



**HAL**  
open science

# Benchmarking seventeen clustering methods on a text dataset

Martine Cadot, Alain Lelu, Michel Zitt

► **To cite this version:**

| Martine Cadot, Alain Lelu, Michel Zitt. Benchmarking seventeen clustering methods on a text dataset.  
| [Research Report] LORIA. 2018. hal-01532894v6

**HAL Id: hal-01532894**

**<https://hal.science/hal-01532894v6>**

Submitted on 23 Apr 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# BENCHMARKING SEVENTEEN CLUSTERING METHODS ON A TEXT DATASET

Additional material for the article

*Bibliometric delineation of scientific fields*

*(Michel Zitt, Alain Lelu, Martine Cadot, Guillaume Cabanac)*

*published in*

*Springer Handbook of quantitative Science and Technology Indicators*

*W. Glänzel, H.F. Moed, U. Schmoch, M. Thelwall (Eds.), 2019*

Martine Cadot<sup>1</sup>, Alain Lelu, Michel Zitt

UPDATE 21/04/2019

---

<sup>1</sup> Corresponding author - Martine CADOT, LORIA, Equipe MULTISPEECH - <martine.cadot@wanadoo.fr>

# Table of Contents

1 - Introduction.....	3
2 - The test data.....	3
3 - Comparison criteria.....	5
4 - Overview of the methods.....	5
4.1 - Hierarchical methods for building clusters.....	5
a) Tool family 1: similarity measures.....	5
b) Tool family 2: aggregating method.....	7
c) The hierarchical methods under test.....	8
4.2 - The factor-oriented methods .....	9
4.3 - Hybrid factorial/clustering methods.....	10
a) Probabilistic Latent Semantic Analysis (pLSA).....	10
b) Latent Dirichlet Allocation (LDA).....	11
c) A fuzzy clustering method: Fuzzy C-Means (FCM).....	11
4.4 - Neighborhood methods.....	12
a) A density clustering method: DBSCAN [19][Ester et al. 1996].....	12
b) Graph clustering methods.....	12
Louvain.....	12
Affinity Propagation.....	12
Spectral Clustering.....	12
InfoMap.....	13
Density Peaks.....	13
5 - Experimental comparisons and conclusions: .....	13
5.1 - Deterministic methods.....	13
a) Correspondence Analysis.....	13
b) Hierarchical methods.....	14
c) DBSCAN.....	14
d) Louvain.....	14
e) InfoMap.....	14
f) Density Peaks.....	14
5.2 - Local optima methods:.....	15
5.3 - Voluntarily biased initializations and settings.....	16
5.4 - Tentative recommendations.....	16
5.5 - Perspectives.....	17
Table of the main results.....	17
6 - References:.....	20
Annex: raw data for establishing table 2 and result referred to in section 4.1.c.....	22

# 1 - Introduction

We examine a few classical methods of clustering texts in a corpus - whether direct or through the extraction of "communities" of words or documents derived from these texts - from a user perspective: that of researchers confronted with the delimitation of scientific domains out of standard bibliographic databases. We won't investigate corpuses of informal online discussions on fora, neither corpuses of SMSs nor of Web pages.

For each method, we limit ourselves to the following questions:

- Are their results reproducible? In the context of non-deterministic methods, the algorithms have an objective function giving an intrinsic measure of the results' quality. Does the optimization of this function bring them closer to the "ground truth", extrinsic by definition?
- Could they be considered a "gold standard" approaching the real structure present in the data?
- Especially, are they able to detect unbalanced structures, namely the coexistence of large classes and small classes (according to the terminology below)?

The ambitions are therefore limited. In the discussion section (see also the text of the chapter), some other relevant issues frequently addressed in benchmarking exercises are not addressed, such as computing complexity, memory efficacy, robustness in the presence of noise or fluctuations, linear vs. non-linear class separability.

We will indifferently call "cluster" or "topic" the grouping of elements (documents, terms, etc.) by those methods under test. In contrast "class" will designate the categories of items manually tagged upstream. Machine learning is the branch of data processing that seeks to generalize the attribution of these categories to documents, not included in a training set, that have not been subject to this costly human labeling. Up until now, the canonical process has only been used to a limited extent to solve the problem of delimitation, whereas core-periphery schemes for example are quite common. This might change in the future. Hence we will address the sole process of "clustering" (or "topic extraction") rather than the "classification" one.

We will also reserve the term "embedded mapping" for those methods which intrinsically incorporate this mapping process, namely factorial methods and Kohonen maps (a.k.a. SOM, Self-organizing maps). The maps obtained from the other methods which provide topics, are derived from a secondary process placing the topics in relation to each other, usually in two dimensions. Their *document X topics* output tables may become the input of bi-dimensional placement algorithm for example a Principal Component Analysis or a multidimensional scaling (MDS) representation of their topics columns.

## 2 - The test data

Among the corpora of public access texts written in standard English and labeled, we selected the one that came closest to our concerns i.e. Reuter's "21 578 news reports" [1][Lewis et al. 2004] in its refined version "ModApté Split" [2][Apté et al. 1994]. In order to make the results reproducible, and to avoid linguistic and/or statistical pre-processing, always difficult to specify unambiguously, we have opted for the documents X words matrix directly available to the public on the site associated to [3]

[Cai et al. 2005], even if the lexical processing is rather basic (no compound terms, all numbers considered as words ...). We chose to keep only the words of total occurrences greater than 15, to limit the requirements for memory space and computation time. This process resulted in a *documents X words* matrix of size 6829 X 3244, showing the occurrence count of each word in each document.

A particular challenge posed by these data lies in the imbalance between the class sizes assigned by the Reuters indexers. To limit the difficulty, and although it is common practice in the machine learning community to select the first ten classes, we will limit ourselves to the first six. The first two account for 84% of the corpus in terms of number of documents and the following four share equally the rest. Here are the sizes, concise titles and glimpses of the contents of these classes of documents (i.e. news reports):

- (3713) [earn]: investment opportunities.
- (2055) [acq]: corporate mergers and acquisitions.
- (321) [money fix]: exchange rates.
- (298) [crude]: crude oil prices.
- (245) [trade]: national and international trade.
- (197) [interest]: bank interest rates.

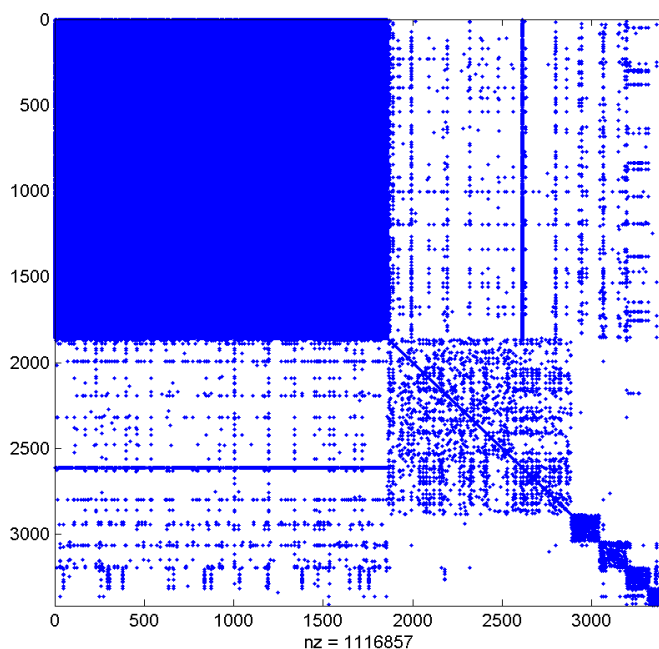


Figure 1: Cosines between Reuter's document vectors. Dark points represent cosines greater than 0.5. The order of the documents is that of the Reuter's classes. For the sake of legibility only one out of two documents has been represented.

The authority-based clustering reached by Reuters indexers – a distinct advantage of this data set – as a faithful representation of a real structure is corroborated by the cosine table between vector-documents, thresholded at 0.5 (visualization Figure 1) and the density table (Table 2): a very homogeneous big class [earn], another [acq] less dense and linked to the first, 3 small classes [money

fix] [crude] and [trade] homogeneous and related to [earn], but not to [acq], and a last small class [interest] related to [trade]. The disparities in size and density between classes are likely to be representative of many situations met in information related data, including retrieval and scientometrics.

We expect an efficient unsupervised method will retrieve by and large this structure.

	Cl.1	Cl.2	Cl.3	Cl.4	Cl.5	Cl.6
<b>Cl.1 : earn</b>	<b>.35</b>					
<b>Cl.2 : acq</b>	.07	<b>.13</b>				
<b>Cl.3 : money fix</b>	.06	.06	<b>.13</b>			
<b>Cl.4 : crude</b>	.04	.04	.05	<b>.18</b>		
<b>Cl.5 : trade</b>	.06	.05	.05	.06	<b>.15</b>	
<b>Cl.6 : interest</b>	.04	.06	.05	.05	<b>.11</b>	<b>.25</b>

*Table 1: intra and inter-class densities Reuter's Mode Apté split (6 classes)*

### 3 - Comparison criteria

To compare the extracted clusters with the reference classes, we chose two well-established partition comparison indicators, namely the Normalized Mutual Information (NMI)[4][Cover, Thomas 1991] and the Adjusted Rand Index (ARI) [5][Rand 1971] that require neither a one-to-one correspondence between clusters and classes, nor a strictly equal number of each. Their value is zero when the two partitions are independent or if one of them is the trivial partition (one class only). Their value is 1 when they are identical. The following will show differences in behavior between these two indicators.

### 4 - Overview of the methods

There are many methods for aggregating clusters, reviewed in a huge literature. For the sake of clarity we divide them into five categories: hierarchical, factorial, hybrid, probabilistic, and neighborhood methods. This list simplifies the categories mentioned in our Chapter: "density" and "graph clustering" are grouped into the "neighborhood methods" category; as for the K-means family, the member we consider most suited to text clustering has been included in the "hybrid" category.

#### 4.1 - Hierarchical methods for building clusters

Clusters are groups of objects which on defined criteria are more similar to each other than they are to objects in other groups. There are many traditional ways to build them, all of which require the choice of at least two tools: a similarity measure between two objects arising from their features on some criteria, and a method of grouping objects into subsets (the clusters) starting from the matrix of their pairwise similarities.

##### a) Tool family 1: similarity measures

For example, if the objects are the texts of a corpus and their features are the presence/absence of words out of a list of N words, the indices of similarity between two texts i and j are classically defined by a formula involving the four following counts:

- a, the number of words simultaneously present in the 2 texts
- b, the number of words present in the first text and lacking in the second
- c, the number of words present in the second text and lacking in the first
- d, the number of words simultaneously lacking in the 2 texts

We can set that  $a + b + c + d = N$ ,  $a + b = \text{occ}(i)$ ,  $a + c = \text{occ}(j)$  where  $\text{occ}(k)$  denotes the number of occurrences in  $k$ , i.e. the number of words in the text  $k$ .

Many similarity indexes have been proposed in the literature. In Table 2, we give the values of some of the most common ones for the associations of two documents (which can be considered as 2-itemsets) among four E, F, G and H specified in the Annex, where we note the presence/absence of  $N = 50$  words. The rank of the associations when ordered by decreasing values of each indicator is provided in red.

i	j	Occ (i)	Occ (j)	a	b	c	d	Support	MaxInc	Jaccard	Ochiai	$p$	Specializati		Simple
													on	Ochiai2	matching
E	F	20	10	5	15	5	25	5 (3)	0.50 (3)	0.20 (3)	0.35 (3)	1.25 (3)	0.22 (3)	0.26 (2)	30 (3)
E	G	20	30	9	11	21	9	9 (1)	0.45 (4)	0.22 (1)	0.37 (1)	0.75 (5)	-0.30 (5)	0.14 (5)	18 (5)
E	H	20	7	4	16	3	27	4 (4)	0.57 (2)	0.17 (5)	0.34 (5)	1.43 (2)	0.34 (2)	0.25 (3)	31 (2)
F	G	10	30	6	4	24	16	6 (2)	0.60 (1)	0.18 (4)	0.35 (4)	1.00 (4)	0	0.20 (4)	22 (4)
F	H	10	7	3	7	4	36	3 (5)	0.43 (5)	0.21 (2)	0.36 (2)	2.14 (1)	0.64 (1)	0.31 (1)	39 (1)
G	H	30	7	2	28	5	15	2 (6)	0.29 (6)	0.06 (6)	0.14 (6)	0.48 (6)	-0.60 (6)	0.07 (6)	17 (6)
Rank 1 2-itemsets								EG	FG	EG	EG	FH	FH	FH	FH
Rank 2 2-itemsets								FG	EH	FH	FH	EH	EH	EF	EH

**Table 2: values of some association indices between 2 document vectors out of 4 (i.e. E, F, G, H).**

- *Support*: the number  $a$  of cooccurrences of words between the 2 texts, which some authors normalize as  $a / N$ ;
- *MaxInc*:  $\text{Max}(a / (a + b); a / (a + c))$ , the of maximum ratio cooccurrences of the 2 texts to the occurrences of each text;
- *Jaccard*:  $a / (a + b + c)$
- *Ochiai*:  $a / \sqrt{(a + b) * (a + c)}$
- *Specialization*:  $(p^2 - 1) / (p^2 + 1)$ , where  $p = (a * N) / ((a + b) * (a + c))$  (= ratio to assumption of independence of rows and columns, transformed into similarity, see e.g. Grupp). Correlates: various specialization measures such as Balassa index, Grupp revealed advantage index, probabilistic affinity in square matrixes, used e.g. by (Zitt et al., 2000).
- *Ochiai2*:  $(a * d) / \sqrt{(a + b) * (a + c) * (b + d) * (c + d)}$
- *SimpleMatching*:  $a + d$ , or  $(a + d) / N$

We notice that the ranks differ depending on which of these nine indices is used. The last four indices ( $p$ , Specialization, Ochiai2 and SimpleMatching) are considered less "local" than the

others because of the importance they give to the number of words simultaneously lacking in the two texts, and criticized on this ground. The indices  $p$  and Specialization take into account a law of probability (here the Chi2 of independence). The Support and SimpleMatching consider only raw numbers, while the other indices introduce some normalization by a variable size. This is just a hint of the growing list (not far from a hundred indices at the present time) of these indices whose purpose is to quantify the link between two texts according to the presence/absence of words - for an extensive presentation see for example [6][Choi et al. 2010].

Till now we have limited our scope to binary features (1: presence of a word versus 0: absence), whereas one could consider numerical characteristics or "weights" (e.g. number of repetitions of each word in a text, TF-IDF or BM25/Okapi weighting, etc.). However, the abundance of similarity indices is not as important as in the binary case, as it is mostly limited to variants of the Bravais-Pearson correlation coefficient (i.e. Spearman correlation, biserial correlation) or to a transformation into similarities of the dissimilarities that the classical distances (city-bloc, Euclidian, etc.) implement. One can even have a combination of various types of features. Hence the choice for weighting the features in the formulas increases the variety of resulting multicriteria arbitrations. In complex cases, the recodification of quantitative features into binary ones often enlarges the scope of practicable methods, e.g. by taking into account non-linear effects.

## b) Tool family 2: aggregating method

To build hierarchies, one can proceed upward or downward. In an ascending way, one starts from the similarity (or dissimilarity) matrix between the  $p$  objects taken in pairs. In step 1, we consider each object as a group, and in the next step we merge the 2 groups with the smallest dissimilarity into a single group, giving  $p-1$  groups. The dissimilarities between the merged group and the remaining  $p-2$  groups are then calculated using a "link" formula combining the dissimilarities of each element of the new group with the remaining  $p-2$  groups. This merging step is repeated until a desired number of groups is reached, or else one sole group.

Updating dissimilarities at each merger can be done using one of various linkage formulas. For example, taking as a dissimilarity index the complement to one of the Ochiai2 index, along with the Max link, and limiting to two groups:

	<table style="border-collapse: collapse; text-align: center;"> <tr><th>E</th><th>F</th><th>G</th><th>H</th></tr> <tr><th>E</th><td>0</td><td>0.74</td><td>0.87</td><td>0.75</td></tr> <tr><th>F</th><td></td><td></td><td>0.80</td><td>0.69</td></tr> <tr><th>G</th><td></td><td></td><td></td><td>0.93</td></tr> <tr><th>H</th><td></td><td></td><td></td><td></td><td>0</td></tr> </table>	E	F	G	H	E	0	0.74	0.87	0.75	F			0.80	0.69	G				0.93	H					0		<table style="border-collapse: collapse; text-align: center;"> <tr><th>E</th><th>G</th><th>FH</th></tr> <tr><th>E</th><td>0</td><td>0.87</td><td><b>0.75</b></td></tr> <tr><th>G</th><td></td><td></td><td>0</td><td>0.93</td></tr> <tr><th>FH</th><td></td><td></td><td></td><td></td><td>0</td></tr> </table>	E	G	FH	E	0	0.87	<b>0.75</b>	G			0	0.93	FH					0	$=\text{Max}(\text{dis}(\text{EF}),\text{dis}(\text{EH}))=\text{Max}(0.74;0.75)$ $=\text{Max}(\text{dis}(\text{GF}),\text{dis}(\text{GH}))=\text{Max}(0.80;0.93)$
E	F	G	H																																												
E	0	0.74	0.87	0.75																																											
F			0.80	0.69																																											
G				0.93																																											
H					0																																										
E	G	FH																																													
E	0	0.87	<b>0.75</b>																																												
G			0	0.93																																											
FH					0																																										

**Table 3: Creating 2 clusters with the aggregation algorithm**

Two clusters (EFH) and (G) result from this process.

The order of aggregation can be strongly different amongst methods and distance metrics, with biases and shortcomings reported in an abundant literature (see the chapter's text). The chain effect undermining the efficient single linkage without specific control is well known. In our toy example, it is easy to check that the Jaccard metrics results in the (FH) and (EG) clusters, instead of (EFH) and (G), regardless of the aggregation method used.



## Note on the limitations of distance-based methods

Clustering methods, as well as factorial and hybrid methods, explicitly use distances, defined between pairs of words or documents, in the framework of linear algebra and matrix calculus. Probabilistic methods can also be expressed as matrix decompositions [14][Hoffman 1999], i.e. as hybrid methods implicitly using distances. In any case one must never forget that distance is a measure of global association between two complex entities. It ignores the information of interaction brought by the sequential character of the succession of words in a document, and more basically by the  $n$  to  $n$  non-ordered associations, called  $n$ -itemsets.

Clearly, the  $n$ -level itemset approach is more informative than the  $m$ -level if  $n > m$ . For example, triple combinaisons of terms are richer than co-occurrences. [30][Cadot, Lelu 2012] have shown that binary relations have to be complemented by interactions for reconstructing relations of all kinds. To illustrate this point, we have given in the Annex the counter-example of two *documents X words* raw tables, i.e. two distribution instances for 50 words in the 4 documents E, F, G, H producing the same binary association indices of table 2 but different supports for the 3-itemsets EFG and EFH. This advantage comes at the cost of computer requirements, which may be alleviated by usual means such as thresholding on low-information terms. Graph representations, being sets of binary relations, are also prone to this limitation, which may be overcome using hypergraphs.

To create "relief" and to partially introduce the interaction into the distance-based methods, it is possible to code all or part of the word  $n$ -grams as new variables,  $n$  being small. A "natural" and less cumbersome well-known solution is to take into account the compound expressions in full or in part, which amount to word  $n$ -grams selected by use. Needless to say, this method is blind to the many couples, triplets, etc. of non-consecutive words that co-occur in abstracts dealing with the same scientific theme.

Another possibility for introducing a dose of interaction is to operate a reference shift, i.e. to a data space greatly increased by new variables, functions of the original variables, for example by polynomial functions of these variables (in the case of binary coding). This is what the "kernel trick", well-known in the Machine Learning community, provides. It underlies the good performance of Support Vector Machines: there is no need to explicitly calculate the coordinates of the documents in the augmented space, as their similarity stems from a simple dot product using a pre-defined "kernel". The resulting matrix of similarity, called Gram matrix, can then be used by any clustering method, for example  $k$ -means in the case of Kernel  $K$ -means [29] [Girolami 2002]. But it is no longer possible to explain the "reason why" of the groupings - a black box effect admitted in the supervised learning context, but not always welcome in delineation exercises.

### c) The hierarchical methods under test

We have tested three methods corresponding to three types of aggregation, the first two quite classical in scientometric uses:

- Group average link: the average link between two clusters consists in computing the average distance between individuals of each cluster

- Ward D2: Ward distance aims to maximize inter-class inertia, via  $\text{dist}(C1, C2) = (n1 * n2 / (n1 + n2)) \text{dist}(G1, G2)$  where  $n1$  and  $n2$  are the sizes of clusters  $C1$  and  $C2$ ,  $G1$  and  $G2$  their respective centers of gravity.

- Mc Quitty: when examining the inter-cluster distance matrix, pairs of clusters in a reciprocal neighborhood situation (i.e.  $C1$  is the nearest neighbor of  $C2$ , while  $C2$  is the nearest neighbor of  $C1$ ) are merged.

## 4.2 - The factor-oriented methods

The usual factor-oriented methods (mainly PCA : *Principal Component Analysis*, CA : *Correspondence Analysis*, LSA: *Latent Semantic Analysis*) are based on the decomposition of a matrix of  $r$  rows (documents) and  $c$  columns (terms) into three smaller matrices, respectively of size  $(r, k)$  ( $k, k$ ) and  $(c, k)$ , where  $k$  is the number of factors extracted, called singular value decomposition (SVD), which is a basic operation of linear algebra:

$$X \sim U \Delta V'$$

$\Delta$  is a diagonal matrix, called matrix of the eigenvalues, and the matrices  $U$  and  $V$  ("eigenvectors") have orthogonal columns, with Euclidean norm one<sup>2</sup>.  $U$  and  $V$  represent the projections of the documents (or terms) on the  $k$  factor axes, which give rise to a direct visualization of these elements in two-dimensional maps, usually featuring the first two factors. While the English-speaking world often uses the "nonlinear unfolding" variant of this type of maps, namely the *Multidimensional Scaling* (MDS<sup>3</sup>), in Latin Europe and the Netherlands the *Correspondence Analysis* (CA) continues to stimulate theoretical interest<sup>4</sup> and practice since half a century. But these methods are suitable for data of small or medium size, and the ergonomics of planes collections for visualization exceeding two dimensions is inconvenient.

The *Latent Semantic Analysis* (also known as Latent Semantic Indexing) [Deerwester et al. 1988], i.e. direct application of the SVD to large *texts X words* matrices, breaks with any desire for visualization or individual interpretation of factors. It merely offers a space of "reduced" dimensions (generally a few hundreds factors for data-tables of minor dimension smaller than a few thousand elements) in which more relevant distances can be computed, than in the original space. In particular documents without common words but of the same semantic field are identified as close to one another.

An undeniable advantage of factorial methods is to be deterministic, i.e. to obey the "one data set, one method, one single result" principle.

*Independent Component Analysis* (ICA) [8][Hérault, Ans 1984] is widely used in signal processing, where it solves the so-called "cocktail party" problem (unravel  $n$  conversations from  $n$  microphones dispersed in the room). It imposes an independence constraint on the resulting components, more

---

<sup>2</sup>Older factor methods, mainly used by psychologists, have been somewhat forgotten in our "big data" era, some of which lead to oblique factors (see Varimax or Oblimax rotations, which maximize "simple structure", i.e. interpretability by minimizing the number of salient items for each factor). Principal Component Analysis (PCA), on the other hand, creates orthogonal factors and requires a matrix of centered-reduced variables, making it unsuitable for processing large amounts of data. More recent algorithms take advantage of the "sparse" nature of most of big data, to reduce both memory and computing power requirements.

<sup>3</sup>Based on a different principle: minimize "stress", which measures the global difference between the "true" distances in the multidimensional space and the distances in a 2D representation.

<sup>4</sup>The transformed array of which the AFC makes the SVD has links with the Laplacian graph we will mention below when addressing spectral clustering [7bis][Von Luxburg 2007].

stringent than the constraint of factor orthogonality required by most factor methods. It starts by transforming the data space into the space of the first eigenvectors, defined by the matrix documents  $X$  topics, all the columns of whose are of variance one. It is necessary to specify initially the number of dimensions of this "spherical" space, as well as the number of independent components that one wishes to obtain. Its *FastICA* variant [9][Hyvärinen 1999] is able to process *document X word* matrices, but is still little used in the field.

### 4.3 - Hybrid factorial/clustering methods

*Non-negative Matrix Factorization* (NMF) [10][Lee, Seung 1999] is based on the principle of avoiding the negative projections inherent to the factor analyses, in order to characterize the descriptors and the described objects by positive or null indicators of "salience". In image analysis for example, negative coefficients have no meaning. The common practice in NMF is to perform a decomposition into two matrices only  $X = HV$  where the sole columns of the matrix  $V$  are normalized, the matrix  $H$  being the equivalent to the product  $U \Delta$  seen above. The non-negativity constraint of  $H$  and  $V$  is respected thanks to a multiplicative update algorithm for  $V$  and  $H$ , unlike the factorial methods where the update corrections are additive. The results show a "clustering effect" with  $V$  and  $H$  values either close to zero or to the maximum. In fact, the axes defined by the columns of these matrices point to areas of high data density, and NMF can be considered as a clustering method producing fuzzy and overlapping clusters. The axes are usually oblique, forming angles less than or equal to 90°.

An older method, closer to clustering per se, is *Axial K-Means (AKM)* [11][Lelu 1994]. It leads to a similar result., i.e. positive "typicity" (or "centrality") coefficients in the 0 to 1 interval characterizing not only documents in a cluster (here explicitly defined), but also documents that are not part of it, which makes it possible to interpret this structure in terms of fuzzy and overlapping clusters. Each cluster axis is the principal axis, in the sense of Spherical Factor Analysis [12], of the sub-cloud representing this cluster on the surface of the unit sphere. Like all K-means-inspired algorithms, this method is fast, memory-sparing and suitable for large datasets.

These two methods are non-deterministic: they require, and are sensitive to, initialization values. Their algorithms allow them to converge towards local optima only. We will return to this problem below, from an experimental point of view.

In the neural-inspired model of *Self-Organizing Map (SOM)*, a geometric arrangement framework for the topics needs to be specified, in most of the cases in the form of a square or rectangular grid, in addition to the number of topics and to a random initialization seed. The advantage is to directly get a relevant 2D visualization of the topics in relation to each other. The disadvantage is that, depending on the initialization seed, disturbing edge effects may occur, especially if a central topic happens to be located at the perimeter of the grid. Different initialization seeds can create very different maps, in hardly recognizable configurations, although the main "strong forms" remain by and large recognizable from one map to another.

#### a) Probabilistic Latent Semantic Analysis (pLSA)

To overcome the indiscriminate nature of the LSA factors and the impossibility to interpret the extracted axes individually, while offering a statistical basis for this type of method, [14][Hoffman 1999] created the pLSA, based on the decomposition:

$$P(d,w) = \sum_z P(z) P(d/z) P(w/z)$$

where  $P(d, w)$  is the joint probability of the document  $d$  and the word  $w$  which models the number of occurrence of  $w$  in  $d$  divided by the total number of word occurrences in the corpus – one specific normalization of the data matrix –,  $P(d/z)$  is the conditional probability of the document given the value of the categorical variable  $z$  (for all the documents these probabilities sum to 1); same principle for the word  $w$ .  $P(z)$  is the probability of each category (or topic)  $z$ . The number  $Z$  of categories is fixed,  $D$  and  $W$  are respectively the numbers of documents and words.

This decomposition is expressed  $\mathbf{P} = \mathbf{U}\mathbf{D}\mathbf{V}^T$  in matrix notation, in accordance to the same scheme seen above, but with non-orthogonal columns for  $\mathbf{U}$  as well as  $\mathbf{V}$ . Each column of these matrices is interpreted as a "topic", in the same way as NMF.

Hoffman models  $P(d/z)$  and  $P(w/z)$  as multinomial laws, with a number of parameters<sup>5</sup>  $D.Z$  for the first,  $W.Z$  for the second – these are considerable numbers, and we will return to this fact below when dealing with the Influence of the initialization. Starting from an initialization at random of the matrices  $P(z)$ ,  $P(d/z)$ ,  $P(w/z)$ , he uses the Expectation Maximization (EM) algorithm [15] [Dempster et al. 1977] to converge, via updating the "stack" of intermediate matrices  $P(z/d, w)$ , to a local optimum of the objective function optimized by this algorithm, namely the log-likelihood [16][Fisher 1912] of the data with respect to the multinomial laws obtained at each iteration step. In contrast, linear algebra-based methods such as SVD optimize another criterion, the sum of the squared differences between reconstituted and original data, based on the concept of variance, not log-likelihood.

## b) Latent Dirichlet Allocation (LDA)

The LDA [Blei et al. 2003] was designed starting from a criticism of the statistical model of pLSA: the latter is based on a mixture of multinomial laws whose number of parameters linearly increases with the number of documents, which can thus grow indefinitely, unlike the number of words; this is called "overfitting"<sup>6</sup>. One would prefer to model the documents by a "generative" law, with few parameters. This is why LDA drastically reduces the number of parameters by modeling the  $Z$  distributions of documents conditionally to the topics, using a Dirichlet law - approximation, with  $Z$  parameters, of a set of  $Z$  multinomial distributions. The point of the  $W.Z$  parameters of words remains. Interestingly, the LDA has given way to many variants and refinements: taking into account the word  $N$ -grams, or subdivisions of texts, even dynamic aspects, ...

## c) A fuzzy clustering method: Fuzzy C-Means (FCM)

This is a historical method [18][Dunn 1973] that for each document produces a probability of belonging to each topic. These probabilities sum to 1, whereas for pLSA and LDA the probabilities of each document to be in one topic sum to one, and for NMF the squared typicality coefficients of each document belonging to one topic sum to one. The AKM method exhibits for each document indicators of centrality in the topic, within the range 0 to 1, with no constraint on their sums or sums of squared values - hence a possible interpretation as a fuzzy and overlapping cluster structure for documents whether belonging to the topic or not.

---

<sup>5</sup>The parameters of these multinomial laws consist of the probabilities of occurrence of each of the  $D$  categories for each of the  $Z$  subpopulations, for example to die of a certain cause when one is a man or one is a woman. In this case, there are  $D.Z$  parameters, some of which are contrasted (breast cancer, prostate cancer, etc.), and others not. They sum to 1 (alas!) for each subpopulation.

<sup>6</sup>We will see below what is concretely meant by the notion of overfitting, which is quite common in the context of supervised learning.

FCM needs the number of desired topics to be specified beforehand. The resulting structure optimizes a criterion based on the sum of these probability values at power  $\alpha$ , where  $\alpha$  is greater than 1 and also has to be initialized.

## 4.4 - Neighborhood methods

### a) A density clustering method: DBSCAN [19][Ester et al. 1996]

A radius  $R$  defines the neighborhood of a point (here a document), and the *minpts* parameter sets the minimum number of points in its neighborhood necessary for this point to be considered as the seed of a cluster (or its extension). Otherwise, it is considered as noise. The algorithm works by progressively extending the clusters. Its "naïve" implementation is very slow, but can be greatly accelerated by appropriate data structures.

One of the characteristics of this method is its ability to detect clusters of any shape, not necessarily linearly separable, as well as isolated points ("outliers") and border points between two clusters. This is a clear strength in some applications where contiguity is valued. In bibliometrics, experts are expected to prefer homogeneous aggregates, even slightly biased, easier to evaluate than elongated continuums. Two advantages: the method is deterministic and no number of clusters has to be fixed. Its two parameters are delicate to adjust, and have trouble to deal with clusters of different density.

### b) Graph clustering methods

#### ***Louvain***

This method [20][Blondel et al. 2008] optimizes a global "modularity" indicator for the graph, comparing to the graph with the same global distribution of the links, but whose values of the edges are computed under the assumption of the nodes being independent (Null model). It has the merit of having no parameter nor number of topics to adjust, but it is established since [22][Lancichinetti, Fortunato 2011] that its criterion of modularity embeds the problem of the "resolution limit": the more extended is the graph, the lesser the number of clusters - which is problematic in the context of Big Data in general and bibliometrics.

Several algorithms inspired by Louvain attempt to overcome this limitation by introducing a "resolution parameter", for example *Smart Local Moving Algorithm* [21][van Eck et al. 2010]. But it has also been shown in [22][Lancichinetti, Fortunato 2011] that these approaches also involve difficulties in taking into account clusters of different sizes and densities.

#### ***Affinity Propagation***

This method [Dueck, Frey 2007] relies on a "messages passing" principle, and successive updates of two matrices: one called "responsibility" quantifies the ability of each item to serve as an exemplary type, a "model", to each other; the second called "availability" quantifies the capacity of each item to take as exemplary another one, taking into account all the preferences for it. It does not impose to specify a number of clusters, but has a resolution parameter, called "preference".

#### ***Spectral Clustering***

This method [Meila, Shi 2000] is based on the use of K-means in a transformed space, that of the K first non trivial eigenvectors of the Laplacian matrix of the graph, the one that CA also relies on, see

[25] [Lelu, Cadot 2010]. In this latent semantic space of reduced dimensions clusters of any shape may emerge - which is not necessarily an advantage for bibliometric applications, as we have seen above. In addition to the need for specifying the number  $K$  of clusters, it is subject to the hazard of  $K$ -means initialization. For an answer to the question: what is the number  $K^*$  of clusters significantly present in the graph (significant in the statistical sense), and not constituted by noise, [26][Lelu, Cadot 2013] provides an answer based on Monte Carlo simulations.

### ***InfoMap***

This method [27][Rosvall, Bergstrom, 2007] is soundly grounded in information theory. It quantifies the "density landscape" of a directional (or not) graph by assigning binary codes to the nodes : the more often are they traversed by "random walkers" browsing through the graph, the shorter they are . As a result the description of the graph is compressed, and the graph is partitioned into fuzzy (or not) modules: each module is assigned a code in the same way, i.e. the more frequented by random walkers, the shorter. The shortest overall description of the graph provides the number and composition of the resulting clusters, along with the degree of centrality of each node in its cluster. The method does not require any parameterization nor specification of a desired number of clusters.

### ***Density Peaks***

The originality and interest of this method[28] [Rodriguez, Laio 2014] mainly lie, in addition to its deterministic nature, in the possibility offered to the user to choose the cluster seeds in a graph called "decision graph" giving for each entity its density and its minimum distance to a denser entity, thus resolving the problems of "rough" density landscapes in the process of detecting density peaks. Several methods for computing density are possible, and a resolution parameter called "neighbor rate" is necessary. It operates from the matrix of inter-entity distances. We will see below if this appealing supervision capability fulfils its promises.

## **5 - Experimental comparisons and conclusions:**

### **5.1 - Deterministic methods**

#### **a) Correspondence Analysis**

If we are looking for few topics, say six at most, a simple way to proceed is to perform a correspondence analysis of the *documents X words* matrix (it is a matrix of counts) limited to the first  $K$  non trivial eigenvectors, and then to deduce clusters by looking, at each document, for the axis on which its projection is of greater modulus and of the same sign as the majority<sup>7</sup>. This old classical method exhibits appealing features: one single, reproducible pass, easy interpretation via the projections of the words, possibility of dealing with important corpora as long as the vocabulary size does not exceed a few tens thousand words. In this way six factors were extracted from our test set in less than one second with a Pentium 6core i7, 3.33 GHz CPU, with honorable results : ARI = .39 and NMI = .41. We checked that this latter value was maximum for  $K = 6$ .

---

<sup>7</sup>The projections with largest module usually happen on the same side of the axis, either positive or negative. The Alceste method [Reinert 1986] is a more rigorous method for extracting a limited number of clusters in the same space of the first  $K$  factors of CA: it performs a greedy hierarchical descending partition on these axes.

## **b) Hierarchical methods**

To our surprise, these methods proved to be among the best in our benchmark, with Ward link at the top: the latter surpasses its competitors in terms of NMI (.53 vs. .51 for the best non-hierarchical method). McQuitty and Group Average respectively rank second and third in terms of NMI.

## **c) DBSCAN**

The choice of the two parameters "radius of the neighborhood" and "minimal number of neighbors" led us to many rounds. Many of them resulted in a "one dominant cluster plus a dust of quasi-individual cluster" structure, or in rejecting the majority of points as noise, except a few small clusters. The "least bad" compromise was found for  $R = 1.1$  and  $\text{minpts} = 5$ , where 63% of the corpus is spread over 3 clusters of unbalanced sizes. The resulting  $\text{ARI} = .46$  value has no significance since 37% of the corpus was poured into the common pot of rejected documents. DBSCAN's inability to take into account clusters of different sizes and densities is heavily penalizing, at least on datasets of this type.

## **d) Louvain**

This method of graph partition operates on an adjacency matrix between documents, which can be defined in many ways - here we chose the inter-document euclidean cosines. The cosine matrix between document vectors generated 3 clusters and an honorable NMI of .42 and ARI of .28. A cosine threshold at .1 divided the size of the data by a factor of more than two while producing an equivalent NMI of 0.42. A cosine threshold at .5 divided the size of the data by an order of magnitude but produced a mediocre ARI of 0.15 and four similarly-sized clusters. Though interesting as they arise from a method that does not require fixing a number of clusters nor any other parameter, these results confirm the often pointed disadvantage [27] of the modularity criterion optimized here, namely its "resolution limit": the larger the corpus, the lower its sensitivity to the existence of clusters, hence the small number of extracted clusters.

## **e) InfoMap**

Although this method introduces randomness at the deepest and lowest level of its algorithm (when generating a large number of random walks in a graph), it can be termed as stable and reproducible, in the sense that its best results - measured by the description length of the graph decomposed into modules - in about ten passages seem very close. As far as our test corpus is concerned, it resulted in 5 clusters of unbalanced sizes. The resulting value of the ARI indicator (.24) is only within the average range of all methods. But the value of the NMI indicator (.44) is in the high range, a result all the more remarkable for this method that it does not need, alone with Louvain, any parameter to adjust or number of clusters to specify. Here this method correctly individualizes the small class 5, and even splits the small class 6.

## **f) Density Peaks**

This method proved poor in dealing with clusters of different sizes and densities, and it was necessary to look for cluster primers that were not obvious to detect on the "decision graph", as well as to parameterize a very low "neighborhood rate" (0.08%) with the eventual disappointing values of ARI (.26) and NMI (.40).

## 5.2 - Local optima methods:

A common feature of NMF, KMA, pLSA, LDA, ICA and Spectral Clustering is to input a prerequisite number  $K$  of desired topics, and to output  $K$  vectors (values of the documents for each topic). CA does this too, in a framework of global optimization, and its vectors are orthogonal, which is not generally the case for the other methods, which could be gathered under the banner of "oblique factors methods". Although interpretations in terms of fuzzy and overlapping clusters are still possible from this general structure, it is easier in practice - and easier to comment on - to derive crisp clusters by simply identifying the maximum factor value for each document.

Each method in this family has its own objective function, to which state-of-the-art algorithms can only converge towards a local optimum. Hence the importance of initialization procedures for the quality of the final result. These always incorporate a random draw of initial values that directly or indirectly generate the *document X topics* output matrix. Unless the lines of code and the seed of initialization are specified, the results are not reproducible. It is necessary to reiterate the runs with different initialization seeds to approach the optimal values of the objective function, corresponding to partitions which may diverge a lot if the number of clusters desired is high (say, a few tens at least). Most often, one or two tens of repetitions are needed for reaching optimal values that peak at a few hundredths or thousandths near the best value of the objective function over a few thousand repetitions.

A major additional problem is that an optimum of an objective function does not necessarily correspond to a satisfactory partition according to our user criteria (or according to numerical criteria that would prove to be close). We could mention again NMI, which seems closer to our "natural" judgment criteria and is widely adopted in the recent literature. We have explored this problem with the LDA and KMA methods. LDA has provided the highest values of the ARI and NMI<sup>8</sup> criteria measuring the similarity between the obtained partition and the reference partition, over one or two tens passes. Table 3 below shows that the maximum ARI (or NMI) is far from coinciding with that obtained when the objective function is at its best value<sup>9</sup>. The chances of stumbling upon an initialization leading to a humanly satisfactory partitioning optimum are therefore quite low.

Affinity Propagation and Smart Local Moving Algorithm can also be placed in the same category as the methods with a fixed number  $K$  of clusters because they have one or two resolution parameters whose adjustment leads to the same result. Nor are they deterministic.

Compared to other methods, LDA proved the least flawed from this point of view. It results, at its minimum of perplexity, in a partition in three clusters, in practice two of which correspond more or less to classes 1 and 2, and the third to the four small remaining classes, mixed with elements of classes 1 and 2 - which is not really the desired result. The other methods do not behave better: KMA tends to create even more balanced clusters, and is the second "least bad" of the lot. Beyond these two methods yielding ARIs above .4, two other methods, NMF and ICA, produce ARIs around .3. Finally pLSA brings up the rear with a .13 ARI.

---

<sup>8</sup>After comparing all the methods presented here to the reference partition in Reuter's classes using the ARI criterion, we found that the NMI criterion, which varies in approximately the same direction, better reflected the similarities between clusters and small classes.

<sup>9</sup>Depending on the methods, a maximum of the objective function (pLSA, ICA, KMA) or a minimum (NMF, LDA) is sought.



### 5.3 - Voluntarily biased initializations and settings

In order to provide extra means of comparisons between these performances, it is interesting to know which maximum level of ARI can be obtained<sup>10</sup> over one or two dozen runs which, depending on the methods, can be initialized at random, or be subject to adjustment of their parameters: here too, the Group average link method clearly stands out with remarkable ARI and NMI values (resp. .71 and .64), optimal for a 10-cluster cut, which recreates a few small classes; Smart Local Moving Algorithm also stands out with ARI and NMI values (resp. .6019 and .5484) among the best in the whole test – but also low values when we do not adjust its parameters... Then comes LDA (.55), followed by ICA (.54) in the space of the first 10 eigenvectors, KMA (.48), and Spectral Clustering (.41) in the space of the first 7 eigenvectors. Eventually NMF, Fuzzy C-means and Affinity Propagation give values lower than .4.

All these elements show that

- 1) apart from Group average link, most of the objective functions of the tested methods are not suited to the proposed class structure, whose skewed distribution of cluster size is not likely to be exceptional.
- 2) only InfoMap and Louvain are getting close to the "true" structure of the data, with no parameters to adjust nor number of clusters to specify.
- 3) there may exist more appropriate objective functions, under a particular setting, as suggested by Smart Local Moving Algorithm, and therefore ample space for new research.. Research on these topics has been constantly active for more than half a century, and seems to be ever increasing and ever more complex (see the development of Deep Learning).

### 5.4 - Tentative recommendations

The present experiment on typical methods in the "mapping-clustering-community detection" continuum relies on a single dataset – which nevertheless exemplifies some typical difficulties encountered in practical applications. It supports the warnings such as [22][Lancichinetti, Fortunato 2011] that the current state of the art is unsatisfactory. It would be foolhardy to draw peremptory conclusions about the absolute superiority of a particular method. The comparison is especially difficult for methods requiring initialization and parameterization procedures. The case of deterministic methods of hierarchical clustering is somewhat different, since the tree is unique but asks for a cut-off parameter to obtain a partition under various constraints. Only InfoMap and Louvain do not require any of these elements, and Infomap is likely to provide plausible human-friendly categories at first glance, without the need for prior expert knowledge, but what about posterior confrontation with the understanding of experts and users involved in validation and appropriation of the outcomes? This remains an open question in the general case. We can say that the human categorization capacities have to do with the density fluctuations in a multidimensional data space, and that they are uncomfortable with too many categories. Experts and users may also favor some particular scale, or several scales, which are independent of the mathematical optimum of methods. Deterministic hierarchical methods have the advantage of reaching a global unique – though unstable when faced with data fluctuations – solution (the tree) where the cut-off may be sought by combining human decision on the approximate level (s) of scale judged desirable for analysis and discussion, and mathematical optimization of the cut-off.

---

<sup>10</sup> A case which cannot happen, by definition, in the framework of non-supervision, as ARI needs class labels.

While waiting for new methodological breakthroughs, InfoMap partition, on the one hand, can provide a sound basis that Fast Local Moving Algorithm, or LDA, together with expert knowledge of the domain, may correct and refine. Group average linkage, amongst hierarchical methods, remains appealing, alone or in combination – in the range of computing requirements – with a deterministic global outcome with eventual setting of the scale and local optimization.

As far as indicators are concerned, NMI seems the closest at hand to our judgment criteria, for it seems to take into account the correctness of reconstitution of small as well as large classes.

## 5.5 - Perspectives

We are aware that the work presented here has been realized in inevitable constraints of time and means. We will enjoy it to be continued by others, and will continue it ourselves:

- We insisted in presenting results obtained from publicly available data, results that anybody can verify or challenge. These data are issued from a single method of indexing the Reuters 21578 corpus. It will be interesting to examine the influence of other modes of indexing, whether more basic (stemming), or on the contrary more elaborate from a linguistic point of view - taking into account, for example, compound terms or the grammatical categories of words, not to mention 2- or 3- itemsets.

- We used public versions of the chosen algorithms, mainly written in Matlab, Octave, or Scilab code, or being published as Windows executables. It will be interesting to check if other implementations lead to the same results.

- Countless variants of the methods presented here have not been mentioned, especially kernel versions: testing the most promising ones will enrich the debate.

- We have deliberately measured the quality of unsupervised algorithms with a supervised learning methodology. We consider that this perspective is necessary, though not sufficient for the task of delimiting scientific fields - a task undertaken with the paradoxical injunction of detecting partially or totally hidden emergences, to the actors' eyes, while confirming or marginally correcting the apprehension that these scientific actors and institutions have of their own place. More specifically, delineation studies may include, with precaution, some predictive aspects, as recalled in the chapter's text. However, the detection of emergences cannot be fully covered with these techniques, especially when limiting the approach to texts. Rapidly growing citations on a particular topic, at a relatively small scale in the beginning, are amongst early warning signs which may anticipate creation of significant clusters at large scale.

## Table of the main results

For the ARI indicator (respectively NMI) :

a - the left column displays the intrinsic performances, without knowledge of the reference partition targeted,

b - the right one displays the opposite, therefore a voluntarily biased search for parameter(s) maximizing the similarity to this partition.

<i>Used algorithm</i>	<i>Comments</i>	<i>Objective function</i>	<i>ARI</i>		<i>NMI</i>	
			<i>a</i>	<i>b</i>	<i>a</i>	<i>b</i>
CA	for K=6 best ARI (for K=5)	Deterministic	.3934	.4052	.4082	.4072
CAH / Group average link	cutoff value for 6 clusters cutoff value for 10 clusters	Deterministic	<b>.6292</b>	.7097	.3270	.4440
CAH / Mc Quitty	cutoff value for 6 clusters cutoff value for 13 clusters	Deterministic	<b>.4979</b>	.5719	<b>.4639</b>	.5511
CAH / Ward D2	cutoff value for 6 clusters cutoff value for 12 clusters	Deterministic	.3176	.2371	<b>.5267</b>	.5267
DBSCAN	ARI <sub>max</sub> , for 6 clusters, $R=1.1$ , $minpts=5$	Deterministic		.4601		n.a.
NMF	best objective function (10 passes) for K=6 ,4480=NMI for best ARI	.7743 .7744	.3095	.3863	n.a.	.4480
AKM	best objective function (20 passes) for K=6 best ARI "ex-post" initialization	.2614 .2558 .2573	.4349 [.8265]	.4795	<b>.5079</b> [.6410]	n.a.
pLSA	random initialization for K=6 "ex-post" initialization best around "ex-post" initialization	-22 173 -14 690 -14 384	.1344 [1.000] [.6768]		.0673 [1.000] [.6357]	
LDA	best objective function (=min, 20 pass.) for K=6 best ARI "ex-post" initialization	503.80 523.50 480.63	<b>.4625</b> [.6184]	.5460	.4052 [.7333]	.5345
ICA (=ACI)	best objective function (for 7 eigenvectors) best ARI (for 10 eigenvectors)		.2818	.5390	n.a.	n.a.
Fuzzy C-Means (FCM)	best ARI ( $power=1.082$ ) for K=6	4300.90		.3337		n.a.
Louvain	COS (filling ratio : 99.66%) COS threshold: 0.1 (filling ratio : 43.30%) COS threshold: 0.5 (filling ratio : 9.58%)		.2794 .2750 .1506		.4212 .4230 n.a.	

Spectral clustering	best ARI (for 7 eigenvectors)	N.C.	.4178		n.a.
Affinity Propagation	forr 6 clusters, <i>Preference</i> =-18.2		.1955		n.a.
Smart Local Moving Algorithm	best ARI (for <i>Resolution</i> =1000, <i>minpts</i> =30)		<b>.6019</b>		<b>.5484</b>
InfoMap	COS threshold: 0.1		.2420		.4359
Density Peaks	best ARI (with Gaussian kernel and <i>neighbor rate</i> = 0.08%)		.2624		.4018

**Table 4 of the main results**

## 6 - References:

- [1][Lewis et al. 2004] Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr), 361-397. <http://www.daviddlewis.com/resources/testcollections/rcv1/>
- [2] [Apté et al. 1994] Apté, Chidanand, Fred Damerau, and Sholom M. Weiss. "Automated learning of decision rules for text categorization." *ACM Transactions on Information Systems (TOIS)* 12.3 (1994): 233-251.
- [3][Cai et al. 2005] Cai, Deng, Xiaofei He, and Jiawei Han. "Document clustering using locality preserving indexing." *IEEE Transactions on Knowledge and Data Engineering* 17.12 (2005): 1624-1637. <http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html>
- [4][Cover, Thomas 1991] Cover, Thomas M., and Thomas, Joy A. "Entropy, relative entropy and mutual information." *Elements of information theory2* (1991): 1-55.
- [5][Rand 1971] Rand, William M. "Objective criteria for the evaluation of clustering methods." *Journal of the American Statistical association* 66.336 (1971): 846-850.
- [Zitt et al. 2000] Zitt, Michel, Elise Bassecoulard, and Yoshiko Okubo. "Shadows of the past in international cooperation: Collaboration profiles of the top five producers of science." *Scientometrics* 47.3 (2000): 627-657.
- [6][Choi et al. 2010] Choi, Seung-Seok, Sung-Hyuk Cha, and Charles C. Tappert. "A survey of binary similarity and distance measures." *Journal of Systemics, Cybernetics and Informatics* 8.1 (2010): 43-48.
- [7] [Deerwester et al. 1988] Scott Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, Beck L.: Improving information retrieval with latent semantic indexing, Proceedings of the 51st Annual Meeting of the American Society for Information Science 25, 36–40 (1988)
- [7bis][Von Luxburg 2007] Von Luxburg, Ulrike. "A tutorial on spectral clustering." *Statistics and computing* 17.4 (2007): 395-416.
- [8][Hérault, Ans 1984] J. Hérault and B. Ans, "Réseau de neurones à synapses modifiables: décodage de messages sensoriels composites par apprentissage non supervisé et permanent", *Comptes Rendus de l'Académie des Sciences Paris, série 3*, 299: 525-528, 1984.
- [9][Hyvärinen 1999] Hyvärinen, Aapo. "Fast and robust fixed-point algorithms for independent component analysis." *IEEE Transactions on Neural Networks* 10.3 (1999): 626-634.
- [10][Lee, Seung 1999] Lee, Daniel D., and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." *Nature* 401.6755 (1999): 788-791.
- [11][Lelu 1994] Alain Lelu: Clusters *and* factors: Neural algorithms for a novel representation of huge and highly multidimensional data sets. In: *New Approaches in Classification and Data Analysis*, ed. by Edwin Diday, Yves Lechevallier, Martin Schader, Patrice Bertrand (Springer, Berlin 1994) pp.241–248
- [12][Domengès, Volle 1979] Domengès, Dominique, and Michel Volle. "Analyse factorielle sphérique: une exploration." *Annales de l'INSEE*. Institut national de la statistique et des études économiques, 1979.
- [13][Kohonen 1998] Kohonen, Teuvo. "The self-organizing map." *Neurocomputing* 21.1 (1998): 1-6.

- [14][Hofman 1999] Thomas Hofmann: Probabilistic latent semantic indexing, SIGIR'99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, Berkeley, CA 1999, ed. by Fredric Gey, Marti Hearst, Richard Tong (ACM, New York, NY 1999) 50–57
- [15][Dempster et al. 1977] A.P. Dempster, N.M. Laird et Donald Rubin, « Maximum Likelihood from Incomplete Data via the EM Algorithm », *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no 1, 1977, p. 1–38
- [16][Fisher 1912] Ronald Fisher, « On an absolute criterion for fitting frequency curves », *Messenger of Mathematics*, no 41, 1912, p. 155-160
- [17][Blei et al. 2003] David M. Blei, Andrew Y. Ng, Michael I. Jordan: Latent Dirichlet Allocation, *Journal of Machine Learning Research* 3, 993–1022 (2003)
- [18][Dunn 1973] Dunn, J. C. (1973-01-01). "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters". *Journal of Cybernetics*. 3 (3): 32–57
- [19][Ester et al. 1996] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu: A densitybased algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise, KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, OR 1996, ed. by Evangelos Simoudis, Jiawei Han, Usama Fayyad (AAAI, Palo Alto 1996) 226–231
- [20][Blondel et al. 2008] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre: Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment* 2008(10), P10008 (2008)
- [21][van Eck et al. 2010] Nees Jan van Eck, Ludo Waltman: Software survey: VOSviewer, a computer program for bibliometric mapping, *Scientometrics* 84(2), 523–538 (2010)
- [22][Lancichinetti, Fortunato 2011] Lancichinetti, Andrea, and Santo Fortunato. "Limits of modularity maximization in community detection." *Physical review E* 84.6 (2011): 066122.
- [23][Dueck, Frey 2007] Frey, Brendan J., and Delbert Dueck. "Clustering by passing messages between data points." *Science* 315.5814 (2007): 972-976.
- [24] [Meila, Shi 2000] Marina Meila, Jianbo Shi: Learning Segmentation by Random Walks, NIPS'00: Proceedings of the Neural Information Processing Systems Conference, Denver, CO 2000, ed. by Todd K. Leen, Thomas G. Dietterich, Volker Tresp (MIT Press, Cambridge, MA 2000) 873–879
- [25] [Lelu, Cadot 2010] Alain Lelu, Martine Cadot. Espace intrinsèque d'un graphe et recherche de communautés. Frédéric Amblard. Première conférence sur les Modèles et l'Analyse des Réseaux : Approches Mathématique et Informatique - MARAMI 2010, Oct 2010, Toulouse, France. pp.1, 2010. HAL Id: hal-00516865
- [26][Lelu, Cadot 2013] Alain Lelu, Martine Cadot. A Proposition for Fixing the Dimensionality of a Laplacian Low-rank Approximation of any Binary Data-matrix. The Fifth International Conference on Information, Process, and Knowledge Management - eKNOW 2013, Feb 2013, Nice, France. IARIA, pp.70-73, 2013. ⟨hal-00773436⟩
- [27][Rosvall, Bergstrom 2007] Martin Rosvall, Carl T. Bergstrom: An information-theoretic framework for resolving community structure in complex networks, *Proceedings of the National Academy of Sciences* 104(18), 7327–7331 (2007)

[28][Rodriguez, Laio 2014] A. Rodriguez, A. Laio: Clustering by fast search and find of density peaks, *Science* 344(6191), 1492–1496 (2014)

[29][Girolami 2002] M. Girolami. Mercer kernel-based clustering in feature space. *IEEE Transactions on Neural Networks*, 13(3):780–784, 2002.

[30][Cadot, Lelu 2012] Martine Cadot, Alain Lelu. Combining Explicitness and Classifying Performance via MIDOVA Lossless Representation for Qualitative Datasets. *International Journal On Advances in Software, IARIA*, 2012, 5 (1&2), pp.1-16. [⟨hal-00596718⟩](#)

[31][Benzécri 1973] Jean-Paul Benzécri: *L'analyse des correspondances*, *Analyse des données*, Vol.2 (Dunod, Paris 1973)

[32][Robertson et al. 1994] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline HancockBeaulieu, Mike Gatford: *Okapi at TREC-3, TREC'94: Proceedings of the 3rd Text REtrieval Conference*, Gaithersburg, MA 1994, ed. by Donna K. Harman (NIST, Gaithersburg, MA 1994) 109–126

[33][Reinert 1986] Max Reinert: *Un logiciel d'analyse lexicale*, *Les cahiers de l'analyse des données* 11(4), 471–481 (1986)

## **Annex: raw data for establishing table 2 and result referred to in section 4.1.c**

The following two tables display the presence of 50 words in the 4 documents E, F, G and H mentioned in table 2. The last line displays the supports of the 1-itemsets E, F, G, H, of the 2- itemsets EF , EG , ..., GH, as well as for the 3- itemsets and 4- itemset EFGH. The point is that the sole supports of EFG and EFH differ in the two tables: {4, 0} and {1, 1} respectively. The other supports are identical. This counter-example shows that the only binary relations between entities are not enough to completely establish their links.

itemset	E	F	G	H	EF	EG	EH	FG	FH	GH	EFG	EFH	EGH	FGH	EFGH
#word															
1	1	1	1	0	1	1	0	1	0	0	1	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
9	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0
10	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0
11	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0
12	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0
13	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0
14	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0
15	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0
16	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0
17	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	1	1	1	0	0	0	1	1	1	0	0	0	1	0
22	0	1	1	1	0	0	0	1	1	1	0	0	0	1	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
30	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
31	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
32	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
33	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
34	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
35	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
36	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
37	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
38	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
39	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
40	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
41	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
42	1	1	1	0	1	1	0	1	0	0	1	0	0	0	0
43	1	1	1	0	1	1	0	1	0	0	1	0	0	0	0
44	1	1	1	0	1	1	0	1	0	0	1	0	0	0	0
45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
48	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0
49	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0
50	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0
Support	20	10	30	7	5	9	4	6	3	2	<b>4</b>	<b>0</b>	0	2	0



itemset	E	F	G	H	EF	EG	EH	FG	FH	GH	EFG	EFH	EGH	FGH	EFGH
#word															
1	1	1	1	0	1	1	0	1	0	0	1	0	0	0	0
2	1	1	0	1	1	0	1	0	1	0	0	1	0	0	0
3	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0
4	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0
5	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0
6	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0
7	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0
8	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0
9	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0
10	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0
11	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0
12	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0
13	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0
14	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0
15	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0
16	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0
17	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	1	1	1	0	0	0	1	1	1	0	0	0	1	0
22	0	1	1	1	0	0	0	1	1	1	0	0	0	1	0
23	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0
24	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0
25	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0
26	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
30	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
31	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
32	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
33	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
34	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
35	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
36	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
37	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
38	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
39	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
40	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
41	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
49	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Support	20	10	30	7	5	9	4	6	3	2	1	1	0	2	0