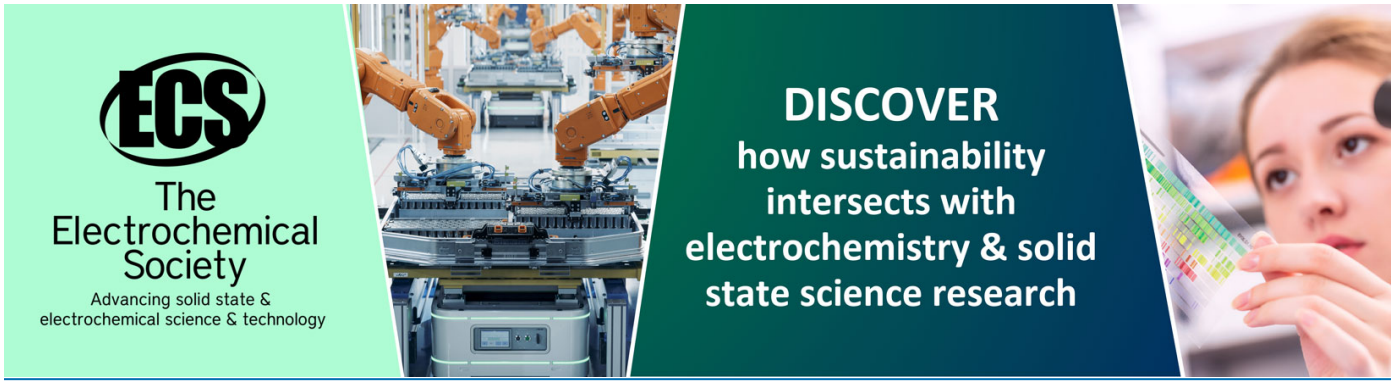
What is this painting about? Experiments on Unsupervised Keyphrases Extraction algorithms



IOP Conf. Series: Materials Science and Engineering **364** (2018) 012050 doi:10.1088/1757-899X/364/1/012050

What is this painting about? Experiments on Unsupervised Keyphrases Extraction algorithms

M T Artese, and I Gagliardi

IMATI – CNR, Via Bassini 15, 20133 Milan, Italy
email: {teresa, isabella}@mi.imati.cnr.it

corrisponding author email: isabella@mi.imati.cnr.it

Abstract. A large number of cultural heritage archives are freely available on the web: they can be in Linked Open data format, or in any other format, such as databases, collections or archives, with some information for each object. To be really enjoyed and enjoyable by the users on the web, a set of scored keywords need to be associated with each item, manually or automatically. The overall problem here addressed is the automatic, unsupervised extraction of keywords/keyphrases from the items of cultural heritage archives, in different languages (English and Italian). The problem is very actual and in literature many papers are devoted to this topic and several approaches have been defined: we present here a work-in-progress, an experimentation done with the aim of automatically associating scored keywords/keyphrases to a painting archive. We have therefore tested five different methods present in literature, such as tf-idf, RAKE, TextRank, …, on two datasets, in English and in Italian, and evaluated the results - using recall and precision@n as the evaluation metrics.

1. Introduction

A large number of cultural heritage archives are freely available on the web: they can be in Linked Open data format, or in any other format, such as databases, collections or archives with pictures and some textual information for each object. To be really enjoyed and enjoyable on the web, users should be able to interact and navigate with simple and easy-to-use tools, using both the pictorial and textual information. Multimedia systems on the web, usually, provide facilities for searching, browsing, clustering and visualizing different kinds of visual data and related information. Many of their innovative tools are related to image searching and visualization functionalities in the context of cultural heritage web sites. To be effectively and efficiently searched, retrieval, and browsed, textual data need to be assigned (scored) keywords/keyphrases, manually or automatically.

Keywords / key phrases are single words or groups of words that characterize the content of the document and are useful for identifying the documents relevant to a given query and / or for "suggesting" something in some way related.

Keywords/keyphrases should, therefore, satisfy the following characteristics:

- have a good coverage: in that, all the keywords/keyphrases of an item are able to describe all its aspects;
- be relevant: general enough to represent more than a single item, as well as specific not to represent the whole set of items.

As well known, manually associating keywords to documents requires time, skills, specialized personnel, and more steps to ensure that the chosen terms are consistent, adequate, relevant, with a good coverage, sufficiently general and timely. To automate this activity, much work has been done over the years and researchers have explored both supervised and unsupervised techniques to address the problem of automatic keyphrase extraction. The problem is still very actual and in literature, many papers are devoted to this topic and different approaches have been defined. Furthermore, most of the studies and experiments have been conducted on texts in English, while there are few experiments concerning other languages, such as Italian.

In this paper we will consider the problem of automatic unsupervised extraction of keywords on texts, by presenting a work-in-progress, an experimentation done on a corpus of texts (in English and in Italian) related to paintings, to be integrated into a Multimedia Information system on the web.

We have applied five methods/algorithms present in literature, such as TextRank, RAKE, tf-idf, … to texts in English and Italian, to better understand the different characteristics of these methods and their use in a cultural heritage context in different languages, and present here the results, evaluating them in the context of our experimentation.

The paper is organized as follows: after the presentation of the overview of works related to automatic unsupervised keyword extraction, the data used and the experimentation runs are presented, together with the details of the solutions tested in the runs. The results obtained are then presented, compared and discussed.

2. State-of-the Art: Unsupervised Keyphrases Extraction algorithms

Although the problem of extracting keywords, keyphrases able to represent the content of pieces of text in natural language has been faced since the first Information Retrieval systems appeared, the advent of new tools and techniques makes it still very actual: in literature, many papers are devoted to this topic [1][2][3][4] and several approaches have been defined [5][6][7][8][9].

In the following, we present different methods that have shown to work well and have been tested in different domains, and that we have used in our experimentation.

2.1. TextRank

TextRank [7] is a graph-based ranking model for text processing, and is based on the same assumptions of PageRank [10][11], that is: i) important pages are linked by important pages, and ii) the PR values of a page is essentially based on the probability of a user visiting that page.

TextRank models a text document as a graph where each word (or sentence, according to granularity chosen) in the document is represented as a node, and the semantic relation between each node is represented by an edge. A scoring algorithm will then be run on the graph to assess the value of every word, which will be used to rank them, then top ranking words are selected as keywords. Keywords that are adjacent in the document are combined to form multi-word keywords. The authors [7] report that TextRank achieves its best performance when only nouns and adjectives are selected as potential keywords.

2.2. RAKE

Rapid Automatic Keyword Extraction (RAKE) [8] is an unsupervised, domain-independent, and language-independent method for extracting keywords from individual documents.

The algorithm is composed of the following steps: i) it extracts a set of candidate keywords (single words or sequences of contiguous words), ii) assigns a score to these candidate keywords, based on the

co-occurrence graph: the score is based on the frequency of the words and their degree, iii) adjoins keywords and iv) extracts the top T scoring keywords.

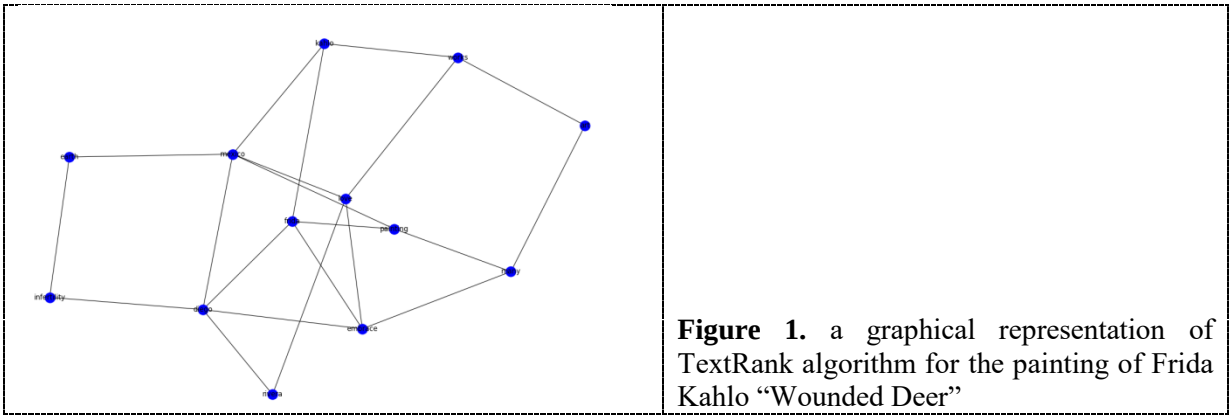


Figure 1. a graphical representation of TextRank algorithm for the painting of Frida Kahlo "Wounded Deer"

2.3. Tf-idf

Tf–idf [6][12], short for term frequency–inverse document frequency, is widely used in Information Retrieval and is used to measure the importance of a word in a document, within a collection or corpus. At a high level, a tf-idf score finds the words that have the highest ratio of occurring in the current document vs the frequency of occurring in the larger set of documents.

2.4. Latent Semantic Analysis or Latent Semantic Indexing

LSI [13] is a technique able of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms.

LSI assumes that words that are close in meaning will occur in similar pieces of text. A term-document matrix, as that created by Tf-idf model, is the starting point for the technique used to reduce the number of rows in the matrix, while preserving the similarity structure among columns, thus converting terms and documents to points in a lower-dimension space.

2.5. Other methods

In literature, many other methods have been proposed and tested as, among others, SingleRank [14] and ExpandRank [15], that are essentially a TextRank expansion, or cluster-based approaches as KeyCluster [16].

3. Dataset

For our experimentations, we started from the 'paintings 91' [17][18] database, which consists of well-known paintings by famous artists ranging from Vermeer to Chagall, from Bosch to Matisse. The data set consists of 4,214 paintings from 91 different artists; is composed of images, and some basic information such as author, artist categorization, and style classification for each painting. No title nor painting description is available in the original data. We first associated to each painting its title (in Italian or in English), and then, in a completely automated way, we retrieved from the web a description of the painting, whose provenience is mainly from Wikipedia or the first result of a google search based on title + author query, in Italian and/or in English. To ensure that the webpages retrieved are semantically related to the paintings, a similarity score based on the co-occurrence of words in the web page title and painting title has been computed and assigned.

Table 1. Characteristics of Dataset used.

Dataset	**No. items**	**Type of data**	**Language**	**Average length, max, min**
Painting91 (original dataset)	4214	Authors, artist categorization, style	Title in Italian or English	-

		classification and title		
English	3794	Authors, Title and content	English	6348 chars (max 25926, min 500)
Italian	3088	Authors, Title and content	Italian	4047 chars (max 19436, min 502
Other languages	718	-	German, French, Dutch, Polish, …	-

Characteristics of the data:

- language: while data are available in different languages, such as German, French, Dutch, Polish, in this paper we present the results only for English and Italian data;
- length: informative content of each item in English language has an average length of 6348 characters, varying from 500 to almost 26000; while in Italian, texts are shorter, with an average of 4000 characters, (from 500 to less than 20000); in both languages, contents of less than 500 lengths were discarded;
- topics: all data refers to paintings, from Durer to Winslow, from Tintoretto to Boccioni, from Botticelli to Salvador Dalì, with a time span from 1400 to 2000, and with baroque, surrealism, pop art styles.

4. The experiments

With the aim of testing these methods and of identifying if (and, in case, which) method produces better results on these data with a set of test queries, and if there is any noticeable difference between English and Italian, a standard Information Retrieval approach [6] has been followed, shown in Fig. 2.

In the processing phase, the following steps, rather common, were carried out, with the aim to identify the most relevant keywords, to be used as an aid to perform a query and innovative ways to navigate the data:

- Pre-process data: remove stopwords from the text, perform stemming, tokenize, … ,
- Extract a list of candidate keywords /keyphrases using some heuristics,
- Score each candidate keywords/keyphrases, according to different criteria and methods,
- Select the first m keywords/keyphrases.

We run the following Unsupervised Keyphrases Extraction algorithms, described in section 2:

- TextRank
- Tf-idf
- RAKE
- Latent Semantic Indexing (LSI)

and the following python implementation:

- Automatic Keyword Extraction (AKE) (using gensim python package [21]): this method, based on the TextRank algorithm, has been applied “as is” with the default values as parameters, that is windows size = 2 and considering only nouns and adjectives. Because in this implementation pre-processing and stemming algorithms only works for English, we have disabled them, using texts as presented, in both languages.

We tested Tf-idf both using 1-gram and n-grams basis for word extraction; for LSI we performed tests with a different number of topics: here we report the results for 20 and 100 topics.

Some algorithms have to be performed document by document (TextRank, RAKE, AKE), while others have to be run on the whole collection of documents (tf-idf, LSI).

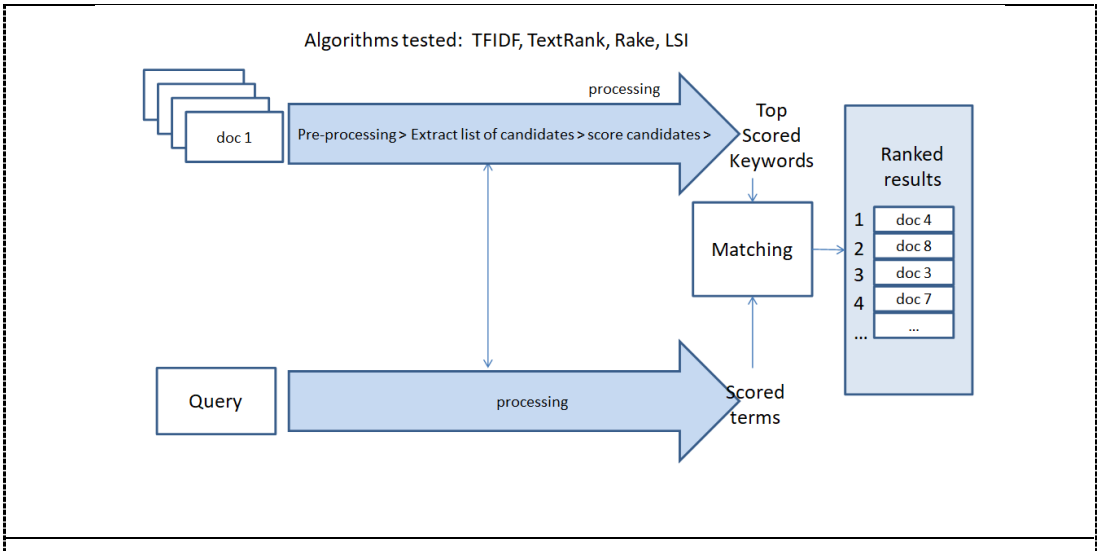


Figure 2. Information Retrieval architecture of our approach.

According to the diagram in Figure 2, the keywords / key phrases automatically extracted are then exploited for indexing documents, in a vector space model and the cosine similarity is the chosen metric used to retrieve and rank documents in response to the query.

Each run on an algorithm/dataset (English/Italian) gives rise to a different set of (weighted) keywords/keyphrases, associate to each document: so that the retrieved document list can be different both in the documents retrieved and in the order of the retrieval.

Table 2 shows information and some statistics on the results obtained from the processing phase, applying the 5 algorithms on Italian and English datasets, with tf-idf runs for 1-gram and n-grams, and LSI with 20 topics and 100 topics. In this phase the same parameters were applied to the 2 datasets, modifying, where necessary, stopword lists and stemming algorithms.

Table 2. keyphrases stats for the 5 algorithms tested.

Algorithm	No. different keyphrases	*Average length of keyphrases*	No. keyphrases with more than 10 occurrences (keyphrases10)	*Average length of keyphrases10*	No. keyphrases with more than 100 occurrences (keyphrases100)	*Average length of keyphrases100*
TextRank						
English	27.439	2,011	2237	1,581	133	1,083
Italian	14.674	1,755	958	1,280	63	1,015
Tf-idf						
English 1-gram	11.781	1	663	1	0	-
English n-grams	15.931	2,398	550	1,627	0	-
Italian 1-gram	11.599	1	486	1	0	-
Italian n-grams	16.254	1,755	1,28	1,280	1	2,0
RAKE	s					
English	8.511	1,963	549	1,828	16	1,375
Italian	6.300	1,838	405	1,476	15	1,133
LSI 20 topics						
English	98	1	89	1	39	1
Italian	98	1	95	1	21	1
LSI 100 topics						
English	314	1	286	1	125	1

Italian	330	1	320	1	66	1
AKE						
English	13362	1,106	1882	1,026	137	1
Italian	15562	1,355	1406	1,102	1333	1

Some considerations about the results:

- The number of different keyphrases varies greatly on the basis of the algorithm tested: from 98 of LSI with 20 topics to over 27.000 of TextRank.
- Except for LSI and AKE, the other algorithms consistently give lower values for Italian than English, due to the lower number of documents and shorter length of texts. Higher values in AKE are related to the lack of appropriate stemming algorithms for Italian;
- TextRank produces a relatively large number of keywords, using the recommended values: setting a threshold value, as the maximum number of keywords extracted per document or the weight of the nodes must be further tested and evaluated;
- TextRank, RAKE and AKE extract keyphrases longer than 1. Both in English and Italian. We tested tf-idf both for 1-grams and n-grams: in the latter case, the average length of the words extracted is more than 2 for English, and it is around 1.7 for Italian.
- LSI basically works as a clustering algorithm, associating each cluster with a set of keywords that represent it. The results reported were obtained by choosing 20 and 100 as the number of topics.

We tested the implemented algorithms on the two datasets, using the queries of Table 3. Since these methods and algorithms have already been tested and have proven to work well, in order to evaluate the results for our goal, i.e. the integration into a Multimedia Information system, we have assessed the results in terms of recall and precision. Recall is the fraction of relevant instances that have been retrieved over the total amount of relevant instances and its value is 1 when all the relevant instances are retrieved, decreasing up to 0 when none relevant instance is retrieved. Precision is the fraction of relevant instances among the retrieved instances, and the its value varies from 1 to 0, with 1 if only relevant documents are returned, and 0 if none relevant document is retrieved. In this paper, we use P@n, precision measured on the first n topmost results returned, with n = 10, 20 and 100.

Table 3. Queries used to test the algorithms

	Queries
English	art, artist, painter, braque, circle, man, people, landscape, nature, vermeer, green, museum, lissitzky, woman, rome, portrait, death, italian, paintings, mother, women, black, Picasso, Pablo Picasso, new york
Italian	arte, artista, pittore, braque, cerchio, uomo, persone, paesaggio, natura, vermeer, verde, museo, lissitzky, donna, roma, ritratto, morte, italiano, quadri, madre, donne, nero, picasso, Pablo Picasso, new york

TextRank, RAKE and AKE, another implementation of TextRank, index each document only with top ranked terms and when a query is performed, the matching algorithm retrieves and scores only those documents that have in common terms with the query; the other documents are scored 0; on the other hand tf-idf and (in some measure) LSI algorithms give a weight to each term in the document: thus in essence, documents are sorted according to the weight of the terms in the query.

For this reason, the accuracy of TextRank, RAKE and AKE is always very high and around 1, thus confirming the high quality of the algorithm results. On the other hand, recall values can be low because only those documents to which the searched terms have been associated are retrieved.

For the query 'woman', the dataset contains 149 paintings with woman in the title: TextRank retrieves 288 paintings, RAKE 52, and AKE 148. Rake has thus a recall value of 0,5 at the best.

Query 'Picasso': in the dataset are present 48 paintings by Pablo Picasso, for English TextRank retrieves 150 items, Rake 25 and AKE 73; and for Italian, the results are almost the same. Also in this case the recall of RAKE is about 0,5. No noticeable change for query 'Pablo Picasso'. In general, it can be noted that AKE results, for English, are comparable to TextRank, while for Italian, due to the lack of appropriate stemming algorithm, are less performing.

We evaluate the precision only for tf-idf and LSI. Tf-idf and LSI 100 give precision@10 and P@20 value almost close to 1, with lower precision on 100 items. Tests on LSI 20 have shown that 20 clusters for these 4000 items are too few: in fact, LSI 20 inserts documents that are very different from each other in the same clusters, so it fails to extract the keywords correctly.

Precision does not vary much comparing results for tests on Italian and English data sets, because the algorithms work on a single language at a time, thus identifying structure and words, adopting different stoplists when they process the two datasets.

The experimental setup has been implemented in Python 2.7, using NLTK, gensim, newspaper, ski-kit packages, together with some experimental packages in GitHub [23].

5. Conclusion and Future Works

We have presented here a work in progress for the automatic unsupervised extraction of keywords/keyphrases for a corpus of texts (in English and in Italian) related to paintings, to be integrated into a Multimedia Information system on the web. Five algorithms present in literature have been tested on two different datasets and the results of the experimentation have been reported here, to evaluate which algorithm works best, in terms of recall and precision, with respect to query tests. Preliminary results show that TextRank, RAKE methods have been proven to obtain the best values for precision, always around 1, while the recall, especially for RAKE is rather low. AKE results, for English are comparable to TextRank, while for Italian, due to the lack of appropriate stemming algorithm, are less performing. Tf-idf and LSI 100 provides good results, while LSI 20, in these tests fails to extract correct keywords.

Concluding, we can say that TextRank, tf-idf and LSI are the algorithms that provide the best results; TextRank and tf-idf do not need any adaptation, while, for LSI, the choice of the number of clusters is decisive and several tests may be necessary to obtain the best results.

The results of queries on English and Italian datasets perform almost the same, although English benefits from more studies, experimentations, and algorithms. Further tests, with real users, will be performed once these tools will be integrated in the Multimedia Information system on the web.

The results have shown that further works or investigations should be oriented towards:

- although some algorithms should automatically identify synonyms and homonyms, WordNet [19][20][24] could be used to further reinforce these aspects;
- test the algorithms using some controlled set of keyphrases against which match our results as the title of Wikipedia[25] pages or title of books or journal paper;
- test the algorithms on other multimedia archives;
- test other algorithms, such as for example SingleRank, ExpandRank or Keycluster.

References

[1] Hasan, K. S., & Ng, V. (2010, August). Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (pp. 365-373). Association for Computational Linguistics

[2] Hasan, K. S., & Ng, V. (2014, June). Automatic Keyphrase Extraction: A Survey of the State of the Art. In ACL (1) (pp. 1262-1273)

[3] Merrouni, Z. A., Frikh, B., & Ouhbi, B. (2016, October). Automatic keyphrase extraction: An overview of the state of the art. In Information Science and Technology (CiSt), 2016 4th IEEE

International Colloquium on (pp. 306-313). IEEE.
[4] Siddiqi, S., & Sharan, A. (2015). Keyword and keyphrase extraction techniques: a literature review. International Journal of Computer Applications, 109(2).
[5] Gollapalli, S. D., Caragea, C., Li, X., & Giles, C. L. (2015). Proceedings of the ACL 2015 Workshop on Novel Computational Approaches to Keyphrase Extraction. In Proceedings of the ACL 2015 Workshop on Novel Computational Approaches to Keyphrase Extraction.
[6] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA.
[7] Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing.
[8] Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. Text Mining: Applications and Theory, 1-20.
[9] Hulth, A., & Megyesi, B. B. (2006, July). A study on automatically extracted keywords in text categorization. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (pp. 537-544). Association for Computational Linguistics.
[10] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Stanford InfoLab.
[11] Brin, S., & Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks*, *56*(18), 3825-3833.
[12] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), 513-523.
[13] Landauer, T. K. (2006). Latent semantic analysis. John Wiley & Sons, Ltd.
[14] Wan, X., & Xiao, J. (2008, July). Single Document Keyphrase Extraction Using Neighborhood Knowledge. In *AAAI* (Vol. 8, pp. 855-860).
[15] Wan, X., & Xiao, J. (2008, August). CollabRank: towards a collaborative approach to single-document keyphrase extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1* (pp. 969-976). Association for Computational Linguistics.
[16] Liu, Z., Li, P., Zheng, Y., & Sun, M. (2009, August). Clustering to find exemplar terms for keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1* (pp. 257-266). Association for Computational Linguistics.
[17] Khan, F. S., Beigpour, S., Van de Weijer, J., & Felsberg, M. (2014). Painting-91: a large scale database for computational painting categorization. Machine vision and applications, 25(6), 1385-1397.
[18] Artese, M. T., Ciocca, G., & Gagliardi, I. (2017). Evaluating perceptual visual attributes in social and cultural heritage web sites. Journal of Cultural Heritage.
[19] Fellbaum, C. (1998). *WordNet*. John Wiley & Sons, Inc..
[20] Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.

Websites (last consulted on 16 January 2018)
[21] Gensim topic modelling for humans https://radimrehurek.com/gensim/
[22] Newspaper: News, full-text, and article metadata extraction in Python
https://github.com/codelucas/newspaper
[23] github https://github.com/
[24] WordNet: https://wordnet.princeton.edu/
[25] Wikipedia http://www.wikipedia.org