Annals of Library and Information Studies
Vol. 70, March 2023, pp. 33-40
DOI: 10.56042/alis.v70i1.71272

Automatic extraction of significant terms from the title and abstract of scientific papers using the machine learning algorithm: A multiple module approach

Bhaskar Mukherjee[a] and Debasis Majhi[b]

[a]Professor, Department of Library & Information Science, Banaras Hindu University, Varanasi,
Email: mukherjee.bhaskar@gmail.com

[b]Junior Research Fellow, Department of Library & Information Science, Banaras Hindu University, Varanasi,
Email: debasismajhidlis@gmail.com

Received: 09 February 2023; *accepted: 27 March 2023*

Keyword extraction is the task of identifying important terms or phrase that are most representative of the source document. Although the process of automatic extraction of keywords from title is an old method, it was mainly for extraction from a single web document. Our approach differs from previous research works on keyword extraction in several aspects. For those who are non-expert of the scientific fields, understating scientific research trends is difficult. The purpose of this study is to develop an automatic method of obtaining overviews of a scientific field for non-experts by capturing research trends. This empirical study excavates significant term extraction using Natural Language Processing (NLP) tools. More than 15000 titles saved in a .csv file was our dataset and scripts written in Python were our process to compare how far significant terms of scientific title corpus are similar or different to the terms available in the abstract of that same scientific article corpus. A light-weight unsupervised title extractor, Yet Another Keyword Extractor (YAKE) was used to extract the results. Based on our analysis, it can be concluded that these algorithms can be used for other fields too by the non-experts of that subject field to perform automatic extraction of significant words and understanding trends. Our algorithm could be a solution to reduce the labour-intensive manual indexing process.

Keywords: Data mining, Title extraction, Natural Language Processing, YAKE, NLTK, Keyword Extraction-NLP

Introduction

This paper is an attempt to extract significant keywords from the title and abstract of a scientific paper by using machine learning process to understand the trends in research on Bioinformatics. While the title of a research paper reveals the thought contents of the overall research, the abstract represents the gist findings of the scientific research. Often, authors choose title carefully to express the research correctly, however, number of times during reviews, the reviewers or the editors suggest alternative title and ask author to change the title without changing the intention of the paper. The reason behind such suggestions is mainly because of compliance with the journal's guidelines. Other reasons may be to keep those terms in the title that searchers usually approach on the online platforms so that the research paper can be accessed by the searcher. It may also be to avoid any such words that are not part of the research or confusing in nature.

On web, search engines play an important role in searching precise results. The search engine bot or spider crawls hundreds of billion pages using their web crawler. In getting search engine optimization (SEO) success, link building between search query and metatags of a research paper is crucial. Among various metatags, title tag is the most important anchor as it is meant to provide a clear and comprehensive idea of what the research paper is all about. Over the past few years, user search behaviour factors were being discussed a lot as logical proof of choosing right title, by contrast authors sometime chose interesting, catchy title to gather attention. More than the author names, the title affects how a paper is perceived by others. Whimsical, amusing or hilarious titles though appealing, may be wrongly categorized in automatic indexing process and may be cited less often[1].

Term extraction from the titles of the scientific research papers is not a complex task in present days. There are so many simple algorithms available through which we can extract single-word terms from the title. However, extraction of multiple terms through machine learning algorithm with correct semantic value is a complex task. Because a single word when combined with other words reveals

different meaning. For example, 'Information Storage and Retrieval' represent a different context than the words like 'information, 'storage' or 'retrieval'. 'Citation metrics' have different connotation than the word 'citation' and 'metrics'. For creating a meaningful network diagram, significant words with correct semantics are more important than creating cluster of significant terms.

Like title, abstract is yet another important element of a research paper to include important words related to methodology, results and findings. For appropriate indexing purpose and for retrieval, abstract plays an important role in actual data representation and interpretation[2]. The title and abstract of a research paper are often freely available to the readers and easily accessible through journal webpages, indexing databases or search engine result page (SERP). In this paper, we take a machine learning approach to address the problem and compare the output by developing frameworks using two modules of python language. The key issues are to define a better specification on machine extraction by comparing the results of title and abstract extraction of research papers.

With the advancement of science, more and more interdisciplinary research is being carried out throughout the world. Biological research nowadays generates large volume of biomolecular data. The open sharing of these data is crucial towards ground-breaking scientific achievements[3]. The reason behind choosing bioinformatics as a field to test our frameworks was that this recently emerging interdisciplinary field has wide application in human endeavour which collects, stores, analyses and disseminates data on DNA and amino acid sequence using computer technology and statistics.

Review of literature

Xue[4] conducted a study for HTML text extraction of web pages by using supervised machine learning approach. They found that the use of both extracted titles and title fields is always better than use of title field alone. And the use of extracted titles is particularly helpful in the task of name page findings. In contrast, Gali[5] observed that HTML text extraction from service-based web pages fails. Because advertisements in the service-based web pages often give more visual emphasis than main headings. Therefore, they proposed a novel method that combines statistical features, linguistic knowledge, and text segmentation. In another study Uzun[6] presented a hybrid approach that contains two steps. The first step discovers informative contents using Decision Tree Learning of machine learning method and creates rules from the results. The second step extracts informative contents using rules obtained from the first step.

There are few other studies which deals with title extraction process. For example, Giuffrida et al.[7] extracted contents from PostScript files using postotext tool, another rule-based system PDFX is described by Constatin[8]. Tkaczyk[9] developed a comprehensive open-source system for extracting metadata and parsed bibliographic references from scientific articles in born digital format.

Rinartha and Kartika[10] suggested a model for keyword extraction by Rapid Automatic Keyword Extraction (RAKE). They suggested that word frequency and RAKE can be combined to get more accurate results in extracting keywords from title. Wand[11] developed an algorithm that can automatically extract keywords from the meta-information of each article and generate the basic data for review articles by implementing Rapid Automatic Keyword Extraction algorithm on the title and abstract of each article. Finally, they classified all extracted keywords into class by calculating Levenshtein distance between each of them.

Nakaijima[12] performed an analysis using 12 years of articles on high-temperature superconductors and developed a method by examining research article, review literature and co-citation among research articles. In another study, Gunawan[13] applied TextRank algorithm to extract the keywords and then generate keyphrase based on the rank on the vertices. Their results explained that the number of assigned keywords play a crucial role in determining the recall value. The recall value of TextRank algorithm raises from 38.4% (5 keywords) to 55.26% (10 keywords) and 61.54% (15 keywords). Jiang[14] suggested a model of automatic extraction keywords in the field third-generation semiconductor materials based on entity recognition (ER) and relation extraction (RE) techniques. Based on this ER and RE, a neural network using domain knowledge (DKNet) is proposed to improve ER performance. Adjusting keyword sequence of each entity type as prior knowledge, adds a dedicated embedding to encode entity categories, then combines prior knowledge and encoded vectors with the context through a gated

information fusion module to assist recognition. By applying this process, they have suggested a multi-aspect attention-based network model (MANet) to enhance the attention to relation-indicative words.

Objectives of the study

- To understand the text-characteristics in scientific titles and their abstracts for a multidisciplinary field like bioinformatics; and
- To compare the output of N-gram text extraction results by developing algorithm through YAKE module.

Methodology

This research is based on Natural Language Processing (NLP) applications for automatic text extraction from scientific articles. For comparing the results of different modules of python compatible modules for text extraction, we needed raw data. We downloaded 15000 records from Web of Science (WoS) database on bioinformatics for the period 2021 to 2022. We used Pandas, a library in python that has functions like analysing, cleaning, exploring, and processing data. Along with pandas, we have used different modules of Natural Language Toolkit (NLTK) like PorterStemmer, WordNetLemmatizer, Yake, CounterVector modules for title extraction. The WordNetLemmatizer was used to group together the different inflected forms of a word so that they can be considered as a single word (eg. Rocks->Lemmatization -> rock; corpora-> Lemmatization -> corpus; Caring -> Lemmatization -> Care etc.). On the other hand, PorterStemmer was used to stem the word so that not only noting but chopping the word into its root form could be achieved (eg. "python", "pythoner", "pythoning", "pythoned" ->PorterStemmer -> python).

Stop words or bag of insignificant words are in-built feature of NLTK which can eliminate commonly used English stop words from title. But on analysing those words, we observed that the lists of such words are insufficient. We therefore developed a dictionary of stop words and run the same in python library. Our newly developed stop wordlist also includes those words which have no significance in scientific titles (eg. 'named', 'formerly', 'accordingly', 'across', 'actually' etc.).

By using above mentioned process, we converted the text into small segments, called tokens, consisting of words. Most of these words are unigram or a single word. As most of the word thus segmented do not serve our needs, we attempted to use tools for bigrams and trigrams. By fixing the lower and upper boundary of n-values, different n-grams were generated using YAKE and CounterVector modules. Yake is a light-weight unsupervised automatic keyword extraction of NLTK of Pandas which relies on statistical text features extracted from a single document to select most relevant keywords of a text[15]. On the other hand, CounterVector is another model of text classification where text data have been vectorized and total occurrence of word can be calculated. It is needless to mention here that NLP models do not understand the textual data, therefore, it is essential to convert such data to vector/numbers.

Results

Figure 1 displays the results of our program and Table 1 analyses the result. It is seen that research article titles on an average consists of 14 words (Sd 4.2), out of which almost 70% words are significant (Sd 2.9) and remaining 30% are stop words. The average number of words per abstract is 220 and the average number of significant words is 133.

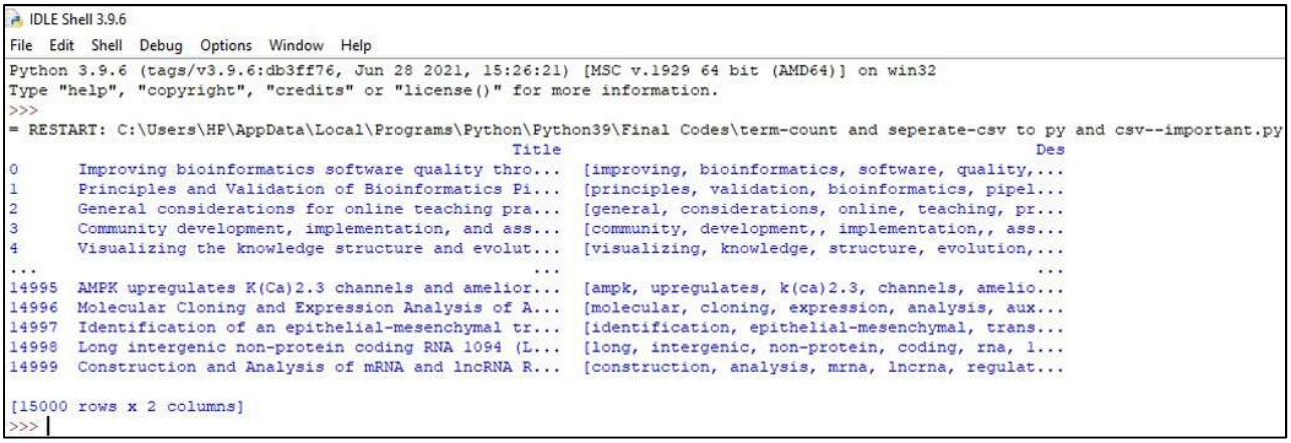

Fig. 1 — Text extraction of significant words excluding stop words

Although a script can be written for identifying the common and uncommon words of a title but the output results we got by applying those script did not fulfil our requirements, and therefore was discarded.

As seen in Fig. 1, although titles were broken into single significant words but the same does not reveal the actual context of the subject. For example, in the line 0 we got a result for the title "Proteomics analysis of lipid droplets in mouse neuroblastoma cells" as ['proteomics', 'analysis', 'lipid', 'droplets', 'mouse', 'neuroblastoma', 'cells']. The term lipid and droplets as single term have no significance until we got our results as 'lipid droplets'. Similarly for the title "The expression and role of tubulin polymerization-promoting protein 3 in oral squamous cell carcinoma" extraction of each significant word does not serve our purpose if we do not get the results like 'squamous cell carcinoma' as a single significant term. Furthermore, term extraction in this process does not provide the overall frequency of significant terms to understand the trends in research.

To solve the problem, we have deployed N-gram technique. In natural language processing, N-grams are continuous sequence of n-number of neighbouring words or tokens in a document. There are several ways N-gram can be deployed. For this study, we have tested N-gram using two different modules of python: one for counting the frequency and another for measuring the most and least important terms. CounterVector is a module which helps in transforming a given text into vector based on its frequency. Scikit-learn's CounterVector with N-gram features of panda module have been deployed here to extract significant keywords from a corpus of text and count its frequency. In Figure 2, the extracted results of our dataset are shown. Here, results show bi-grams like 'hepatocellular carcinoma', 'breast cancer', 'cell carcinoma' along with their frequency in the whole corpus of titles. Along with that, tri-grams like 'squamous cell carcinoma' was also extracted from the text corpus. It is seen that 'bioinformatics analysis' (1084 times), 'hepatocellular carcinoma' (460) and 'breast cancer' (445 times), are

Table 1—Frequency of significant words in titles and abstracts of scientific articles

	Title Characters	Title words	Significant words (Title)	Abstract words	Significant words (Abstract)
Average Frequency	113.23	14.39	10.34	219.94	133.39
Sd	32.19394	4.203348	2.942147	50.88	32.71

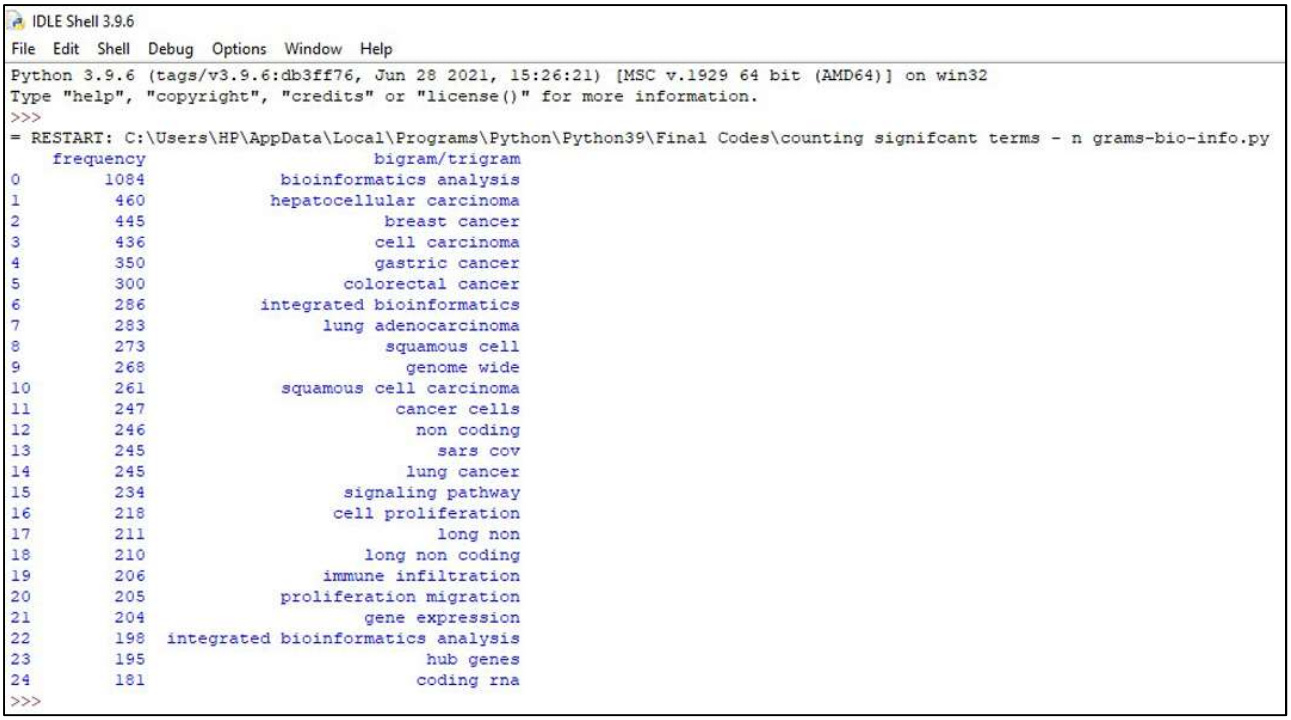

Fig. 2 — Text extraction of significant words from Title using CounterVector

the top three significant words in our dataset. To make a cross-check, we have also tallied these frequencies with the actual dataset available in .csv format and found that the extracted terms and frequency remain same in both.csv files and extracted results of our program.

Further, we have attempted to apply the same code for extraction of frequently used terms in abstracts of the titles. Figure 3 shows the result. Comparing the rank of abstract terms with the title terms, it was seen that there is difference in their rank. The term 'bioinformatics analysis' ranks first in title but it ranks second in abstract. Similarly the top ranked term 'gene expression' in the abstract is the 13th ranked term in title.

Only 36% (9 out of top 29) terms were similar in both title and abstract extracted terms and remaining 64% terms are unique in both title and abstract. And some of the terms in abstract were also repeatedly expressed, like 'protein-protein', 'protein interaction' and 'protein-protein interaction', however such repetition is not seen in title extracted terms.

Numbers in brackets are the frequency

Next, we have attempted to examine how far the algorithm generated keywords are like author generated keywords. From the results as shown in the table 2, algorithm generated keywords are quite accurate and occurrence of such keywords are in

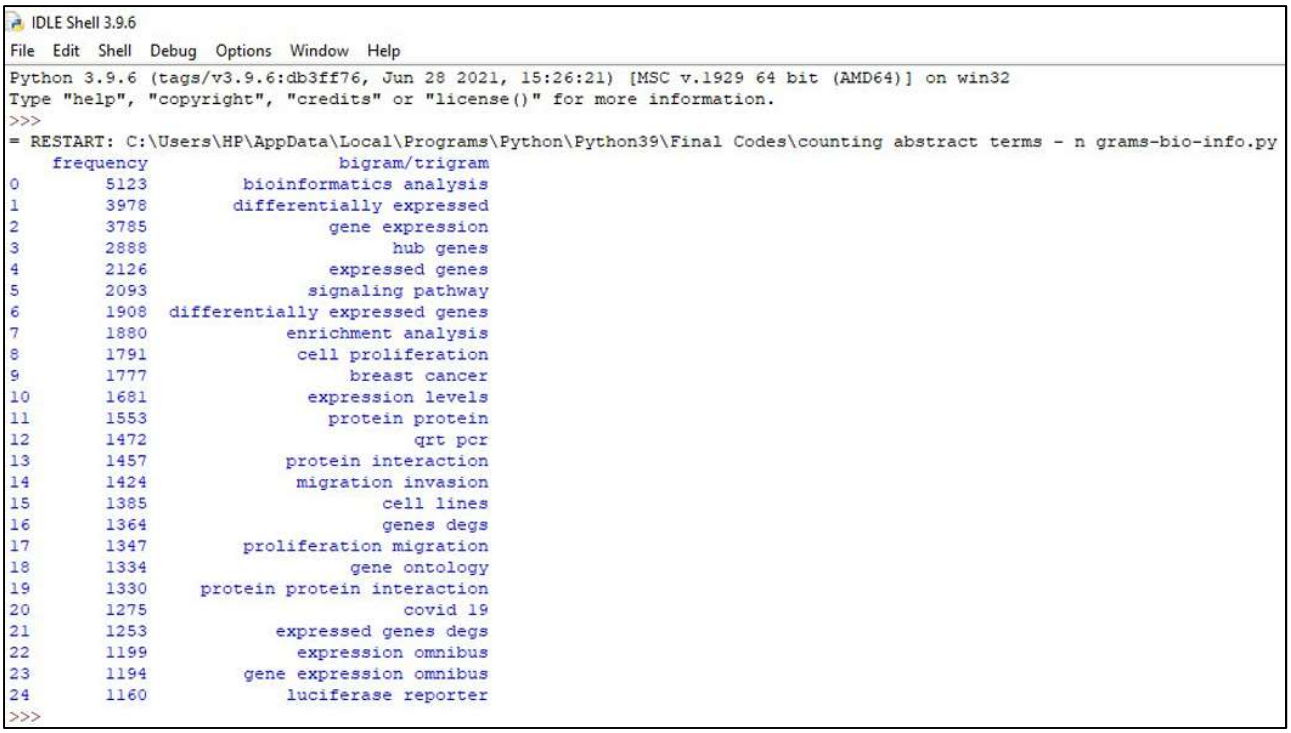

Fig. 3 — Keywords extraction from Abstract through CounterVector (top 25 keywords)

Table 2—Comparison between algorithm-generated title & abstract keywords with author-generated keywords

Author produced keywords	Title extracted keywords	Abstract extracted keywords
Bioinformatics (2596)	bioinformatics analysis (1084)	bioinformatics analysis (5123)
bioinformatics analysis (910)	hepatocellular carcinoma (460)	differentially expressed (3978)
Prognosis (805)	breast cancer (445)	gene expression (3785)
Biomarker (428)	cell carcinoma (436)	hub genes (2888)
Proliferation (329)	gastric cancer (350)	expressed genes (2126)
Proteomics (313)	colorectal cancer (300)	signaling pathway (2093)
hepatocellular carcinoma (284)	integrated bioinformatics (286)	differentially expressed genes (1908)
machine learning (272)	lung adenocarcinoma (283)	enrichment analysis (1880)
breast cancer (270)	squamous cell (273)	cell proliferation (1791)
Immune infiltration (265)	genome wide (268)	breast cancer (1777)

accordance with the author generated keywords. This justifies the validity of our algorithm.

Although CounterVectorize clarifies the most frequent words in the text corpus but unable to identify the more important or less important (significant) keywords of scientific research titles. To overcome the drawbacks, we have written a python script using YAKE (Yet Another Keyword Extraction) module of panda that uses statistical features of a single document to get better results. Figure 4 shows our output.

On analyzing the results of N-gram as generated through YAKE, we observed that N-grams created by the YAKE are more appropriate than others, like SpaCy, RAKE, Gensim. In our result the term 'cell lung cancer' scored2.344861670606526e-06 followed by 'integrated bioinformatics analysis' with score 2.3901992076133613e-06. In our dataset of 15000 titles, these two terms came 126 and 74 times respectively. In YAKE, lower the score, the more relevant keyword is.

To explore to what extent the title extracted keywords are similar with the abstract extracted keywords, we have deployed the abstracts of the titles in YAKE module. As shown in Figure 5, abstracts produced a different set of words than title. Only 24% terms have similarity in both title and abstract. Remaining 76% terms are different in these two sets of results. It is interesting to note here that the term 'cell lung cancer' received highest attention in terms from title extraction process but does not come under top 25 abstract extracted words. Similarly, variation in results have been observed for the term 'gene expression'.

Discussion

Proper indexing of research paper titles is essential for make it searchable through search engines. However, for humans, it is quite difficult to handle such a huge volume of data and index it accurately. This research is an attempt to understand whether it is possible by machine to extract correct keywords that represent the actual content of the scientific titles.

In this work, we present an algorithm that generate data to be further used for identifying significant areas of a research field of the titles by leveraging automatic keyword extraction and similarity calculation. We used three different tools of natural language processing viz, simple, moderate and sophisticated, for this purpose and compared the results of automatic extraction from title and abstract of research articles. The reason behind choosing YAKE for keyword extraction was that it followed unsupervised algorithm, and there was no need to train the corpus. It does not need any pre-defined

Fig. 4 — Keywords extraction from the Title using YAKE (top 25 keywords)

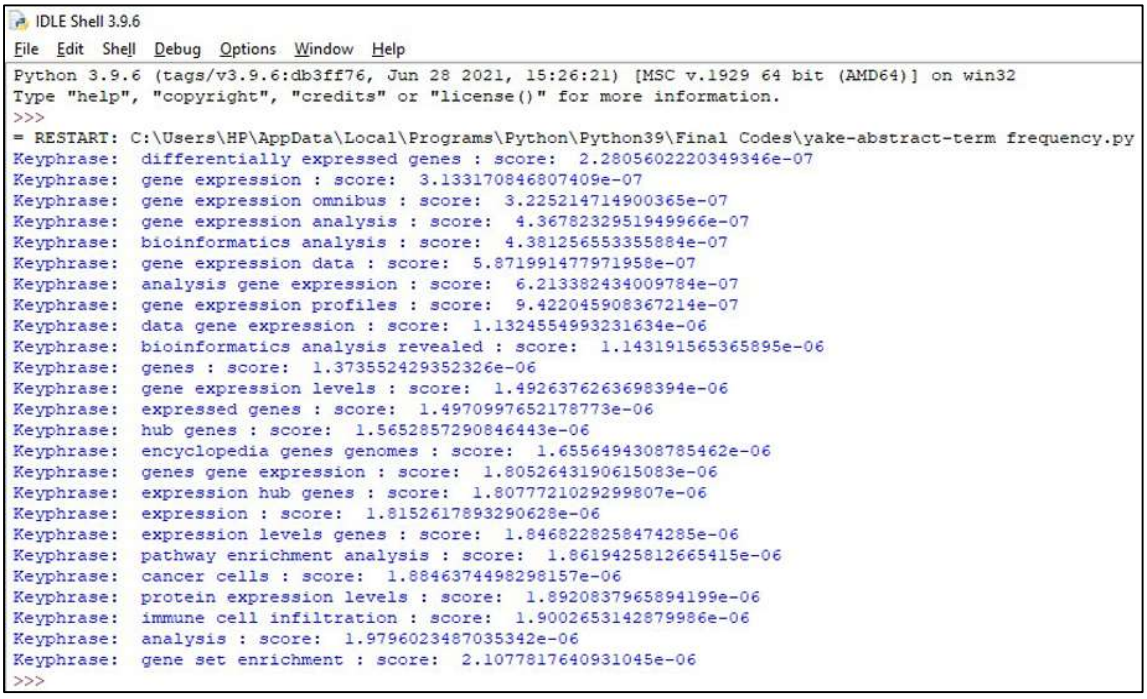

Fig. 5 — Keywords extraction from the Abstract using YAKE (top 25 keywords)

rules, dictionary or thesaurus. It uses statistical features from the text and can extract keywords from a large dataset without re-training. Other procedure of NLP techniques such as part-of-speech tagging (PoS), name-entry recognition (NER), normalization, linguistic parsing or steaming requires specialized tools[15]. For counting the score, YAKE judges the 5 features of a word: casing, word position, word frequency, word relatedness to context, and word different sentence. These scores gained from the 5 features combined into a single score for each keyword. This process was assumed to be better than the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm.

Natural Language Processing although allows machine to interpret human language, however for retrieval, common words (like articles, propositions, pronouns, conjunctions etc.) never serve any purpose. Therefore, text pre-processing is essential for developing effective machine extraction. We observed that scientific research articles usually contain ± 4 such common words. While the readily available stop words list that are derived from non-technical resources are available in NLP domain, rigorously identified subject specific insignificant, uninformative stop words list is needed for effective text extraction process.

For the present research, although we have developed an additional list of almost 535 such stop words which are insignificant in scientific field, however we observed there is no universal rule for some terms like ‘analysis’, ‘differential’ ‘invasion’, ‘network’ ‘computation’ etc. These words can seriously harm topic formation because they create spurious co-occurrence of un-related words. They may probably occupy the top positions of the multiple topics, consequently resulting in inscrutable topic of a domain. Therefore, sophisticated statistical metrics from term frequency is needed for better precision than recall.

Our analysis also shows that machine extracted keywords are quite similar with the author-provided keywords and sometimes machine extracted keywords are more accurate with the actual words of scientific titles. How far these techniques are useful for subjects like humanities or social science by maintaining the thematic structure of the titles needs to be excavated.

We have used the YAKE modules of NLTK to extract the significant words from the titles and abstracts of a corpus of scientific titles. Although our results shows a fair picture to understand the trend of a subject but this technique is inappropriate for extraction of keyword at micro level of granularity (line-by-line) for a multiple row based file. Although the keyword-based approach has the advantage of identifying the uses of technologies for large corpus of data without expert knowledge of the domain, but

it cannot easily represent how technologies are used in their area. New techniques like semantic relations between technologies through a combination of linguistic patterns and statistical methods and Subject-Action-Object (SAO) network analysis have been evolved, considerable number of research are yet to conducted with large scale real dataset.

Conclusion

With rapidly growing research literature in every field, indexing keywords manually is an impossible task. The CounterVector algorithm used in this paper can be used for generating keywords in other fields. The algorithm could be a solution to reduce the labour-intensive manual indexing process. Our algorithm has certain limitations which requires human experts to illustrate significance of some keywords.

References

1 Bavdekar S B, Formulating the right title for a research article, *Journal of the Association of Physicians of India*, 64 (2016) 53–6.

2 Alexandrov AV and Hennerici MG, Writing good abstracts, *Cerebrovascular Discovery*, 23 (2007) 256–59.

3 Vamathevan J, Apweiler R and Birney E, Biomolecular Data Resources: Bioinformatics Infrastructure for Biomedical Data Science, *Annual Review of Biomedical Data Science*, 2 (2019) 199–222.

4 Xue Y, Hu Y, Xin G, Song R, Shi S, Cao Y, Lin CY and Li H, Web page title extraction and its application, *Information Processing & Management*, 43 (2007) 1332–47. https://doi.org/10.1016/j.ipm.2006.11.007.

5 Gali N, Content-Based Title Extraction from Web Page. *12th International Conference on Web Information Systems and Technologies*, 2016, pp. 204–10. https://doi.org/ 10.5220/0005794102040210.

6 Uzun E, Agun HV and Yerlikaya T, A hybrid approach for extracting informative content from web pages, *Information Processing & Management*, 49(4) (2013) 928–44. https://doi.org/10.1016/j.ipm.2013.02.005.

7 Giuffrida G, Shek E C and Yang J, Knowledge-based metadata extraction from postscript files. *Proceedings of the fifth ACM conference on Digital Libraries,* (2000) 77–84. https://doi.org/10.1145/336597.336639

8 Constantin A, Pettifer S and Voronkov A, PDFX: fully-automated pdf-to-xml conversion of scientific literature. *ACM Symposium on Document Engineering*, (2013) 177–180.

9 Tkaczyk P, Szostek P, Dendek J, Fedoryszak M and Bolikowski L, CERMINE -- Automatic Extraction of Metadata and References from Scientific Literature, *11th IAPR International Workshop on Document Analysis Systems, Tours, France* (2014) 217-221. doi: 10.1109/ DAS.2014.63.

10 Rinartha, Komang and Kartika L G S, Rapid Automatic Keyword Extraction and Word Frequency in Scientific Article Keywords Extraction. *3rd International Conference on Cybernetics and Intelligent System (ICORIS)*, (2021) 1–4. *IEEE Xplore*, https://doi.org/10.1109/ICORIS52787.2021. 9649458.

11 Wang J, Su G, Wan C, Huang X and Sun L, A keyword-based literature review data generating algorithm—analyzing a field from scientific publications, *Symmetry*, 12(6) (2020) 903,https://doi.org/10.3390/sym12060903.

12 Nakajima R and Nobuyuki M, Topic Extraction to Provide an Overview of Research Activities: The Case of the High-Temperature Superconductor and Simulation and Modelling, *Journal of Information Science*, 47(5) (2021) 590–608. https://doi.org/10.1177/0165551520920794.

13 Gunawan D, Purnamasari F, Ramadhiana R and Rahmat RF, Keyword extraction from scientific articles in Bahasa Indonesia using TextRank algorithm, In *4th International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM),* (2020) 260–64.

14 Jiang X, He K and Yang B, Automatic information extraction in the third-generation semiconductor materials domain based on DKNet and MANet, *IEEE Access*, 10 (2022) 29367–76.

15 Campos R, Mangaravite V, Pasquali A, Jorge A, Nunes C and Jatowt A, YAKE! keyword extraction from single documents using multiple local features, *Information Sciences*, 509 (2020) 257–89.