Word embedding models

Daniel Hsu

COMS 6998-7 Spring 2025

Latent semantic indexing

Latent semantic indexing

(Deerwester, Dumais, Furnas, Landauer, Harshman, 1990)

- <u>Term-document matrix</u> $A \in \{0,1\}^{V \times N}$ (simplified)

$$A_{i,j} = 1\{\text{term } i \text{ appears in document } j\}$$

documents

terms

	document 1	document 2	document 3	…
aardvark	1	1	1	
abacus	0	0	1	
abalone	0	1	0	
…				

- <u>Latent Semantic Indexing (LSI)</u>: low-rank SVD factorization $A \approx BC$
  - Terms represented by rows of B, documents represented by columns of C

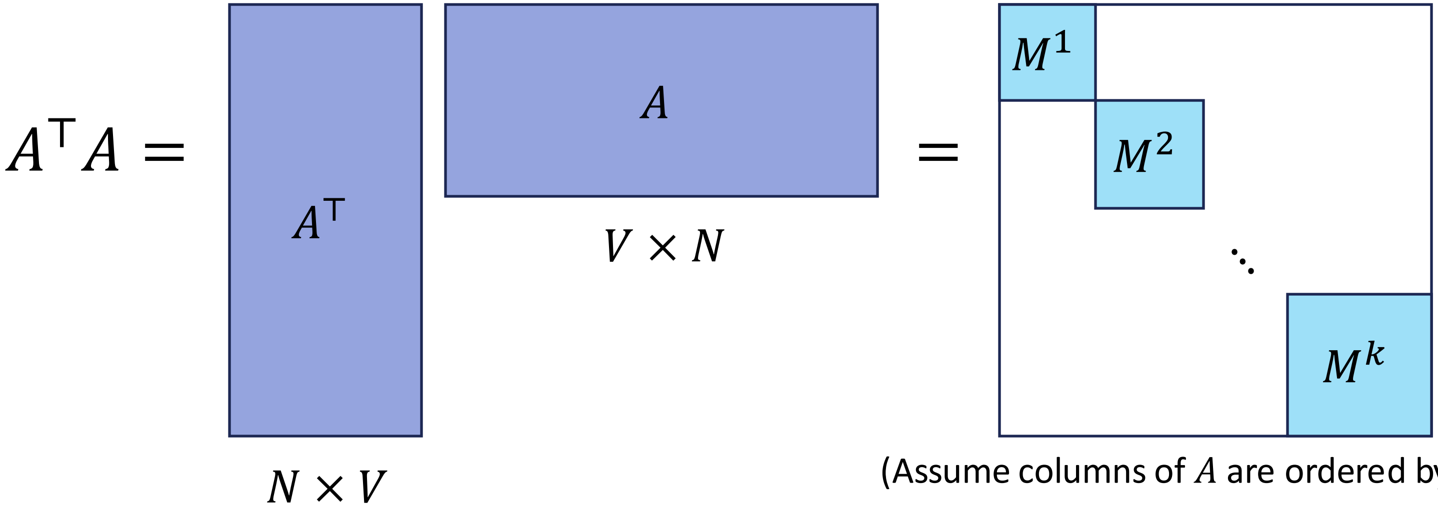
Problems LSI is intended to address

- Old ideas:
  - document = bag-of-words
  - word = meaning is derived from what documents use it
- Problems:
  - Synonymy: some documents about "cars" only use "automobile"
  - Polysemy: documents about both water sports and internet use "surfing"

Probabilistic analysis of LSI

(Papadimitriou, Raghavan, Tamaki, Vempala, 2000)

- Modeling assumptions for analysis:
  - documents are **clustered** by topic   (k = number of topics)
  - terms used by topic t documents are **disjoint** from terms used by topic t' documents



(Assume columns of A are ordered by topic)

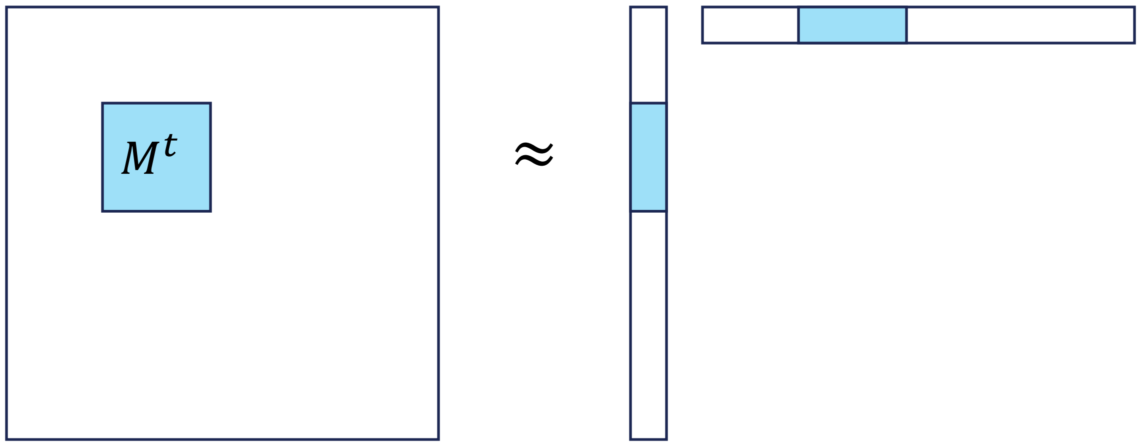
- Eigendecomposition of $A^\top A$ is determined by eigendecompositions of the blocks $M^1, M^2, \ldots, M^k$

Using spectral graph theory

- Within cluster t: if documents are sufficiently "well-connected", then

$$\lambda_1(M^t) \gg \lambda_2(M^t)$$

  - Top eigenvector: positive constant on documents in cluster t
  - Additional assumptions ensure well-connectedness is likely!



- Conclusion: top k eigenvectors of $A^\top A$ are ≈ characteristic vectors for the k clusters

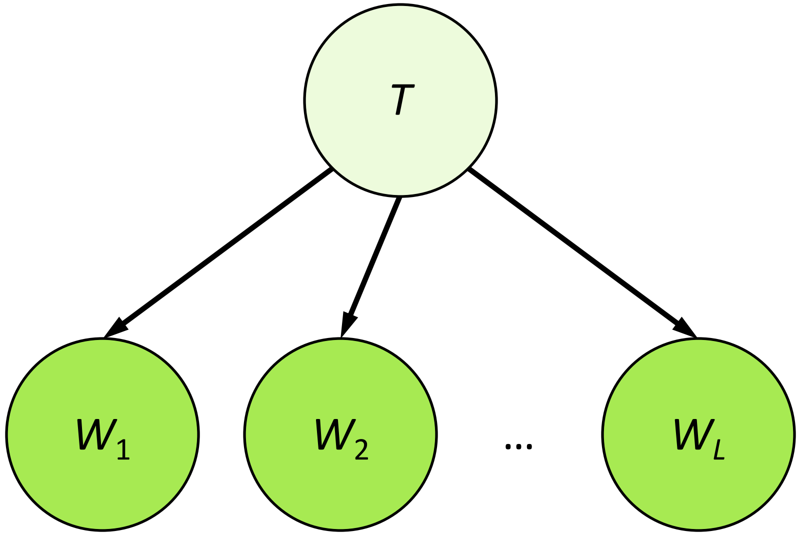
Speed-up using random projection

- Document representation obtained by projecting to k-dimensional subspace of $\mathbb{R}^V$ (so document is represented by k coefficients)
- Suggested speed-up:
  - Step 1: project documents to *random* ℓ-dimensional subspace of $\mathbb{R}^V$:
$$R^\top A \in \mathbb{R}^{\ell \times N}$$
  - Step 2: apply LSI to $R^\top A$
- Main result: If $\ell \gg k + \log N$, then best rank-$2k$ approximation to $RR^\top A$ is about as good as best rank-k approximation to A

- Modern version of this: "sketch-and-solve methods" from randomized numerical linear algebra

Latent "bag-of-words" model

John Rupert Firth: "You shall know a word by the company it keeps!"



T = topic of a length-L document (hidden variable)
$W_1, W_2, \ldots, W_L$ = the L words of the document

Assume: $W_1, W_2, \ldots, W_L$ conditionally independent given T

$$P(w_1, w_2) = \sum_{t=1}^{k} P(w_1|t)P(t)P(w_2|t)$$
$$= (UDU^\top)_{w_1, w_2}$$

where

$$U_{w,t} = P(w|t), \qquad D_{t,t} = P(t)$$

Super-simplified analysis

- Each document has $L = 2$ words
  - A is term-document matrix based on first word
  - $\tilde{A}$ is term-document matrix based on second word
- Look at $V \times V$ matrix $A\tilde{A}^\top$

$$\mathbb{E}\left[A\tilde{A}^\top\right] \propto RDR^\top$$