Modelling Text Semantics

Márius Šajgalík*
Institute of Informatics, Information Systems and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovičova 2, 842 16 Bratislava, Slovakia
marius.sajgalik@stuba.sk

Abstract

In the dissertation, we focus on modelling text semantics. We identify two sub-goals, which aims at modelling abstract text semantics. While the first sub-goal is oriented on modelling the general text semantics, the second sub-goal is focused on the discriminative semantics, which can be of more information value. Besides proposing new methods to fulfil these sub-goals, we also examine a practical application of our proposed method of discriminative keyword extraction.

Our contribution can be split into three parts. First, we propose a method to model abstract text semantics via key-concepts and show how it improves over standard keyword extraction methods. As a second contribution, we propose a method to model discriminate abstract text semantics, which is based on categorised text documents. We show how better representation of text semantics can improve over state-of-the-art methods in text categorisation even with traditional keywords. Finally, we propose an approach to modelling user interests using our method of discriminative keyword extraction, which is evaluated on real-world noisy data in diverse domains.

Categories and Subject Descriptors

G.2.2 [**Graph Theory**]: Graph algorithms; H.3.1 [**Content Analysis and Indexing**]: Abstracting methods, Linguistic processing; H.3.3 [**Information Search and Retrieval**]: Relevance feedback, Retrieval models, Selection process; H.3.4 [**Systems and Software**]: User profiles and alert services; I.2.7 [**Natural Language Processing**]: Language parsing and understanding, Text analysis

Keywords

text semantics, key-concept extraction, keyword extraction, user modelling, discriminative representation, distributed representation, text categorisation

*Recommended by thesis supervisor: Prof. Mária Bieliková

1. Introduction

As one of the oldest mediums of communication in computers, text is still widely used in communication. Text is the primary medium of communication in most of newspapers, blogs, law documents, product manuals, etc. Although we can easily communicate via phones and various applications for voice or video communication (e.g., Skype), text has still multiple advantages over other means of communication. It is much easier to skim through a larger portion of text to grasp some basic information hidden in it than to listen to a longer recording, even if trying to skip some parts of it. It is also easier to communicate by writing some short text replies when multitasking so that we do not need to keep our attention only to one thing at a time. That is the reason why people on social networks like Facebook or Twitter or widely used messaging services like Google Hangouts or Slack communicate mostly by sending text messages to each other. There are also multiple enhancements (e.g., word suggestions built into software keyboards in mobile devices or more intelligent reminder suggestions and message suggestions), which simplify the text communication even more.

1.1 Open Problems and Dissertation Goals

The main goal of our dissertation is to enhance automatic understanding of text semantics. Despite the extensive research done for several decades, it still remains an open problem. The main cause is that natural language is not exact and the expressed ideas are much more abstract, sometimes even latent for ourselves. Yet still, we as humans can easily "get the point" in majority of cases.

The problem is that the notion of understanding text semantics has rather broad scope. It is certainly not clear at first glance what formal steps should be taken in order to achieve this goal. To investigate the problem, we need to look at the main shortcomings of existing methods which attempt to solve this same problem. We want to break down the big goal, whose accomplishment is rather vague and ambiguous, to smaller sub-goals that could be formulated more clearly and deterministically and thus could be achieved more easily as well.

Among common shortcomings of many existing statistical and rule-based methods is the absence of understanding word meaning. Mostly, they treat words as separate and individual units. They are not aware of any relations between similar words. If we are to understand the meaning of a whole text, first we should understand the meaning of individual words which comprise it. We should be aware of multiple relations between words, so that we are able to

infer indirect presence of more abstract concepts behind the simple words present in text. That leads us to our first goal, which aims at resolving this issue.

> ***Dissertation goal 1 - Enhance automatic understanding of the abstract text semantics.*** *We want to understand the meaning of text. That includes understanding of the core topics discussed in the text. These topics are not necessarily directly mentioned. We need to understand individual words, which are just expressions of abstract concepts, as well as latent relations between them. Therefore, an abstract representation is needed, which would enable modelling the latent relations between words and topics of such abstract concepts. Our goal is to represent even the abstract concepts concisely and exactly in order to provide a simple description of modelled text semantics.*

Fulfilling the first goal should provide us with deeper understanding of text and the latent abstract concepts, which are described or mentioned indirectly. From other point of view, however, we are not always interested in such generic and concise representation. Another common shortcoming of existing methods is that they do not consider the working domain of information space. Most methods seek only a general importance of modelled semantics and do not consider the specifics of different areas of interests. Having text documents in multiple categories, we care preferably about the discriminative properties of the document representation, not the general ones, which are common for all of them. We formulate the following second goal.

> ***Dissertation goal 2 - Rationalize models of the abstract text semantics.*** *Our goal is to create models of the abstract text semantics which are both discriminative and justifiable. We want to adapt the presentation of modelled text semantics to the respective audience. The presented semantics needs to have the highest information value possible. Considering common information retrieval tasks like organising and searching documents, users within a particular domain care preferably about the document topics which are discriminative within the domain. The topics which are general within the domain are usually considered to have lower information value, since within the domain of interest, they do not help to better organise or search documents. On the other hand, we want to be able to justify our model. For example, we want to justify the predictions of a text classifier, or justify the personalised recommendation in a user model.*

Given its definition, the second dissertation goal can be more practical than the first one, since it requires modelling of abstract semantics in order to model its discriminativeness and rationale. It may be easier to optimise directly a concrete objective than trying to build a generic model which should be universally applicable.

2. Approaches to Modelling Text Semantics

Although keyword-based text representation was developed relatively a long time ago [7], it is still widely used [2, 5, 14] due to its simplicity. One of the best known models is *tf-idf* [7], which computes the word importance by combining local frequency with global document probability. The computation of *tf-idf* model is based on the word frequency statistics and there has been multiple different extensions of it for various purposes like user modelling [6]. Besides the absolute word frequency, we can also consider its relative frequency [13] or the context graph of words [9].

Many researchers focus on keyword extraction to categorise text documents, which enabled development of multiple discriminative metrics. In most cases, the discriminative metrics are based on the absolute word frequency statistics [8, 12, 3], which can captured by four variables A, B, C and D (see Table 1). These variables represent the frequency of a word W given the category CAT. The metrics based on the variables A, B, C and D are summarised in Table 2, where N is the sum of all four variables.

Table 1: Frequency table given word W and category CAT.

Frequency	of word W	of other words
in category CAT	A	B
in other categories	C	D

Table 2: Frequency table given word W and category CAT.

Metric name	Function expressed in terms of ABCD statistics
rf	$log(2 + \frac{A}{max(C,1)})$
tds	$\frac{A/(A+B)}{(A+C)/N}$
ig	$N \times$ $\frac{A}{N} \times \frac{log(A \times N)}{(A+C) \times (A+B)} \times$ $\frac{B}{N} \times \frac{log(B \times N)}{(B+D) \times (A+B)} \times$ $\frac{C}{N} \times \frac{log(C \times N)}{(A+C) \times (C+D)} \times$ $\frac{D}{N} \times \frac{log(D \times N)}{(B+D) \times (C+D)}$
gr	$-ig /$ $(\frac{A+B}{N} \times log(\frac{A+B}{N}) +$ $\frac{C+D}{N} \times log(\frac{C+D}{N}))$
χ^2	$N \times \frac{(A \times D - B \times C)^2}{(A+D)(B+C)(A+B)(C+D)}$
idf	$log(\frac{N}{A+C})$

3. Modelling Abstract Text Semantics

To fullfil our first dissertation goal, we developed a method of key-concept extraction. Instead of words, we extract concepts (mapped to WordNet synsets), which are unambiguously defined and carry higher information value than simple words. The key part of our method is PageRank algorithm, which is used for both word sense disambiguation and the computation of concept importance. The proposed method can be divided into the following steps:

1. Choose candidate terms.
2. Build concept graph as a subgraph of WordNet vertices reachable from the candidate terms.
3. Calculate PageRank values of the graph vertices to obtain the most probable word senses.
4. Enrich the concept graph with collocation edges.
5. Calculate PageRank values of the graph vertices to obtain the locally important concepts.
6. Calculate the key-concepts by considering the information content of the locally important concepts.

In the first step, we select nouns (including compound nouns) as candidate terms based on part-of-speech tags. Subsequently, we take all the noun synsets in WordNet, which contain at least one of the candidate terms. These synsets are recursively enriched by hypernym synsets. That enables us to get to more abstract concepts (WordNet synsets), which represent more abstract topics discussed in the text documents.

Next, we perform first pass of PageRank algorithm and combine the PageRank values with the global concept probabilities to get the most probable concepts for the words in the document. To compute the importance of concepts, we utilise the idea of TextRank algorithm [9] and enrich the concept graph with edges connecting the concepts of the respective collocated words.

However, PageRank values do not consider the global importance of concepts, just the local one within the respective document. Because of that, we also consider the global probability and information content of concepts for the final computation of concept importance. The information content is analogous to *idf* factor (inverse document frequency) used for words.

3.1 Results

As we can see in Table 3, the use of key-concepts performs better than *tf-idf* baseline, even when using as little as only 3 key-concepts. The key-concepts represent very efficient representation of document content - concise and exact, yet precise enough, capturing also the abstract concepts discussed in the document. In contrast with words, which are ambiguous, concepts have clear interpretation. Each concept is mapped to a WordNet synset, which contains a definition and relations to other synsets, e.g. hypernyms, hyponyms, holonyms, meronyms, synonyms, etc.[1].

The use of concepts has also its disadvantages. In computational semantics, we often need to compute a simple word similarity value and the use of concept graph like WordNet, which contains only hypernym-hyponym or synonym relations, is rather complicated. Moreover, the handmade WordNet is still not perfect, both the content and the structure. Sometimes, it may be impossible to identify the correct sense of a word. According to [4], the word can have multiple senses even if placed in a context. All these facts suggest that there may exist a better representation of text semantics, which we address in the next section.

[1]The complete list of relations is available at `http://wordnet.princeton.edu/wordnet/man/wninput.5WN.html`

Table 3: The classification accuracy for different document representations.

Document representation	Classification accuracy
20 key-concepts	40.77
15 key-concepts	40.73
10 key-concepts	41.48
5 key-concepts	40.49
3 key-concepts	38.74
1 key-concepts	29.47
tf-idf	36.95

4. Modelling Discriminative Abstract Text Semantics

The late boom of deep learning resulted in development of various unsupervised methods, which can learn word semantics out of raw text [1]. These methods map words into a multidimensional latent feature space, which captures the word meaning. In contrast with the ontologies, we do not need to disambiguate the exact meaning of a word, since the feature vector can capture multiple senses. Moreover, such distributed representation is highly scalable, as we can also map the meaning of sentences or documents into the same feature space.

We utilise the distributed representation in our method of discriminative keyword extraction, which can be also used to model the feature vectors of whole documents as well. The advantage of such vectors is the simplicity of measuring word (or document) similarity using the cosine distance, or the possibility of doing vector operations like vector addition [11].

In our method we focus on categorised documents. We utilise discriminative metrics to propagate words which are more discriminative in the context of document categories. This way, we ignore words which are generally important, but have very low information value given the actual domain of interest.

The proposed method consists of the following steps:

1. Extraction of candidate phrases.
2. Computation of vector representation of the candidate phrases.
3. Substitution of candidate phrases by the most similar words.
4. Computation of a discriminative metric.
5. Selection of the keywords.
6. Computation of document vector.

The candidate phrase extraction is similar to the approach used in [15] while limiting the maximal length of phrases.

To use a discriminative metric, we compute a vector representation for each phrase and substitute it with most similar words. Then we can use any discriminative metric (see Table 2) to rank words. To compute a vector representation of the whole document, we can sum the feature vectors of the respective keywords.

4.1 Experimental Evaluation

We can see the influence of different metrics used in our method to the performance of the classification (Figures 1 and 2). In Figure 1, we can see that the variants using *ig,rf* and *tds* are among the best performing. We can also see that *tf-chi-squared* and *chi-squared* metrics secure the stable improvement with increasing number of extracted keywords. As the 20newsgroups dataset has well balanced categories, the difference between the micro-averaged and macro-averaged F1 score is negligible.

However, in case of Reuters-21578 dataset, the results are more interesting (Figure 2). As around 70% of all documents are contained in only two categories, we need to discriminate the topics on lower level of granularity. We can see that metrics with *tf* factor help to improve the score, as the *tf* factor has a tendency of propagating the smaller topics.

5. Using Discriminative Abstract Semantics to Model User Interests

To demonstrate the manifold of possible applications of our method of discriminative keywords extraction, we used it for modelling user interests. The problem of modelling user interests is very popular nowadays, since practically every website hones to provide a personalisation of its content to attract more users.

We use a user model similar to [10]. The user interest model is defined as a tripartite graph G (see Equation 1), where vertices V is a joint set of three disjoint sets U, V and D denoting sets of users, words and documents, respectively and E is a set of hyperedges. The hyperedge (u, w, d) denotes that the user u has a collection of documents D_u and word w is a keyword of document $d \in D_u$.

$$G = (V, E); \quad V = U \cup W \cup D$$
$$E = \{(u, w, d) | u \in U, w \in W, d \in D\} \quad (1)$$

To fill the content of the user model, we use our method of discriminative keyword extraction. Based on several assumptions, we map users to categories bijectively. Thus, our method can extract personalised keywords for each document in user's collection.

In Tables 4 and 5, we can see the examples of the extracted personalised keywords. We focus on two different domains - digital libraries (Annota dataset) and wild web (Brumo dataset), with different strengths of interest relations. Despite the noise, we can see that the extracted keywords capture the main topics discussed in the documents.

6. Conclusions

The proposed method of key-concept extraction addresses the first dissertation goal. We move from words to more abstract units - concepts. The main advantage of using concepts instead of words or latent topics is their concise and exact representation. The conciseness is advantageous for the automated processing systems due to its memory efficiency, while exact meaning of concepts can be leveraged to provide better presentation of semantics to the end users.

The proposed method of discriminative keyword extraction addresses the second dissertation goal. In contrast with the first dissertation goal, now we do not care only about some general topics discussed in the document, but we are more interested in the discriminative ones instead. The proposed method makes use of discriminative statistics and focuses on putting more weight on discriminative terms rather than terms that are just generally important. At the same time, our model can also provide a rationale behind all its decisions, as embedding words, documents and categories into the same feature vector space creates a joint model, which can be queried by words, documents, categories, or practically anything embedded into the same feature vector space, interchangeably.

We believe in great potential of our method of discriminative keyword extraction. As we showed, our method is capable of extracting very small number of keywords to describe a document and still categorise the document very accurately. This means we can use our method of discriminative keywords extraction to extract few words to be presented to a human, e.g., to rationalize why a particular document is recommended, or what other users are likely to be interested in it.

To validate if our method is truly applicable in wide range of such real-world problems, we have evaluated it on the task of modelling user interests in two different domains. We show that it can be used not just for the standard task of text categorisation, but can be applied in various real-world scenarios like modelling user interests on "wild Web" as well as modelling interests of a researcher in digital libraries. We also show that our method is capable of modelling user interests even if the interest relations are weak and not explicit.

Moreover, we make another contribution by showing that the evaluation of categorisation of text documents can be used as an automatic quantitative evaluation technique, which can speed up the research progress in the domain of user modelling substantially.

Acknowledgements. This work was partially supported by grants No. APVV-0208-10, VEGA 1/0675/11, VEGA 1/0971/11, KEGA 009STU-4/2014, VEGA 1/0646/15 and APVV-15-0508.

References

[1] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, Mar. 2003.

[2] T. Bohne, S. Rönnau, and U. M. Borghoff. Efficient keyword extraction for meaningful document perception. In *Proceedings of the 11th ACM symposium on Document engineering*, DocEng '11, pages 185–194, New York, NY, USA, 2011. ACM.

[3] F. Debole and F. Sebastiani. Supervised term weighting for automated text categorization. In *Proceedings of the 2003 ACM Symposium on Applied Computing*, SAC '03, pages 784–788, New York, NY, USA, 2003. ACM.

[4] K. Erk. What is word meaning, really?: (and how can distributional models help us describe it?). In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, GEMS '10, pages 17–26, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

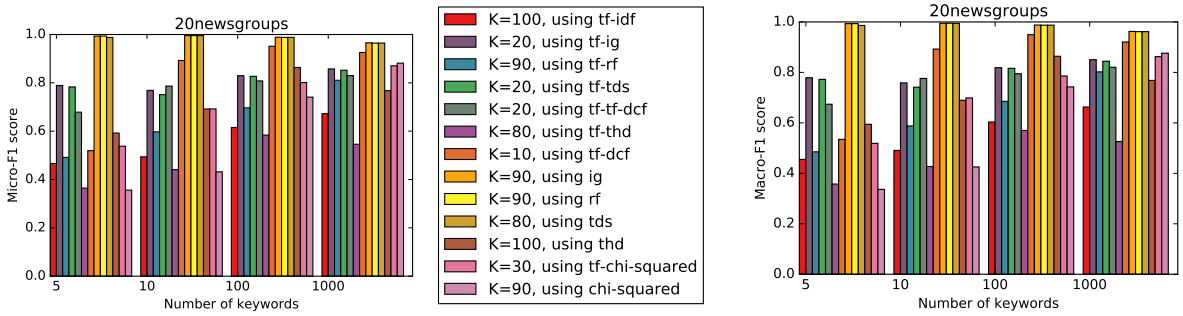

Figure 1: Comparison of different statistical metrics used in ranking words evaluated on 20newsgroups dataset.

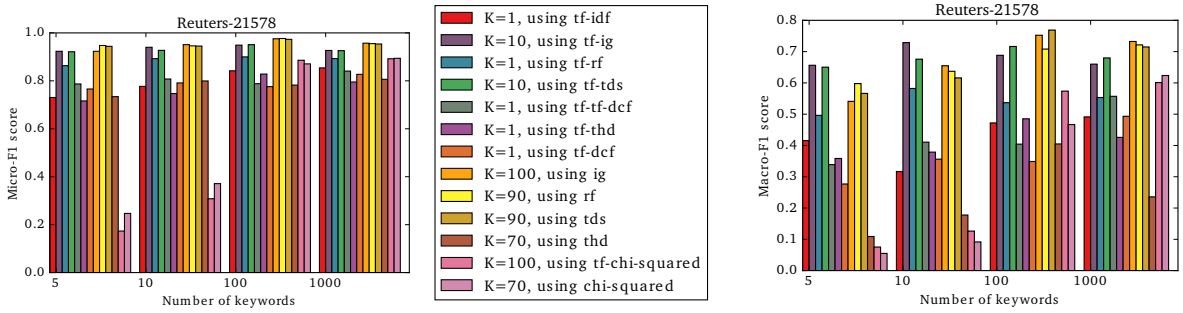

Figure 2: Comparison of different statistical metrics used in ranking words evaluated on Reuters-21578 dataset.

[5] J. Feng, F. Xie, X. Hu, P. Li, J. Cao, and X. Wu. Keyword extraction based on sequential pattern mining. In *Proceedings of the Third International Conference on Internet Multimedia Computing and Service*, ICIMCS '11, pages 34–38, New York, NY, USA, 2011. ACM.

[6] Q. Gao, F. Abel, G.-J. Houben, and K. Tao. Interweaving trend and user modeling for personalized news recommendation. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '11, pages 100–103, Washington, DC, USA, 2011. IEEE Computer Society.

[7] K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 1972.

[8] M. Lan, C.-L. Tan, and H.-B. Low. Proposing a new term weighting scheme for text categorization. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, pages 763–768. AAAI Press, 2006.

[9] R. Mihalcea and P. Tarau. Textrank: Bringing order into texts. In D. Lin and D. Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411. Association for Computational Linguistics, July 2004.

[10] P. Mika. Ontologies are us: A unified model of social networks and semantics. *Web Semant.*, 5(1):5–15, Mar. 2007.

[11] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[12] Y. Park, S. Patwardhan, K. Visweswariah, and S. C. Gates. An empirical analysis of word error rate and keyword error rate. In *INTERSPEECH*, pages 2070–2073, 2008.

[13] M.-S. Paukkeri and T. Honkela. Likey: Unsupervised language-independent keyphrase extraction. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 162–165, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[14] T. Uherčík, M. Šimko, and M. Bieliková. Utilizing microblogs for web page relevant term acquisition. In P. Emde Boas, F. Groen, G. Italiano, J. Nawrocki, and H. Sack, editors, *SOFSEM 2013: Theory and Practice of Computer Science*, volume 7741 of *Lecture Notes in Computer Science*, pages 457–468. Springer Berlin Heidelberg, 2013.

[15] T. Vu, A. T. Aw, and M. Zhang. Term extraction through unithood and termhood unification. In *Proceedings of the Third International Joint Conference on NLP*, pages 631–636, 2008.

Selected Papers by the Author

M. Šajgalík, M. Barla, and M. Bieliková. Searching for Discriminative Words in Multidimensional Continuous Feature Space. *In Computer Speech and Language*, Elsevier, ISSN: 0885-2308. Indexed in Current Contents. Impact factor: 1.324. Submitted revised manuscript.

M. Šajgalík, M. Barla, and M. Bieliková. Modelling User Interests in Latent Feature Vector Space based on Topical Discriminativeness. *In User Modeling and User-Adapted Interaction*, Springer, ISSN: 0924-1868. Indexed in Current Contents. Impact factor: 2.292. Submitted.

M. Šajgalík, M. Barla, and M. Bieliková. Exploring Multidimensional Continuous Feature Space to Extract Relevant Words. *In Statistical language and speech processing : second international conference, SLSP 2014, Grenoble, France, October 14-16, 2014. Proceedings*. Springer, 2014, pp. 159-170. ISBN 978-3-319-11397-5. Indexed in Scopus: 2-s2.0-84921933352.

M. Šajgalík, M. Barla, and M. Bieliková. From ambiguous words to key-concept extraction. *In DEXA 2013, 24th International Workshop on Database and Expert Systems Applications, 26-29 August 2013 Prague, Czech Republic*. - Los Alamitos : IEEE Computer Society, 2013. - ISBN 978-0-7695-5070-1. - pp. 63-67. Indexed in Scopus.

M. Šajgalík, M. Barla, and M. Bieliková. Efficient Representation of the Lifelong Web Browsing User Characteristics. *In CEUR Workshop Proceedings Vol. 997 LLUM 2013 : 3rd International Workshop on LifeLong User Modelling, June 10, 2013, Rome, Italy in conjuction with UMAP 2013 Conference*. CEUR-WS, Vol. 997, 2013. - ISBN 1613-0073. - on-line, pp. 21-30

Table 4: Examples of personalised keywords extracted from selected research articles in Annota dataset.

Discovering value from community activity on focused question answering sites: a case study of stack overflow *question, answer, ask, answering, yes, query, ponder, rephrase, clue, wonder*
Cloud Computing *software, desktop, computer, virtualization, computing, server, multiserver, technology, multithread, mainframe*
5-ALA mediated photodynamic therapy induces autophagic cell death via AMP-activated protein kinase. *polypeptide, transduction, postsynaptic, autophagic, porphyrin, oligomeric, oxidase, modulatory, esterase, histidine*
Algae Energy: Algae as a New Source of Biodiesel *biomass, energy, coal, renewable, biodiesel, gas, hydrogen, electricity, fuel, gasification*
Hybrid Web Recommender Systems *recommend, propose, autocompletion, recommended, predefine, recommendation, consider, websearch, inferencing, recommender*
Context-aware query classification *contextualisation, contextualization, relevance, disambiguate, contextualise, contextual, contextualized, context, disconfirm, contextualised*
A community question-answering refinement system *answer, question, explanation, answering, ask, query, clue, clarification, reply, yes*

Table 5: Examples of personalised keywords extracted from selected web pages in Brumo dataset.

http://www.sei.cmu.edu/security/ *technology, interface, software, cryptography, cryptographic, academia, networked, interactional, community, computing*
http://en.wikipedia.org/wiki/Eduardo_Frei_Montalva *dictatorship, democratization, syllogistic, emancipatory, manifestation, monistic, regime, presidency, posteriori, historiographical*
http://en.wikipedia.org/wiki/Programming_paradigm *language, idiom, dialect, phonology, dialectal, orthography, orality, phonemic, phonological, prosody*
http://help.coursera.org/customer/portal/articles/1164685-can-i-access-the-course-content-after-a-course-ends- *course, access, process, layout, curriculum, parameterization, curricular, configuration, slope, material*
http://help.coursera.org/customer/portal/articles/1164685-can-i-access-the-course-content-after-a-course-ends- *course, access, process, layout, curriculum, parameterization, curricular, configuration, slope, material*
http://www.imdb.com/title/tt0102926/ *murderous, demonic, vengeful, sinister, psychopath, demented, villain, villainous, psychopathic, evil*
http://www.techrepublic.com/blog/10things/10-techniques-for-gathering-requirements/287 *standard, guideline, requirement, criterion, rigorous, stringent, standardization, prerequisite, minimum, criteria*
http://grooveshark.com/ *music, internet, video, digital, audio, entertainment, wireless, iTunes, multimedia, streaming*
http://www.optasports.com/en.aspx *championship, tournament, opener, squad, season, game, team, match, qualifier, club*