In « Proceeding of the JADT 18 Conference » D.F.Iezzi et al. (eds). UniversItalia, Roma, 436-443.

Looking for *topics*: a brief review

Ludovic Lebart[1]

[1]Telecom-ParisTech – ludovic@lebart.fr

Abstract

This paper presents a brief review of several endeavors to identify latent variables (axes or clusters). When dealing with textual data, these latent variables (clusters or axes) are sometimes designated *ex ante* by the term "topic". The first attempts to identify interpretable latent variables dates back to factor analysis at the beginning of last century. Recent years have witnessed a series of algorithmic attempts such as non-negative matrix factorization (NMF) or Latent Dirichlet Allocation (LDA). In the meantime, latent variables are also identified through several hybridizations and synergies of principal axes methods and clustering techniques. A single medium-sized classical corpus (Shakespeare's 154 Sonnets) will serve as a benchmark to sketch and compare in a compact way some characteristic features of several methods.

Keywords: Topic Modelling, NMF, LDA, Correspondence Analysis, Factor Analysis, Clustering.

1. Introduction

There is a profusion of new disciplines around the industrial applications involving texts, with subsequent proliferations of tools and disparities of terminologies. There are also disparities in the attitude towards the texts, sometimes influenced by the availability and the user-friendliness of software. The problems entailed by huge sets of newsgroup or tweets are quite different from those encountered when dealing with literature, political discourses, psychological surveys. Because they are well known, translated in almost every language, deeply studied and commented, we will use Shakespeare's Sonnets as a benchmark to briefly compare the ability of several techniques to recognize topics in a corpus.

2. An outline of the contents of Shakespeare's sonnets

The 154 sonnets of William Shakespeare deal with themes such as love, friendship, effects of time, beauty, treason, lust, death. Note that the definition of topics in Text Mining is a pragmatic one, and may also recover the concepts of theme and motif.

2.1. Theme, Topic, Motif

Usually, a topic is an objective explanation of the subject matter, whereas a theme represents the deeper underlying message. A motif is simply a recurring idea or pattern used to reinforce the main theme. Schematically, topics answer the questions:, "What's the story about? Who? What? How?" and themes answer: "Why was the story written?". Topics in literature are easier to identify than themes.

Three main contiguous series of sonnets are generally recognized as three dominant themes:

Sonnets 1 to 17: (*Procreation*). These sonnets celebrate the beauty of a young man who is urged by the poet to marry so as to perpetuate that beauty.

Sonnets 18 to 126: (*Young man*). This longest sequence concerns the same young man (not definitively identified), the destructive effect of time, the force of love, friendship and poetry.

Sonnets 127 to 154: (*Dark Lady*).These sonnets are mostly addressed to a dark haired woman, not without some irony and cynicism (the two last sonnets 153 and 154 are specific epigrams in an ancient style; they should deserve in fact a specific category).

2.2. Eight themes derived from expert commentaries

The themes *Young man* and *Dark lady* could contain five sub-themes. While the first theme (*Procreation*) remains untouched, the new *Young man* and *Dark lady* themes will comprise only those sonnets which were not assigned to the five new categories below (*Absence, Storm, Rivalry, Death, Eternal poetry*).

Table 1. List of eight a priori *themes/topics with the corresponding sonnets numbers*

Procreation	1 - 17
YoungMan	20-25, 33-38, 40-42, 46, 47, 49, 53-55, 59-60,62-70, 75-77, 88-106, 108-112, 115-125,
DarkLady	127-136, 139, 140, 143-146, 153,154
Absence	26-32, 39, 43-45, 48, 50-52, 56-58, 61, 113-114
Storm	141,142,147-152
Rivalry	78-87
Death	71-74
Etern_poetry	18,19,81

The partition of sonnets given in Table 1 is inspired by the works of Alden (1913) and Paterson (2010) but not explicitly mentioned by these authors. Figure 1 shows however that, after a blind correspondence analysis ignoring these themes, most of their locations are statistically significant on the principal plane of visualization.

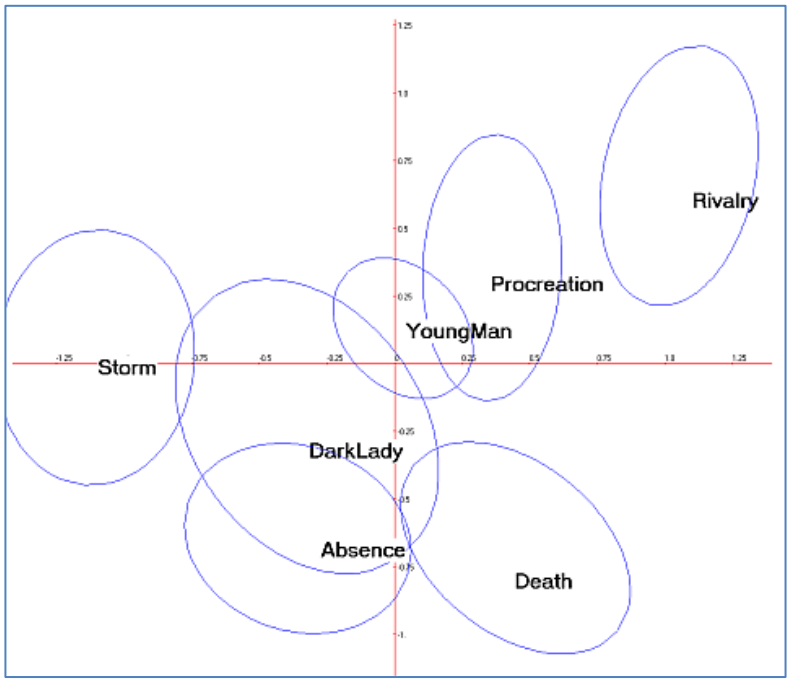

Figure 1. Locations of 7 themes/topics in the principal plane of the correspondence analysis of the lexical table (154 sonnets x 173 words, [min. frequency = 10]) as supplementary categorical variables. Conservative bootstrap confidence ellipses [drawing with replacement of sonnets] show the significant distances between several pairs of a priori *themes. Note that the theme "Eternal Poetry", too much overlapping with others, is missing in this graphical display.*

Evidently, the following attempts to find topics into the corpus of sonnets will ignore that *a priori* partition into themes. We do not expect either to retrieve automatically these themes.

However, the knowledge of these themes issued from literary criticism can provide us with a template for reading and interpreting the results more easily. Note that statistical tools mainly based on frequencies detect almost indifferently topics, themes or motifs.

3. Six selected methods for topic research

Among the six procedures selected in the present paper, four (RFA, FCA, LOA, LSA) make use of the Singular Values Decomposition (SVD). The remaining two methods (NMF and LDA), less geometrical, involve a specific model and more complex algorithms.

RFA (Rotated Factor Analysis) is historically the first attempt to identify unobserved "latent factors" (Thurstone, 1947, after the pioneering papers of Spearman,1904, and of Garnet, 1919). RFA involves SVD in one of the most popular algorithms known as Principal Factor Analysis. In this case, the *topics* are the words characterizing each kept factors. Initially conceived for numerical values, it could be adapted to sparse frequency tables. [**R** packages 'psych' and 'GPArotation'].

FCA (Fragmented Correspondence Analysis), in the vein of *ALCESTE* methodology (Reinert, 1986), is based on the CA of fragments of texts [in our case 7 consecutive lines, i.e.: half a sonnet), cf. Lebart (2012). The principal axes of CA serve to cluster these fragments (hybrid clustering using Hierarchical Classification –Ward criterion – and k-means). At the end of the process, the *topics* are defined by the series of words that characterize each cluster (software 'DtmVic').

LOA (LOgarithmic Analysis) (Kazmierczak, 1985) is similar to Spectral Mapping (Lewi, 1976) thanks to a difference of weighting. Both methods, like CA, comply with the principle of distributional equivalence (stability of the results vis-à-vis fusions of similar columns or rows). Applied to contingency or frequency tables, LOA often produces results similar to those of CA, with less sensitivity towards outliers as a consequence of the logarithmic shrinkage. A clustering of sonnets (similar to that of FCA) is then performed. The *topics* are then the words characterizing each cluster (software: 'DtmVic').

LSA (Latent Semantic Analysis (or Indexing), Deerwester *et al.*, 1990) which is basically a SVD of the matrix of Tf.Idf coefficients (Term frequency x Inverse of document frequency). A clustering of sonnets (similar to that of FCA and LOA) is then performed. The *topics* are then the words characterizing each cluster. (**R** package 'lsa' [Fridolin Wild] and 'DtmVic')

In the domain of text analysis, the two following methods belong more specifically to the field of "Topic Modelling".

NMF (non-negative matrix factorization) starts with an equation that reminds Singular Values Decomposition (SVD): Decomposition of a data matrix **A** as the product of two matrices of lower rank, **B** and **C**: **A** = **B** **C**. The marked difference lies in a constraint of positivity of the coefficients of **B** and **C** (those of **A** being already supposed > 0) (Lee and Seung, 1999; Berry *et al*, 2007; after Paatero &Tapper, 1994. See also Gaujoux, 2010). In the topic modeling context, the main output of NMF is a set of topics characterized by list of words (software 'scikit-learn' [Python] by Grisel O., Buitinck L., Yau C.K; In: Pedregosa *et al.*, 2011).

LDA (Latent Dirichlet Allocation) (Blei et al., 2003; Griffiths et al., 2007) is a generative statistical model (involving unobserved topics, words, and document) devised to uncover the underlying semantic structure of a collection of texts (documents, supposed to be a mixture of a small number of topics). The method is based on a hierarchical Bayesian analysis of the texts. (package **R**: 'topicmodels', and software 'scikit-learn' [Python]).

At this stage, we have limited our investigation to six techniques out of a great number of approaches likely to identify topics. Among these approaches let us mention the direct use of CA without fragmentation of the texts, the techniques of clustering (used in FCA and LOA) which contain many more methods and variants, the already mentioned *Alceste* methodology (Reinert, 1986). The present piece of research evidently needs to be extended. In fact, each method involves also a series of parameters (threshold of frequency for the words; preprocessing options such as lemmatization/stop words; size of fragments or context units, number of iterations). The following experiment limited to six methods will be tersely summarized. A thorough investigation would need many more pages.

4. Excerpts from the list of 49 topics (limited to two topics per method)

The number of topics detected by each of the six selected methods varies between six and ten. Only two topics are printed below for each method.

4.1 Rotated Factor Analysis (Rotation Oblimin). *(2 topics out of 6)*

RFA1 eyes see bright lies best form say days
RFA2 beauty false old face black now truth seem

4.2 FCA (Fragmented Correspondence Analysis) *(2 topics out of 7)*

FCA1 beauty truth muse age youth praise old eyes glass long seen lies false time days
FCA2 night day bright see look sight

4.3 Logarithmic Analysis (Spectral mapping) *(2 topics out of 8)*

LOA1 summer away youth sweet state hand seen age rich beauty time hold nature death
LOA2 pen decay men live earth verse muse once life hours make give gentle death

4.4 Latent Semantic Analysis *(2 topics out of 8)*

LSA1 time heart beauty more one eyes eye now myself art still sweet world
LSA2 end grace leave words lie spirit change shame self could ever decay write

4.5 NMF topics *(2 topics out of 10)*

NMF0: love true new hate sweet dear say prove lest things best like ill let know fair soul
NMF1: beauty fair praise art eyes old days truth sweet false summer nature brow black live

4.6 Latent Dirichlet Allocation LDA *(2 topics out of 10)*

LDA0 summer worse praise nature making time like increase flower let copy rich year die
LDA1 sing sweets summer hear love music eyes bear single confounds prove shade eternal

5. A synthesis of produced topics

How to compare the complete lists of topics, since neither the order of topics, nor the order of words within a topic are meaningful? We deal here with real 'bags of words' exemplified by the excerpts of lines in section 4. We will add the eight *a priori* themes defined in table 1. Each *a priori* theme corresponds to a subset of sonnets. That subset will be described by its characteristics words. We can then perform a clustering of these 57 topics/themes (49 + 8). The technique of additive trees (Sattath and Tversky, 1977; Huson and Bryant, 2006) seems to be the most powerful tool for synthesizing in compact form these 57 topics/themes (figure 2). Let us recall one important property of additive trees: the real distance between two points can be read directly on the tree as the shortest path between the two points.
Ideally, we expect to find a tree with as many branches as there are real topics in the corpus, each branch of the additive tree being characterized by seven labels: six labels corresponding to the six methods briefly described above, plus one label corresponding to one *a priori*

theme. Such situation occurs when each method has uncovered the same real topics. The observed configuration is not that good, but we can distinguish between six and nine main branches, which is probably the order of magnitude of the number of different topics. We note also that several different methods often participate in the same branch, which suggest that that branch correspond to a real topic discovered by almost all the six methods. Let us mention that a similar additive tree performed on the 49 topics (not involving the eight *a priori* themes) produces approximately the same branches. Thus, the eight *a priori* themes can be considered here as illustrative elements, serving only as potential identifiers of the branches.

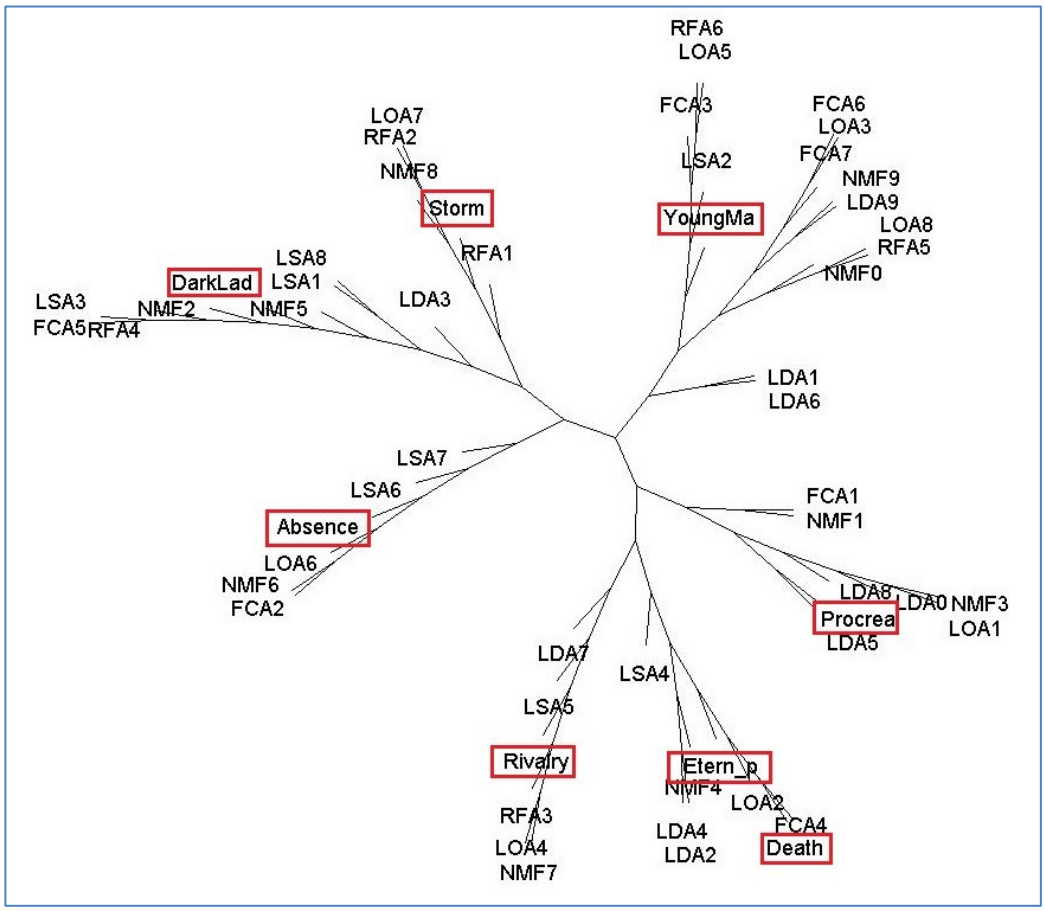

Figure 2. Additive Tree describing the links between the 49 topics provided by the 6 selected methods and the 8 a priori themes. The identifiers are those of section 4 for the 6 selected methods. The 3 first letters indicate the method, followed by the index of the produced topic. The distance between two topics is the chi-square distance between their lexical profiles. Threshold of frequencies for words: 2. The boxed identifiers of the a priori themes are those (possibly shortened) of table 1.

It is remarkable that the eight *a priori* themes (boxed labels) are well distributed over the whole of Figure 2. If we except the branch of the tree located in the upper right part of the display, on the right of the label "Young man", all the main branches have as a counterpart one of the *a priori* themes. As an example of interpretation of figure 2, the branch in the lower center part of figure 2: [NMF7, LOA4, RFA3, LDA7, LSA5] is clearly linked to the *a priori* topic named *Rivalry* (see section 2.2) (concurrence of five methods out of six). Most of the branches of the additive tree could be interpreted likewise. The upper right branch identified by none of the *a priori* themes may represent an unforeseen topic. More research and an

expertise in Elizabethan poetry are required to confirm that we are dealing here with an undetected new theme. To conclude, we can only observe that each of the involved method, be it ancient or modern, may contribute to detect topics… and that exploratory tools are essential to visualize the complexity of the process and assess the obtained results.

References

Alden, R. M. (1913). *Sonnets and a Lover's Complaint*. New York: Macmillan.

Berry M.W., Browne M., Langville Amy N., Pauca V.P., and Plemmons R.J. (2007). "Algorithms and applications for approximate nonnegative matrix factorization". In: *Computational Statistics & Data Analysis* 52.1: 155-173.

Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3: 993—1022.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K.and Harshman R. (1990). Indexing by latent semantic analysis, *J. of the Amer. Soc. for Information Science,* 41 (6): 391-407.

Garnett J.-C. (1919). General ability, cleverness and purpose. *British J. of Psych,* 9, 345-366.

Griffiths T.,L., Steyvers M., and Tenenbaum J.,B. (2007). Topics in Semantic Representation. *Psychological Review*, 114, 2, 211-244.

Huson D. H., Bryant D. (2006) Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution,* 23 (2): 254 - 267. Software available from www.splitstree.org.

Kazmierczak J.-B. (1985). Analyse logarithmique : deux exemples d'application. *Revue de Statist. Appl.*, 33, (1), 13-24.

Lee D.D. and Seung H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature,* 401: 788-791.

Lebart L. (2012). Articulation entre exploration et inférence. In : *JADT_2012*. Dister A., Longree D., Purnelle G., Editors. Presse Universitaire de Liège.

Lewi P.J. (1976). Spectral mapping, a technique for classifying biological activity profiles of chemical compounds. Arzneim. Forsch. in: *Drug Res.* 26, 1295-1300.

Paterson D. (2010). *Reading Shakespeare Sonnets.* Faber & Faber Ltd. London.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot M. and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research* , 12, 2825-2830.

Reinert, M. (1986). Un logiciel d'analyse lexicale: [ALCESTE]. *Cahiers de l'Analyse des Données*, 4, 471–484.

Sattath S. and Tversky A. (1977). Additive similarity trees. *Psychometrika,* vol. (42), 3: 319-345.

Shakespeare, W. (1901). *Poems and sonnets: Booklover's Edition*. Ed. The University Society and Israel Gollancz. New York: University Society Press. Shakespeare Online. Dec. 2017.

Spearman C. (1904). General intelligence, objectively determined and measured. *Amer. Journal of Psychology,* 15, 201-293.

Gaujoux R.*et al.* (2010). A flexible R package for nonnegative matrix factorization. In: BMC Bioinformatics 11.1 (2010): 367.

Thurstone L. L. (1947). *Multiple Factor Analysis.* The Univ. of Chicago Press, Chicago.