Data Mining and Matrices (FSS 2019) - Assignment 3 - NMF

Philipp Jamscikov
Matriculation Number: 1531375
pjamscik@mail.uni-mannheim.de

1 Topic modelling with NMF

We will first apply NMF to the normalized newsgroup data to discover latent topics. Start by computing the NMF optimizing the KL divergence for r = 4 factors. The output of this step will be two unnormalized matrices W and H.

Before starting with the analysis of the factorization itself, it is worth to perform an explorative data analysis (EDA) up-front, as this will facilitate the comparison of the obtained factorizations.

The data set consists of 400 rows, each row is a document's Bag-of-Words (BoW) representation, and each entry is a relative frequency such that the data set sums to unity. The data set covers 4 topics, with 100 documents per topic. The topics appear in this order: "Cryptography" (383 different terms over all 100 docs) , "Medicine" (488), "Space" (451) , "Christian Religion" (454). As shown in the accompanying $understanding\_sorting\_order.ipynb$, given "Cryptography", the remaining 3 topics are sorted in descending order regarding their term overlap with the first topic. For a given topic, documents are ordered as follows: given a topic and the set of all terms it contains, determine for each of its documents the number of intersecting terms. Sort the documents in ascending order: i.e. start with the document whose set of terms has the least overlap with all of the topic's terms. This logic explains the monotonically decreasing upper green line we would obtain if we connect the rightmost points of any two consecutive rows in figure 1. Here, sorted documents constitute the Y-axis, words/terms/tokens the X-axis, and the intensities the relative frequencies of term-document-combinations. The black values to the right indicate that these words are not being observed for a given document. Throughout this paper, we suppress frequencies/probabilites below 0.0001% in both, the original data as well as in the reconstructions, in order to obtain better visualizations.

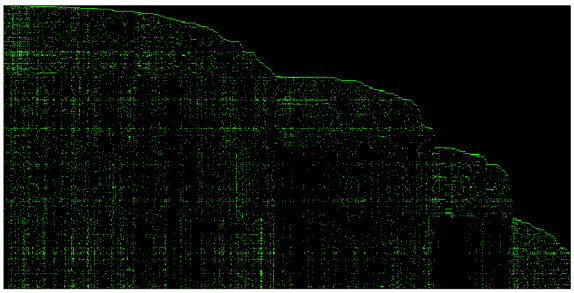

Figure 1: Relative Frequencies of Document-Term Combinations Represented as Pixel Intensities

In order to improve the interpretability of figure 1 and the different reconstructions of the original data matrix, we will modify this image by plotting 4 additional boxes. Each box shows us how the vocabulary (i.e. the set of all terms) is gradually extended by each topic. So, starting in the upper left corner of figure 2 we can interpret the boxes as such: the first topic "Cryptography" introduces 383 terms, "Medicine" extends this vocabulary by 226 terms, thus resulting in an vocabulary of 609 terms overall (second box in figure 2). "Space" extends the shared vocabulary of the first two topic with 109 additional terms, which is what the third box illustrates. The last box shows that the first three topics cover 718 terms, and that 72 additional terms are being in-

troduced by "Christian Religion" to result in the overall vocabulary size of 800. Please refer to $understanding\_sorting\_order.ipynb$ if in need of further clarifications.

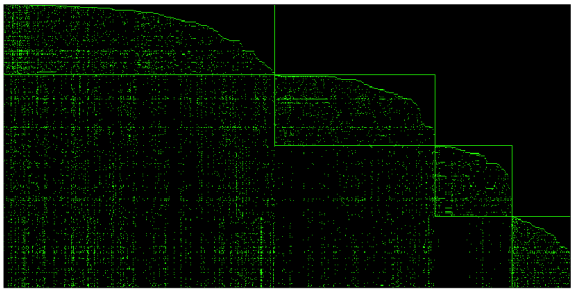

Figure 2: Relative Frequencies of Document-Term Combinations Represented as Pixel Intensities

The representation in figure 2 has the advantage that we now can clearly visually identify faulty reconstructions for any of the first three topics: we simply choose the corresponding box and look whether green or red values appear on the right hand side. If yes, then the reconstruction assigns some term to a document which has in fact never been observed for the given topic/document. Also, investigating what happens within each of the boxes reveals some interesting properties of the reconstructions, that we can interpret as topic-specific imputations. Of course, the reconstructions also affect the other areas, e.g. the left half of figure 2. But as these regions show rather frequent terms shared by many of the documents, visually interpreting the reconstructions in these areas is cumbersome, and hence we focus on what happens to the upper right black area of figure 1.

Study the top-10 terms in the right-hand factor H. Are the results good? To evaluate the results, you must consider the terms associated with each topic and argue why you think they do or do not constitute a meaningful topic (and what that topic is) remember, we understand the term topic very loosely, e.g., words related to sports is a valid topic. Also, the topics are not necessarily only those of the newsgroups. Remember also that the terms are stemmed.

Please note that I relate to one possible factorization. We obtain different results if we re-run the code several times since NMF is not unique. For example, topics appear in a different order, and also the composition of the top-10 terms may change. (Although most of the time I observed the same 8-10 top terms per topic).
As we know the four ground truth topics, we are able to judge whether our rank-4 decomposition of the original data set captures each of the topics in a meaningful way. Ideally, each of topics is described by a set of relatively few and distinctive terms that are not being shared with other topics. From the majority of the high-weight tokens in the first row of $\mathbf{H}$, we can derive the topic "Christian Religion" very well, since they are very characteristic (e.g. "god", "christian", "church"). Also "people" seems to substantially contribute to the topic, but as it occurs also among the top candidates for two of the other topics ("Cryptograpy", "Medicine"), we have to consider it as rather generic and less expressive. (We should keep this observation in mind with regard to Exercise 4 - i.e. we may obtain a better clustering through removing generic tokens. For this purpose we may also use TFIDF-weights.) From the second row of $\mathbf{H}$ and its the top-weighted tokens such as "space", "launch", and "orbit", we can easily deduce the topic "Space". The token "system" is shared with the top-10 tokens of row 4 and topic "Cryptography". This is another candidate that may disappear when using TFIDF weights to represent the document collection. Row 3 of $\mathbf{H}$ and terms such as "disease", "doctor", and "medic" allow us to defer that this row has to represent the topic "Medicine". Also the results for the fourth row and the last topic "Cryptography" may be considered meaningful, with tokens such as "encrypt", "system" and "secur". The top-token "kei" could maybe be explained with more domain knowledge, or as a second guess may represent the word key and be an artefact of the applied stemming.

b) Study the reconstructed matrix. Does it look like you would have expected? Which aspects are covered well? Which are not?

The reconstructed matrix is displayed in figure 3. We see that the area on the right of each of the

upper 3 boxes is largely black which indicates a good reconstruction. Nonetheless, if you zoom in and look e.g. on the right of the second box, you see some bright green pixels which indicates that here are terms falsely being reconstructed for the second topic and some of its documents. As these green pixels are above the third box, and bearing in mind that the third box is very characteristic for the third topic (as the area below the third box is largely black, and hence there is only few vocabulary overlap with the last topic), we can conclude that the reconstruction has certain difficulties to fully "separate" the second topic "Medicine" and the third topic "Space". This observation is also conveyed in the confusion matrix in figure 9.

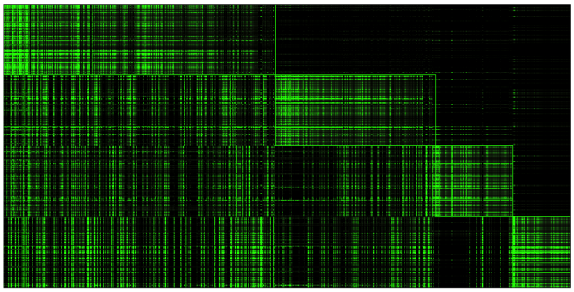

Figure 3: NMF Reconstruction ($r = 4, KL$)

The left half of figure 3 shows us how the vocabulary that is shared by all 4 topics is being reconstructed. It is a good sign that all terms are being reconstructed for all 4 topics and that although there is a large term overlap between the topics, the pattern varies between each of the four topics. This means the reconstruction succeeds in finding topic-specific combinations of those terms shared by all topics.

Lastly, we'll discuss what is happening within each of the four boxes. We will select the fourth box, but the claims made generalize to any of the four boxes. If we compare the lower right box from figure 1 with the one from figure 3, we observe that the first may be divided into two regions: the lower left corner is filled with a greenish triangle and the the upper right corner with a black one. The latter however is "filled" throughout with green pixels and hence resembles as a fully filled green box. This is not by chance, but the NMF detects which terms are characteristic for a given topic, and then imputes these terms for all of a topic's documents when reconstructing the original matrix.

c) Take the rank-4 truncated SVD of the data and study the decomposition along the lines mentioned above. Compare!

In the case of the SVD, we observe that now terms can also contribute in a negative way to a given topic. This behaviour is illustrated in figure 4, which displays columns 50-60 of **V**. We observe that the term "drug" contributes positively and negatively to two topics each.

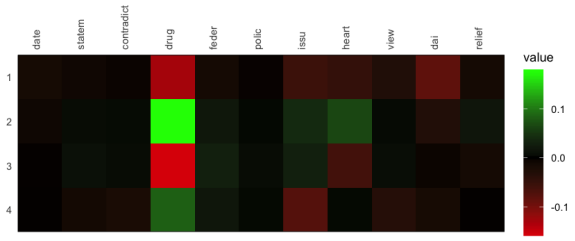

Figure 4: Columns of SVD's **V** Matrix ($r = 4$)

A possible interpretation is that given a topic, one would almost never expect to see a term with a high negative weight. This way of describing a topic is (at least for humans) counterintuitive: one would always describe an object by the presence of attributes, not by their absence. If we look at the four rows of **V** and are trying to defer the topics by looking at the ten terms/tokens with the largest absolute values, it is impossible to deduce the corresponding topic for all 4 topics in the ground truth. Regarding the second row of **V** and the positive weights of "cancer", "drug", "diseas", and "diet" we may conclude that this row represents the topic "Medicine". It is also quiet likely that either row 3 or row 4 represent "Cryptogra-

phy" since both include high positive values for e.g. "encrypt" and "kei". As we observe in row 3's top-10 more positive-weighted terms related to "Cryptography", we rather opt for row 3 representing this topic. This leaves us with two topics not clearly being assigned: "Space" and "Christian Religion". The reconstruction we obtain via SVD is displayed in figure 5. On the first sight, this reconstruction resembles figure 5, but following the line of argument from above, we see that there are more coloured pixels on the right-hand-side of each of the three upper boxes which we can clearly identify as mistakes made during reconstruction. As we are comparing two different objective functions, it makes sense only to visually judge the obtained reconstructions. We conclude that NMF performs better than SVD. Another mistake particular to the SVD is that values are being reconstructed that are smaller than zero. As we are dealing with frequency counts, we know beforehand that this cannot happen. A countermeasure would be to simply replace them with zero. Again, this shows that NMF also seems to be better theoretically founded for non-negative data.

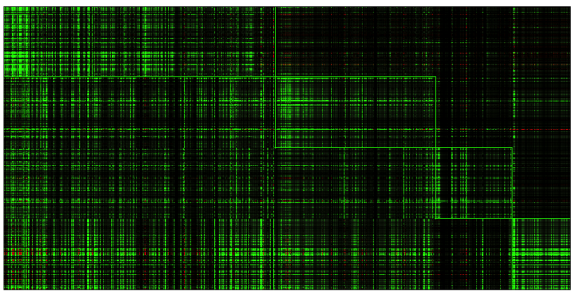

Figure 5: SVD Reconstruction ($r = 4$)

d) Now try different values of k (at least r = 2 and r = 8) and repeat the analysis (for NMF only). How do the results change? Can you name a single best rank?

The rank-2 reconstruction is shown in figure 6. As expected, the NMF fails to detect all four ground truth topics, instead we observe that the NMF reconstructs two prototypical blocks of data, each with a particular distribution of terms. One block is in the middle of figure 6, whereas the other block is split by this middle block and corresponds to the remaining upper and lower rows. If we investigate the top-10 terms of each topic, we see that one topic is a mixture of "Cryptography" and " Christian Religion", whereas the other topic comprises "Medicine" and "Space".

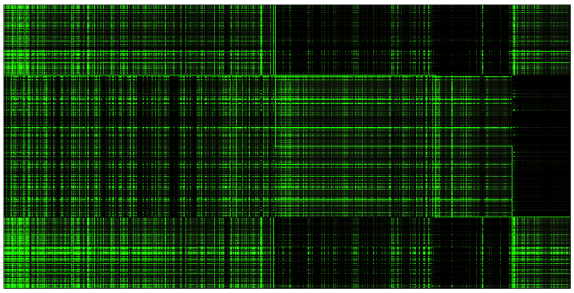

Figure 6: NMF Reconstruction ($r = 2$)

The rank-8 reconstruction is shown in figure 7, in which I find it hard to see any improvements over the reconstruction shown in figure 3. Investigating the top-terms in the 8 rows of $\mathbf{H}$, we observe that some of the topics now appear up to two or three times, but at least one time.

Knowing the ground truth and recalling Occam's Razor, choosing $r = 4$ seems to be the only valid choice for 4 latent topics. Working on a real-world data set and not knowing the ground truth, this procedure would require a much more fine-grained analysis. An approach may be to ask different domain experts for an estimate of the number of latent topics and then investigate factorizations with a rank in the estimated range.

e) Apply Gaussian NMF (i.e., using Euclidean norm). Do the results change? In your opinion, which NMF variant produces better results, if any? Argue!

Here, we discuss the implications of substituting the previously applied Generalized Kullback-Leibler (GKL) by the Euclidean Norm as a cost

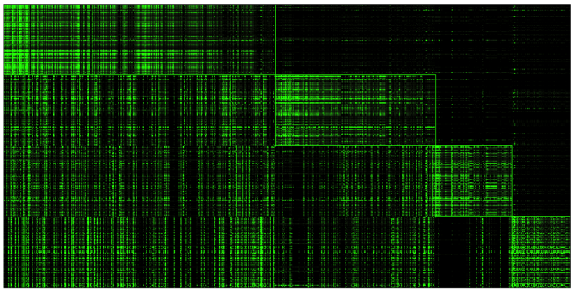

Figure 7: NMF Reconstruction ($r = 8$)

measure to calculate the reconstruction error. Visually, the rank-4 Gaussian NMF' structures does not largely differ from from the rank-4 GKL NMF and hence is not displayed in the report. In the accompanying Jupyter Notebook, we however observe that the reconstruction mistakes made by the Gaussian NMF are not necessarily as those of the GKL NMF.

If we rerun the code producing the Gaussian NMF factorization several times, we observe that this procedure is again non-deterministic and may result in different factorization: the same row of **R** may represent a different topics depending on the initialization of the underlying optimization algorithm.Also, the top-10 terms of each token may differ depending on the initialization: e.g. the token "privaci" appears in some of the runs among the top-candidates for "Cryptography", whereas it is missing in some of the other runs.

Overall, there is a large overlap between each of the top-10 terms of the GKL NMF and the Gaussian NMF. The Frobenius norm of the four unordered components indicates that each component is again roughly equally important, for example (3.891174e-07, 2.939844e-08, 1.435971e-08, 7.097215e-09) for one particular run.With these results at hand, I find it hard to decide for or against any of the variants.

2 PLSA

The R script provides two functions, which factor the unnormalized output W and H of the NMF into three non-negative matrices:

a) nmf.lsr produces an m $\times$ r matrix $\mathbf{L}'$, an r $\times$ r diagonal matrix $\mathbf{\Sigma}'$, and an r $\times$ n matrix $\mathbf{R}'$ such that $\tilde{\mathbf{L}}\tilde{\mathbf{R}} = \mathbf{L}'\mathbf{\Sigma}'\mathbf{R}'$, and the columns of $\mathbf{L}'$ as well as the rows of $\mathbf{R}'$ sum to one.

b) nmf.slr produces an m $\times$ m diagonal matrix $\mathbf{\Sigma}''$, an m $\times$ r matrix $\mathbf{L}''$, and an r $\times$ n matrix $\mathbf{R}''$ such that $\tilde{\mathbf{L}}\tilde{\mathbf{R}} = \mathbf{\Sigma}''\mathbf{L}''\mathbf{R}''$, and the rows of $\mathbf{L}''$ as well as the rows of $\mathbf{R}''$ sum to one.

Run NMF with KL divergence and r = 4 and factor the resulting decomposition using each of the two functions nmf.lsr and nmf.slr. Study the result. Which information is contained in each of the three matrices? What can you say about the sum of the entries in each matrix? Can you give a probabilistic interpretation of the result (i.e., each entry (i,j) of each matrix)?

Let's recall first what the normalized matrix **P** represents, as this will help to better understand the resulting decompositions. **P** represents the full joint probability distribution of two discrete random variables (and hence its entries sum to one). One random variable represents the distribution of documents $p(D = d)$, the other the distribution of words $p(W = w)$. As such, an entry in **P** is the probability of seeing a particular value of both random variables at the same time $P(W = w, D = d) = p(d, w)$.

Both nmf.lsr and nmf.slr are trying to reconstruct this full joint probability distribution, and both additionally introduce for this purpose a discrete random variable representing the distribution of (latent) topics $p(Z = z)$. Also, both make the assumption that documents and words are conditionally independent given a topic which leaves us with this formulation of the full joint: $p(w, d) = \sum_z p(w, d, z) = \sum_z p(z) * p(d|z) * p(w|z)$. This makes intuitively sense: in order to obtain the full joint probability of seeing the combination of the i-th document and the j-th word, we have to combine all of their joint conditional probabilities, i.e. the probabilities for a given topic.

The last term $p(w|z)$ is covered in each of the two decompositions by the row-stochastic $\mathbf{R} =$

$\mathbf{R}' = \mathbf{R}''$ matrix. We can interpret an entry in a given row as in the following example: for the topic "Medicine", we expect to see the token "diseas" with a chance of 0.23%. Also, we might say that terms with a high value strongly contribute to a given topic.

This leaves us with two terms of the full joint: $p(z)$ and $p(d|z)$. $P(z)$ corresponds to the entries of the r-dimensional diagonal matrix $\mathbf{\Sigma}'$, where each entry is divided by the L1-norm of the matrix in order to produced a probability distribution. We can interpret each entry as the probability of a given topic. We know the ground truth and that each topic should occur with the same probability of 25%. Having rerun the code several times, I obtained different results which can be attributed to the non-uniqueness of the underlying NMF and different initializations. The probability I observed for any topic was in the range of 17-48%. I had yet to observe any probability vector close to the one expected and with 25% in each element. I assume that this indicates that topics are not monolithic, but consist of several subtopics, and that certain of these subtopics are shared among the 4 topics. Instead of evenly distributing these shared subtopics, they are in some factorization fully assigned to one topic which then again results in a probability substantially larger/smaller than the expected 25%.

$P(d|z)$ is depicted in nmf.lsr's column-stochastic $\mathbf{L}'$ matrix. Hence an entry in any of the $\mathbf{L}'$'s columns represents the probability that we see a document given a topic. A non-probabilistic and arguably more intuitive interpretation would be to interpret an entry in $\mathbf{L}'$ as the influence that a document has on a topic. Documents that have a low value in a given column of the left matrix contribute little to the topic, whereas documents with a high value contribute more.

Lastly, we focus on the second decomposition nmf.slr and its matrices $\mathbf{\Sigma}''$ and $\mathbf{L}''$. We analyzed the third matrix $\mathbf{R}''$ already above. Again, it makes sense to recall the overall objective of the factorization: we want to to reconstruct the full joint probability distribution $p(w,d) = \sum_z p(w,d,z) = \sum_z p(z) * p(d|z) * p(w|z)$. We might reorder the last equation's terms as $p(w,d) = p(z) * p(d|z) * \sum_z p(w|z)$. We observe that $p(z) * p(d|z) = p(z,d) = p(d) * p(z|d)$ by the product rule. This leads to the reformulation of the full joint of the i-th document and the j-th term as $p(w,d) = \sum_z p(w,d,z) = \sum_z p(d) * p(z|d) * p(w|z)$.

The first term $p(d)$ is covered by the diagonal matrix $\mathbf{\Sigma}''$ which sums to unity after normalizing it by its L1-norm. The probability vector for the documents is simply obtained by concatenating its diagonal entries. As above and in the case of $\mathbf{\Sigma}'$ it may make more sense not to interpret the entries as probabilities but as relative contributions a given document has to reconstruct the original matrix. This contribution varies substantially among documents. An interesting experiment would be to see, how this quantity changes if we include one or more duplicates of the document in the collection.

$P(z|d)$ corresponds to the 400*4-dimensional $\mathbf{L}''$, which is now row-stochastic. We can directly read off its rows to which topic we would assign a given document by selecting the column with the maximum value.

3 Clustering

The documents in the data came from four newsgroups. Your task is to cluster the documents in such a way that the clusters correspond to the newsgroups (which we can think of as topics). Note that you are not allowed to use the class labels during clustering, i.e., we pretend that we are in an unsupervised setting. To evaluate the quality of the clustering, we treat cluster identifiers as predicted labels and consider the accuracy (fraction of correctly predicted labels) and the confusion matrix. Examples can be found in the provided R code. Cluster the normalized newsgroup data into 4 clusters using each of the methods below and study the results. Also look at the clusters manually. Which clustering(s) perform well, which do not? Why?

a) k-means

b) k-means on $\mathbf{U}_4\mathbf{\Sigma}_4$ (i.e., the first two factor matrices of rank-4 truncated SVD)

c) k-means on the W matrix of the NMF (using KL divergence and r = 4)

d) k-Means on the $\mathbf{L}'$ matrix of factorization $\mathbf{L}'\mathbf{\Sigma}'\mathbf{R}'$

A naive baseline to assess the clustering with

four centroids, is the accuracy obtained by assigning all the documents to one single cluster except for randomly chosen 3 items that are being assigned to the remaining three clusters. As our data set is balanced, this is guaranteed to result in an accuracy of roughly 25%.

As all of the approaches a) - d) fail in a similar way and do not perform significantly better than the naive baseline, we discuss them jointly. Please note that we may obtain different clustering results depending on a different random initialization of the cluster centroids, as well as a result of the non-uniqueness of possible NMF factorizations for a given data set and loss function. This is of course not the case for b) as a matrix' SVD provides us with a unique solution.

If we apply the clustering a), c), d) (several times), the documents all are being assigned to one cluster, except for 1-3 documents that are being assigned to the remaining three topics (most of the time incorrectly). This behaviour strongly resembles the baseline approach described above. The accuracy I observed during different runs varied between 26-30%.

As the result of b) is deterministic if we fix a random seed, we have a look at the corresponding confusion matrix in figure 8.

	3	4	2	1
Space (3)	99	89	97	97
Chr. Rel. (4)	0	11	2	3
Med. (2)	0	0	1	0
Crypt. (1)	0	0	0	1

Figure 8: Confusion Matrix of K-Means on $\mathbf{U}_4\mathbf{\Sigma}_4$ (Acc.: 27.75%)

We see that almost all documents get assigned to the topic "Space". In the case of "Christian Religion", the clustering algorithm correctly identifies 11 documents and does not predict any other clusters. This is the only result I would consider meaningful in a)-d).

e) k-Means on the $\mathbf{L}''$ matrix of factorization $\mathbf{\Sigma}''\mathbf{L}''\mathbf{R}''$

Clustering the topics according to $P(z|d)$, which corresponds to the row-stochastic L", works very well. The best result I obtained during several runs is reported in figure 9 with an overall accuracy of 96%.

	1	4	3	2
Crypt. (1)	97	0	3	2
Chr. Rel. (4)	2	97	0	2
Space (3)	1	2	96	2
Med. (2)	0	1	1	94

Figure 9: Confusion Matrix of K-Means on $\mathbf{\Sigma}''\mathbf{L}''\mathbf{R}''$ (Acc.: 96%)

We see that this performance generalizes to all topics/classes having a comparable, high accuracy. If we investigate the rows of $\mathbf{L}''$, we observe that the factorization is in most cases very confident of a particular topic conditioned on the document, which serves as an explanation of the accurate clustering.

4 Beat the NMF

Experiment with preprocessing methods and alternative clustering methods to see if you can find a method that works better than the best result obtained in the previous task

The goal of this section is to arguably overfit on the data set, and correctly assign each document to its document. I fully succeed with the following approach. (1) Remove all terms that appear in all of the four topics. (2) Calculate relative frequencies ($p(d, w)$). (3) Run SVD. (4) Run k-means on $\mathbf{U}_4\mathbf{\Sigma}_4$.

What I consider the more important take-away from this exercise, is however the empirical nature of finding a suitable factorization and clustering. Different pre-processing approaches, different factorizations, non-uniqueness of certain factorizations (NMF), and different post-processing techniques (clustering algorithms) lead to different results that can only be reliably found via cross-validation and in an experimental setup. As such the non-deterministic results (NMF) presented in the remainder have to be considered with a critical mind.

I conducted three experiments and applied the clustering b), c), and e) from section 3. First, I transferred the data set from frequencies to Boolean matrix where each entry simply indicates the presence or absence of a term in a given document. Second, I filtered out all terms that occurred in all topic sections, thus remaining with a a vocabulary of 646 terms. Third, I combined the first

two approaches (first filtering, then Boolean masking). The results are shown in figure 10 which reports the accuracy for each of the approaches. Equal or better results are in bold.

	1	2	3
$\mathbf{U}_4\mathbf{\Sigma}_4$	0.41	**1**	0.41
$\mathbf{W}$	0.43	0.33	0.47
$\mathbf{L}''$	**0.96**	0.45	0.45

Figure 10: Clustering Performance of Different Factorizations and Preprocessing Approaches