Intelligent Data Analysis

Week 5 – Lecture 9

Latent Semantic Analysis (LSA)

Martin Russell

Objectives

- To understand, intuitively, how **Latent Semantic Analysis (LSA)** can
  - Discover latent **topics** in a corpus and represent them in terms of words
  - Achieve dimension reduction for document vectors
  - Represent words in terms of topics
- To understand the relationship between LSA and PCA applied to a set of document vectors

Vector Notation for Documents

- Suppose that we have a set of documents

  $$D = \{\boldsymbol{d_1}, \boldsymbol{d_2}, \ \dots \ , \boldsymbol{d_N}\}$$

  think of this as the corpus for IR
- Suppose that the number of **different** words in the **whole corpus** is $\boldsymbol{V}$ (vocabulary size)
- Now suppose a document $\boldsymbol{d}$ in D contains $\boldsymbol{M}$ different terms: $\{\boldsymbol{t_{i(1)}}, \boldsymbol{t_{i(2)}}, \dots, \boldsymbol{t_{i(M)}})$
- Finally, suppose term $\boldsymbol{t_{i(\mathrm{m})}}$ occurs $\boldsymbol{f_{i(\mathrm{m})}}$ times

Vector Notation

- The **vector representation** $\boldsymbol{vec(d)}$ of $\boldsymbol{d}$ is the $\boldsymbol{V}$ dimensional vector:

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

$i(1)^{\text{th}}$ place

$i(2)^{\text{th}}$ place

$i(M)^{\text{th}}$ place

Notice that this is the <u>weighting</u> – i.e. the <u>term frequency</u> times the <u>inverse document frequency</u> $w_{i(1)} = f_{i(1)\boldsymbol{d}} \times \boldsymbol{IDF(i(1))}$ from text IR

- $vec(d)$ is the **document vector** for d

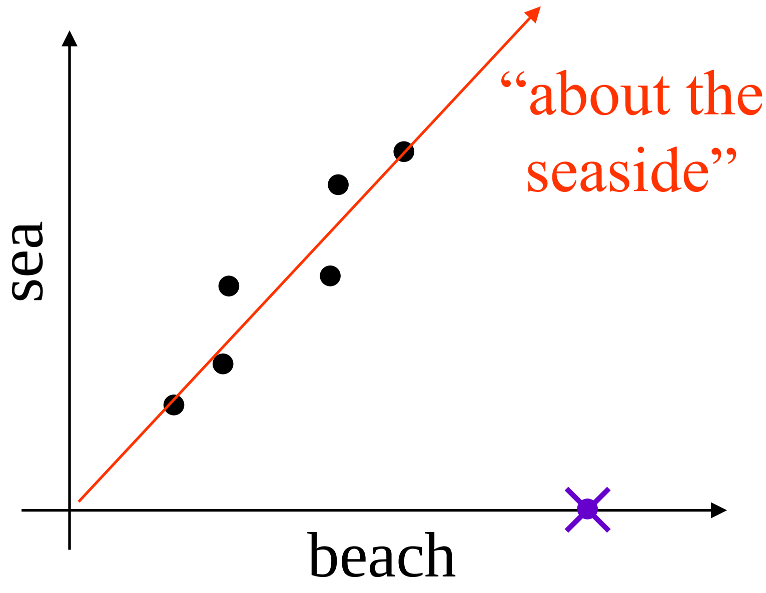
UNIVERSITY OF BIRMINGHAM

Latent Semantic Analysis (LSA)

- Suppose we have a real corpus with a large number of documents
- For each document $\boldsymbol{d}$ the dimension of the vector $\boldsymbol{vec}(\boldsymbol{d})$ is potentially several thousands
- Let's focus on just 2 of these dimensions, corresponding, say, to the words 'sea' and 'beach'
- Intuitively, often, when a document $\boldsymbol{d}$ includes 'sea' it will also include 'beach'

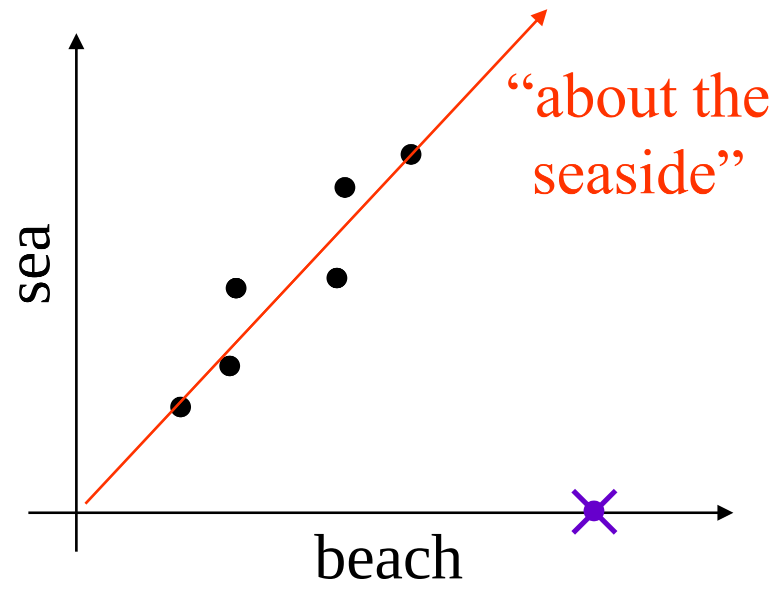
LSA continued

- Equivalently, if $\boldsymbol{vec}(\boldsymbol{d})$ 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 words **automatically**, from **data**
- In the example we have found an equivalence set of 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 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)`

Singular Value Decomposition

- Recall **eigenvector decomposition**
- An eigenvector of a square matrix A is a vector e such that $Ae = \lambda_e e$, where λ is a scalar.
- For certain matrices A we can write $A=UDU^T$, where U is an **orthogonal matrix** and D is **diagonal (Spectral theorem)**
  - 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 the $N \times V$ matrix whose nth row is the document vector for the nth document

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

Weighting for term t_m in d_n

Singular Value Decomposition (SVD)

$$A = USV^T$$

Direction of most significant correlation

$$A = \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1N} \\ u_{21} & u_{22} & \dots & u_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ u_{N1} & u_{N2} & \dots & u_{NN} \end{bmatrix} \begin{bmatrix} \boxed{s_1} & 0 & 0 & \dots & \dots & 0 \\ 0 & s_2 & 0 & \dots & \dots & 0 \\ \vdots & \vdots & \dots & 0 & \vdots & \vdots \\ 0 & 0 & \dots & s_N & 0 & 0 \end{bmatrix} \begin{bmatrix} \boxed{v_{11} \quad v_{21} \quad \dots \quad v_{m1} \quad \dots \quad v_{V1}} \\ v_{12} \quad v_{22} \quad \dots \quad v_{m2} \quad \dots \quad v_{V2} \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ v_{1m} \quad v_{2m} \quad \dots \quad v_{mn} \quad \dots \quad v_{Vm} \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ v_{1V} \quad v_{2V} \quad \dots \quad v_{mV} \quad \dots \quad v_{VV} \end{bmatrix}$$

'Strength' of most significant correlation

$$m = \frac{1}{N} \sum_{n=1}^{N} vec(d_i)$$

↑ average document vector

$$W = W - m$$

$W^T W$: covariance matrix

$$W = USV^T$$

In PCA,

$$C = VDV^T$$

$$D \Leftrightarrow S^2$$

$$C = W^T W = (USV^T)^T (USV^T)$$
$$= (VSU^T) \cdot USV^T$$
$$= V \cdot S^2 \cdot V^T$$

Interpretation of LSA

- The matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are **orthogonal matrices**
  - 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 **singular values** $s_1,\ldots, s_N$ are positive and satisfy $s_1 \geq s_2 \geq \ldots \geq s_N$
- The off-diagonal entries of $\boldsymbol{S}$ are all zero

UNIVERSITY OF BIRMINGHAM

Interpretation of LSA (continued)

- Focusing on $\boldsymbol{V}$:
- The columns of $\boldsymbol{V}$ $\{v_1,\ldots,v_V\}$ are V dimensional unit vectors orthogonal to each other
- They form a new **orthonormal basis** (coordinate system) for the document vector space
- Each column of $\boldsymbol{V}$ is a document vector corresponding to a **semantic class** (**topic**) in the corpus

Super vector

- The importance of the topic v_n is indicated by the magnitude of the singular vector 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 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.

Interpretation of LSA (continued)

- Now consider U
- It is easy to show that

$$Av_n = USV^T v_n = s_{nn} u_n$$

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

UNIVERSITY OF BIRMINGHAM

Topic-based representation

- Columns of 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 of $vec(d)$ in the direction of v_n
- ..the component of $vec(d)$ corresponding to topic n
- Hence the vector $top(d) = \begin{bmatrix} vec(d)\cdot v_1 \\ vec(d)\cdot v_2 \\ \vdots \\ vec(d)\cdot v_n \\ \vdots \\ vec(d)\cdot v_V \end{bmatrix}$

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

UNIVERSITY OF BIRMINGHAM

Topic-based dimension reduction

- Since the singular value s_n indicates the importance of topic v_n, we can choose to truncate the vector $top(d)$ when s_n becomes small:

$$top(d) \approx \begin{bmatrix} vec(d) \cdot v_1 \\ vec(d) \cdot v_2 \\ \vdots \\ vec(d) \cdot v_n \end{bmatrix} = V_{(n)}^T vec(d)$$

  where $\boldsymbol{V}_{(n)}$ is the $\mathrm{V} \times n$ matrix comprising the first n columns of $\boldsymbol{V}$

- $top(d)$ is a **reduced ($\boldsymbol{n}$) dimensional vector representation** of document d

Topic-based word representation

- Suppose w is the i^{th} word/term in the vocabulary
- The **one-hot vector** $h(w)$ is the vector:

$$h(w) = \begin{bmatrix} 0 \\ . \\ . \\ . \\ 0 \\ 1 \\ 0 \\ . \\ . \\ 0 \end{bmatrix} \longleftarrow i^{\text{th}} \text{ entry}$$

- $h(w)$ is like the document vector for a document consisting of just the word w

Topic-based word representation

- We can use $h(w)$ to obtain a vector $top(w)$ that describes w in terms of the topics that it contributes to:

h(w) · v

$$top(w) = \begin{bmatrix} h(w) \cdot v_1 \\ h(w) \cdot v_2 \\ \vdots \\ \vdots \\ h(w) \cdot v_n \end{bmatrix} = V_{(n)}^T top(w)$$

Words in similar direction
=> represent similar topics

- where $V_{(n)}$ is the $\mathrm{V} \times n$ matrix comprising the first n columns of $\boldsymbol{V}$

Topic-based word representation

- Intuitively, if two words v and w are synonyms they will contribute in a similar way to similar topics and the vectors $top(v)$ and $top(w)$ will point in similar directions
- A vector like $top(w)$ is sometimes referred to as a **word embedding**
- We will re-visit this idea later

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)

Topic based document analysis

- There are other approaches to identifying a set of topics that represent a collection of documents – **Latent Dirichlet Alocation (LDA)**

topics ~ distribution of words

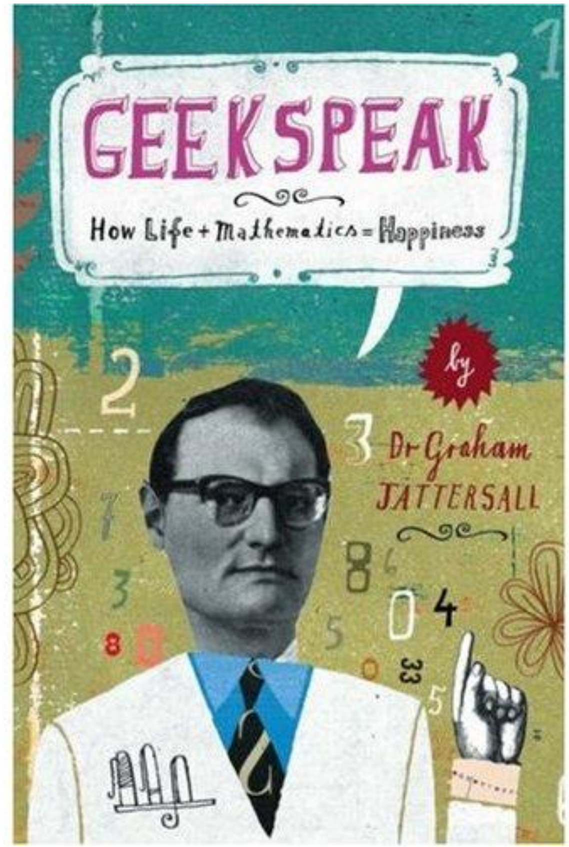
UNIVERSITY OF BIRMINGHAM

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

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'



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

Relationship between LSA and PCA

- What is the relationship between LSA and PCA?

Summary

- Latent Semantic Analysis
- Interpretation of LSA
- Topic-based representation of documents
- Topic-based dimension reduction

Homework – this is challenging!

- Find the C program doc2vec.c on the course webCT page
- Use this to convert the BEng project specs from Laboratory 1 into a matrix of document vectors - this is the Word-Document matrix A from the notes
- Load this matrix into MATLAB and perform SVD
- Can you interpret the resulting Singular Vectors (columns of V)?