Data Mining
Lecture 11: Semantic Spaces

Jo Grundy

ECS Southampton

21st March 2022

Semantic Spaces - Introduction

Distributional Semantics - Hypothesis:

Words that have similar distributions have similar meanings

"Words that occur in similar contexts have similar meanings" Wittgenstein 1953

"A word is characterised by the company it keeps" Firth 1958

We can exploit this to uncover *hidden meanings*

Semantic Spaces - Introduction

E.g. What does "dhuaihf" mean?
"The dhuaif was hot and fresh, made a good meal"

"Small shiny dhuaif swam in the water near my boat"

Semantic Spaces - Introduction

Semantic Spaces:

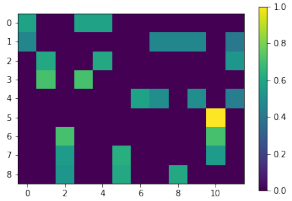
- represent word meanings as vectors that keep track of the words distributional history
- focus on semantic similarity
- similarity measured using geometrical methods

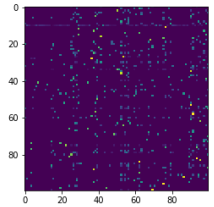
e.g. Cosine similarity between PC and Windows = 0.77
Cosine similarity between PC and window = 0.13
In Japanese, A. Utsumi, IEEE SMC 2010

Semantic Spaces - Construction

Matrix Construction:
Consider a term-document matrix which describes occurrences of terms in documents

- Sparse
- Weighted (e.g. TF.IDF)





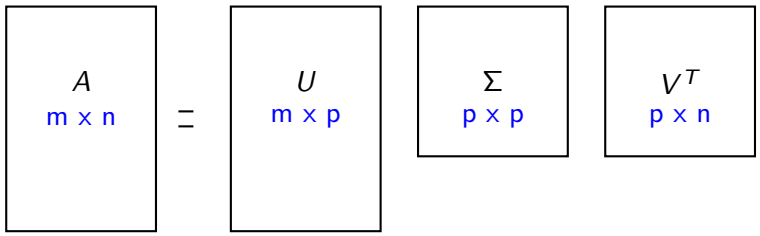
Semantic Spaces - Latent Semantic Analysis

Latent Semantic Analysis (LSA) makes a low-rank approximation
It assumes the term-document matrix:

- is noisy, and should be de-noised
- is more sparse than it should be

Semantic Spaces - Recap SVD

$$A = U\Sigma V^T$$

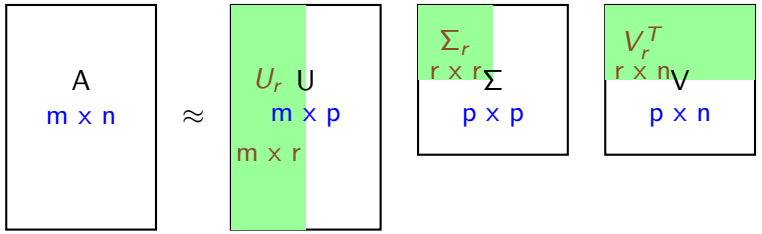


Where p is rank of matrix A

U called *left singular vectors*, contains the eigenvectors of AA^T,
V called *right singular vectors*, contains the eigenvectors of A^TA
Σ contains square roots of eigenvalues of AA^T and A^TA

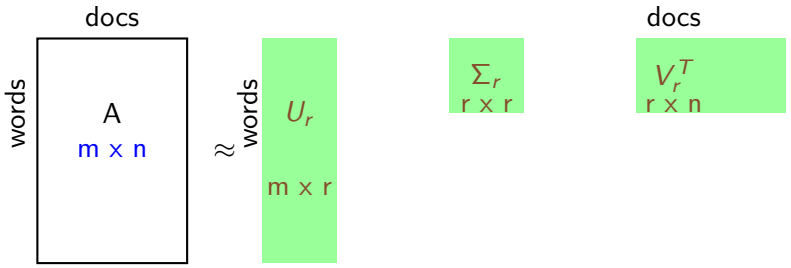
If A is matrix of mean centred featurevectors, V contains principal components of the covariance matrix

Semantic Spaces - Recap Truncated SVD



Uses only the largest r singular values (and corresponding left and right vectors)
This can give a *low rank approximation* of A, $\tilde{A} = U_r\Sigma_r V_r$
This has the effect of minimising the Frobenius norm of the difference between A and $\tilde{A}$

Semantic Spaces - LSA



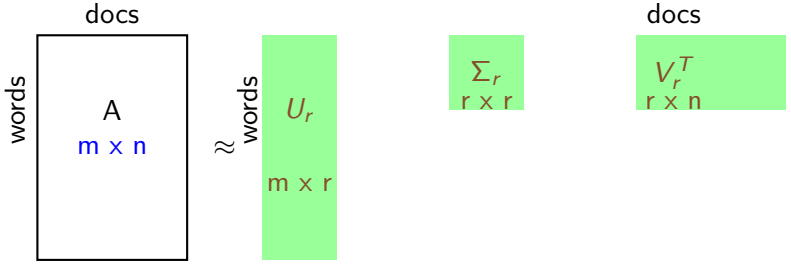
Each row of V_r corresponds to an eigenvector of $M^T M$

- This means it is proportional to the covariance or correlation between the documents
- These are the *concepts*

Each row of U_r describes a term as a vector of weights with respect to r *concepts*

Each column of V_r describes a document as a vector of weights with respect to r concepts

Semantic Spaces - LSA



Term concepts and document concepts have the same dimensionality, but represent different spaces.

Semantic Spaces - LSA

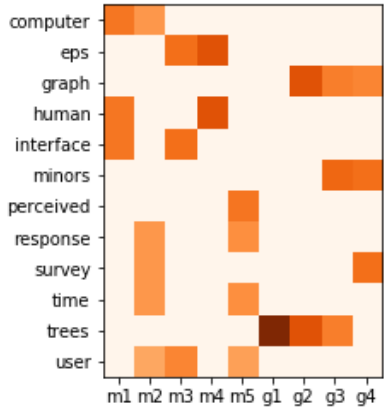
Example:
a set of strings:

m1	"Human machine interface for ABC computer applications"
m2	"A survey of user opinion of computer system response time"
m3	"The EPS user interface management system"
m4	"System and human system engineering testing of EPS"
m5	"Relation of user perceived response time to error measurement"
g1	"The generation of random, binary, ordered trees"
g2	"The intersection graph of paths in trees"
g3	"Graph minors IV: Widths of trees and well-quasi-ordering"
g4	"Graph minors: A survey"

Semantic Spaces - LSA

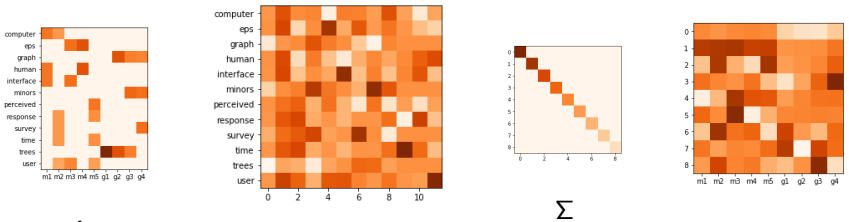
calculate TF.IDF

0.58	0.46	0.	0.	0.	0.	0.	0.	0.
0.	0.	0.6	0.71	0.	0.	0.	0.	0.
0.	0.	0.	0.	0.	0.	0.71	0.55	0.52
0.58	0.	0.	0.71	0.	0.	0.	0.	0.
0.58	0.	0.6	0.	0.	0.	0.	0.	0.
0.	0.	0.	0.	0.	0.	0.	0.63	0.6
0.	0.	0.	0.	0.58	0.	0.	0.	0.
0.	0.46	0.	0.	0.49	0.	0.	0.	0.
0.	0.46	0.	0.	0.	0.	0.	0.	0.6
0.	0.46	0.	0.	0.49	0.	0.	0.	0.
0.	0.	0.	0.	0.	1.	0.71	0.55	0.
0.	0.4	0.52	0.	0.43	0.	0.	0.	0.



Semantic Spaces - LSA

SVD: $A = U\Sigma V^T$

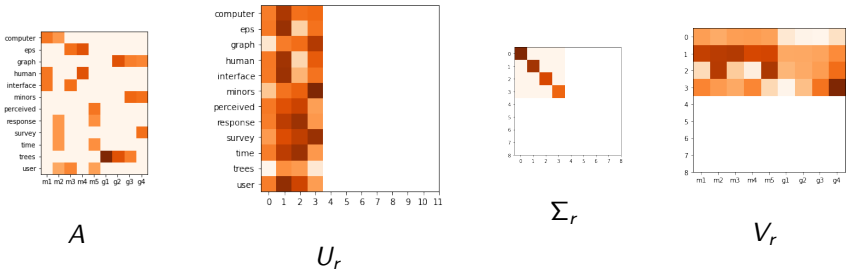


A U Σ V

We then reduce the dimensionality by choosing only the first few eigenvalues (in Σ) and the corresponding columns in U and V.

Semantic Spaces - LSA

SVD: $A \approx U_r \Sigma_r V_r^T \quad r = 4$



A U_r Σ_r V_r

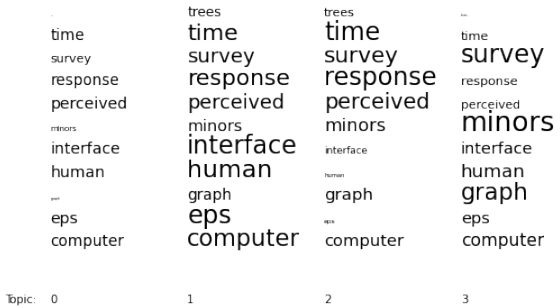
Each row of U_r describes a word as a vector of weights with respect to r *concepts*

Each column of V_r describes the title as a vector of weights with respect to r concepts

Semantic Spaces - LSA

What do these 'concepts' mean?
U_r gives us the weighting of the words for each concept

We can show that weighting for each concept:

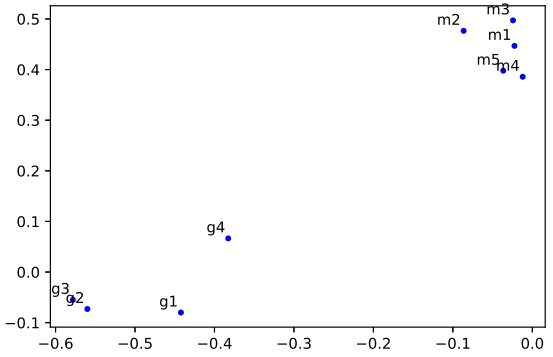


This shows us perhaps that the concepts are not always particularly meaningful.

Semantic Spaces - LSA

Cosine similarity of document vectors can be compared ($r = 2$)

- Vectors for "m1" and "m2" give cosine similarity = 0.93
- Vectors for "g1" and "g2" give cosine similarity = 0.83
- Vectors for "g1" and "m1" give cosine similarity = 0.18

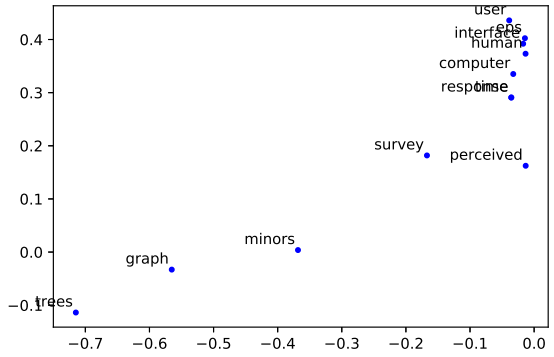


Clustering algorithms can be used on the vectors

Semantic Spaces - LSA

Cosine similarity of word vectors can be compared ($r = 2$)

- Vectors for "human" and "interface" give cos similarity = 0.95
- Vectors for "human" and "user" give cos similarity = 0.11
- Vectors for "graph" and "minor" give cos similarity = 0.90



Clustering algorithms can be used on the vectors

Semantic Spaces - LSI

Latent Semantic Indexing (LSI) LSA can be used for document retrieval

- Given Query: view as query vector $\boldsymbol{q}$
- Project $\boldsymbol{q}$ in to document space
- Compare with document vectors, find closest

Results work mathematically
However, results may not be easy to interpret in terms of natural language.

Semantic Spaces - LSA

Problems?

- Polysemious words - with multiple meanings - aren't captured
  - The vector representation averages all meanings of the word
  - e.g. 'fit' is an adjective and a verb
- Word order is ignored (use n-grams?)
- LSA assumes words and documents form a joint Gaussian distribution, however a Poisson distribution is observed

Semantic Spaces - LSA

Web search in one language will not normally give relevant results in another language, as the words will not match.
The search engine could index translated documents, but:

- Automatic translation is far from perfect
- Manual translation is very slow and expensive

Semantic Spaces - LSA

To use two languages, you can make a word-document matrix using documents from both languages.
E. g. in Canada, parliamentary records are kept in both French and English. They are direct translations of each other.

- "Mr Speaker, we are in constant touch with our consular officials in Libya."
- "Monsieur le Président, nous somme en communication constante avec nos représentants consulaire en Libye."

The two documents would be preprocessed separately (stemming etc.) then concatenated before making the word-document matrix using TF.IDF
Words that are direct translations of each other are close together in word space.

Semantic Spaces - LSA

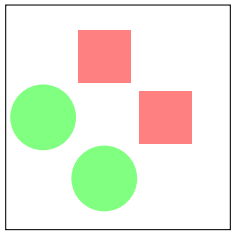
This would clearly only be of use to documents that have a direct translation.
However, documents without a translation can be projected in to the same space

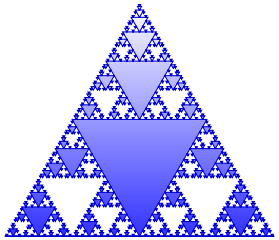
- Make a word-document matrix using a collection of monolingual documents: $\bar{B}$
- All the rows for the words from other language will have 0
- Use $\bar{V}^T = \Sigma^{-1} U^T \bar{B}$

Where Σ is the diagonal and U is the term representation matrix from the SVD on the bilingual corpus
This gives a projection of the new monolingual documents in to the bilingual space.
A search can then be accomplished by encoding a query from the other language and projecting it in to the same space the same way, and measuring the cosine similarity in that bilingual space

Semantic Spaces - LSA

So far: *bag of words* (BOW) from natural language.
However the maths should work for compositions of occurrences in any unit.
For example, in image search we might want to search for other images with circles.
The image could be encoded with the number of different shapes it has.





Semantic Spaces - LSA

Need to make a large multidimensional space in which images, keywords and visual terms can be placed
In training:

- Learn how images and keywords are related
- Place images and keywords close together in the space

Unannotated images can be placed in the space based on the visual terms they contain

- Images can be placed based on their visual terms in the space
- They should lie near the keywords that describe them

Semantic Spaces - LSA

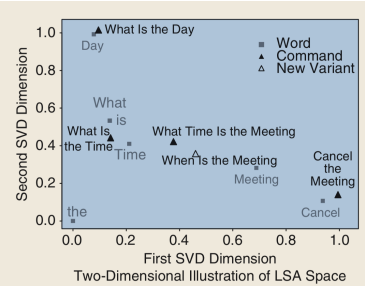
This lower dimensional space can be used to:

- Find Images using similar words
- Find images with similar images
- Return possible key words for an image
- Find relationships between words, and between words and visual terms
- Image segmentation

Semantic Spaces - LSA

LSA can also be used for:

- Language modelling 'item command-based speech recognition
- spam filtering
- Pronunciation modelling
- e.t.c..



Two-Dimensional Illustration of LSA Space

Semantic Spaces - Summary

LSA is a powerful application of truncated SVD
Unfortunately, it has problems with:

- Words with more than one meaning
- Abstract concepts

LSA has had come success, but there are newer techniques.