Data Mining and Machine Learning

Latent Se (LSA)

Peter Jančovič

Objectives

- To understand, intuitively, how Latent Semantic Analysis (LSA) can discover latent topics in a corpus

Vector Notation

- The vector representation ***vec(d)*** of ***d*** is the ***V*** dimensional vector:

$$(0,\ldots,0, w_{i(1),d}, 0,\ldots,0, w_{i(2),d}, 0,\ldots,0, w_{i(M),d}, 0,\ldots,0)$$

$i(1)^{th}$ place

$i(2)^{th}$ place

$i(M)^{th}$ place

Notice that this is the weighting – i.e. the term frequency times the inverse document frequency

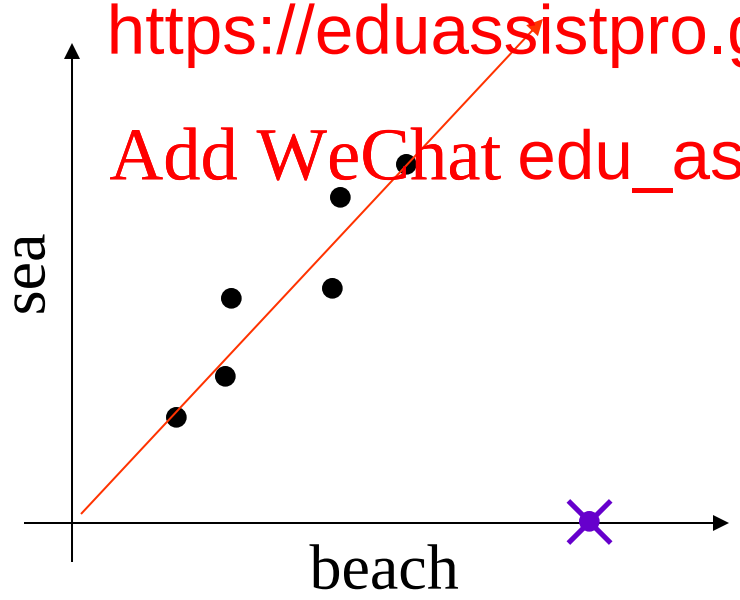
$w_{i(1),d} = f_{i(1),d} \times IDF(i(1))$ from text IR

Latent Semantic Analysis (LSA)

- Suppose we have a real corpus with a large number of documents
- For each document ***d*** the dimension of the vector ***vec***(***d***) will t s of) thousands
- Let's focus on just 2 of the ions, corresponding, say, to the ' and 'beach'
- Intuitively, often, when a document ***d*** includes 'sea' it will also include 'beach'

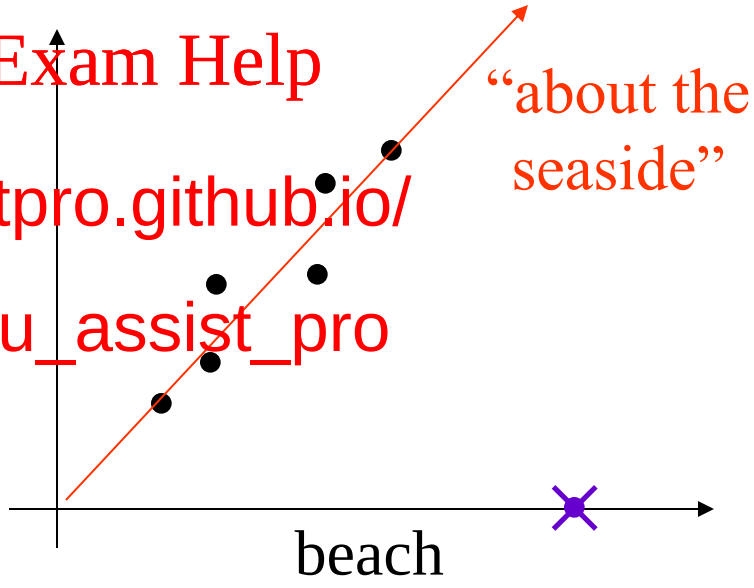
LSA continued

- Equivalently, if ***vec*(*d*)** has a non-zero entry in the 'sea' component, it will often have a non-zero entry in the 'beach' component



Latent Semantic Classes

- If we can detect this type of structure, then we can discover relationships between wor **automatical**
- In the example we have found an equivalence set o terms, including ‘beach’ and ‘sea’, which is ‘about the seaside’



Finding Latent Semantic Classes

- LSA involves some advanced linear algebra – the description here is just an outline
- First construct the 'word-document' matrix *A*
- Then decom Value Decomposition (SVD)
  - SVD is a standard techniqu x algebra
  - Packages such as MATLAB have SVD functions:

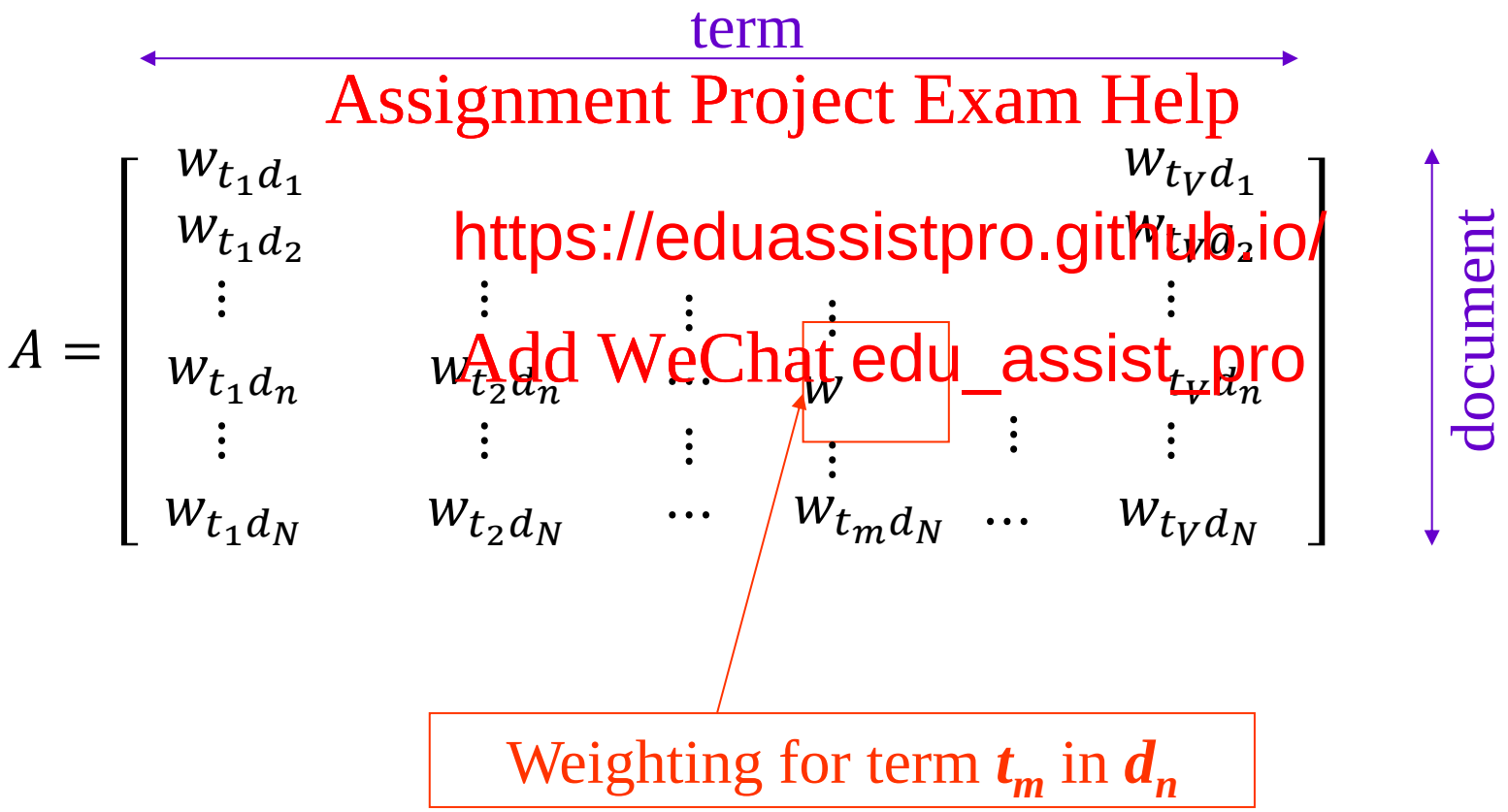
    >>[U,S,V]=svd(A)

Singular Value Decomposition

- Remember eigenvector decomposition?
- An eigenvector of a square matrix A is a vector e such that $Ae = \lambda e$, where λ is a scalar
- For certain $A = UDU^T$, where U is an **orthogonal** tation) and D is **diagonal**
  - The elements of D are the eigenvalues
  - The columns of U are the eigenvectors
- You can think of SVD as a more general version of eigenvector decomposition, which works for general matrices

Word-Document Matrix

- The Word-Document matrix is a N x V matrix whose n^{th} row is $vec(d_n)$

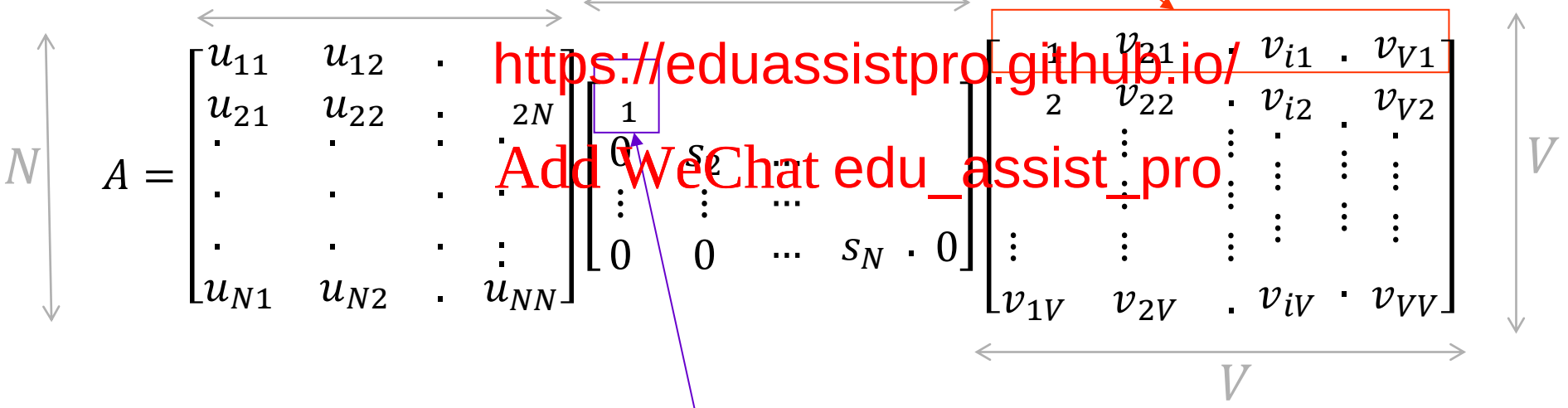


Singular Value Decomposition (SVD)

N=number of docs, V=vocabulary size

$$A = USV^T$$

Direction of most significant correlation



'Strength' of most significant correlation

Interpretation of LSA

- The matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices
  - Their entries are real numbers
  - $\boldsymbol{U}$ is N x ocuments) and $\boldsymbol{V}$ is V x size)
  - They satisfy $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I} =$ $^T = \boldsymbol{I} = \boldsymbol{V}^T\boldsymbol{V}$
- The singular values $s_1,..., s_N$ are positive and satisfy $s_1 \geq s_2 \geq ... \geq s_N$
- The off-diagonal entries of $\boldsymbol{S}$ are all zero

Interpretation of LSA (continued)

- Focussing on $\boldsymbol{V}$:
  - The columns of $\boldsymbol{V}$, $\{v_1,\ldots,v_V\}$ are unit vectors and orthogonal to each other
  - They for asis (coordinate system) f r space
  - Each column of $\boldsymbol{V}$ is a ector corresponding to a semantic class (topic) in the corpus
  - The importance of the topic corresponding to v_n is indicated by the size of the singular value s_n

Interpretation of LSA (continued)

- Since v_n is a document vector, its j^{th} value corresponds to TF-IDF weight for j^{th} term in the vocabulary for the corresponding document/topic
- This can be pic corresponding to v_n – a larg that the j^{th} term in the vocabulary is signifi e topic

Interpretation of LSA (continued)

- Now consider U
- It is easy to show that

$$Av_n = USV^T v_n = s_n u_n$$

- While v_n des combination of terms/words, mbination of documents

Topic-based representation

- Columns of $\boldsymbol{V}$, $v_1,\ldots,v_V$ are an **orthonormal basis** (coordinate system) for the document vector space
- If d is a document, $vec(d)\cdot v_n$ is the magnitude of the component o n of v_n
- ..the compon nding to n
- Hence the vector $top(d) = \begin{bmatrix} vec(d)\cdot v_1 \\ vec(d)\cdot v_2 \\ \cdot \\ \cdot \\ vec(d)\cdot v_V \end{bmatrix} = V^T\, vec(d)$

is a **topic-based representation** of d in terms of $v_1,\ldots,v_V$

More information about LSA

- See:

  Landauer, T.K. and Dumais, S.T., "A solution to Platos problem: The Latent Semantic Analysis theory of the and representation of knowled *eview* *104(2), 211-240* (1997)

Thoughts on document vectors

- Once *d* is replaced by *vec*(*d*) it becomes a point in a vector space
- How does the structure of the vector space reflect the properties of it?
- Do clusters of vectors correspond to semantically related documents?
- Can we partition the vector space into semantically different regions?
- These ideas are a link between IR and Data Mining

For an alternative perspective…

- Chapter 14: “The cunning fox”
- Application of LSA to ‘dating agency’ personal adverts
- LSA suggest of a personal advert can b expressed as a weighted combination of a few basic ‘concepts’

Dr Graham Tattersall, “Geekspeak: How life + mathematics = happiness”, 2007

Summary

- Latent Semantic Analysis
- Interpretatio