3	0	0.5	0.5	0
D4	0	0	0	1
D5	0.8	0	0	.2

H

Word

Topic	W1	W2	W3	W4
T1	0.5	0.1	0.8	0
T2	0	0.3	0	0
T3	0	0.5	0.5	0
T4	0	0	0	1.5

W

Quiz:

Predict The Number of Times W1 shows up in D5

Topic

Document	T1	T2	T3	T4
D1	0.2	0	0.8	0
D2	0	1	0	0
D3	0	0.5	0.5	0
D4	0	0	0	1
D5	0.8	0	0	.2

Word

Topic	W1	W2	W3	W4
T1	0.5	0.1	0.8	0
T2	0	0.3	0	0
T3	0	0.5	0.5	0
T4	0	0	0	1

Question:

How Would You Predict The Number of Times W1 shows up in D5

	T1	T2	T3	T4
D1	0.2	0	0.8	0
D2	0	1	0	0
D3	0	0.5	0.5	0
D4	0	0	0	1
D5	0.8	0	0	.2

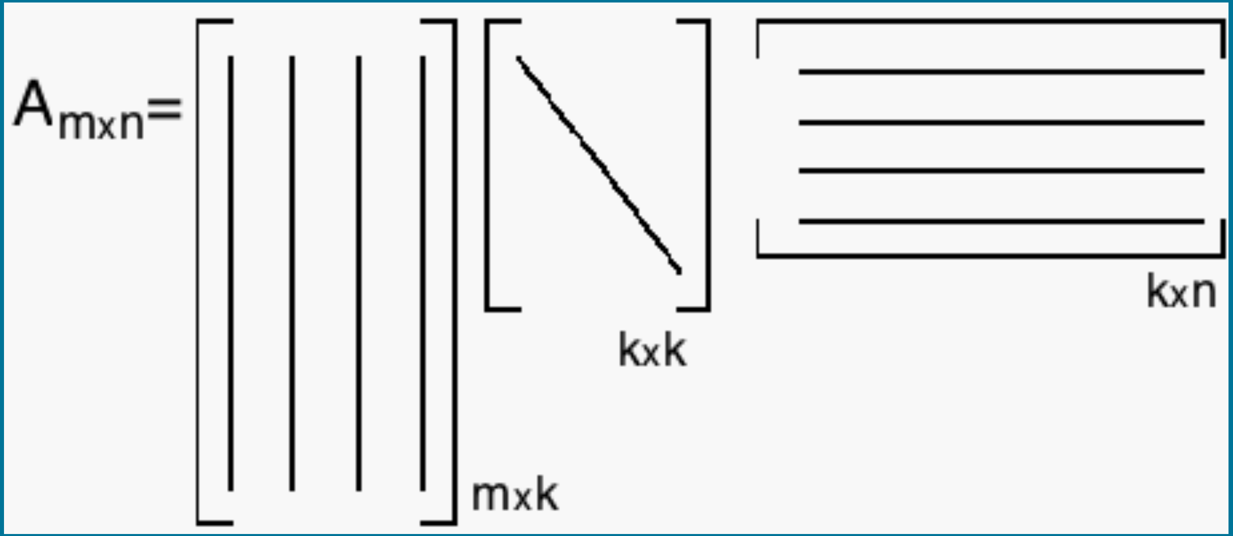
	W1	W2	W3	W4
T1	1	0.1	0.8	0
T2	0	0.3	0	0
T3	0	0.5	0.5	0
T4	0	0	0	1

	W1	W2	W3	W4	W5	W6
D1	0	0	1	0	1	0
D2	1	1	0	2	0	1
D3	0	3	1	0	2	1
D4	0	0	0	1	0	1
D5	1	0	0	0	1	0

The Path to Better Matrices

- Changing H or W will change
  - Predictions
  - Errors
- Some H and W combinations are more consistent with data than other H and W combinations
- We can optimize

Matrix Factorization – You've Seen It Before



- Trivial to convert SVD to matrix shapes we want
- Result would fail a sanity check

Ideal Matrix Characteristics

What Restrictions Might We Want For H (Doc) Matrix

- ❑ Integers
- ❑ Symmetric
- ❑ No negative Values
- ❑ Topic values for each document sum to 1
- ❑ Document values for each topic sum to 1

Alternating Least Squares Algorithm

Initialization

1. Create document term matrix
   - Optionally apply TF-IDF
2. Initialize H and W matrices with random values

Optimization

3. Optimize H (holding W fixed)
   - Set negative values in H to 0
4. Optimize W (holding H fixed)
   - Set negative values in W to 0
5. Repeat steps 3 and 4 until convergance

Details on H (and W) Optimization

$$\min_{w}(Y - HW)^2$$

Details on H (and W) Optimization

$$\min_{w}(Y - HW)^2$$

$$2H'(Y - HW) = 0$$

Details on H (and W) Optimization

$$\min_w (Y - HW)^2$$

$$2H'(Y - HW) = 0$$

$$H'HW = H'Y$$

Details on H (and W) Optimization

$$\min_{w}(Y - HW)^2$$

$$2H'(Y - HW) = 0$$

$$H'HW = H'Y$$

$$W = (H'H)^{-1}H'Y$$

Details on H (and W) Optimization

$$W = \underbrace{\boxed{(H'H)^{-1}}}_{KxK}\ \underbrace{\boxed{H'}}_{KxD}\ \underbrace{\boxed{Y}}_{DxW}$$

Closing TIps

- Remove least and most common words
- TF-IDF (usually) worthwhile
- Topic count can depend on application