Dimensionality Reduction

Automatically acquiring lexical semantics from text

Tim Van de Cruys

University of Groningen

Statistics Seminar
May 30, 2007

Introduction

Two main reasons for performing dimensionality reductions:

- Intractable computations
  - When number of elements and number of features is too large, similarity computations may become intractable
  - reduction of the number of features makes computation tractable again
- Generalization capacity
  - the dimensionality reduction is able to describe the data better, or is able to capture intrinsic semantic features
  - dimensionality reduction is able to improve the results (counter data sparseness and noise)

Latent Semantic Analysis: Introduction 1/2

- Application of a mathematical/statistical technique to simulate how humans learn the semantics of a word
- LSA finds 'latent semantic dimensions' according to which words and documents can be identified
- Words (and passages) are represented as high-dimensional vectors in this semantic space
- Goal: counter data sparseness (poverty of the stimulus) and get rid of noise

Latent Semantic Analysis: Introduction 2/2

What is Latent Semantic Analysis technically speaking?

- The application a **singular value decomposition**
- to a **term-document matrix**
- to improve **vector space measures**

Bag-of-word semantics

- LSA represents 'bag-of-word' semantics
- Idea that meaning of a passage equals the sum of the meaning of its words
- Meaning = an unordered set of word tokens, syntax is not taken into account
- Done by representing several passages in a **term-document matrix**

Term-document matrix 1/2

Consider two documents:

- België is een koninkrijk in het midden van Europa, met als hoofdstad **Brussel**. **Brussel** heeft een Nederlandstalige en een Franstalige universiteit, maar de grootste studentenstad is **Leuven**. **Leuven** telt 27.000 studenten.
- Nederland is een West-Europees land aan de Noordzee. De hoofdstad van Nederland is **Amsterdam**. **Amsterdam** telt twee universiteiten. **Groningen** is een belangrijke studentenstad. In **Groningen** studeren 37.000 studenten.

Term-document matrix 2/2

$$\begin{bmatrix} & B & NL \\ Groningen & 0 & 2 \\ Leuven & 2 & 0 \\ Amsterdam & 0 & 2 \\ Brussel & 2 & 0 \end{bmatrix}$$

Note: Values for words are transformed to values that represent their **importance** in the passage (entropy, mutual information)

Matrix

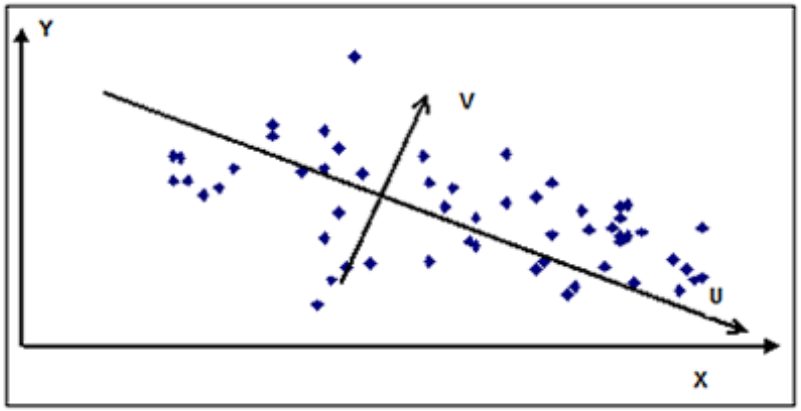
$$\begin{bmatrix} & B & NL \\ Groningen & 0 & 2 \\ Leuven & 2 & 0 \\ Amsterdam & 0 & 2 \\ Brussel & 2 & 0 \end{bmatrix}$$

Apply cosine similarity measure

- $cos(\overrightarrow{x}, \overrightarrow{y}) = \frac{\overrightarrow{x} \cdot \overrightarrow{y}}{|\overrightarrow{x}||\overrightarrow{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$
- Examples:
  - $cos(Groningen, Amsterdam) = \frac{4}{\sqrt{16}} = 1$
  - $cos(Groningen, Brussel) = \frac{0}{\sqrt{16}} = 0$

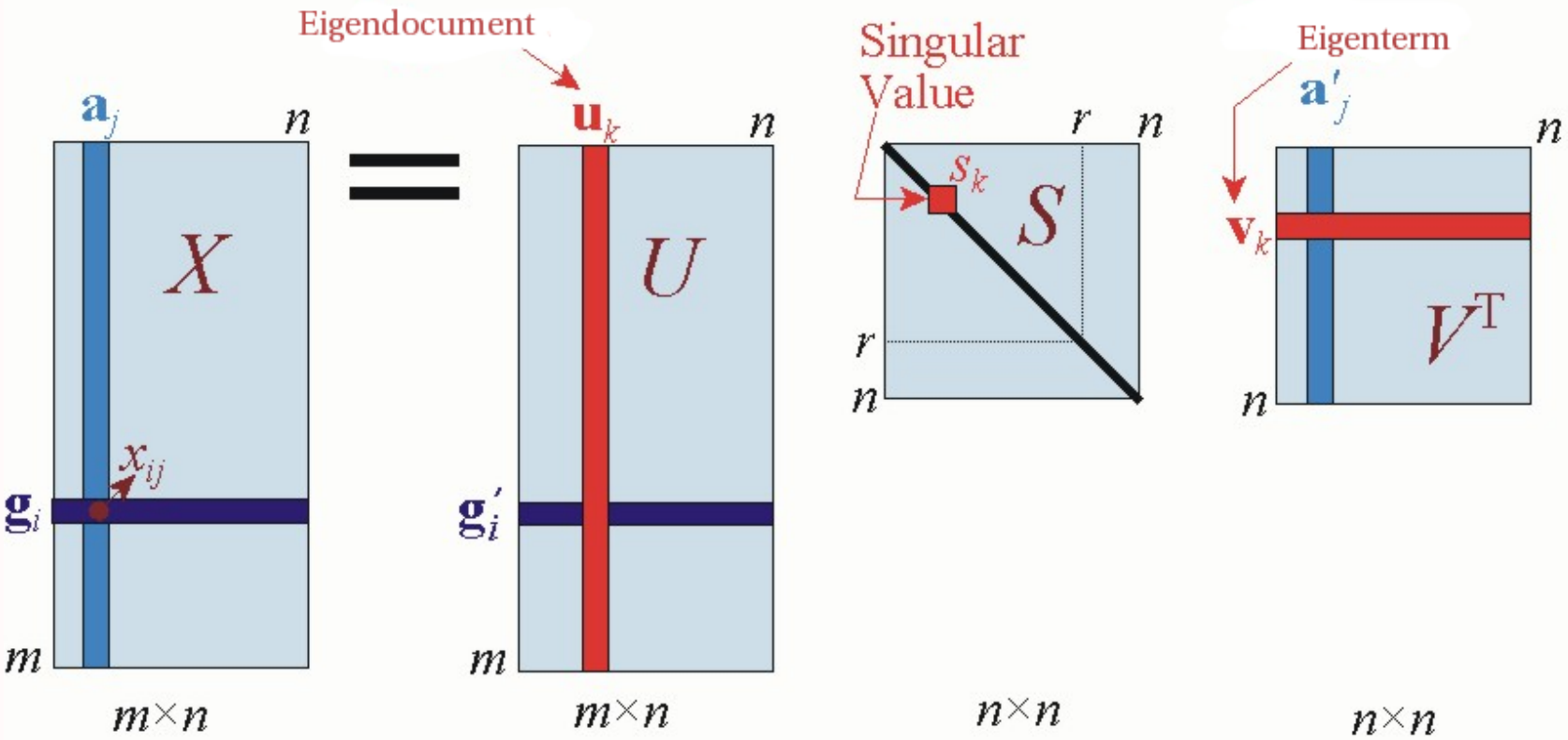
Singular Value Decomposition

- Mathematical/statistical technique closely related to principal component analysis, factor analysis
- Find the dimensions that explain most variability by finding **eigenvectors** of matrix
- Only keep the n most important dimensions (n=50-1000)



SVD: three matrices

$$X = USV^{\mathrm{T}}$$



Example 1a

$$A \begin{bmatrix} & B & NL \\ Groningen & 0 & 2 \\ Leuven & 2 & 0 \\ Amsterdam & 0 & 2 \\ Brussel & 2 & 0 \end{bmatrix} =$$

$$U \begin{bmatrix} 0.00 & 0.71 \\ -0.71 & 0.00 \\ 0.00 & 0.71 \\ -0.71 & 0.00 \end{bmatrix} * S \begin{bmatrix} 2.83 & 0 \\ 0 & 2.83 \end{bmatrix} * V^T \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$$

Example 1b

$$\begin{bmatrix} & B & NL & B \\ Groningen & 0 & 2 & 0 \\ Leuven & 2 & 0 & 0 \\ Amsterdam & 0 & 2 & 0 \\ Brussel & 2 & 0 & 1 \end{bmatrix} =$$

$$U \begin{bmatrix} 0.00 & -0.71 & \emptyset \\ -0.66 & 0.00 & \emptyset \\ 0.00 & -0.71 & \emptyset \\ -0.75 & 0.00 & \emptyset \end{bmatrix} * S \begin{bmatrix} 2.92 & 0 & \emptyset \\ 0 & 2.83 & \emptyset \\ \emptyset & \emptyset & \emptyset \end{bmatrix} *$$

$$V^T \begin{bmatrix} -0.97 & 0.00 & 0.26 \\ 0 & -1.00 & 0.00 \\ \emptyset & \emptyset & \emptyset \end{bmatrix} \cong A' \begin{bmatrix} 0.0 & 2.0 & 0.0 \\ 1.9 & 0.0 & 0.5 \\ 0.0 & 2.0 & 0.0 \\ 2.1 & 0 & 0.6 \end{bmatrix}$$

Methodological remarks

- LSA in CLEF-corpus: 4 years of Dutch newspaper texts (Algemeen Dagblad, NRC Handelsblad)
- terms = nouns
  documents = articles
- 20.000 terms * 100.000 documents matrix
- reduced to 300 latent semantic dimensions

Dimension 37

contingent_periode	0.55974	verslagweek	0.51471
NBC's	0.321241	courant	0.31842
week_staat	0.226316	dag_geld_rente	0.192096
kas_reserve	0.151812	belening	0.131637
belasting_afdracht	0.127845	geldmarkt_tarief	0.109959
verkrapping	0.10905	beleningen	0.0904901
geldmarkt	0.0747359	hoofde	0.0622487
DNB	0.0586182	storting	0.0567692
contingent	0.0542929	voorschot	0.0505641
omloop	0.0499995	mutatie	0.0472168
benutting	0.0324582	procentpunt	0.0323419
voorschot_rente	0.030187	schatkist	0.029555
ambtenaren_salaris	0.0289534	beroep_contingent	0.0286949
bankbiljet	0.0264982	basispunt	0.0243368
bankwezen	0.0237	liquiditeiten	0.0233299

Dimension 64

23u	0.105716	Sneak	0.0904175
16u15	0.0882233	preview	0.0798345
On deadly ground	0.0636528	Cool runnings	0.0633079
22u30	0.0602807	17u30	0.0547169
Trois couleurs	0.0408388	Sister	0.0407604
ardilla	0.0395078	roja	0.0395078
The snapper	0.0364226	Mrs. Doubtfire	0.0335517
21u15	0.0328192	Monk	0.0325375
Intersection	0.024875	14u45	0.02288
spirits	0.022065	euro	0.021874
The piano	0.0212932	Aladdin	0.016921
Desmet	0.0166517	Ace Ventura	0.0162278
Mr. Jones	0.0149521	The three musketeers	0.0138651
La	0.0125177	rocker	0.01078
Philadelphia	0.0104788	15u30	0.0104311

Clustering with LSA 1/2

- azijn bieslook bleekselderij blokje bosuitje boter bouillon champignon citroen citroensap crème deciliter deeg deksel dressing eetlepel folie garnaal gehakt gram ham hoofdgerecht keukenpapier knoflook koekepan komkommer kook lepel mossel mosterd olijf olijfolie oven pan pannetje paprika peper peterselie plak plakje prei reepje room salade saus slagroom spinazie takje theelepel Tip tomaat ui vocht voor_gerecht vruchtvlees vulling warmtebron zeef zout
- artillerie Banja Luka Belgrado Bihac directrice enclave ex-Joegoslavië her_opening Knin Krajina Kroaat Kroatië luchtmachtbasis massagraf Milosevic Montenegro Oost-Slavonië Radovan Karadzic Servië Serviër Tudjman UNPROFOR VN-soldaat vredesplan Zagreb

Clustering with LSA 2/2

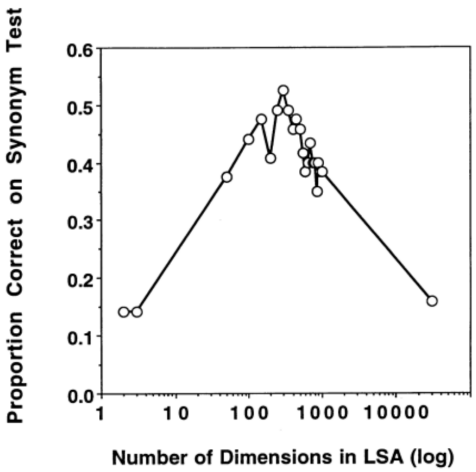
- aantasting afbraak broeikaseffect CO dikte halfrond katalysator KNMI kooldioxyde kristal meting Montreal Nature oceaan ozon_laag stabilisatie straling verbranding vulkaan waarneming zonlicht zuurstof

Clustering with syntactic relations (distributional similarity)

- bieslook bosuitje citroensap deciliter eetlepel eierdooier gember gram kaneel knoflook koriander mosterd peper peterselie theelepel tijm
- bak container doos kist koffer kom pan pot schaal zak
- Albanië Armenië Georgië Kazachstan Kroatië Macedonië Oekraïne Oezbekistan Wit-Rusland
- Hutu jood Kroaat moslim Palestijn Serviër

LSA & synonym tests

- LSA trained on Grolier Encyclopedia, and given synonym test (TOEFL test).
- LSA scores 65%, identical to the average score of a large sample of students applying for college entrance in the United States from non-English speaking countries.



LSA & essay grading

- LSA was used to assign scores to essay questions
- Trained on instructional text read by students, and a few example essays (or pre-graded essays)
- LSA scores found to be as good as expert assigned scores
- Moreover: LSA correlated significantly better with individual expert graders than one expert correlated with another
- LSA grades essays as good as expert **and** more objective

Technique 1/2

- Given a non-negative matrix V, find non-negative matrix factors W and H such that:

$$V_{nxm} \approx W_{nxr} H_{rxm} \tag{1}$$

- Choosing $r \ll n, m$ reduces data
- Constraint on factorization: all values in three matrices need to be *non-negative values* (≥ 0)
- Constraint brings about a *parts-based* representation: only additive, no subtractive relations are allowed

Technique 2/2

- Different kinds of NMF's that minimize different cost functions:
  - Square of Euclidean distance (L1-norm)
  - Kullback-Leibler Divergence (L2-norm)
    $\Rightarrow$ better suited for language phenomena
- To find NMF is to minimize $D(V \| WH)$ with respect to W and H, subject to the constraints $W, H \geq 0$
- This can be done with *update rules*

$$H_{a\mu} \leftarrow H_{a\mu} \frac{\sum_i W_{ia} \frac{V_{i\mu}}{(WH)_{i\mu}}}{\sum_k W_{ka}} \qquad W_{ia} \leftarrow W_{ia} \frac{\sum_\mu H_{a\mu} \frac{V_{i\mu}}{(WH)_{i\mu}}}{\sum_v H_{av}} \tag{2}$$

- these update rules find a *local minimum* in the minimization of KD distance

Results

- Context vectors (10K nouns * 2K co-occurring nouns) extracted from CLEF corpus
- NMF is able to capture semantic dimensions (much more obvious and clear than LSA)

Example 1/3

in	de	smaak	vallen	⟷	in	de	put	vallen
		*geur					kuil	
		*voorkeur					krater	
		*stijl					greppel	

- exploit the non-compositionality of the MWE
- verb – PP combinations in which the verb prefers *only one single noun* in a cluster are good MWE candidates

Example 2/3

- **smaak**: idioom, karakter, persoonlijkheid, stijl, temperament, thematiek, uiterlijk, uitstraling, voorkomen

MWE candidate	$P_{v,n}$	$R_{v,n,c}$	MWE?
val#in smaak	.12	1.00	yes
val#in karakter	.00	.00	no
val#in stijl	.00	.00	no

Example 3/3

- **put**: gaatje, gat, kloof, krater, kuil, lek, scheur, valkuil

MWE candidate	$P_{v,n}$	$R_{v,n,c}$	MWE?
val#in put	.00	.04	no
val#in kuil	.00	.11	no
val#in kloof	.00	.01	no
val#in gat	.04	.72	both