Data Mining and Machine Learning

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

Latent Semantic Analysis (LSA)

Peter Jančovič

Objectives

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

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

Vector Notation

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

Assignment Project Exam Help

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

https://powcoder.com

Add WeChat powcoder

$i(1)^{th}$ 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 typically be several (tens of) thousands
- Let's focus on just 2 of these dimensions, corresponding, say, to the words 'sea' and 'beach'
- Intuitively, often, when a document ***d*** includes 'sea' it will also include 'beach'

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

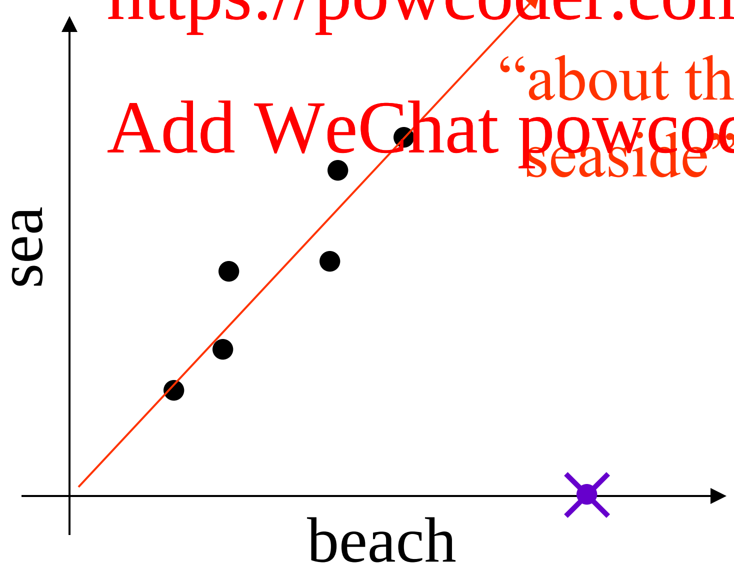
LSA continued

- 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

Assignment Project Exam Help

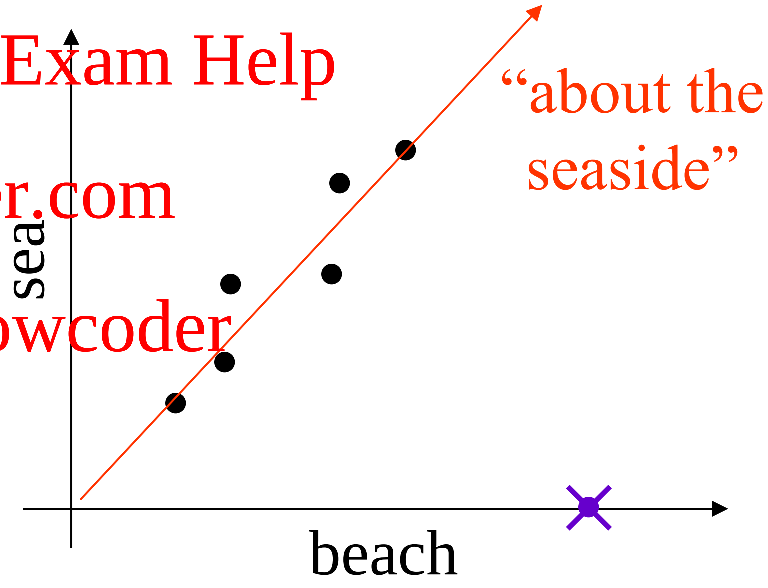
https://powcoder.com

Add WeChat powcoder



Latent Semantic Classes

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



Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

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 decompose *A* using Singular Value Decomposition (SVD)
  - SVD is a standard technique from matrix algebra
  - Packages such as MATLAB have SVD functions:

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

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 matrices A we can write $A=UDU^T$, where U is an **orthogonal matrix** (rotation) 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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Word-Document Matrix

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

term

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

$$A = \begin{bmatrix} w_{t_1 d_1} & w_{t_2 d_1} & \cdots & w_{t_m d_1} & \cdots & w_{t_V d_1} \\ w_{t_1 d_2} & w_{t_2 d_2} & \cdots & w_{t_m d_2} & \cdots & w_{t_V d_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{t_1 d_n} & w_{t_2 d_n} & \cdots & \boxed{w_{t_m d_n}} & \vdots & w_{t_V d_n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{t_1 d_N} & w_{t_2 d_N} & \cdots & w_{t_m d_N} & \cdots & w_{t_V d_N} \end{bmatrix}$$

document

Weighting for term $\boldsymbol{t_m}$ in $\boldsymbol{d_n}$

Singular Value Decomposition (SVD)

N=number of docs, V=vocabulary size

$$A = USV^T$$

Direction of most significant correlation

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

$$A = \overbrace{\begin{bmatrix} u_{11} & u_{12} & \cdot & u_{1N} \\ u_{21} & u_{22} & \cdot & u_{2N} \\ \cdot & \cdot & \cdot & \vdots \\ \cdot & \cdot & \cdot & \vdots \\ \cdot & \cdot & \cdot & \vdots \\ u_{N1} & u_{N2} & \cdot & u_{NN} \end{bmatrix}}^{N} \overbrace{\begin{bmatrix} \boxed{s_1} & 0 & \cdots & 0 & \cdot & 0 \\ 0 & s_2 & \cdots & 0 & \cdot & 0 \\ \vdots & \vdots & \cdots & \vdots & \cdot & \vdots \\ 0 & 0 & \cdots & s_N & \cdot & 0 \end{bmatrix}}^{V} \underbrace{\begin{bmatrix} \boxed{v_{11} \quad v_{21} \quad \cdot \quad v_{i1} \quad \cdot \quad v_{V1}} \\ v_{12} \quad v_{22} \quad \cdot \quad v_{i2} \quad \cdot \quad v_{V2} \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ v_{1V} \quad v_{2V} \quad \cdot \quad v_{iV} \quad \cdot \quad v_{VV} \end{bmatrix}}_{V}$$

‘Strength’ of most significant correlation

Interpretation of LSA

- The matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are <u>orthogonal matrices</u>
  - Their entries are real numbers
  - $\boldsymbol{U}$ is N x N (N is the number of documents) and $\boldsymbol{V}$ is V x V (V is the vocabulary size)
  - They satisfy $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I} = \boldsymbol{U}^T\boldsymbol{U}$, $\boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{I} = \boldsymbol{V}^T\boldsymbol{V}$
- The <u>singular values</u> $s_1, \ldots, s_N$ are positive and satisfy $s_1 \geq s_2 \geq \ldots \geq s_N$
- The off-diagonal entries of $\boldsymbol{S}$ are all zero

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Interpretation of LSA (continued)

- Focussing on ***V***:
  - The columns of ***V***, $\{v_1,\ldots,v_V\}$ are unit vectors and orthogonal to each other
  - They form a new orthonormal basis (coordinate system) for the document vector space
  - Each column of ***V*** is a document vector 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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

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 used to interpret the topic corresponding to v_n – a large value of v_{nj} indicates that the j^{th} term in the vocabulary is significant for the topic

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Interpretation of LSA (continued)

- Now consider U
- It is easy to show that

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

Assignment Project Exam Help

- While v_n describes the n^{th} topic as a combination of terms/words, u_n describes it as a combination of documents

https://powcoder.com
Add WeChat powcoder

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, $vec(d).v_n$ is the magnitude of the component of $vec(d)$ in the direction of v_n

Assignment Project Exam Help
https://powcoder.com

- ..the component of $vec(d)$ corresponding to topic n

Add WeChat powcoder

- 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 acquisition, induction and representation of knowledge", *Psychological Review 104(2), 211-240* (1997)

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

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 the documents in 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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

For an alternative perspective…

- Chapter 14: "The cunning fox"
- Application of LSA to 'dating agency' personal adverts
- LSA suggests that the meaning of a personal advert can be expressed as a weighted combination of a few basic 'concepts'

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

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

Summary

- Latent Semantic Analysis
- Interpretation of LSA

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder