A Comparison of Latent Semantic Analysis and Correspondence Analysis of Document-Term Matrices

Qianqian Qi
Utrecht University
q.qi@uu.nl

David J. Hessen
Utrecht University
d.j.hessen@uu.nl

Tejaswini Deoskar
Utrecht University
t.deoskar@uu.nl

Peter G. M. van der Heijden
Utrecht University and University of Southampton
P.G.M.vanderHeijden@uu.nl

Abstract: Latent semantic analysis (LSA) and correspondence analysis (CA) are two techniques that use a singular value decomposition (SVD) for dimensionality reduction. LSA has been extensively used to obtain low-dimensional and dense vectors that capture relationships among documents and terms. In this article, we present a theoretical analysis and comparison of the two techniques in the context of document-term matrices. We show that CA has some attractive properties as compared to LSA, for instance that effects of margins arising from differing document-lengths and term-frequencies are effectively eliminated, so that the CA solution is optimally suited to focus on relationships among documents and terms. A unifying framework is proposed that includes both CA and LSA as special cases. We empirically compare CA to various LSA based methods on two tasks, a document classification task in English and an authorship attribution task on historical Dutch texts, and find that CA performs significantly better. We also apply CA to a long-standing question regarding the authorship of the Dutch national anthem *Wilhelmus* and provide further support that it can be attributed to the author Datheen, amongst several contenders.
Keywords: Latent semantic analysis; Correspondence analysis; Singular value decomposition; Authorship attribution; Text classification.

1 Introduction

Latent semantic analysis (LSA) is a well-known method used in computational linguistics that uses Singular Value Decomposition (SVD) for dimensionality reduction in order to extract contextual and usage-based representations of words from textual corpora. Amongst many other tasks, LSA has been used extensively for information retrieval, by using associations between documents and terms (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Dumais, 1991; Dumais, Furnas, Landauer, Deerwester, & Harshman, 1988). The exact factorization achieved via SVD has been shown to achieve solutions comparable in some ways to those obtained by modern neural network based techniques (Levy & Goldberg, 2014; Levy, Goldberg, & Dagan, 2015), commonly used to obtain dense word representations from textual corpora.

Correspondence analysis (CA) is a popular method for the analysis of contingency tables (Greenacre, 1984, 2017). More specifically, it provides a graphical display of dependence between rows and columns of a two-way contingency table (Greenacre & Hastie, 1987). Like LSA, CA is a dimensionality reduction method. The methods have much in common as both use a singular value decomposition (SVD). In both cases, after dimensionality reduction, many natural language processing or text mining tasks, such as text

clustering or document classification, amongst many others, may be performed in the reduced dimensional space rather than in the higher dimensional space provided by the raw document-term matrix.

While a few empirical comparisons of LSA and CA, with mixed results, can be found in the literature, a comprehensive theoretical comparison is lacking. For example, Morin (1999) compared the two methods in the automatic exploration of themes in texts. Séguéla and Saporta (2011) compared the performance of CA and LSA with several weighting functions in a document clustering task, and found that CA gave better results. On the other hand, Séguéla and Saporta (2013) compared the performance of CA and LSA with TF-IDF on a recommender system, but found that CA performs less well.

The present article presents a theoretical comparison of the two techniques, and places them in a unifying framework. We show that CA has some favourable properties over LSA, such as a clear interpretation of the distances between documents and between terms of the original matrix, and a clear relation to statistical independence of documents and terms. Second, we empirically evaluate and compare the two techniques, by applying them to two tasks in two languages. The first is an authorship attribution task in Dutch, where we evaluate the two techniques on a large set of historical Dutch texts written by six well-known Dutch authors of the 16th century. Here, we additionally use CA to determine the unknown authorship of *Wilhelmus*, the national anthem of the Netherlands, whose authorship is controversial: CA attributes *Wilhelmus* to the author Datheen, out of the six contemporary contenders. To the best of our knowledge, this is the first application of CA to the *Wilhelmus*. The second task is a document classification task in English, using the BBCSport dataset. In both cases, we find that CA performs better.

The rest of the article is organized as follows. In section 2, we introduce terminology and illustrate relevant properties of interest of a document-term matrix on a toy dataset. Section 3 and Section 4 elaborate on the techniques LSA and CA in turn and apply them to the toy dataset. A unifying framework is proposed in Section 5. Section 6 evaluates the performance of LSA and CA for authorship attribution of documents where the author is known, and of the *Wilhelmus*, whose author is unknown. In Section 7 a second study is described that concerns document classification of the BBCSport dataset. The article ends with a conclusion.

2 A Toy Dataset

In areas of computational linguistics, information retrieval, and text mining, a *document-term* matrix is commonly used to represent documents and terms. [1] Here we start with a discussion of its properties, using a toy dataset for illustration. We discuss the types of information available in such matrices so that, later on, when we introduce LSA and CA,

[1] A document-term matrix is similar to a word-context matrix, commonly used to represent word meanings, in the sense that it is also a matrix of counts. However, in the context of word-context matrices the ways in which the counts are transformed are different from the way they are transformed for document-term matrices, and therefore, due to space limitations, we defer a comparison of CA and LSA of word-context matrices to future work.

it is better appreciated what information is actually analysed by these methods. For LSA this discussion is valuable, as for LSA different solutions can be found depending on the way the original matrix is transformed before it is analysed. For CA there is only one type of analysis but in this analysis specific aspects of the matrix are ignored.

Suppose $\boldsymbol{F} = [f_{ij}]$ is a document-term matrix of size $m \times n$, in which rows correspond to documents, columns are associated with terms, and an element is the frequency of occurrence of a term in a particular document. The matrix $\boldsymbol{P} = [p_{ij}]$ is the matrix of joint observed proportions, where the element in the ith row and the jth column is given by $p_{ij} = f_{ij}/f_{++}$ and $\sum_i \sum_j p_{ij} = 1$. The marginal proportions are denoted by r_i and c_j for the ith row and jth column respectively, where $r_i = \sum_{j=1}^{n} p_{ij}$, $c_j = \sum_{i=1}^{m} p_{ij}$.

For row i, the vector of n conditional proportions p_{ij}/r_i is called a row profile. Similarly, for column j the vector of m conditional proportions p_{ij}/c_j is the column profile. The vector with elements c_j is called the *average row profile*, and similarly, the vector with elements r_i is the *average column profile*.

Table 1: A document-term matrix $\boldsymbol{F}$: size 6×6

	lion	tiger	cheetah	jaguar	porsche	ferrari
doc1	2	2	1	2	0	0
doc2	2	3	3	3	0	0
doc3	1	1	1	1	0	0
doc4	2	2	2	3	1	1
doc5	0	0	0	1	1	1
doc6	0	0	0	2	1	2

Consider the 6×6 document-term matrix $\boldsymbol{F}$ in Table 1 containing 6 documents and 6 terms, with the frequency of occurrence of terms in each document (Aggarwal, 2018). Based on term-frequencies in each document, the first three documents can be considered to primarily refer to *cats*, the last two primarily to *cars*, and the fourth document to both. The fourth term, *jaguar*, is polysemous because it can refer to either a cat or a car.

Table 2 shows the matrix $\boldsymbol{P}$ of joint observed proportions. Here, the average column profile is $\boldsymbol{r} = [0.171, ..., 0.122]^T$ and the average row profile is $\boldsymbol{c} = [0.171, ..., 0.098]^T$. Table 3 shows the matrix $\boldsymbol{E} = [r_i c_j]$ of expected proportions under independence. Under independence, all row profiles are identical as $e_{ij}/r_i = r_i c_j / r_i = c_j$, and similarly for the column

Table 2: The matrix $\boldsymbol{P}$ of joint observed proportions

	lion	tiger	cheetah	jaguar	porsche	ferrari	total
doc1	0.049	0.049	0.024	0.049	0.000	0.000	0.171
doc2	0.049	0.073	0.073	0.073	0.000	0.000	0.268
doc3	0.024	0.024	0.024	0.024	0.000	0.000	0.098
doc4	0.049	0.049	0.049	0.073	0.024	0.024	0.268
doc5	0.000	0.000	0.000	0.024	0.024	0.024	0.073
doc6	0.000	0.000	0.000	0.049	0.024	0.049	0.122
total	0.171	0.195	0.171	0.293	0.073	0.098	1.000

Table 3: The matrix $\boldsymbol{E}$ of expected proportions under independence

	lion	tiger	cheetah	jaguar	porsche	ferrari	total
doc1	0.029	0.033	0.029	0.050	0.012	0.017	0.171
doc2	0.046	0.052	0.046	0.079	0.020	0.026	0.268
doc3	0.017	0.019	0.017	0.029	0.007	0.010	0.098
doc4	0.046	0.052	0.046	0.079	0.020	0.026	0.268
doc5	0.012	0.014	0.012	0.021	0.005	0.007	0.073
doc6	0.021	0.024	0.021	0.036	0.009	0.012	0.122
total	0.171	0.195	0.171	0.293	0.073	0.098	1.000

profiles. Comparing the joint proportions in Table 2 with these expected proportions in Table 3 reveals how documents are related to terms. For instance, the joint proportions for document 1 and *lion* and document 1 and *tiger* are (0.049, 0.049) and are higher than their expected proportions (0.029, 0.033), which means that the terms *lion* and *tiger* appear more often than average in document 1. However, for document 1 and *porsche* and document 1 and *ferrari*, the joint proportions are 0.000 and 0.000 and are lower than their expected joint proportions (0.012,0.017), which indicates that the terms *porsche* and *ferrari* appear less often than average in document 1.

Table 4: Row profiles of $\boldsymbol{F}$

	lion	tiger	cheetah	jaguar	porsche	ferrari	total
doc1	0.286	0.286	0.143	0.286	0.000	0.000	1.000
doc2	0.182	0.273	0.273	0.273	0.000	0.000	1.000
doc3	0.250	0.250	0.250	0.250	0.000	0.000	1.000
doc4	0.182	0.182	0.182	0.273	0.091	0.091	1.000
doc5	0.000	0.000	0.000	0.333	0.333	0.333	1.000
doc6	0.000	0.000	0.000	0.400	0.200	0.400	1.000
average row profile	0.171	0.195	0.171	0.293	0.073	0.098	1.000

Table 4 shows the row profiles of the original matrix $\boldsymbol{F}$. Here, the *average row profile* shows which terms are more and which are less often used over *all documents*: *jaguar* is used the most and *porsche* is used the least over all the documents. Differences in row profiles between documents can be interpreted by comparing the elements of their row profiles with the average row profile. For example, document 1 has proportions (0.286, 0.286) for (*lion*, *tiger*) and these are higher than the averages (0.171, 0.195); on the other hand, document 5 has proportions (0.000, 0.000) for (*lion*, *tiger*) lower than the averages (0.171, 0.195). This shows that the terms *lion* and *tiger* appear more often than average in document 1 but less often than average in document 5. For a matrix of column profiles, a similar interpretation can be made.

After establishing this dataset, the various matrices of interest, and the terminology used, we next examine LSA and CA in the following two sections, in turn.

3 Latent Semantic Analysis

Latent Semantic Analysis (LSA) has been extensively used for improving information retrieval by using the associations between documents and terms (Deerwester et al., 1990; Dumais et al., 1988), amongst many other tasks. Since individual terms provide incomplete and unreliable evidence about the meaning of a document, in part due to synonymy and polysemy, individual terms are replaced with derived underlying (latent) semantic factors. Although LSA is a very well-known technique, we first present a detailed analysis of the mathematics involved in LSA here as this is usually not found in the literature, and in a later section, it will help in making the comparison between LSA and CA explicit. We start with LSA of the raw document-term matrix and then discuss LSA of weighted matrices, namely a matrix with row-normalised elements with L1, with L2, and a matrix that is transformed by TF-IDF. The discussion is illustrated using the toy data from §2, with the aim to present a clear view of the properties of the dataset captured by the LSA analysis.

3.1 LSA of Raw Document-Term Matrix

LSA is an application of the mathematical tool SVD, and can take many forms, depending on the matrix analysed. We start our discussion of LSA with the SVD of a raw document-term matrix (Berry, Dumais, & O'Brien, 1995; Deisenroth, Faisal, & Ong, 2020). SVD can be used to decompose $\boldsymbol{F}$ into a product of three matrices: $\boldsymbol{U}^f$, $\boldsymbol{\Sigma}^f$, and $\boldsymbol{V}^f$, namely

$$\boldsymbol{F} = \boldsymbol{U}^f \boldsymbol{\Sigma}^f (\boldsymbol{V}^f)^T \tag{1}$$

Here, assuming $\boldsymbol{F}$ has size $m \times n$ and $n > m$ and $\boldsymbol{F}$ has full rank, $\boldsymbol{U}^f$ is a $m \times m$ matrix with orthonormal columns called left singular vectors (that is, $(\boldsymbol{U}^f)^T \boldsymbol{U}^f = \boldsymbol{I}$), $\boldsymbol{V}^f$ is a $n \times m$ matrix with orthonormal columns called right singular vectors (that is, $(\boldsymbol{V}^f)^T \boldsymbol{V}^f = \boldsymbol{I}$), and $\boldsymbol{\Sigma}^f$ is a $m \times m$ diagonal matrix with singular values on the diagonal in descending order.

We denote the first k columns of $\boldsymbol{U}^f$ as the $m \times k$ matrix $\boldsymbol{U}^f_k$, the first k columns of $\boldsymbol{V}^f$ as the $n \times k$ matrix $\boldsymbol{V}^f_k$, and the k largest singular values on the diagonal of $\boldsymbol{\Sigma}^f$ as the $k \times k$ matrix $\boldsymbol{\Sigma}^f_k$ ($k \leq m$). Then $\boldsymbol{U}^f_k \boldsymbol{\Sigma}^f_k (\boldsymbol{V}^f_k)^T$ provides the optimal rank-k approximation of $\boldsymbol{F}$ in a least-squares sense. That is, $\boldsymbol{X} = \boldsymbol{U}^f_k \boldsymbol{\Sigma}^f_k (\boldsymbol{V}^f_k)^T$ minimizes the matrix in Equation (2) amongst all matrices $\boldsymbol{X}$ of rank k:

$$||\boldsymbol{F} - \boldsymbol{X}||^2 = \sum_i \sum_j (f_{ij} - x_{ij})^2 \tag{2}$$

The idea is that the matrix $\boldsymbol{U}^f_k \boldsymbol{\Sigma}^f_k (\boldsymbol{V}^f_k)^T$ captures the major associational structure in the matrix and throws out noise (Dumais, 1991; Dumais et al., 1988). The total sum of squared singular values is equal to $\text{tr}((\boldsymbol{\Sigma}^f)^2)$, where tr is the sum of elements on the main diagonal of a square matrix. The proportion of the total sum of squared singular values explained by the rank k approximation is $\text{tr}((\boldsymbol{\Sigma}^f_k)^2)/\text{tr}((\boldsymbol{\Sigma}^f)^2)$.

SVD can also be interpreted geometrically. As $\boldsymbol{F}$ is of size $m \times n$, each row of $\boldsymbol{F}$ can be represented as a point in an n-dimensional space with the row elements as coordinates, and each column can be represented as a point in an m-dimensional space with the column elements as coordinates. In a rank-k approximation, where $k < (m, n)$, each of the original m documents and n terms are approximated by only k coordinates. Thus SVD projects the sum of squared Euclidean distances from these row (column) points to the origin in the n (m)-dimensional space as much as possible to a lower, a k-dimensional space. The Euclidean distances between the rows of $\boldsymbol{F}$ are approximated by the Euclidean distances between the rows of $\boldsymbol{U}_k^f \boldsymbol{\Sigma}_k^f$ from below, and the Euclidean distances between the rows of $\boldsymbol{F}^T$ are approximated by the Euclidean distances between the rows of $\boldsymbol{V}_k^f \boldsymbol{\Sigma}_k^f$ from below.

The choice of k is crucial in many applications (Albright, 2004). A lower rank approximation cannot always express prominent relationships in text, whereas the higher rank approximation may add useless noise. How to choose k is an open issue (Deerwester et al., 1990). In practice, the value of k is selected such that a certain criterion is satisfied, for example, the proportion of explained total sum of squared singular values is at least a pre-specified proportion. Also, the use of a scree plot, showing the decline in subsequent squared singular values, can be considered.

As it turns out, the raw document-term matrix $\boldsymbol{F}$ in Table 1 does not have full rank; its rank is 5. The SVD of $\boldsymbol{F}$ in Table 1 is

$$
\begin{aligned}
\boldsymbol{F} &= \boldsymbol{U}^f \boldsymbol{\Sigma}^f (\boldsymbol{V}^f)^T \\
&= \begin{bmatrix}
-0.411 & 0.175 & 0.825 & 0.252 & -0.239 \\
-0.646 & 0.314 & -0.562 & 0.301 & -0.279 \\
-0.232 & 0.127 & 0.034 & -0.099 & 0.503 \\
-0.562 & -0.203 & 0.044 & -0.603 & 0.333 \\
-0.099 & -0.456 & -0.024 & -0.404 & -0.672 \\
-0.186 & -0.778 & -0.034 & 0.556 & 0.223
\end{bmatrix}
\begin{bmatrix}
8.425 & 0 & 0 & 0 & 0 \\
0 & 3.261 & 0 & 0 & 0 \\
0 & 0 & 0.988 & 0 & 0 \\
0 & 0 & 0 & 0.574 & 0 \\
0 & 0 & 0 & 0 & 0.272
\end{bmatrix} \\
&\quad \begin{bmatrix}
-0.412 & 0.214 & 0.655 & -0.344 & 0.486 \\
-0.488 & 0.311 & 0.087 & 0.180 & -0.540 \\
-0.440 & 0.257 & -0.748 & -0.259 & 0.339 \\
-0.611 & -0.369 & 0.039 & 0.366 & -0.148 \\
-0.101 & -0.441 & -0.014 & -0.783 & -0.426 \\
-0.123 & -0.679 & -0.048 & 0.186 & 0.392
\end{bmatrix}^T .
\end{aligned}
\tag{3}
$$

For the raw matrix, LSA-RAW in Table 5 shows the singular values, the squares of the singular values, and the proportions of explained total sum of squared singular values (denoted as PSSSV). Together, the first two dimensions account for $0.855 + 0.128 = 0.983$ of the total sum of squared singular values. Therefore, the documents and the terms can be approximated adequately in a two dimensional representation using $\boldsymbol{U}_2^f \boldsymbol{\Sigma}_2^f$ and $\boldsymbol{V}_2^f \boldsymbol{\Sigma}_2^f$ as coordinates. As the Euclidean distances between the documents and between the terms in the two-dimensional representation, i.e., between the rows of $\boldsymbol{U}_2^f \boldsymbol{\Sigma}_2^f$ and the rows of $\boldsymbol{V}_2^f \boldsymbol{\Sigma}_2^f$, approximate the Euclidean distances between rows and between columns of the original matrix $\boldsymbol{F}$, such a two dimensional representation simplifies the interpretation of the matrix considerably.

On the other hand, it is somewhat more difficult to examine the relation between a

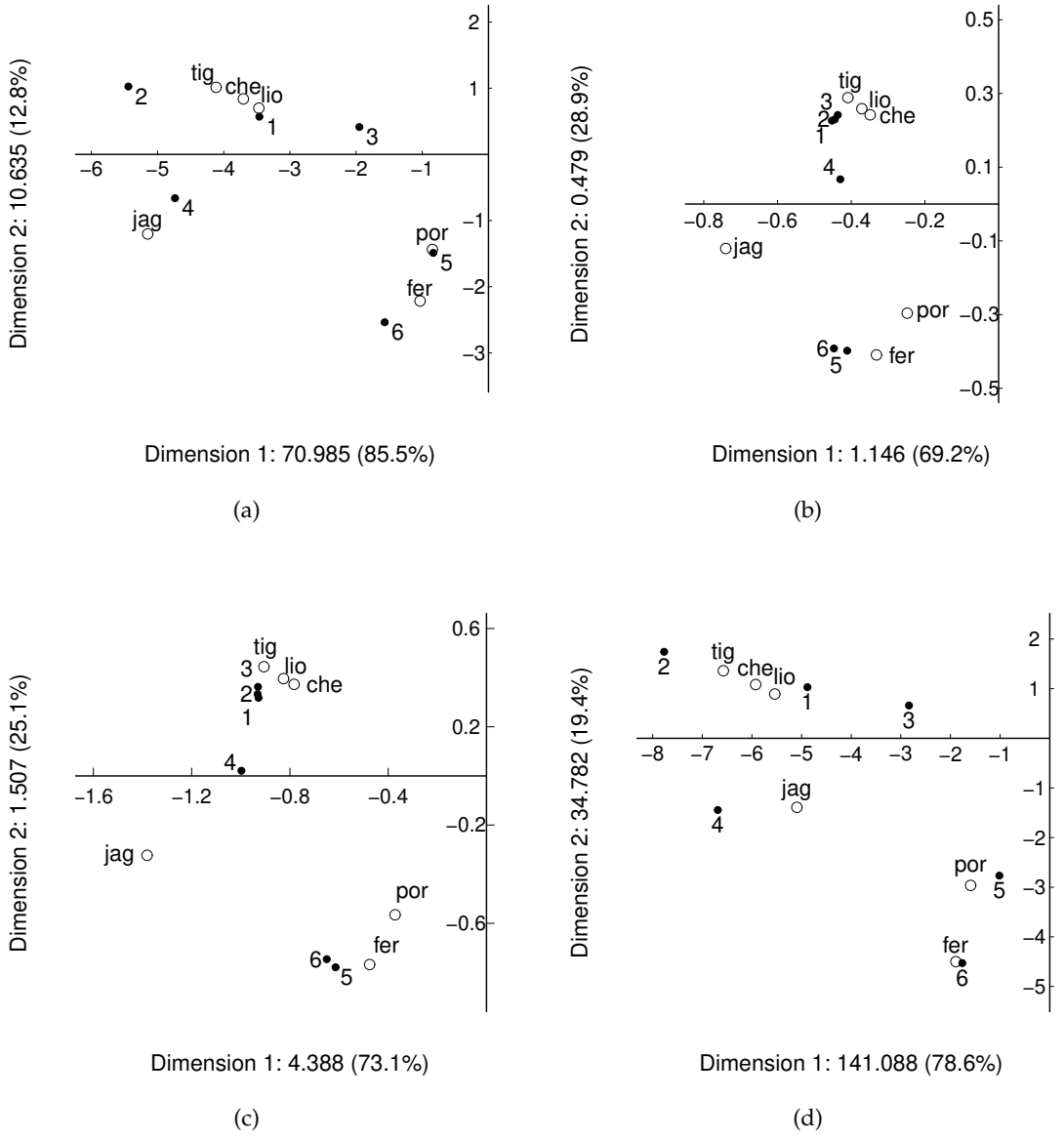


Figure 1: A two-dimensional plot of documents and terms (a) for raw matrix $\boldsymbol{F}$; (b) for row-normalized data $\boldsymbol{F}^{L1}$; (c) for row-normalized data $\boldsymbol{F}^{L2}$; (d) for matrix $\boldsymbol{F}^{\text{TF-IDF}}$.

Table 5: The singular values, the squares of singular values, and the proportion of explained total sum of squared singular values (PSSSV) for each dimension of LSA of $\boldsymbol{F}$, of $\boldsymbol{F}^{L1}$, of $\boldsymbol{F}^{L2}$, and of $\boldsymbol{F}^{\text{TF-IDF}}$.

methods	items	dim1	dim2	dim3	dim4	dim5
LSA-RAW	singular value	8.425	3.261	0.988	0.574	0.272
	square of singular value	70.985	10.635	0.976	0.330	0.074
	PSSSV	0.855	0.128	0.012	0.004	0.001
LSA-NROWL1	singular value	1.070	0.692	0.123	0.114	0.046
	square of singular value	1.146	0.479	0.015	0.013	0.002
	PSSSV	0.692	0.289	0.009	0.008	0.001
LSA-NROWL2	singular value	2.095	1.228	0.239	0.198	0.092
	square of singular value	4.388	1.507	0.057	0.039	0.009
	PSSSV	0.731	0.251	0.009	0.007	0.001
LSA-TFIDF	singular value	11.878	5.898	1.565	1.017	0.449
	square of singular value	141.088	34.782	2.451	1.034	0.202
	PSSSV	0.786	0.194	0.014	0.006	0.001

document and a term. The reason is that, by choosing a Euclidean distance-representation both for the documents and for terms, the singular values are used *twice* in the coordinates $\boldsymbol{U}_2^f\boldsymbol{\Sigma}_2^f$ and $\boldsymbol{V}_2^f\boldsymbol{\Sigma}_2^f$, and the inner product of coordinates of a document and coordinates of a term does not approximate the corresponding value in $\boldsymbol{F}$. Directions from the origin can be interpreted, though, as the double use of the singular values only leads to relatively reduced coordinates on the second dimension in comparison to the coordinates on the first dimension.

The two-dimensional representation of LSA-RAW is shown in Figure 1(a). From Figure 1(a), we see that documents 5 and 6 are close and therefore they appear to be similar. There is an order of 5, 6, 3, 1, 4, and 2 on the first dimension. This order is related to the row margins of Table 1, where 2 and 4 have the highest frequencies and therefore are further away from the origin. Overall the two-dimensional representation of the documents reveals a mix of the row margins and the profiles of terms used by the documents, namely, the row profiles of Table 1. This mix makes the graphic representation difficult to interpret. Similarly, *porsche* and *ferrari* are lower left but close to the origin, *tiger*, *cheetah*, and *lion* are upper left and further away from the origin, and *jaguar* is far away at the lower left. Also here there is a mix of the column margins and the column profiles of Table 1. The terms *porsche* and *ferrari* are related to documents 5 and 6 as they have the same position w.r.t. the origin, and similarly for *tiger*, *cheetah*, and *lion* to documents 1, 2, and 3, and *jaguar* to document 4.

Although the first dimension accounts for 85.5% of the total sum of squared singular values, it provides little information about the relations among documents and terms. In particular, from Table 1 we expect that documents 1 to 3 are similar, documents 5 and 6 are similar, and document 4 is in-between; term *jaguar* is between cat terms (*tiger*, *cheetah*, and *lion*) and car terms (*porsche* and *ferrari*), but we cannot see that from the first dimension. This is because the margins of Table 1 play a dominant role in the first dimension.

3.2 LSA of Weighted Document-Term Matrix

Weighting can be used to prevent differential lengths of documents from having differential effects on the representation, or be used to impose certain preconceptions of which terms are more important (Deerwester et al., 1990). The frequencies f_{ij} in the raw document-term matrix $\boldsymbol{F}$ can be transformed with the aim to provide a better approximation of the interrelations between documents and terms (Nakov, Popova, & Mateev, 2001). The weight w_{ij} for term j in document i is normally expressed as a product of three components (Ab Samat, Murad, Abdullah, & Atan, 2008; Kolda & O'leary, 1998; Salton & Buckley, 1988)

$$w_{ij} = L(i,j) \times G(j) \times N(i) \tag{4}$$

where the local weighting $L(i,j)$ is the weight of term j in document i, the global weighting $G(j)$ is the weight of the term j in the entire document set, and $N(i)$ is the normalization component for document i.

When $L(i,j) = f(i,j)$, $G(j) = 1$, and $N(i) = 1$, the weighted $\boldsymbol{F}$ is equal to $\boldsymbol{F}$. In matrix notation, Equation (4) can be expressed as $\boldsymbol{W} = \boldsymbol{NLG}$, where $\boldsymbol{N}$ is a diagonal matrix with diagonal elements $N(i)$ and $\boldsymbol{G}$ is a diagonal matrix with diagonal elements $G(j)$. Notice that pre- or post-multiplying by a diagonal matrix leaves the rank of the matrix $\boldsymbol{L}$ intact.

We examine two common ways to weight f_{ij}. One is row normalization (Ab Samat et al., 2008; Salton & Buckley, 1988) with L1 and L2. The other is term frequency-inverse document frequency (TF-IDF) (Dumais, 1991).

3.2.1 SVD of Matrix with Row-Normalized Elements with L1

In row-normalized weighting with L1, we use Equation (4) with $L(i,j) = f_{ij}$, $G(j) = 1$, and $N(i) = 1/\sum_{j=1}^{n} f_{ij}$, and apply an SVD to this transformed matrix that we denote as $\boldsymbol{F}^{L1}$, which consists of the row profiles of $\boldsymbol{F}$ shown in Table 4.

We perform LSA of Table 4 and find Table 5, part LSA-NROWL1. This shows that a rank 2 matrix approximates the data well as $0.692 + 0.289 = 0.981$ of the total sum of squared singular values is explained by these two dimensions. The first two columns of LSA of $\boldsymbol{F}^{L1}$ can be used to approximate $\boldsymbol{F}^{L1}$, see Equation (5).

$$\begin{aligned}
\boldsymbol{F}^{L1} &\approx \boldsymbol{U}_2^{L1}\boldsymbol{\Sigma}_2^{L1}(\boldsymbol{V}_2^{L1})^T \\
&= \begin{bmatrix} -0.423 & 0.327 \\ -0.415 & 0.332 \\ -0.408 & 0.349 \\ -0.401 & 0.097 \\ -0.384 & -0.575 \\ -0.417 & -0.567 \end{bmatrix} \begin{bmatrix} 1.070 & 0 \\ 0 & 0.692 \end{bmatrix} \begin{bmatrix} -0.347 & 0.374 \\ -0.382 & 0.417 \\ -0.326 & 0.350 \\ -0.692 & -0.174 \\ -0.232 & -0.428 \\ -0.310 & -0.592 \end{bmatrix}^T
\end{aligned} \tag{5}$$

Documents and terms can be projected on a two dimensional space using $\boldsymbol{U}_2^{L1}\boldsymbol{\Sigma}_2^{L1}$ and $\boldsymbol{V}_2^{L1}\boldsymbol{\Sigma}_2^{L1}$ as coordinates, see Figure 1(b). In this representation documents 1, 2, and 3 are

quite close, and so are 5 and 6. Also, the terms *ferrari* and *porsche* are close and related to 5 and 6, *tiger*, *lion*, and *cheetah* are close and related to 1, 2, and 3.

Although the first dimension accounts for 69.2% of the total sum of squared singular values, this dimension does not provide information about different use of terms by the documents as all documents have a similar coordinate. This is caused by the same marginal value 1 for each of the documents in Table 4, which leads to almost the same distance from the origin. Also, we would expect *jaguar* to be in between cat terms (*tiger*, *cheetah*, and *lion*) and car terms (*porsche* and *ferrari*), but on the first dimension it appears as a separate, third group. This is caused by the high values in its column in Table 4, which lead to a larger distance from the origin.

3.2.2 SVD of Matrix with Row-Normalized Elements with L2

In row-normalized weighting with L2, we use Equation (4) with $L(i,j) = f_{ij}$, $G(j) = 1$, and $N(i) = 1/\sqrt{\sum_{j=1}^{n} f_{ij}^2}$. The transformed matrix, denoted as $\boldsymbol{F}^{L2}$, is shown in Table (6). We then perform LSA on (6). Table 5, part LSA-NROWL2, indicates that a rank 2 matrix approximates the data well, as the sum of the PSSSV of the first two dimensions $0.731 + 0.251 = 0.982$ contributes to 98.2% of the total sum of squared singular values. The first two columns of LSA of $\boldsymbol{F}^{L2}$ can be used to approximate $\boldsymbol{F}^{L2}$, see Equation (6).

$$
\begin{aligned}
\boldsymbol{F}^{L2} &\approx \boldsymbol{U}_2^{L2}\boldsymbol{\Sigma}_2^{L2}(\boldsymbol{V}_2^{L2})^T \\
&= \begin{bmatrix} -0.443 & 0.259 \\ -0.445 & 0.271 \\ -0.444 & 0.295 \\ -0.476 & 0.017 \\ -0.293 & -0.635 \\ -0.310 & -0.608 \end{bmatrix} \begin{bmatrix} 2.095 & 0 \\ 0 & 1.228 \end{bmatrix} \begin{bmatrix} -0.394 & 0.323 \\ -0.432 & 0.362 \\ -0.374 & 0.304 \\ -0.659 & -0.263 \\ -0.178 & -0.460 \\ -0.227 & -0.625 \end{bmatrix}^T
\end{aligned} \tag{6}
$$

Table 6: A row-normalised document-term matrix $\boldsymbol{F}^{L2}$

	lion	tiger	cheetah	jaguar	porsche	ferrari
doc1	0.555	0.555	0.277	0.555	0.000	0.000
doc2	0.359	0.539	0.539	0.539	0.000	0.000
doc3	0.500	0.500	0.500	0.500	0.000	0.000
doc4	0.417	0.417	0.417	0.626	0.209	0.209
doc5	0.000	0.000	0.000	0.577	0.577	0.577
doc6	0.000	0.000	0.000	0.667	0.333	0.667

Documents and terms can be projected on a two dimensional space using $\boldsymbol{U}_2^{L2}\boldsymbol{\Sigma}_2^{L2}$ and $\boldsymbol{V}_2^{L2}\boldsymbol{\Sigma}_2^{L2}$ as coordinates, see Figure 1(c). In this representation documents 1, 2, and 3 are quite close, and so are 5 and 6. Also, the terms *ferrari* and *porsche* are close and related to 5 and 6, *tiger*, *lion*, and *cheetah* are close and related to 1, 2, and 3.

Although the first dimension accounts for 73.1% of the total sum of squared singular values, and so, a major portion of the information in the matrix, we do not find the impor-

tant aspect in the data that document 4 should be in between documents 1-3 on the one hand and documents 5-6 on the other hand on this dimension. This is caused by the high values in the row for doc4 in Table 6, which lead to a larger distance from the origin than the other documents have. Also, we would expect *jaguar* to be in between cat terms (*tiger*, *cheetah*, and *lion*) and car terms (*porsche* and *ferrari*), but on the first dimension it appears as a separate, third group. This is caused by the high values in its column in Table 6, which lead to a larger distance from the origin.

3.2.3 SVD of the Term Frequency-Inverse Document Frequency Matrix

Table 7: A document-term matrix $\boldsymbol{F}^{\text{TF-IDF}}$

	lion	tiger	cheetah	jaguar	porsche	ferrari
doc1	3.170	3.170	1.585	2	0	0
doc2	3.170	4.755	4.755	3	0	0
doc3	1.585	1.585	1.585	1	0	0
doc4	3.170	3.170	3.170	3	2	2
doc5	0.000	0.000	0.000	1	2	2
doc6	0.000	0.000	0.000	2	2	4

TF-IDF is one commonly used vector space representation of text data. We use Equation (4) with $L(i,j) = f_{ij}$, $G(j) = 1 + \log(\frac{n\text{docs}}{df_j})$, and $N(i) = 1$, one form of TF-IDF, where $n$docs is the number of documents in the set and df_j is the number of documents where term j appears, and then apply an SVD to this transformed matrix that we denote as $\boldsymbol{F}^{\text{TF-IDF}}$, see Table 7. As is common in the literature, here we choose 2 as the base of the logarithmic function.

We perform LSA of Table 7 and find Table 5, part LSA-TFIDF. This shows that a rank 2 matrix approximates the data well as $0.786 + 0.194 = 0.980$ of the total sum of squared singular values is explained by these two dimensions. The matrix $\boldsymbol{F}^{\text{TF-IDF}}$ in Table 7 is approximated in the first two dimensions as follows:

$$\begin{aligned}\boldsymbol{F}^{\text{TF-IDF}} &\approx \boldsymbol{U}_2^{\text{TF-IDF}}\boldsymbol{\Sigma}_2^{\text{TF-IDF}}(\boldsymbol{V}_2^{\text{TF-IDF}})^T \\ &= \begin{bmatrix} -0.411 & 0.175 \\ -0.654 & 0.296 \\ -0.239 & 0.112 \\ -0.563 & -0.245 \\ -0.086 & -0.469 \\ -0.148 & -0.768 \end{bmatrix} \begin{bmatrix} 11.878 & 0 \\ 0 & 5.898 \end{bmatrix} \begin{bmatrix} -0.466 & 0.151 \\ -0.554 & 0.231 \\ -0.499 & 0.184 \\ -0.429 & -0.236 \\ -0.134 & -0.502 \\ -0.159 & -0.763 \end{bmatrix}^T \end{aligned} \tag{7}$$

Figure 1(d) is a two-dimensional plot of the documents and terms using $\boldsymbol{U}_2^{\text{TF-IDF}}\boldsymbol{\Sigma}_2^{\text{TF-IDF}}$ and $\boldsymbol{V}_2^{\text{TF-IDF}}\boldsymbol{\Sigma}_2^{\text{TF-IDF}}$ as coordinates for the 6×6 sample document-term matrix $\boldsymbol{F}^{\text{TF-IDF}}$. The configuration of documents in Figure 1(d) is very similar to that in Figure 1(a). The configuration of terms in Figure 1(d) is different from that of terms in Figure 1(a). In Figure 1(d), there is an order of *porsche*, *ferrari*, *jaguar*, *lion*, *cheetah*, and *tiger* on the first dimension,

whereas in Figure 1(a), there is an order of *porsche*, *ferrari*, *lion*, *cheetah*, *tiger*, and *jaguar* on the first dimension. Compared with Figure 1(a), the first dimension of Figure 1(d) shows that *jaguar* is in between cat terms (*tiger*, *cheetah*, and *lion*) and car terms (*porsche* and *ferrari*).

3.3 Conclusions regarding LSA of Different Matrices

The relationships among the documents and terms in the raw document-term matrix can be blurred due to differences in margins arising from differing document-lengths and term-frequencies. LSA of the raw matrix leads to a mix of relationships among documents and terms, and margins. In order to provide a better approximation of the interrelations between documents and terms, the weighting schemes can be used.

We conclude that normalizations of the documents have a beneficial effect. Yet, the properties of the frequencies that are evident from Table 1 where we expect, for example, that *jaguar* lies in between *porsche* and *ferrari* on the one hand and *tiger*, *cheetah*, and *lion* on the other hand, are not fully represented on the first dimension. This is due to the fact that the column margins of Tables 4 and 6 still play a role on the first dimension. The TF-IDF matrix also has a positive effect. LSA is not successful, for example, in representing the expected relationships between documents on the first dimension that documents 1 to 3 are similar, 5 and 6 are similar, and document 4 is in-between. This is due to the fact that the row margins of Table 7 still play a role on the first dimension. We can try to repair this aspect as well, by applying a transformation of the rows and columns of Table 1 simultaneously. However, the transformations appear ad hoc. Instead we present in the next section a different technique, which better fits the properties of the data: CA.

4 Correspondence Analysis

CA provides a low-dimensional representation of the interaction or dependence between the rows and columns of the contingency table (Greenacre & Hastie, 1987), which can be used to reveal the structure in the data (Hayashi, 1992). CA has been proposed multiple times, apparently independently, emphasizing different properties of the technique. Some important contributions are provided in the Japanese literature, by Hayashi (1956, 1992), who emphasizes the property of CA that it maximizes the correlation coefficient between the row and column variable by assigning numerical scores to these variables; in the French literature, by Benzécri (1973), who emphasizes a distance interpretation, where Greenacre (1984) expressed Benzécri's work in a more conveniential mathematical notation; and in the Dutch literature, by Gifi (1990); Michailidis and De Leeuw (1998), who emphasize optimal scaling properties. We present CA here mainly from the French perspective.

The aim of CA as developed by Benzécri is to find a representation of the rows of frequency matrix $\boldsymbol{F}$ in such a way that Euclidean distances between the rows in the representation correspond to so-called χ^2-distances between rows of $\boldsymbol{F}$ (Gifi, 1990). As in Section 2, we work with $\boldsymbol{P}$ with elements $p_{ij} = f_{ij}/f_{++}$. In the χ^2-distance profiles play an

important role. The squared χ^2-distance between the kth row profile with elements p_{kj}/r_k and the lth row profile with elements p_{lj}/r_l is

$$\delta_{kl}^2 = \sum_j \frac{(p_{kj}/r_k - p_{lj}/r_l)^2}{c_j} \tag{8}$$

Thus the difference between the jth elements of the two profiles is weighted by column margin of Table 2, c_j, so that this difference plays a relatively more important role in the χ^2-distance if it stems from a column having a small value c_j.

A representation where Euclidean distances between the rows of the matrix are equal to χ^2-distances is found as follows. In matrix notation, the matrix whose Euclidean distances between the rows are equal to χ^2-distances between rows of $\boldsymbol{F}$ is equal to $\boldsymbol{D}_r^{-1}\boldsymbol{P}\boldsymbol{D}_c^{-\frac{1}{2}}$, where $\boldsymbol{D}_r$ is a diagonal matrix with r_i as diagonal elements and $\boldsymbol{D}_c$ is a diagonal matrix with c_j as diagonal elements. Suppose we take the SVD of

$$\boldsymbol{D}_r^{-\frac{1}{2}}\boldsymbol{P}\boldsymbol{D}_c^{-\frac{1}{2}} = \boldsymbol{U}^{sp}\boldsymbol{\Sigma}^{sp}(\boldsymbol{V}^{sp})^T \tag{9}$$

Here $\boldsymbol{D}_r^{-\frac{1}{2}}\boldsymbol{P}\boldsymbol{D}_c^{-\frac{1}{2}}$ is a matrix with standardized proportions, hence the superscripts sp on the right hand side of the equation. Then, if we pre-multiply both sides of the Equation (9) with $\boldsymbol{D}_r^{-\frac{1}{2}}$, we get:

$$\boldsymbol{D}_r^{-1}\boldsymbol{P}\boldsymbol{D}_c^{-\frac{1}{2}} = \boldsymbol{D}_r^{-\frac{1}{2}}\boldsymbol{U}^{sp}\boldsymbol{\Sigma}^{sp}(\boldsymbol{V}^{sp})^T \tag{10}$$

Thus a representation using the rows of $\boldsymbol{D}_r^{-\frac{1}{2}}\boldsymbol{U}^{sp}\boldsymbol{\Sigma}^{sp}$ as row coordinates leads to Euclidean distances between these row points being equal to χ^2-distances between rows of $\boldsymbol{F}$. Similar to Equation (8) we can also define χ^2-distances between the columns of $\boldsymbol{F}$, and in matrix notation this leads to the matrix $\boldsymbol{D}_r^{-\frac{1}{2}}\boldsymbol{P}\boldsymbol{D}_c^{-1}$. Then, in a similar way as for the χ^2-distances for the rows, Equation (9) can be used as an intermediate step to go to a solution for the columns. Post-multiplying the left and right hand sides in Equation (9) by $\boldsymbol{D}_c^{-\frac{1}{2}}$ provides us with the coordinates for a representation where Euclidean distances between the column points (the rows of $\boldsymbol{D}_c^{-\frac{1}{2}}\boldsymbol{V}^{sp}\boldsymbol{\Sigma}^{sp}$ as coordinates for these columns) are equal to χ^2-distances between the columns of $\boldsymbol{F}$. Notice that Equation (9) plays the dual role of an intermediate step in going to a solution both for the rows and the columns.

The matrices $\boldsymbol{D}_r^{-\frac{1}{2}}\boldsymbol{U}^{sp}\boldsymbol{\Sigma}^{sp}$ and $\boldsymbol{D}_c^{-\frac{1}{2}}\boldsymbol{V}^{sp}\boldsymbol{\Sigma}^{sp}$ have a first column being equal to 1, a so-called artificial dimension. This artificial dimension reflects the fact that the row margins of the matrix $\boldsymbol{D}_r^{-1}\boldsymbol{P}$ with the row profiles of Table 1 are 1 and the column margins of the matrix $\boldsymbol{P}\boldsymbol{D}_c^{-1}$ with the column profiles of Table 1 are 1. This artificial dimension is eliminated by not taking the SVD of $\boldsymbol{D}_r^{-\frac{1}{2}}\boldsymbol{P}\boldsymbol{D}_c^{-\frac{1}{2}}$ but of $\boldsymbol{D}_r^{-\frac{1}{2}}(\boldsymbol{P}-\boldsymbol{E})\boldsymbol{D}_c^{-\frac{1}{2}}$, where the elements of $\boldsymbol{E}$ are defined in Table 3 as the product of the margins r_i and c_j. Due to subtracting $\boldsymbol{E}$ from $\boldsymbol{P}$, the rank of $\boldsymbol{D}_r^{-\frac{1}{2}}(\boldsymbol{P}-\boldsymbol{E})\boldsymbol{D}_c^{-\frac{1}{2}}$ is $m-1$, which is 1 less than the rank of $\boldsymbol{F}$. Notice that the elements of $\boldsymbol{D}_r^{-\frac{1}{2}}(\boldsymbol{P}-\boldsymbol{E})\boldsymbol{D}_c^{-\frac{1}{2}}$ are standardized residuals under the independence model, and the sum of squares of these elements yields the so-called total inertia, which is equal to the Pearson χ^2 statistic divided by sample size f_{++}. By taking the SVD of the

matrix of standardized residuals, we get

$$\boldsymbol{D}_r^{-\frac{1}{2}}(\boldsymbol{P}-\boldsymbol{E})\boldsymbol{D}_c^{-\frac{1}{2}} = \boldsymbol{U}^{sr}\boldsymbol{\Sigma}^{sr}(\boldsymbol{V}^{sr})^T \tag{11}$$

and

$$\boldsymbol{D}_r^{-1}(\boldsymbol{P}-\boldsymbol{E})\boldsymbol{D}_c^{-1} = \boldsymbol{\Phi}^{sr}\boldsymbol{\Sigma}^{sr}(\boldsymbol{\Gamma}^{sr})^T \tag{12}$$

where $\boldsymbol{\Phi}^{sr} = \boldsymbol{D}_r^{-\frac{1}{2}}\boldsymbol{U}^{sr}$ and $\boldsymbol{\Gamma}^{sr} = \boldsymbol{D}_c^{-\frac{1}{2}}\boldsymbol{V}^{sr}$. We use the abbreviation sr for the matrices on the right hand side of Equation (11) to refer to the matrix of standardized residuals on the left hand side of the equation. CA simultaneously provides a geometric representation of row profiles and column profiles of Table 1, where the effects of row margins and column margins of Table 1 are eliminated. $\boldsymbol{\Phi}^{sr}$ and $\boldsymbol{\Gamma}^{sr}$ are called standard coordinates of rows and columns respectively. They have the property that their weighted average is 0 and weighted sum of squares is 1:

$$\mathbf{1}^T\boldsymbol{D}_r\boldsymbol{\Phi}^{sr} = \mathbf{0}^T = \mathbf{1}^T\boldsymbol{D}_c\boldsymbol{\Gamma}^{sr} \tag{13}$$

and

$$(\boldsymbol{\Phi}^{sr})^T\boldsymbol{D}_r\boldsymbol{\Phi}^{sr} = \mathbf{I} = (\boldsymbol{\Gamma}^{sr})^T\boldsymbol{D}_c\boldsymbol{\Gamma}^{sr} \tag{14}$$

Equation (13) reflects the fact that the row and column margins of $\boldsymbol{P}-\boldsymbol{E}$ vanish (Van der Heijden, De Falguerolles, & De Leeuw, 1989).

We can make graphic displays using $\boldsymbol{\Phi}_k^{sr}\boldsymbol{\Sigma}_k^{sr}$ and $\boldsymbol{\Gamma}_k^{sr}\boldsymbol{\Sigma}_k^{sr}$ as coordinates, which has the advantage that Euclidean distances between the points approximate χ^2-distances both for the rows of $\boldsymbol{F}$ and for the columns of $\boldsymbol{F}$, but it has the drawback that $\boldsymbol{\Sigma}_k^{sr}$ is used twice. We can also make graphic displays using $\boldsymbol{\Phi}_k^{sr}\boldsymbol{\Sigma}_k^{sr}$ and $\boldsymbol{\Gamma}_k^{sr}$, or $\boldsymbol{\Phi}_k^{sr}$ and $\boldsymbol{\Gamma}_k^{sr}\boldsymbol{\Sigma}_k^{sr}$. Thus, from Equation (12), this has the advantage that the inner product of the coordinates of a document and the coordinates of a term approximates the corresponding value in $\boldsymbol{D}_r^{-1}(\boldsymbol{P}-\boldsymbol{E})\boldsymbol{D}_c^{-1}$.

If we choose $\boldsymbol{\Phi}^{sr}\boldsymbol{\Sigma}^{sr}$ for the row points and $\boldsymbol{\Gamma}^{sr}$ for the column points, then CA has the property that the row points are in weighted average of the column points, where the weights are the row profile values. Actually, $\boldsymbol{\Gamma}^{sr}$ can be seen as coordinates for the extreme row profiles projected onto the subspace. The extreme row profiles are totally concentrated into one of the terms. For example, $[0,0,1,0,0,0]$ represents the row profile of a document that is totally concentrated into *cheetah*. At the same time, if we choose $\boldsymbol{\Phi}^{sr}$ for the row points and $\boldsymbol{\Gamma}^{sr}\boldsymbol{\Sigma}^{sr}$ for the column points, column points are in weighted average of row points, where the weights are the column profile values. In a similar way as for the rows, $\boldsymbol{\Phi}^{sr}$ provide coordinates for the extreme column profiles projected onto the subspace. The relationship between these row points and column points can be shown by rewriting Equation (11) and using Equation (13) as

$$\boldsymbol{D}_r^{-1}\boldsymbol{P}\boldsymbol{\Gamma}^{sr} = \boldsymbol{\Phi}^{sr}\boldsymbol{\Sigma}^{sr} \tag{15}$$

and

$$\boldsymbol{D}_c^{-1}\boldsymbol{P}^T\boldsymbol{\Phi}^{sr} = \boldsymbol{\Gamma}^{sr}\boldsymbol{\Sigma}^{sr} \tag{16}$$

These equations are called the transition formulas. In fact, this is one of the ways in which the solution of CA can be obtained: starting from arbitrary values for the columns, one first centers and standardizes the column coordinates so that the weighted sum is 0 and the weighted sums of squares is 1, next places the rows in the weighted average of the columns, then places the columns in the weighted average of the rows, and so on, until convergence. This is known as reciprocal averaging (Hill, 1973, 1974).

The origin in the graphic representation for the rows stands for the average row profile, which can be seen as follows. Let $\boldsymbol{D}_r^{-1}\boldsymbol{P}\boldsymbol{D}_c^{-\frac{1}{2}}$ be the matrix where Euclidean distances between the rows are χ^2-distances between rows of $\boldsymbol{F}$. Assume we plot the rows of this matrix using the n elements of each row as coordinates. Then, eliminating the artificial dimension in $\boldsymbol{D}_r^{-1}\boldsymbol{P}\boldsymbol{D}_c^{-\frac{1}{2}}$ leads to the subtraction of the average row profile from each row, as $\boldsymbol{D}_r^{-1}\boldsymbol{E}$ is a matrix with the average row profile in each row. In other words, the cloud of row points is translated to the origin, with the average row profile being exactly in the origin (compare Equation (13): $\mathbf{0}^T = \mathbf{1}^T\boldsymbol{D}_c\boldsymbol{\Gamma}^{sr}$). When two row points are departing in the same way from the origin, they depart in the same way from the average profile, and when two row points are on opposite sides of the origin, they depart in opposite ways from the average profile. If the documents and terms are statistically independent, then $p_{ij}/r_i = c_j$, and all document profiles would lie in the origin. Thus comparing row profiles with the origin is a way to study the departure from independence and to study the relations between documents and terms (see Section 2). Similarly, the origin in the graphic representation for the columns stands for average column profile.

We now analyse the example discussed in the LSA section. Table 8 shows the matrix $\boldsymbol{D}_r^{-\frac{1}{2}}(\boldsymbol{P}-\boldsymbol{E})\boldsymbol{D}_c^{-\frac{1}{2}}$ of standardized residuals (in lower-case notation, the elements of the matrix are $(p_{ij}-e_{ij})/\sqrt{e_{ij}}$).

Table 8: The matrix $\boldsymbol{D}_r^{-\frac{1}{2}}(\boldsymbol{P}-\boldsymbol{E})\boldsymbol{D}_c^{-\frac{1}{2}}$ of standardized residuals

	lion	tiger	cheetah	jaguar	porsche	ferrari
doc1	0.115	0.085	-0.028	-0.005	-0.112	-0.129
doc2	0.014	0.091	0.128	-0.019	-0.140	-0.162
doc3	0.060	0.039	0.060	-0.025	-0.084	-0.098
doc4	0.014	-0.016	0.014	-0.019	0.034	-0.011
doc5	-0.112	-0.119	-0.112	0.020	0.260	0.204
doc6	-0.144	-0.154	-0.144	0.069	0.164	0.338

We perform an SVD of $\boldsymbol{D}_r^{-\frac{1}{2}}(\boldsymbol{P}-\boldsymbol{E})\boldsymbol{D}_c^{-\frac{1}{2}}$ in Table 8 and find Table 9. Due to subtracting $\boldsymbol{E}$ from $\boldsymbol{P}$, the rank of the matrix in Table 8 is 4, which is 1 less than that in Table 1. The proportion of the total inertia explained by only the first dimension accounts for 0.932 of the total inertia. The matrix $\boldsymbol{D}_r^{-\frac{1}{2}}(\boldsymbol{P}-\boldsymbol{E})\boldsymbol{D}_c^{-\frac{1}{2}}$ in Table 8 is approximated in the first two

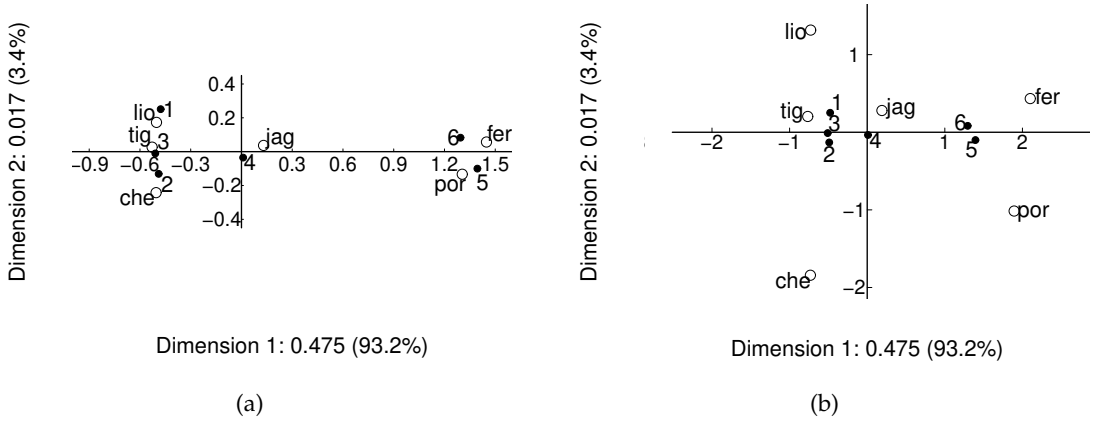
Table 9: The singular values, the inertia, and the proportions of explained total inertia for each dimension of CA.

	dim1	dim2	dim3	dim4
singular value	0.689	0.131	0.124	0.044
inertia	0.475	0.017	0.015	0.002
the proportion of inertia	0.932	0.034	0.030	0.004

dimensions as follows:

$$\begin{aligned}\boldsymbol{D}_r^{-\frac{1}{2}}(\boldsymbol{P}-\boldsymbol{E})\boldsymbol{D}_c^{-\frac{1}{2}} &\approx \boldsymbol{U}_2^{sr}\boldsymbol{\Sigma}_2^{sr}(\boldsymbol{V}_2^{sr})^T \\ &= \begin{bmatrix} -0.286 & 0.789 \\ -0.368 & -0.517 \\ -0.231 & -0.025 \\ 0.007 & -0.138 \\ 0.547 & -0.206 \\ 0.656 & 0.220 \end{bmatrix} \begin{bmatrix} 0.689 & 0 \\ 0 & 0.131 \end{bmatrix} \begin{bmatrix} -0.301 & 0.544 \\ -0.338 & 0.090 \\ -0.303 & -0.761 \\ 0.102 & 0.152 \\ 0.512 & -0.275 \\ 0.656 & 0.136 \end{bmatrix}^T \end{aligned} \quad (17)$$

Figure 2(a) is the map with a symmetric role for the rows and the columns, having $\boldsymbol{\Phi}_2^{sr}\boldsymbol{\Sigma}_2^{sr}$ and $\boldsymbol{\Gamma}_2^{sr}\boldsymbol{\Sigma}_2^{sr}$ as coordinates. The larger the deviations from document (term) points to the origin are, the larger the dependence between documents and terms. Looking only at the first dimension and document profiles' positions, we can see that the groups furthest apart are documents 1-3 on the left-hand side, opposed to documents 5-6 on the right-hand side. They differ in opposite ways from the average row profile that lies in the origin. Document 4 lies between documents 1-3 and documents 5-6. For the term points on the first dimension, the cat terms (*tiger*, *cheetah*, and *lion*) lie on the left, and car terms (*porsche* and *ferrari*) on the right. They differ in opposite ways from the average column profile. Importantly, notice that the term *jaguar* lies between cat terms and car terms, unlike all four of the LSA based analyses presented in Figure 1.



(a) (b)

Figure 2: The data of Table 1 using CA for (a) symmetric map; (b) asymmetric map.

Figure 2(b) is the asymmetric map with documents in the weighted average of the terms ($\boldsymbol{\Phi}_2^{sr}\boldsymbol{\Sigma}_2^{sr}$ and $\boldsymbol{\Gamma}_2^{sr}$ as coordinates, notice that the position of the documents is identical as in Figure 2(a)). From this graphic display we can study the position of the documents as they are in the weighted average of the terms, using the row profile elements as weights. For example, document 1 is closer to *lion* and *tiger* than to *porsche* and *ferrari*, because it has higher profile values than average values on terms *lion* and *tiger* (both 0.286 in comparison with the average profile values 0.171 and 0.195) and lower profile values on the terms *porsche* and *ferrari* (both 0.000 in comparison to 0.073 and 0.098), see Table 4. Thus document 1 is pulled into the direction of *lion* and *tiger*.

4.1 Conclusions regarding CA

In CA, an SVD is applied to the matrix $\boldsymbol{D}_r^{-\frac{1}{2}}(\boldsymbol{P}-\boldsymbol{E})\boldsymbol{D}_c^{-\frac{1}{2}}$ of standardized residuals. Due to $\boldsymbol{E}$, in CA the effect of the margins is eliminated—a solution only displays the relationships among documents and terms. In CA all points are scattered around the origin and the origin represents the profile of the row and column margins of $\boldsymbol{F}$.

In comparison, LSA tries to capture the relationships among documents and terms, which is not easy. The reason is that these relations are blurred by the effect of the margins that are also displayed in the LSA solution. CA does not have this property. Therefore it appears that CA is a better tool for computational linguistics, information retrieval, natural language processing, and text mining.

5 A Unifying Framework

Here we present a unifying framework that integrates LSA and CA. This section also serves the purpose of showing their similarities and their differences.

To first summarize LSA (see section 3.2 for details), a matrix is weighted, and the weighted matrix is decomposed. Assume we start off with the document-term matrix $\boldsymbol{F}$, the row weights of $\boldsymbol{F}$ are collected in the diagonal matrix $\boldsymbol{N}$, the column weights in the diagonal matrix $\boldsymbol{G}$, and there may be local weighting of the elements f_{ij} of $\boldsymbol{F}$ leading to a locally weighted matrix $\boldsymbol{L}$. Thus the weighted matrix $\boldsymbol{W}$ can be written as the matrix product

$$\boldsymbol{W} = \boldsymbol{NLG}. \tag{18}$$

Subsequently, in LSA the matrix $\boldsymbol{W}$ is decomposed using SVD into a product of three matrices: the orthonormal matrix $\boldsymbol{U}$, the diagonal matrix $\boldsymbol{\Sigma}$ with singular values in descending order, and the orthonormal matrix $\boldsymbol{V}$, namely

$$\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{19}$$

with

$$\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I} = \boldsymbol{V}^T\boldsymbol{V}. \tag{20}$$

Graphic representations are usually made using as coordinates $\boldsymbol{U\Sigma}$ for the rows and $\boldsymbol{V\Sigma}$ for the columns.

In contrast, in CA (see section 4 for details) we take the SVD of the matrix of standardized residuals. Let $\boldsymbol{P}$ be the matrix with proportions $p_{ij} = f_{ij}/f_{++}$, where f_{++} is the sum of all elements of $\boldsymbol{F}$; let $\boldsymbol{E}$ be the matrix with expected proportions under independence $e_{ij} = r_i c_j$, where r_i and c_j are the row and column sums of $\boldsymbol{P}$ respectively; let $\boldsymbol{D_r}$ and $\boldsymbol{D_c}$ be diagonal matrices with row and column sums r_i and c_j respectively. Thus the matrix of standardized residuals is $\boldsymbol{D_r}^{-\frac{1}{2}}(\boldsymbol{P}-\boldsymbol{E})\boldsymbol{D_c}^{-\frac{1}{2}}$. If we take the SVD of this matrix we get (11),

$$\boldsymbol{D_r}^{-\frac{1}{2}}(\boldsymbol{P}-\boldsymbol{E})\boldsymbol{D_c}^{-\frac{1}{2}} = \boldsymbol{U\Sigma V}^T. \tag{21}$$

In CA the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are further adjusted by

$$\boldsymbol{\Phi} = \boldsymbol{D_r}^{-\frac{1}{2}}\boldsymbol{U}, \boldsymbol{\Gamma} = \boldsymbol{D_c}^{-\frac{1}{2}}\boldsymbol{V} \tag{22}$$

so that we can write

$$\boldsymbol{D_r}^{-1}(\boldsymbol{P}-\boldsymbol{E})\boldsymbol{D_c}^{-1} = \boldsymbol{\Phi\Sigma\Gamma}^T \tag{23}$$

with

$$\boldsymbol{\Phi}^T\boldsymbol{D_r}\boldsymbol{\Phi} = \boldsymbol{I} = \boldsymbol{\Gamma}^T\boldsymbol{D_c}\boldsymbol{\Gamma}. \tag{24}$$

Graphic representations are usually made using $\boldsymbol{\Phi\Sigma}$ and $\boldsymbol{\Gamma\Sigma}$ for as coordinates for the rows and columns respectively.

This brings us to the point where we can formulate a unifying framework. We distinguish the matrix to be analysed and the decomposition of this matrix. For the matrix to be analyzed the weighted matrix defined in (18) can be used by LSA as well as by CA. Equation (18) is sufficiently general for LSA. For CA, using (23), we set $\boldsymbol{N} = \boldsymbol{D_r}^{-\frac{1}{2}}$, $\boldsymbol{L} = (\boldsymbol{P}-\boldsymbol{E})$ and $\boldsymbol{G} = \boldsymbol{D_c}^{-\frac{1}{2}}$. This shows that the matrix decomposed in CA in (21) can be formulated in the LSA framework in (18).

The decomposition used in LSA leads to orthonormal matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ used for coordinates, see (20), whereas in CA the decomposition leads to weighted orthonormal matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Gamma}$, see (24). If we rewrite (20) as $\boldsymbol{U}^T\boldsymbol{IU} = \boldsymbol{I} = \boldsymbol{V}^T\boldsymbol{IV}$, we see this is a difference between using an identity metric $\boldsymbol{I}$ and a metric defined by the margins that are collected in $\boldsymbol{D_r}$ and in $\boldsymbol{D_c}$. The influence of this metrics used in CA is most clearly visible in the definition of the chi-squared distances (8), that makes that, for example, for row profiles i and i', equally large differences between columns j and j' are weighted by the margins of j and j' in such a way that a column with a smaller margin takes a larger part in the chi-squared distance between i and i'.

6 Authorship Attribution using LSA and CA

In this section we examine the performance of LSA and CA on a dataset originally set up for authorship attribution. We first use the dataset to see how well LSA and CA are

able to assign documents with a known author to the correct author. Second, we assign a document with unknown author to one of the known authors.

Authorship attribution is the process of identifying the authorship of a document; its applications include plagiarism detection and resolving of authorship disputes (Bozkurt, Baghoglu, & Uyar, 2007), and is particularly relevant for historical texts, where other historical records are not sufficient to determine authorship. Both LSA and CA have been used for authorship attribution before. For example, Soboroff, Nicholas, Kukla, and Ebert (1997) applied LSA with n-grams as terms to visualize authorship among biblical Hebrew texts. McCarthy, Lewis, Dufty, and McNamara (2006) applied LSA to lexical features to automatically detect semantic similarities between words (Stamatatos, 2009). Satyam, Dawn, and Saha (2014) used LSA on a character n-gram based representation to build a similarity measure between a questioned document and known documents. Mealand (1995) studied the Gospel of Luke using a visualization provided by CA. Mealand (1997) also measured genre differences in Mark by CA. Mannion and Dixon (2004) applied CA to study authorship attribution of the case of Oliver Goldsmith by visualization.

The *Wilhelmus* is the national anthem of the Netherlands and its authorship is unknown and much debated. There is a substantive amount qualitative research attempting to determine the authorship of the *Wilhelmus*, with quantitative or statistical methods being used relatively recently. To the best of our knowledge, the authorship of the *Wilhelmus* was first studied by statistical methods and computational means in Winkel (2015), whose results on authorship attribution were inconclusive. After that, Kestemont, Stronks, De Bruin, and Winkel (2017a, 2017b) studied the question using PCA and the General Imposters (GI) method, attributing the *Wilhelmus* to the writer Datheen. Vargas Quiros (2017) used the data of Kestemont et al. (2017a, 2017b), and applied the KRIMP compression algorithm (van Leeuwen, Vreeken, & Siebes, 2006) and Kullback-Leibler Divergence — they tended to agree with Kestemont et al. (2017a, 2017b), even though the KRIMP attributed the *Wilhelmus* to another author when a different feature selection method was used. Thus, the results were inconclusive, with a tendency to prefer Datheen. Our paper provides further evidence in favour of attributing the authorship to *Datheen*.

6.1 Data and methods

We use a total of 186 documents by six writers, consisting of 35 documents written by Datheen, 46 by Marnix, 23 by Heere, 35 by Haecht, 33 by Fruytiers, and 14 by Coornhert. These documents contain tag-lemma pairs as terms, obtained through part-of-speech tagging and lemmatizing of the texts, and are made publicly available by Kestemont, Stover, Koppel, Karsdorp, and Daelemans (2016); Kestemont et al. (2017a, 2017b). Following Kestemont, we use the 300 most frequent tag-lemma pairs, thus the document term matrix has size 186 x 300. The average marginal frequencies range from 406 for documents by Fruytiers to 545 for documents by Haecht. See Kestemont (2017) for more details regarding the dataset.

We use two approaches to compare LSA and CA. One is visualization, where we use

LSA and CA to visualize documents by projecting them onto two dimensions. The other is to apply LSA and CA, along with distance measures (described in detail in section 6.3). In line with the foregoing sections, we denote the raw document-term matrix by $\boldsymbol{F}$. In the case of LSA we examine four versions: LSA of $\boldsymbol{F}$ (LSA-RAW), LSA of the row-normalized matrices $\boldsymbol{F}^{L1}$ (LSA-NROWL1) and $\boldsymbol{F}^{L2}$ (LSA-NROWL2), and LSA of the TF-IDF matrix $\boldsymbol{F}^{\text{TF-IDF}}$ (LSA-TFIDF). In addition, we also compare performance with the raw document-term matrix, denoted as RAW, where no dimensionality reduction has taken place.

6.2 Visualization

We first examine all documents of two authors Marnix and Dathleen[2], along with the *Wilhelmus* document, using the 300 most frequent tag-lemma pairs. These form a document-term matrix of size 82×300.

Figure 3 shows the results of analysing this document-term matrix using the four LSA based methods (LSA-RAW, LSA-NROWL1, LSA-NROWL2, LSA-TFIDF), and CA. As seen in Figure 3, all four varieties of LSA fail to show a clear separation, while CA separates documents by the two authors clearly, even though the first 2 dimensions for CA ($11.2\% + 8.6\% = 19.8\%$) account for a much smaller percentage of the total sum of squared singular values than the first 2 dimensions for LSA-RAW ($67.6\% + 12.2\% = 79.8\%$), LSA-NROWL1 ($67.6\% + 12.3\% = 79.9\%$), LSA-NROWL2 ($67.8\% + 11.7\% = 79.5\%$), and LSA-TFIDF ($53.8\% + 11.1\% = 64.9\%$). This is because the margins play an important role in the first two dimensions for LSA-RAW, LSA-NROWL1, LSA-NROWL2, and LSA-TFIDF and the relations between documents are blurred by these margins. We also see that the *Wilhelmus* (shown as w, in red) is clearly attributed by CA to Datheen.

Given the effectiveness of CA and the attribution of the *Wilhelmus* to Datheen in the above analysis, we now show visualisations of CA for documents by Datheen and four other authors in turn (Figure 4). For three out of four authors, there is a clear separation between that author and Datheen. In the case Haecht however (sub-figure (b)), there is no clear separation from Datheen. In all three cases where there is a clear separation, *Wilhelmus* is attributed to Datheen, as before.

Finally, we apply all four varieties of LSA and CA to all documents of the six authors, and the *Wilhelmus*, which form a document-term matrix of size 187×300. Figure 5 shows the results of the analysis of this matrix by LSA-RAW, LSA-NROWL1, LSA-NROWL2, LSA-TFIDF, and CA. As seen in the figure, the LSA methods cannot separate the authors of our dataset well, but CA does a reasonably good job. Again we find that, although the percentage of the total sum of squared singular values in the first two dimensions for CA ($8.6\% + 5.7\% = 14.3\%$) is lower than LSA-RAW ($64.1\% + 11.1\% = 75.2\%$), LSA-NROWL1 ($63.5\% + 10.9\% = 74.4\%$), LSA-NROWL2 ($64.1\% + 10.0\% = 74.1\%$), and LSA-TFIDF ($48.7\% + 10.5\% = 59.2\%$), CA separates the documents quite well. For instance,

[2]We chose these two authors specifically, out of our dataset, as they are the two main contenders for the authorship of *Wilhelmus* – Marnix has been the most popular candidate from qualitative analyses, and since the work of Kestemont et al. (2017a, 2017b) Datheen is also a serious candidate.

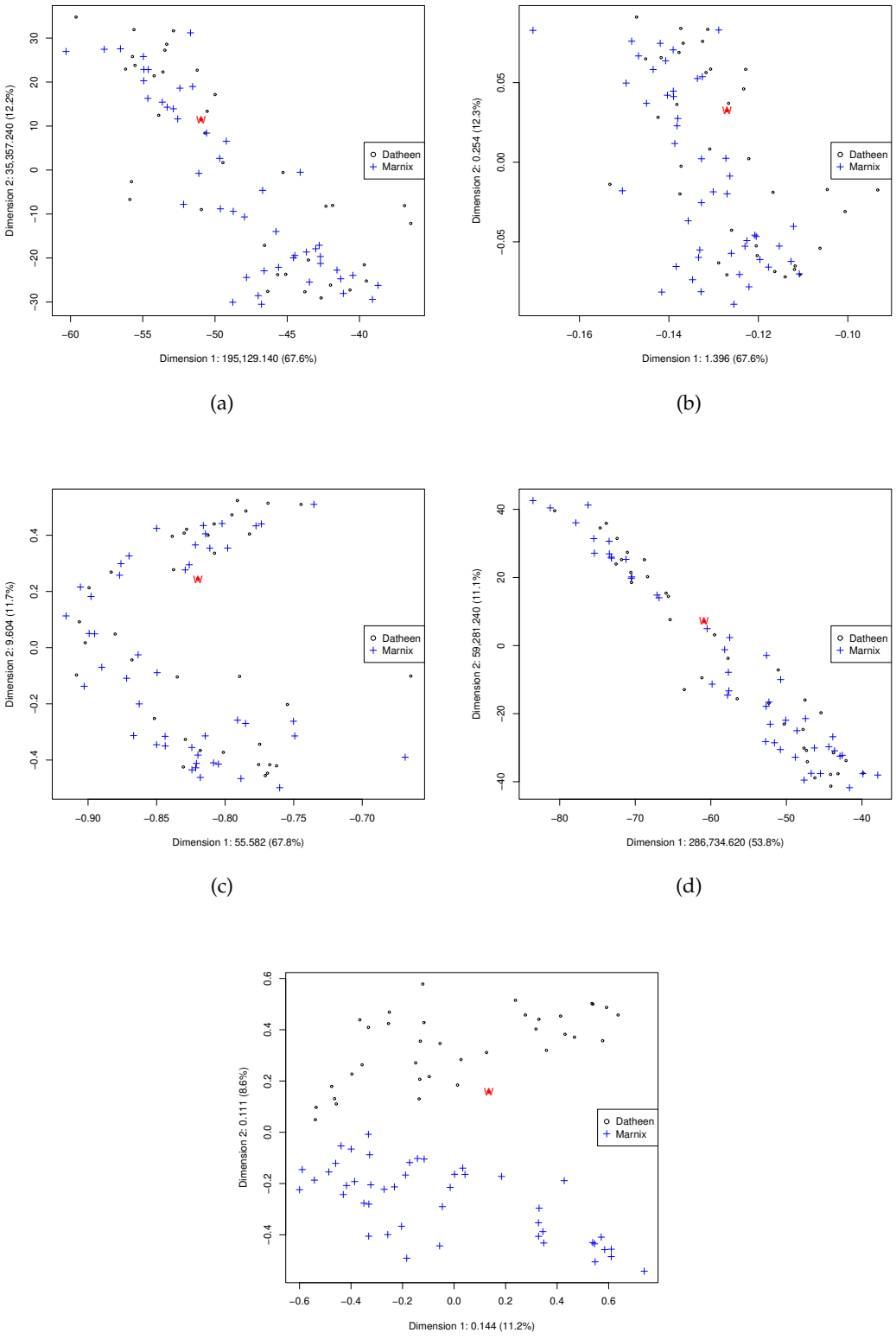
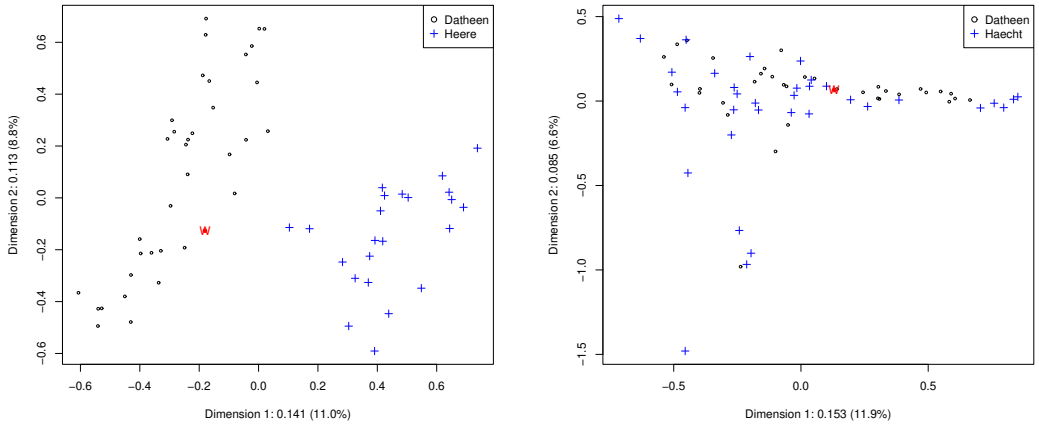
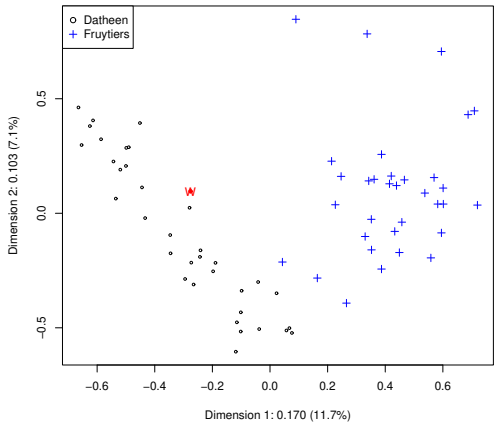


Figure 3: The first two dimensions for each document of author Datheen and author Marnix, and the *Wilhelmus* (in red) by (a) LSA-RAW; (b) LSA-NROWL1; (c) LSA-NROWL2; (d) LSA-TFIDF; (e) CA.

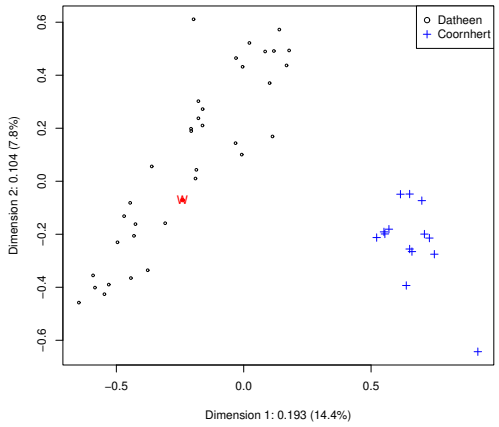


(a)

(b)



(c)



(d)

Figure 4: The first two dimensions for each document of author Datheen and another author, and the *Wilhelmus* (in red) using CA: (a) Heere; (b) Haecht; (c) Fruytiers; (d) Coornhert.

documents written by Marnix are effectively separated from the documents written by other authors. The documents of the other authors also seem to form much more distinguishable clusters, as compared to LSA, except for Datheen and Haecht.

6.3 Distance Measures

In this section, we use distance measures to quantitatively evaluate and compare performance on the authorship attribution problem. We use four different methods based on Euclidean distance for measuring the distance from a document to a set of documents (Guthrie, 2008; Kestemont et al., 2016; Koppel & Seidman, 2013). We choose the Euclidean distance because it plays a central role in the geometric interpretation of LSA and CA (see section 3 and 4).

Centroid Euclidean distance between the document and the centroid of the set of documents. The centroid for a set of documents is calculated by averaging the coordinates across all these documents.

In the other three methods we first calculate the Euclidean distance between the document and every document of the set of documents.

Average average of these Euclidean distances

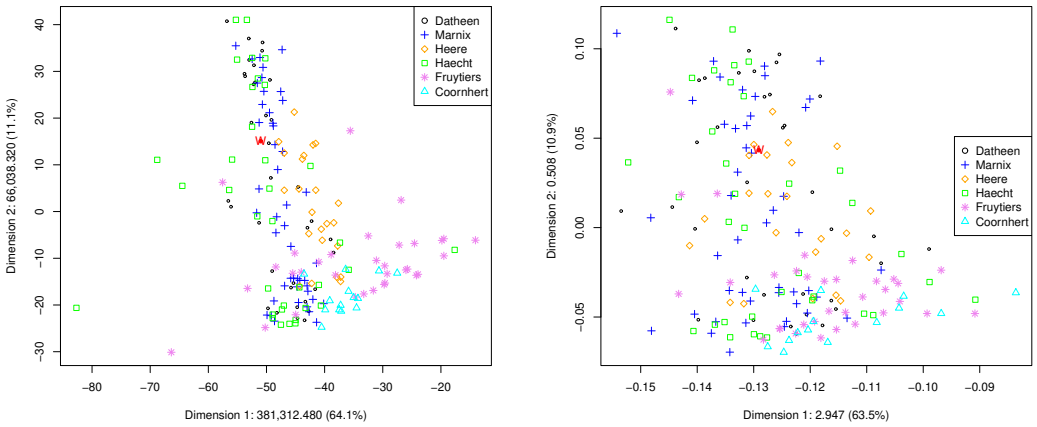
Single the minimum Euclidean distance among the Euclidean distances

Complete the maximum Euclidean distance among the Euclidean distances.

These four methods are similar to the procedures of measuring the distance between clusters in hierarchical clustering analysis, using the centroid, average, single, and complete linkage method respectively (Jarman, 2020).

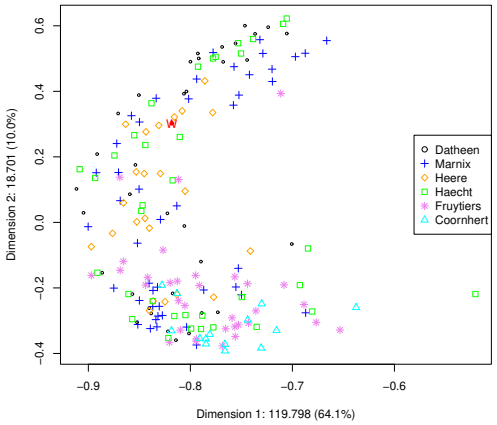
It is crucial to optimise dimensionality for each of the distance measures. For choosing optimal dimensions for each, we use leave-one-out cross-validation (LOOCV) (Kuzi, Shtok, & Kurland, 2016; Wong, 2015) in combination with accuracy. For each distance measure we determine the number of dimensions that provide the highest accuracy in LOOCV. The 186 documents of six authors form a document-term matrix with 186 rows and 300 columns. We perform LSA-RAW, LSA-NROWL1, LSA-NROWL2, LSA-TFIDF, and CA on the document-term matrix to obtain the coordinates of the 186 documents in a lower dimensional space. Using LOOCV, each time we discern the following three steps. At the first step, we select one of the 186 documents. At step two, using the centroid, average, single, and complete linkage method, the distance is computed between the single document and the six author groups of documents. For this single document, the predicted author of the document is the author with the smallest distance. At the final step, we compare the predicted author with the true author of the single document. We repeat this 186 times, once for each single documents. The accuracy is calculated by the ratio: number of times an author is correctly predicted divided by 186.

Table 10 shows the maximum accuracy for LSA-RAW, LSA-NROWL1, LSA-NROWL2, LSA-TFIDF, and CA for the four distance measures, along with the optimal dimensions k

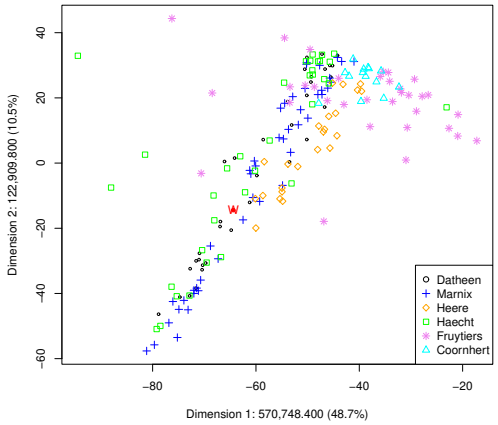


(a)

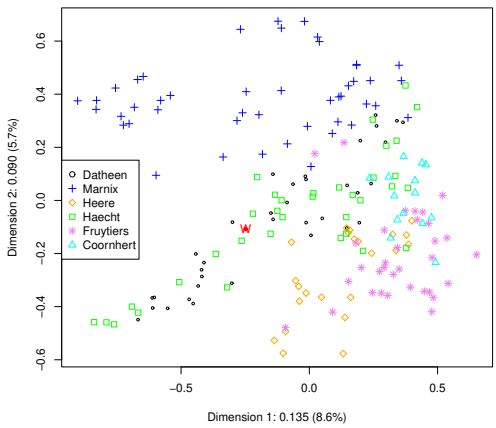
(b)



(c)



(d)



(e)

Figure 5: The first two dimensions for each document of six authors, and the *Wilhelmus* (in red) by (a) LSA-RAW; (b) LSA-NROWL1; (c) LSA-NROWL2; (d) LSA-TFIDF; (e) CA.

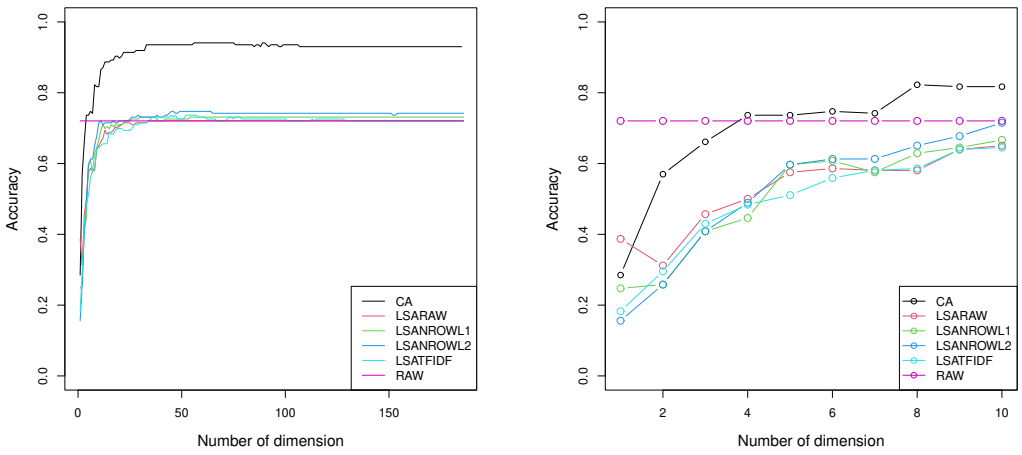
Table 10: The optimal dimensionality k and the accuracy in k for LSA-RAW, LSA-NROWL1, LSA-NROWL2, LSA-TFIDF, and CA, and the accuracy for RAW using different distance measurement methods.

	Centroid		Average		Single		Complete	
	k	Accuracy	k	Accuracy	k	Accuracy	k	Accuracy
RAW		0.720		0.516		0.672		0.177
LSA-RAW	34–186	0.720	60;71–90	0.554	13–15	0.715	1	0.301
LSA-NROWL1	30–37; 54–186	0.731	71–186	0.640	21; 22	0.699	36; 77; 82–85; 87–91	0.220
LSA-NROWL2	45; 46; 49–64	0.747	40	0.704	20	0.699	30; 63–186	0.296
LSA-TFIDF	45; 47; 52–57	0.737	19	0.543	24; 25	0.737	1	0.231
CA	56–75; 89; 90	**0.941**	12	**0.823**	14	**0.780**	7	**0.457**

where this maximum accuracy is reached. First, in the optimal dimensionality, CA yields the maximum accuracy for all distance measurement methods, over the RAW (i.e. without dimensionality reduction) matrix, as well as over all four LSA methods. Second, among all distance measurement methods, the centroid method always has the highest accuracy indicating that the centroid method is preferable over the other distance measurement methods.

6.3.1 Further study of centroid method

In order to further explore the centroid method, Figure 6 shows the accuracy with different numbers of dimensions for CA, RAW, LSA-RAW, LSA-NROWL1, LSA-NROWL2, and LSA-TFIDF. Figure 6(a) displays all dimensions on the horizontal axis, and Figure 6(b) focuses on the first 10 dimensions. CA in combination with the centroid method performs better than the other methods almost irrespective of dimension, except for the very first ones.

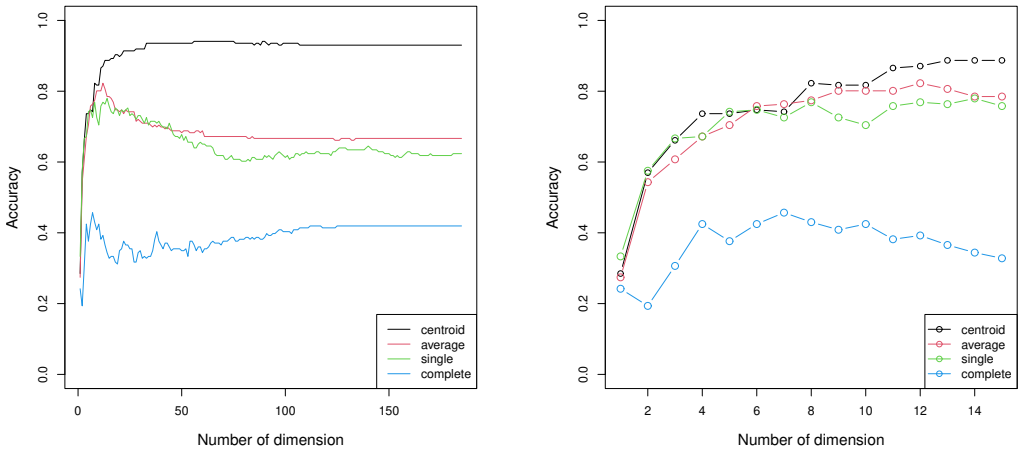


(a) (b)

Figure 6: Accuracy versus the number of dimensions (centroid method) for CA, RAW, LSA-RAW, LSA-NROWL1, LSA-NROWL2, and LSA-TFIDF.

6.3.2 A further study of CA

In order to further explore CA, Figure 7 shows the accuracy with different numbers of dimensions for the four distance measurement methods. In Figure 7(a), the dimensions are ranked from 1 to 185, and in Figure 7(b), we focus on dimensions 1 to 15. We can see that, for CA, the centroid method is best among all distance measurement methods from dimension 8 to dimension 185, where the accuracy of the centroid method is much higher than the maximum accuracy of the other methods starting at dimension 11 and the accuracy is very high over a large range.



(a) (b)

Figure 7: Accuracy versus the number of dimension for CA.

6.4 Authorship attribution of the *Wilhelmus*

Since CA in combination with the centroid method appears to be the best overall, we use them to determine the authorship of the *Wilhelmus*. In the 22 optimal dimensions (dimensions 56–75, 89, and 90), we find that the *Wilhelmus* is attributed to the author Datheen, while Haecht is the second most likely candidate. The distance of the *Wilhelmus* to the centroid of Datheen averaged across 22 optimal dimensions is 0.893, to Haecht is 0.951, to Marnix is 0.998, to Heere is 1.065, to Fruytiers is 1.122, and to Coornhert is 1.303. Thus, CA attributes *Wilhelmus* to Datheen, and provides more weight using an independent statistical technique, to prior results by Kestemont et al. (2017a, 2017b) in resolving this debate.

7 Document Classification: BBCSport

We next perform an evaluation on document classification on an English dataset, BBCSport, described below.

Table 11: The minimum optimal dimensionality k and the accuracy in k for LSA-RAW, LSA-NROWL1, LSA-NROWL2, LSA-TFIDF, and CA, and the accuracy for RAW using different distance measurement methods.

	Centroid		Average		Single		Complete	
	k	Accuracy	k	Accuracy	k	Accuracy	k	Accuracy
RAW		0.904		0.452		0.822		0.137
LSA-RAW	74	0.904	8	0.849	41	0.952	6	0.486
LSA-NROWL1	66	0.952	11	0.945	51	0.959	5	0.658
LSA-NROWL2	83	0.959	62	0.966	22	0.959	6	**0.884**
LSA-TFIDF	100	0.911	7	0.815	25	0.979	6	0.144
CA	146	**0.986**	30	**0.973**	21	**0.986**	1	0.404

7.1 Data

The BBCSport dataset consists of 737 documents and is divided into five categories: "athletics" (101 documents), "cricket" (124 documents), "football" (265 documents), "rugby" (147 documents), and "tennis" (100 documents). The BBCSport dataset (Greene & Cunningham, 2006) have previously been used in the evaluation of text classification, for example Barman and Chowdhury (2020); Bounabi, El Moutaouakil, and Satori (2017, 2018), and is available online at http://mlg.ucd.ie/datasets/bbc.html .

To pre-process this dataset we project all characters to lower case, remove punctuation marks, numbers, and stop words, and apply lemmatization. Subsequently, terms with frequencies lower than 10 are ignored. This gives us a document-term matrix of size 737×2071.

7.2 Visualization

Figure 8 shows the results of an analysis of this document-term matrix by LSA-RAW, LSA-NROWL1, LSA-NROWL2, LSA-TFIDF, and CA. On this dataset as well, we find that the LSA methods do not separate the classes well, but CA does a reasonably good job.

7.3 Distance measures

We use a random seed to divide the BBCSport dataset into 80% training set documents and 20% test set documents to calculate the accuracy of classifying the test set documents correctly. We calculate accuracy on the test set under different dimensions, and chose the maximum accuracy, along with the optimal dimensions. Table 11 shows the maximum accuracy for RAW, the four LSA methods, and CA for the four distance measures, along with the *minimum* optimal dimensions k where this maximum accuracy is reached [3]. For all distance measures except complete, CA yields the maximum accuracy for the optimal dimensionality. CA with centroid and single measure gives the best accuracy overall.

[3]The optimal dimensions that obtain the maximum accuracy are not just one; for reasons of space, we show only the minimum optimal dimension in Table 11.

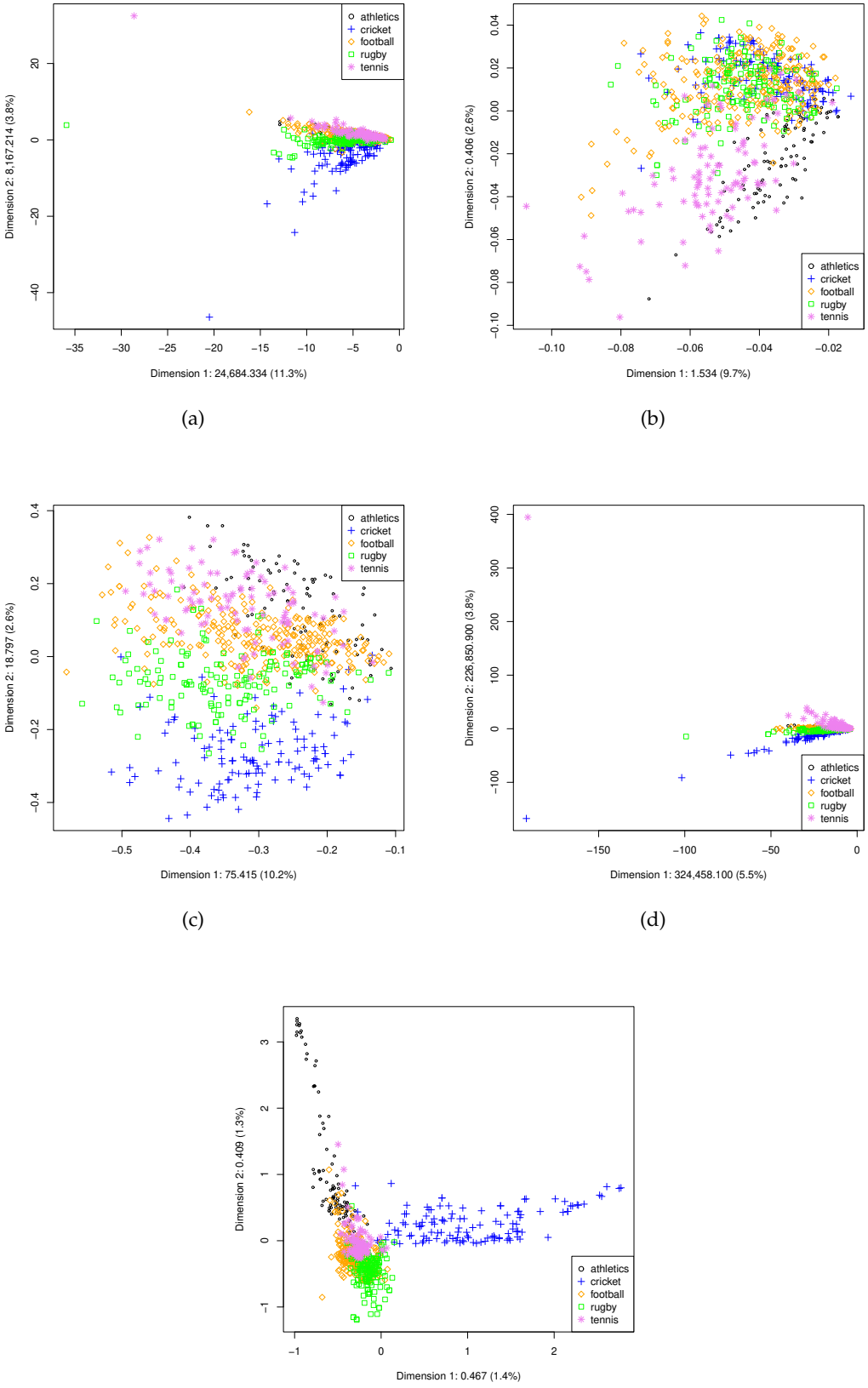


Figure 8: The first two dimensions for each document of BBCSport dataset by (a) LSA-RAW; (b) LSA-NROWL1; (c) LSA-NROWL2; (d) LSA-TFIDF; (e) CA.

8 Conclusion

LSA and CA both allow for dimensionality reduction by the SVD of a matrix; however the actual matrix analysed by LSA and CA is different, and therefore LSA and CA capture different kinds of information. In LSA we apply an SVD to $\boldsymbol{F}$, or to a weighted $\boldsymbol{F}$. In CA, an SVD is applied to the matrix $\boldsymbol{D}_r^{-\frac{1}{2}}(\boldsymbol{P}-\boldsymbol{E})\boldsymbol{D}_c^{-\frac{1}{2}}$ of standardized residuals. The elements in $\boldsymbol{D}_r^{-\frac{1}{2}}(\boldsymbol{P}-\boldsymbol{E})\boldsymbol{D}_c^{-\frac{1}{2}}$ display the departure from the margins, that is, departure from the expected frequencies under independence collected in $\boldsymbol{E}$. Due to $\boldsymbol{E}$, in CA the effect of the margins is eliminated — a solution only displays the dependence between documents and terms. Concluding, in LSA, the effect of the margins as well as the dependence is part of the matrix that is analysed and these margins usually play a dominant role in the first dimension of the LSA solution as usually on the first dimension all points depart in the same direction from the origin. On the other hand, in CA all points are scattered around the origin and the origin represents the profile of the row and column margins of $\boldsymbol{F}$.

In summary, although LSA allows a study of the relations between documents, between terms, and between documents and terms, this study is not easy. The reason is that these relations are blurred by the effect of the margins that are also displayed in the LSA solution. CA does not have this property. Therefore it appears that CA is a better tool for studying the relations between documents, between terms, and between documents and terms. Also, discussed in Section 4, CA has many nice properties like providing a geometric display where the Euclidean distances approximate the χ^2-distances between the rows and between the columns of the matrix, and the relation to the Pearson χ^2 statistic. Overall, from a theoretical point of view it appears that CA has more attractive properties than LSA. Empirically, we evaluated and compared the two methods on two different tasks in two languages, authorship attribution in Dutch, and document classification in English, and found that CA can both separate documents better, and obtain higher accuracies on the tasks as compared to LSA based techniques.

In future work, we would like to extend our analysis to include other transformations, e.g., those based on point-wise mutual information and on word-context matrices, as well as to evaluate the performance of CA and LSA on other natural language processing tasks.

Acknowledgments

The first author is supported by the China Scholarship Council.

References

Ab Samat, N., Murad, M. A. A., Abdullah, M. T., & Atan, R. (2008). Term weighting schemes experiment based on SVD for Malay text retrieval. *IJCSNS*, *8*(10), 357-361.

Aggarwal, C. C. (2018). *Machine learning for text*. Springer.

Albright, R. (2004). Taming text with the SVD. *SAS Institute Inc*.

Barman, D., & Chowdhury, N. (2020). A novel semi supervised approach for text classification. *International Journal of Information Technology*, *12*(4), 1147–1157.

Benzécri, J.-P. (1973). *L'analyse des données* (Vol. 1 and 2). Dunod Paris.

Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM review*, *37*(4), 573–595.

Bounabi, M., El Moutaouakil, K., & Satori, K. (2017). A comparison of text classification methods method of weighted terms selected by different stemming techniques. In *Proceedings of the 2nd international conference on big data, cloud and applications* (pp. 1–9).

Bounabi, M., El Moutaouakil, K., & Satori, K. (2018). A probabilistic vector representation and neural network for text classification. In Y. Tabii, M. Lazaar, M. Al Achhab, & N. Enneya (Eds.), *Big data, cloud and applications* (pp. 343–355). Cham: Springer International Publishing.

Bozkurt, I. N., Baghoglu, O., & Uyar, E. (2007). Authorship attribution. In *2007 22nd international symposium on computer and information sciences* (pp. 1–5).

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391–407.

Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). *Mathematics for machine learning*. Cambridge University Press.

Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers*, *23*(2), 229–236.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 281–285).

Gifi, A. (1990). *Nonlinear multivariate analysis*. Wiley.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. Academic Press.

Greenacre, M. J. (2017). *Correspondence analysis in practice*. CRC press.

Greenacre, M. J., & Hastie, T. (1987). The geometric interpretation of correspondence analysis. *Journal of the American statistical association*, *82*(398), 437–447.

Greene, D., & Cunningham, P. (2006). Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd international conference on machine learning* (pp. 377–384).

Guthrie, D. (2008). *Unsupervised detection of anomalous text* (Unpublished doctoral dissertation). University of Sheffield.

Hayashi, C. (1956). Theory and example of quantification (II). *Proceedings of the Institute of Statistical Mathematics*, *4*, 19–30.

Hayashi, C. (1992). Quantification method III or correspondence analysis in medical science. *Annals of Cancer Research and Therapy*, *1*(1), 17–21.

Hill, M. O. (1973). Reciprocal averaging: an eigenvector method of ordination. *The Journal of Ecology*, *61*(1), 237–249.

Hill, M. O. (1974). Correspondence analysis: a neglected multivariate method. *Applied Statistics*, *23*(3), 340–354.

Jarman, A. M. (2020). *Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method*. Georgia Southern University.

Kestemont, M. (2017). *Who wrote the wilhelmus?* Retrieved July 17, 2021, from https://mikekestemont.github.io/wilhelmus/.

Kestemont, M., Stover, J., Koppel, M., Karsdorp, F., & Daelemans, W. (2016). Authenticating the writings of Julius Caesar. *Expert Systems with Applications*, *63*, 86-96.

Kestemont, M., Stronks, E., De Bruin, M., & Winkel, T. d. (2017a). Did a poet with donkey ears write the oldest anthem in the world? Ideological implications of the computational attribution of the Dutch national anthem to Petrus Dathenus. In *Dh.*

Kestemont, M., Stronks, E., De Bruin, M., & Winkel, T. d. (2017b). *Van wie is het wilhelmus? de auteur van het nederlandse volkslied met de computer onderzocht*. Amsterdam University Press.

Kolda, T. G., & O'leary, D. P. (1998). A semidiscrete matrix decomposition for latent semantic indexing information retrieval. *ACM Transactions on Information Systems*, *16*(4), 322–346.

Koppel, M., & Seidman, S. (2013). Automatically identifying pseudepigraphic texts. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1449–1454).

Kuzi, S., Shtok, A., & Kurland, O. (2016). Query expansion using word embeddings. In *Proceedings of the 25th acm international on conference on information and knowledge management* (pp. 1929–1932).

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 27). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2014/file/feab05aa91085b7a8012516bc3533958-Paper.pdf

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the association for computational linguistics*, *3*, 211–225.

Mannion, D., & Dixon, P. (2004). Sentence-length and authorship attribution: the case of Oliver Goldsmith. *Literary and Linguistic Computing*, *19*(4), 497–508.

McCarthy, P. M., Lewis, G. A., Dufty, D. F., & McNamara, D. S. (2006). Analyzing writing styles with Coh-Metrix. In *Flairs conference* (pp. 764–769).

Mealand, D. L. (1995). Correspondence analysis of Luke. *Literary and linguistic computing*, *10*(3), 171–182.

Mealand, D. L. (1997). Measuring genre differences in Mark with correspondence analysis. *Literary and Linguistic Computing*, *12*(4), 227–245.

Michailidis, G., & De Leeuw, J. (1998). The Gifi system of descriptive multivariate analysis. *Statistical Science*, 307–336.

Morin, A. (1999). *Knowledge extraction in texts: a comparison of two methods.* Retrieved July 17, 2021, from http://www.tfi.99/proceedings/papers/Morin0673.pdf .

Nakov, P., Popova, A., & Mateev, P. (2001). Weight functions impact on LSA performance. *EuroConference RANLP*, 187–193.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513–523.

Satyam, A., Dawn, A. K., & Saha, S. K. (2014). A statistical analysis approach to author identification using latent semantic analysis. *Notebook for PAN at CLEF*.

Séguéla, J., & Saporta, G. (2011). A comparison between latent semantic analysis and correspondence analysis. In *Carme 2011 international conference on correspondence analysis and related methods.*

Séguéla, J., & Saporta, G. (2013). A hybrid recommender system to predict online job offer performance. *Revue des Nouvelles Technologies de l'Information*, *RNTI-E-25*, 177-197. Retrieved from http://halshs.archives-ouvertes.fr/halshs-01126258

Soboroff, I. M., Nicholas, C. K., Kukla, J. M., & Ebert, D. S. (1997). Visualizing document authorship using n-grams and latent semantic indexing. In *Proceedings of the 1997 workshop on new paradigms in information visualization and manipulation* (pp. 43–48).

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, *60*(3), 538–556.

Van der Heijden, P. G. M., De Falguerolles, A., & De Leeuw, J. (1989). A combined approach to contingency table analysis using correspondence analysis and loglinear analysis. *Applied Statistics*, *38*(2), 249–292.

van Leeuwen, M., Vreeken, J., & Siebes, A. (2006). Compression picks item sets that matter. In *European conference on principles of data mining and knowledge discovery* (pp. 585–592).

Vargas Quiros, J. (2017). *Information-theoretic anomaly detection and authorship attribution in literature* (Unpublished master's thesis). Utrecht University.

Winkel, T. d. (2015). *Of Deutsches blood* (Unpublished master's thesis). Utrecht University.

Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, *48*(9), 2839-2846.