EXPLORING NPL: GENERATING AUTOMATIC CONTROL KEYWORDS

Óscar Bernardes, Vanessa Amorim

Porto Accounting and Business School, Polytechnic Institute of Porto (PORTUGAL)

Abstract

Keywords are a tool to help indexers and search engines find relevant papers. Unfortunately, authors use them wrong, unintentionally or to misleading readers into a non-related topic, promoting their articles by using non-representative keywords. Previous scholars (Ansari, 2005; Voorbij, 1998) exposed lack of consistence between abstracts, full-texts and keywords. This is an old but effective practice. An early investigation conducted by Schultz, Schultz and Orr on 1965 matched author keywords to document titles and to indexing terms appointed by subject matter experts, and found out the author supplied keywords matched more closely the terms used by subject matter experts than did the title terms (as cited in Kipp, 2011, p. 249). Fifty-five year after, Terra et al. (2020) suggest seven improvements to keyword parameterization. In fact, author keywords have received relatively little attention in the literature, according to Kipp (2007). Moreover, with the ever-increasing academic data available, finding relevant documents has become more challenging for regular users and library specialists.

The **purpose** of this article is to generate theses keywords using NLP - Natural language processing techniques; NLP is a subfield of linguistics, computer science, and artificial intelligence, taking advance of big data, indexing data while removing human errors and costs (Moskovitch, Martins, Behiri, Weiss & Shahar 2007).

Design/methodology/approach: A 95% sample population of 51.010 master theses population, from the institutional repository of the University of São Paulo, was extracted and selected, representing 48.501 records, then a thematic dictionary was created based on theses major area, subsequently generating the theses' keywords established by the previous dictionary.

Research limitations/implications: The effectiveness of information retrieval is highly dependent on the accurate and complete representation of document content and major area of the theses.

Originality/value: Author keywords have received relatively little attention in the literature (as cited in Kipp, 2011). Not due to lack of importance for all stakeholders, but because of the complexity involved on the task and publisher lack of control. This paper highlights a new method to generate and control author keywords.

Keywords: keyword indexation, NLP keywords, keyword extraction

INTRODUCTION

A keyword(s) is "a word or group of words, possibly in lexicographically standardized form, taken out of a title or of the text of a document characterizing its content and enabling its retrieval" (ISO norm 5963; 1985), and should be representations of a given text that allow readers to recognize its subject in advance (Rose et al., 2010). Keyword extraction methods have the purpose of generating or extracting representative keywords, best describing the subject of a document (Zhou et al., 2013; Litvak et al., 2011, Papagiannopoulou & Tsoumakas, 2019), improving retrieval efficiency. As such effective keywords are a necessity, since keyword is the smallest unit which express meaning of entire document, many applications can take advantage of it such as automatic indexing, classification, clustering, filtering, cataloging, topic detection and tracking (Kaur & Gupta, 2010).

Keyword assignment is a research topic designated to tag document with keywords and can be divided roughly into two categories: keyword generation and keyword extraction, as presented

in Figure 1, the first method keywords are generated or assigned from a controlled dictionary, made from a predefined taxonomy keywords, then the documents are categorized into classes according to their content. On the second method, keywords are extracted from the document and must be, priori declared in text, in this technique words are analyzed in order to determinate and grade the most representative keywords, usually exploring the source properties (Beliga et. al, 2015; Papagiannopoulou & Tsoumakas, 2019).

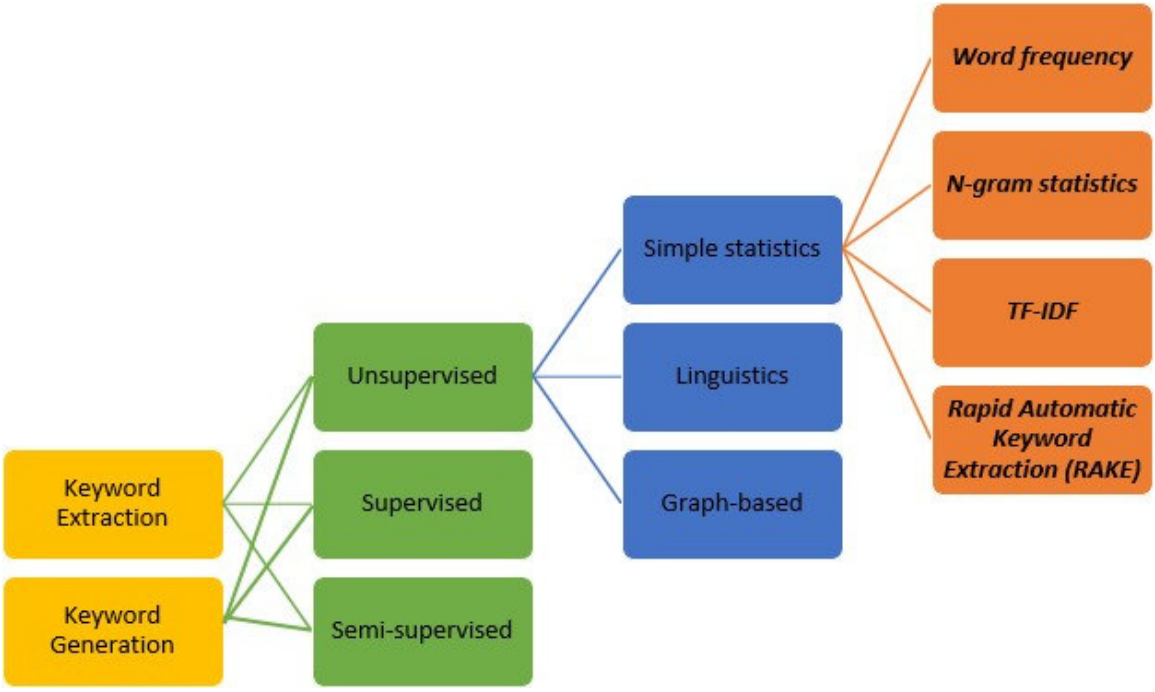

Figure 1. *Automatic keyword indexing. Source: Own elaboration.*

Citing Papagiannopoulou & Tsoumakas (2019, p. 1-2) "unsupervised methods are popular because they are domain independent and do not need labelled training data, i.e. manual extraction of the key phrases, which comes with subjectivity issues as well as significant investment in time and money. Supervised methods on the other hand, have more powerful modelling capabilities and typically achieve higher accuracy than the unsupervised ones according to previous studies". The disadvantage of supervised techniques is the need for a controlled predefined taxonomy keywords, the quality of the training corpus will influence the outputs.

Simple Statistical

This approach is a clean technique for classifying the principal keywords and key expressions within a text, can be based 4 methods: word frequency, N-gram statistics, TF-IDF (stands for term frequency- inverse document frequency), RAKE (Rapid Automatic Keyword Extraction).

Word frequency

Word frequencies perform a list of keyword or combined keyword (phrases) and evaluates the number of findings in a current analysis, based on statistic (count formula). This approach is useful for recognizing persistent terms, and assist categorization established per article or established on a controlled dictionary. Do not pre setup teaching how the computer should isolate and extract the relevant keywords, which is an advantage because master's theses have uncertain keyword location among all text. Conversely, phrases and linguistic can be rejected, which for our propose (keyword extraction) is not so relevant. On the other hand, this approach transforms the text as tokens 'bag of words', removing grammar, structure and phare meaning.

Nevertheless, this can be suitable for a numberless of objectives, locating the major disputes in customer service, highlight topics and interests.

N-gram statistics

An N-gram is an N-character slice of a longer string, we will use the underscore character ("_") to represent space.) consequently, the word "theses" would be composed of the following N-grams: ● bi- grams: _T, TH, HE, ES, SI, IS, S_; ● tri-grams: _TH, THE, HES, ESI, SIS, IS_; ● quad-grams: _THE, THES, HESI, ESIS, SIS_, ... In general, a string of length k, padded with blanks, will have k +1 bi- grams, k +1 tri-grams, k +1 quad-grams, and so on (Cavnar & Trenkle, 2001). N-gram models assigns probabilities to the sequences of words, using big data. Are generally used in speech recognition, statistical natural language processing and sequences of phonemes.

TF-IDF

TF-IDF stands for term frequency-inverse document frequency, is a technique to quantify how important a word is to a document in a collection or text corpus, measuring the weight (importance) for each word in the document. Throughout TF, all tokens (words) are considered equally important, therefore, stop words and certain terms, such as verbs may appear highlighted but have minor importance for keyword generating, consequently we need to rectify the importance, ponder the weigh to down for frequent terms in contrast increase the weight for relevant terms (inverse), by using controlled dictionaries. This will be performed by counting the term frequency of all words and matches those with the inverse word frequency, how unique that token is in the corpus. Multiplying these two measures we obtain a TF-IDF score for that precise word, low results represent not pertinent keywords, higher score point that word(s) as relevant. TD-IDF algorithms have numerous of applications in machine learning, search engines use them to return the "best" query, to score sites and perform rankings of importance.

Rapid Automatic Keyword Extraction (RAKE)

RAKE is an unsupervised, domain-independent, and language-independent method for extracting keywords from individual documents use mutually word frequency and word degree to assign keywords, employing stop words and tokens to analyse candidate keywords. RAKE initiates keyword extraction on a corpus by parsing its text into a set of candidate keywords. First, the document text is split into tokens, then split into sequences of contiguous words at phrase delimiters and stop word positions. Words within a sequence are assigned the same position in the text and together are considered a candidate keyword. After every candidate keyword is classified, a score is calculated for each candidate keyword and defined as the sum of its member word scores, evaluating several metrics for calculating word weight / importance (word frequency or the word degree or the ratio of degree to frequency) for each keyword candidate (Rose et al., 2010; Papagiannopoulou & Tsoumakas, 2019). If two candidate words emerge jointly, more than twice, a different keyword is added to the candidate list, combining that string together and a new weight will be generated.

RAKE is based on assumption that text comprehend multiple punctuation or stop words with minimum lexical meaning, these stop works (e.g.: is, not, that, there, are, can, you, with, of, those, after, all, one) are removed from the corpus and considered to be meaningless. Therefore, is important to analyse deeply the stop words, and how RAKE handle with it, because can misrepresent the candidate keyword by removing the borderer stop word, transforming the brand "just jeans" into "jeans".

Linguistic

This technique uses lexical analysis, syntactic analysis, morphological examination and/or syntactic information (part-of-speech or the associations between words in a dependency grammar representation of sentences) to select which keywords should be designated. Is more computationally intensive and require extended grammar expertise.

Graph-based

A text is a set of multiple sentences (represented in figure 2 by the circle symbol), each sentences have multiple words (represented by square symbol), subsequently it is possible to convert all text into words, and words to points or vectors (word vectorization), by doing this we will map words in space and further we can associate words with similar meanings or representations, using third based solution such as word2vec. The basic concept is to determinate candidates' keywords from the nodes generated by the graph.

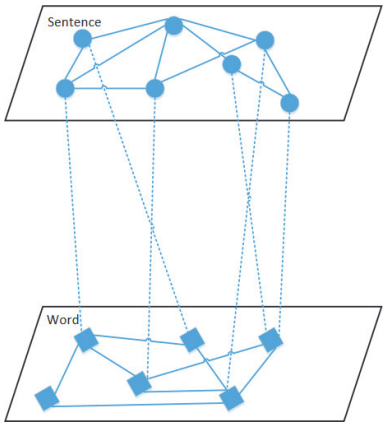

Figure 2. *Graph-based. Source: Ying, Yan, Qingping, Tan, Qinzheng, Xie, Ping, Zeng & Panpan, Li. (2017, p. 250).*

Supervised

The supervised method improves the extraction by controlling the dataset, the most popular supervised techniques were presented by Turney (2000) or Witten et al. (1999), combining control keywords to documents. Some studies have addressed the uses of keywords in users' catalogue searches (Ansari, 2005; Voorbij, 1998) and the effects of using them with control keywords (Strader, 2009). Nevertheless, there is insufficiency research concerning the effect of using unrestricted authors keywords and the efficiency of the retrieval (Gil-Leiva & Alonso-Arroyo, 2007).

For Strader, the discussion involving author keywords and controlled vocabularies is framed how new controlled terms can be indexed, keywords can be inputted as controlled terms instructions, but could affect the maintenance of controlled vocabularies such as Library of Congress Subject Headings (LCSH) and the principles of literary warrant.

Gross and Taylor (in Strader, 2009) worked one a different approach, using users' searches transaction logs of to analyze if controlled vocabulary delivers supplementary keywords and subsequently improves both recall and precision in keyword searches of a catalog, findings indicated an increase of up to 30 percent in the recall of relevant documents using controlled vocabulary.

An obstacle for better precision is related with researchers, and often editorial teams when they assign inaccurate keywords to optimize indexing and retrieval (Gil-Leiva & Alonso-Arroyo, 2007). NLP - Natural language processing could minimize the abuse or replace automatically some keywords.

NPL is a subfield of linguistics and artificial intelligence focus on the interactions between AI and human language, it can calculates distributional semantics based on the assumption that linguistic items with similar distributions have similar meanings, delivery speed and permanent indexing updates, based on current trends or users' searches transaction logs, while removing errors and human analytical costs (Moskovitch, Martins, Behiri, Weiss & Shahar 2007), and has been revealed to be effective in identifying significant documents even in challenging with controlled vocabularies (Yousefi et al., 2019) even though limited application use NLP (Liddy, 2001).

Semi-supervised

Decong Li et al. (2010) defined a semi-supervised technique, called keyphrase, making an assumption that the title of document reproduces the main content of the document and therefore, based on that postulation, keywords should be generated with close semantics to the title. Keyphrase extraction is achieved by estimating the sentence importance in the semantic network, then the weight of the title phrases correlates to the main keywords iteratively. A machine learning approaches can be labelled as semi-supervised approach, depending on the level of training, if researchers use predefined taxonomy or assign controlled text with keywords to their dataset documents. The keyword extraction model can be induced using machine learning algorithms: SVM (Support Vector Machines), C4.5, Naïve Bayes.

METHODOLOGY

The methodology used in the article combined web scraping for obtain the source of the data, then NPL techniques based on semi-supervised approach, combining RAKE with controlled dictionary.

Web scraping is a technique to extract and collect external data in a systematic way, in such process, a web bot agent to reproduce human interaction, collecting data during the browsing process. In 2016, the web traffic originated by web bots constitutes more than a half (51.8%) of the total web traffic, on way to prevent this, is thru recaptcha system where humans need to prove they really humans, identifying match images or decipher hard to read text.

There are available, online, different web scraping code, normally based on third-party libraries which grant access to the client site (data source) by implementing the client side of the HTTP protocol, Glez-Peña (2013) identified libcurl as one of the most popular site access libraries, MechanizeWeb automation module for PERL, the Apache HttpClient package for JAVA and BeautifulSoup for Phyton.

Manual browsing is an option too, but browsing is time-consuming, and is prone to miss valuable details.

For this research, CG enterprise web-scraping software was used, due to is visual performance and because is more dynamic, fast to implement and has strong error handling features. Our scraping cycle went through four stages. First, we designed a web scraping agent to capture the full theses' data, afterwards establishing which information from institutional repository of the University of São Paulo (https://teses.usp.br/teses) should be collected, thus delimiting the authentication rules for each data. Second, we did some debugs to analyse the efficiency and how the agent was performing the crawl, and if all data were structured on the correct fields, after adjustments on XPath, CSS-Selectors and REGEX code, we run the spider for several days to scrap the entire library, extracting only theses from 2001 to 2019 simulating a human behaviour, e.g.: scrolls, mouse click, etc. Finally, the data was stored into a MySQL database and used as source material.

We decided to a designed an approach combining RAKE with a controlled dictionary,

consequently, become a semi-supervised technique, domain restricted and language oriented extracting tool. This strategy took advantage of RAKE features and procedures and decreased uncontrolled outputs.

RESULTS

Results from step 1.1: Scraping master theses

We decided to scrap the digital Library of Theses and Dissertations of the University of São Paulo (https://teses.usp.br), extracting 95% of all master's theses published between 2001 and 2019, and available in the institutional repository (table 1). This task was performed in May of 2019 and collected a sample of 48.501 theses. Over 1 million records were stored, e.g.: dates, keywords, subjects, authors, e-mail, DOI, statistics and PDF.

Table 1. *Theses distribution. Source: Own elaboration.*

year	Thesis	year'	Thesis'
2001	106	2011	3.032
2002	345	2012	3.284
2003	395	2013	3.921
2004	642	2014	3.778
2005	798	2015	4.354
2006	1.485	2016	4.198
2007	2.808	2017	4.019
2008	2.633	2018	4.683
2009	3.268	2019	1.200
2010	3.396	Blank	157
		Total	48.502

Results from step 1.2: Filtering and label

Supervised methods oblige, previous training data, and are frequently dependent on the subject domain. The algorithm must re-learn and establish new correlations after the knowledge are modify (Litvak, 2011). The foremost keywords used on pharmacy theses, should not be related with accountability, for example. Therefore, after obtaining all raw material, the authors filter the data using a SQL Select command to disclose all theses nonrelated, for this stage, the authors could select knowledge area, already provided by University of São Paulo in which aggregate all theses on 850 topics or choose the 74 university departments / schools, or else combine both. We decided to select, the second method only, filtering by departments (excluding Polytechnic School*).

Table 2. *Departments. Source: Own elaboration.*

Archeology and Ethnology Museum	Institute for Energy and Environment	Inter-units in Environmental Science
Bauru College of Dentistry	Institute of Architecture and Urbanism	Inter-units in Latin American Integration
Bioenergy USP, UNESP and UNICAMP	Institute of Astronomy, Geophysics and Atmospheric Sciences	Inter-units in Materials Science and Engineering
Center of Nuclear Energy in Agriculture	Institute of Biomedical Sciences	Inter-units in Museology
College of Agriculture Luiz de Queiroz	Institute of Biosciences	Inter-units in Nursing
College of Animal Husbandry and Food Engineering	Institute of Brazilian Studies	Inter-units in Teaching of Sciences
College of Architecture and Urbanism	Institute of Chemistry	Law College
College of Arts, Sciences and Humanities	Institute of Energetic and Nuclear Research	Lorena College of Engineering
College of Communication and Arts	Institute of Geosciences	Mathematical Modeling in Finance
College of Economics, Administration and Accounting	Institute of International Relations	Museum of Contemporary Art
College of Education	Institute of Mathematics and Computer Science	Polytechnic School*
College of Medicine	Institute of Mathematics and Statistics	Ribeirão Preto College of Dentistry
College of Nursing	Institute of Oceanography	Ribeirão Preto College of Economics, Administration and Accounting
College of Pharmaceutical Sciences	Institute of Physics	Ribeirão Preto College of Medicine
College of Philosophy, Literature and Human Sciences	Institute of Psychology	Ribeirão Preto College of Nursing
College of Physical Education and Sports	Institute of Tropical Medicine	Ribeirão Preto College of Pharmaceutical Sciences
College of Public Health	Inter-units in Aesthetics and History of Art	Ribeirão Preto College of Philosophy, Sciences and Literature
College of Veterinary Medicine and Animal Husbandry	Inter-units in Applied Ecology	Ribeirão Preto College of Physical Education and Sports
Collegy of Dentistry	Inter-units in Applied Human Nutrition	Ribeirão Preto Law College
Collegy of Dentistry, College of Nursing and College of Public Health	Inter-units in Bioengineering	São Carlos Institute of Chemistry
Dante Pazzanese Cardiology Institute	Inter-units in Bioinformatics	São Carlos Institute of Physics
Hospital for the Rehabilitation of Craniofacial Anomalies	Inter-units in Biotechnology	São Carlos Institute of Physics and Chemistry
ICMC and UFSCar Interinstitutional in Statistics	Inter-units in Energy	São Carlos School of Engineering
		Zoology Museum

Results from step 2.1: Create a dictionary of keywords

Using scraping techniques, another time, we created an external list of control keywords to function as dictionary, a different source was selected, to avoid running previous keywords on the original documents. For this proposed, we did use the digital Library of Theses and Dissertations of the University of Federal de Santa Catarina (https://repositorio.ufsc.br), this extraction was made in December of 2020 (table 3).

Table 3. *Dictionary font. Source: Own elaboration.*

Administração	[781]	Ciência e Engenharia de Materiais	[490]	Engenharia de Sistemas Eletrônicos	[10]	Letras/Literatura Brasileira	[65]
Administração Universitária	[219]	Ciências da Engenharia Ambiental	[1]	Engenharia de Transportes e Gestão Territorial	[74]	Linguística	[706]
Agroecossistemas	[391]	Ciências da Reabilitação	[35]	Engenharia e Ciências Mecânicas	[35]	Literatura	[751]
Agroecossistemas	[59]	Ciências Médicas	[245]	Engenharia e Gestão do Conhecimento	[640]	Matemática	[51]
Antropologia Social	[432]	Contabilidade	[223]	Engenharia Elétrica	[1624]	Matemática e Computação Cientifica	[73]
Aquicultura	[469]	Cuidados Intensivos e Paliativos	[66]	Engenharia Mecânica	[1420]	Matemática Pura e Aplicada	[131]
Arquitetura e Urbanismo	[291]	Desastres Naturais	[12]	Engenharia Mecânica	[1]	Métodos e Gestão em Avaliação	[52]
Assistência Farmacêutica	[12]	Design	[75]	Engenharia Química	[786]	Metrologia Científica e Industrial	[89]
Biologia Celular e do Desenvolvimento	[122]	Design e Expressão Gráfica	[128]	Ensino de Biologia	[21]	Nanotecnologia Farmacêutica	[2]
Biologia de Fungos, Algas e Plantas	[91]	Direito	[1480]	Ensino de Física	[61]	Neurociências	[345]
Biologia Vegetal	[137]	Direito	[48]	Ensino de História	[18]	Nutrição	[271]
Bioquímica	[151]	Ecologia	[162]	Estudos da Tradução	[476]	Oceanografia	[38]
Biotecnologia	[174]	Economia	[304]	Farmácia	[469]	Odontologia	[633]
Biotecnologia e Biociências	[138]	Ecossistemas Agrícolas e Naturais	[24]	Farmacologia	[398]	Perícias Criminais Ambientais	[50]
Centro de Ciências Agrárias	[342]	Educação	[1256]	Farmacologia	[33]	Programa de Pós-Graduação Interdisciplinar em Ciências Humanas	[201]
Centro de Ciências da Educação	[268]	Educação Científica e Tecnológica	[445]	Filosofia	[374]	Programa de Pós-Graduação Multicêntrico em Ciências Fisiológicas	[30]
Centro de Ciências da Saúde	[427]	Educação Física	[561]	Física	[305]	Programa de Pós-Graduação Multidisciplinar em Saúde	[28]
Centro de Ciências Físicas e Matemáticas	[504]	Energia e Sustentabilidade	[24]	Geografia	[627]	Propriedade Intelectual e Transferência de Tecnologia para Inovação	[22]
Centro de Comunicação e Expressão	[469]	Enfermagem	[940]	Gestão do Cuidado em Enfermagem	[120]	Psicologia	[811]
Centro de Filosofia e Ciências Humanas	[525]	Engenharia Ambiental	[566]	História	[607]	Química	[752]
Centro Sócio-Econômico	[438]	Engenharia Ambiental	[50]	Informática em Saúde	[28]	Recursos Genéticos Vegetais	[300]
Centro Tecnológico	[2919]	Engenharia Civil	[1071]	Inglês: Estudos Linguísticos e Literários	[183]	Relações Internacionais	[92]
Ciência da Computação	[931]	Engenharia de Alimentos	[367]	Jornalismo	[149]	Saúde Coletiva	[244]
Ciência da Informação	[265]	Engenharia de Automação e Sistemas	[340]	Letras	[35]	Saúde Mental e Atenção Psicossocial	[69]
Ciência dos Alimentos	[335]	Engenharia de Produção	[4536]	Letras/Inglês e Literatura Correspondente	[357]	Saúde Pública	[164]
						Serviço Social	[256]
						Sociologia Política	[572]
						Tecnologias da Informação e Comunicação	[91]
						Urbanismo, História e Arquitetura da Cidade	[130]

Results from step 2.2: Categorizing keywords and associating

Subsequently we associate the keywords (step 2.1) with the departments (step 1.2), restricting the domain, as defined by the literature review. For example, administration keywords (obtained from dc.subject.classification) were used for College of Economics, Administration and Accounting & Ribeirão Preto College of Economics, Administration and Accounting.

Results from step 3.1: Tokenize all text

Tokenization is the act of splitting up a text into stings such as words, keywords, phrases, symbols and other elements called tokens. Tokens can be individual words, phrases or even whole sentences. In the process of tokenization, some characters like punctuation marks are discarded.

Exemple: |The| |Brazilian| |retail| |sector| |is| |undergoing| |a| |major| |transformation| |process|, |in| |an| |environment| |of| |high| |competition| |and| |strong| |trend| |of| |multichannel| |performance|, |with| |the| |emergence| |of| |new| |technologies| |and| |observation| |of|

|changes| |in| |consumer| |behavior|

Results from step 3.2: Converting all words to the lower case

While tokenizing documents, we may encounter similar words but in different cases, eg: upper 'CASE' or lower 'case' or title 'Case'. While the word case is common, different tokens will be generated for them. This increases the size of vocabulary and consequently the dimension of generated word vectors.

Results from step 3.3: Remove special characters and stopwords from the text

Stopwords are the words that do not contain much information about text like 'is', 'a', 'the and many more', includes removing language articles, pronouns and prepositions such as "and", "the" or "to" in English. In this process some very common words that appear to provide little or no value to the NLP objective are filtered and excluded from the text to be processed, hence removing widespread and frequent terms that are not informative about the corresponding text.

Exemple: |brazilian| |retail| |sector| |undergoing| |major| |transformation| |process|, |environment| |high| |competition| |strong| |trend| |multichannel| |performance|, |emergence| |new| |technologies| |observation| |changes| |consumer| |behavior|

Global results from step 4.

We perform the analysis, on ten theses from administration department, and ten random theses from other departments. In each document we determined new candidates for keywords, based on their score, the top five T candidate keywords were selected. In 65% of the thesis the authors keywords matched the candidates.

DISCUSSION & CONCLUSION

In this work, we projected a hybrid (semi-supervised) keyword extraction method for thesis based on web scraping and NPL, combining RAKE with controlled dictionary keywords. We employed recent deep learning models. The present research has some constraints. First, the categorization of the theses was based on the department, therefore we are restricting students' thesis area base on the department main knowledge area. Second, the Portuguese stop words available on github is very limited, therefore we used a personalized version, created for this research, nevertheless is more complete there are, still, many limitations (revealed on the debugs), however, we were constantly upgrading our stop words list, in each round, but is impossible to apply for all documents, in this phase. Third, there isn't a precise control keyword list, so thesis authors mistreat or misrepresent the keyword (Terra et. al, 2020). Fourth, the current hybrid extraction model did not perform full accuracy. It will be impossible, because keywords change with time, nevertheless, according to our analyze based on administrator thesis, it demonstrated high (> 80%) accuracy. Fifth, we should use a key expert panel, per area, to define the accuracy rate.

The proposed semi-supervised keyword extraction model for thesis was validated through performance author and personal comparison, showed a 65% level of matching previous author keyword and has a significant accuracy.

REFERENCES

Ansari, M. (2005). Matching between assigned descriptors and title keywords in medical theses. *Library Review*, *54*(7), 410-414. https://www.emeraldinsight.com/doi/abs/10.1108/00242530510611901. DOI: 10.1108/00242530510611901

Beliga, S., Meštrović, A., & Martincic-Ipsic, S. (2015). An Overview of Graph-Based Keyword Extraction Methods and Approaches. *Journal of Information and Organizational Sciences*, 39, 1-20.

Cavnar, W. & Trenkle, J. (2001). N-Gram-Based Text Categorization. In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*. Retrieved from https://www.let.rug.nl/~vannoord/TextCat/textcat.pdf

Decong L., Sujian L., Wenjie L., Wei W., & Weiguang, Q. (2010). A semi-supervised key phrase extraction approach: learning from title phrases through a document semantic network. In *Proceedings of the ACL 2010 Conference Short Papers* (pp. 296–300).

Gil-Leiva, I. & Alonso-Arroyo, A. (2007). Keywords given by authors of scientific articles in database descriptors. *Journal of the American Society for Information Science and Technology*, 58. 10.1002/asi.20595.

Glez-Peña, Daniel & Lourenco, Anália & López-Fernández, Hugo & Reboiro-Jato, Miguel & Fdez-Riverola, Florentino. (2013). Web scraping technologies in an API world, *Briefings in bioinformatics*, 15(5), 788–797.

International Association for Standardization (ISO). (1985). Documentation. Methods for examining documents, determining their subjects, and selecting indexing terms (ISO 5963:1985). Geneva, Switzerland.

Igal, Z. (2016). Bot Traffic Report. Retrieved from https://www.incapsula.com/blog/bot-traffic-report-2016.html

Kaur, J. & Gupta, V. (2010). Effective Approaches for Extraction of Keywords. *International Journal of Computer Science*, 7.

Kim, Y., Lee, J., & Choi, S. (2020). Validation of deep learning natural language processing algorithm for keyword extraction from pathology reports in electronic health records. *Sci Rep 10* (Article n. 20265). https://doi.org/10.1038/s41598-020-77258-w

Kipp, M. (2007). Tagging Practices on Research Oriented Social Bookmarking Sites. *Proceedings of the Annual Conference of CAIS / Actes du congrès annuel de l'ACSI*. 30. 10.29173/cais223. DOI: 10.29173/cais223

Kipp, M. (2011). Tagging of Biomedical Articles on CiteULike: A Comparison of User, Author and Professional Indexing. *Knowledge Organization, 38*(3). DOI: 10.5771/0943-7444-2011-3-245

Liddy, E. (2001). Natural language processing. In *Encyclopedia of library and information science* (2nd ed., pp. 2126–2136). New York, NY: Marcel Dekker.

Litvak, M., Last, M., Aizenman, H., Gobits, I., & Kandel, A. (2011). DegExt - A Language Independent Graph-Based Keyphrase Extractor. *Advances in Intelligent Web Mastering* - 3, AISC, 86, 121- 130.

Moskovitch, R., Martins, S. B., Behiri, E., Weiss, A., Shahar, Y. A comparative evaluation of full-text, concept- based, and context-sensitive search. *Journal of the American Medical Informatics Association: JAMIA*. 2007 Mar-Apr;14(2), 164-174. DOI: 10.1197/jamia.m1953.

Papagiannopoulou, E. & Tsoumakas, G. (2019). A review of keyphrase extraction. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 10. 10.1002/widm.1339.

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. In *Text mining: applications and theory*, 1 (pp. 1–20).

Strader, C. (2009). Author-Assigned Keywords versus Library of Congress Subject Headings Implications for the Cataloging of Electronic Theses and Dissertations. *Library Resources and Technical Services*, 53, 243-250. DOI: 10.5860/lrts.53n4.243.

Terra, A., Lacruz, C., Bernardes. O, Fujita, M., & Fuente, G. (2020). Subject-access metadata on ETD supplied by authors: A case study about keywords, titles and abstracts in a Brazilian academic repository. *The Journal of Academic Librarianship, 47*(1). https://doi.org/10.1016/j.acalib.2020.102268

Turney, P. (2000). Learning algorithms for keyphrase extraction. *Information Retrieval*, *2*, 303–336.

Voorbij, H. J. (1998). Title keywords and subject descriptors: a comparison of subject search entries of books in the humanities and social sciences. *Journal of Documentation,* 54(4), 466–476. DOI: 10.1108/EUM0000000007178

Witten, I., Paynter, G., Frank, E., Gutwin, C., & Nevill-Manning, C. (1999). Practical Automatic Keyphrase Extraction. In *Proceedings of the 4th ACM Conference of the Digital Libraries*, DL '99 (pp. 254-255), Berkeley, CA, USA.

Ying, Yan, Qingping, Tan, Qinzheng, Xie, Ping, Zeng & Panpan, Li. (2017). A Graph-based Approach of Automatic Keyphrase Extraction. *Procedia Computer Science*, *107*, 248-255. 10.1016/j.procs.2017.03.087.

Yousefi, Zahra & Sotudeh, Hajar & Mirzabeigi, Mahdieh & Nikseresht, Alireza & Mohammadi, Mehdi. (2019). Investigating text power in predicting semantic similarity. *International Journal of Information Science and* Management, 17(1), 17-31.

Zhou, Z., Zou, X., Lv, X., Hu, J. (2013). Research on Weighted Complex Network Based Keywords Extraction. *Chinese Lexical Semantics*, LNCS 8229, 442-452.