ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2134

Species Name Identification for Essential Oils from Biomedical Abstracts Using Text Mining and Natural Language Processing

Species Name of Biomedical Abstracts Related to Essential Oils

Chou-Cheng Chen
Department of Business Management
CTBC Business School
Tainan, Taiwan (R.O.C.)

Abstract:- In recent years, the publication of scientific papers related to essential oils has achieved exponential growth due to the popularity of aromatherapy, although no studies using natural language processing and text mining methods to extract information from scientific articles related to essential oils are currently found. Accordingly, this study is the first to use natural language processing and text mining methods to identify species names appearing in abstracts related to essential oils. We obtained 34,637 abstracts using keywords, "essential oil" to quarry PubMed on 2024/03/15. The 1,081,005 species names of plants and fungi were obtained from Taxonomy FTP on the same day. The nouns from titles of articles related to essential oils were obtained via identification of parts-of-speech and from titles and abstracts extracted within italicized labels. These nouns were used to identify 10,445 plant and fungal species names downloaded from FTP appearing in abstracts related to essential oils with these identification terms being used to detect whether abstracts related to essential oils revealed the species names. 156,371 records contained links between PMID and Taxonomy ID. To the best of our knowledge, our study shows this method can efficiently identify the names of species from abstracts related to essential oil.

Keywords:- *Text Mining; POS; Essential Oil; Species Name.*

I. INTRODUCTION

In recent years, aromatherapy has become increasingly popular, with the use of essential oils becoming more common [1]. Medical research papers related to essential oils that can be found in PubMed have also shown exponential growth since 1999 [2]. We used the term "essential oil" to quarry data from PubMed on 2024/3/15 and obtained 34,637 related scientific articles. In addition, we also quarried keywords from PubMed using terms such as "essential AND oil AND Text AND Mining" and found three related articles [3-5]. Only one of the studies was related to essential oils and text mining, but its main content discussed differences of traditional medicine between various countries [4]. Therefore, this study is the first to use natural language processing and text mining methods to identify the species of plants in scientific articles related to essential oils.

This study used the utility provided by PubMed to download 34,637 relevant articles related to essential oils on 2024/3/15 [6]. On the same day, we also downloaded the taxdmp.zip file from PubMed Taxonomy FTP [7]. After file decompressing, the datasets were treated with natural language processing and text mining methods to identify species names in scientific articles related to essential oils. We found 25,771 articles containing species names with the number of names being 10,445. In this study, we provide the result that links the PMIDs related to essential oils to their species Taxonomy IDs and a total of dataset of 156,371.

II. METHODS

A. Download Abstracts Related to Essential Oil

The scientific abstracts related to essential oils were downloaded using the utility on 2024/3/15, being "https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=essential+oil&retmax=10000&mindate=1833/01/01 &maxdate=2024/03/15" [6, 8]. Since the maximum return PMIDs can only be 10,000 data, "mindate" and "maxdate" was used to adjust the size of return values less than 10,000 [8]. PMIDs were downloaded in batches until all PMIDs were obtained. After downloading the PMID of abstracts related to essential oils, the abstracts were download using the utility containing PMIDs, being "https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&id= 38487964,…..&retmode=xml". "38487964" in the utility is PMID. According to regulations, only 200 PMIDs can be entered at a time. Abstracts were downloaded in batches until all abstracts were obtained.

B. Species Name Data Download

taxdump.tar.gz was downloaded on 2024/3/15, and it has all species names [7]. After unzipping the archive, the datasets including nodes.dmp, names.dmp and division.dmp were used to make dictionary of species names. Nodes.dmp used "tax id" column to link to names.dmp, and nodes.dmp used the "division id" column to link to division.dmp. After linking, the columns including "tax_id", "name_txt", "unique name", "name class" and "division name" were in the same sheet. Species names were filtered using "plants and fungi" of the "division name" column. A total of 1,081,005 pieces of data were obtained, and 632,967 species names were selected

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2134

using data of "class name" column including "blast name", "common name", "equivalent name", "genbank common name", "includes", "in-part", "scientific name" and "synonym". Species abbreviations were created using "scientific name" of "class name" column. To take "Niebla juncosa" as an example: "N. juncosa" was created as "scientific name 1" in "class name" column, "Ni. juncosa" was created as "scientific name 2" in "class name" column, and "Nie. "juncosa" was created as "scientific name 3" in "class name" column, while "name_txt" of the same species were collected in the same data, and each "name_txt" of the same species used the symbol "|" separately. The Taxonomy ID was linked to "name_txt" of the same species, and the final number of species names was tabulated as 490,387.

C. Identification of Species Names in Abstracts

Because the species names include fungi and plants, the program must spend much time identifying species names in abstracts, so we adopted two methods to extract the names in the articles as shown screened from the species name table.

- We used the NuGet package of visual studio to install the Stanford.NLP.CoreNLP package, and used this package for parts-of-speech recognition [9]. The nouns in the title were selected using the tags of "NN", "NNP", and "NNS" after parts-of-speech. "9197967" of PMID was taken as an example; its title is "Allergic contact dermatitis from black cumin (Nigella sativa) oil after topical use." [10]. The result of parts-of-speech recognition was "allergic/JJ contact/NN dermatitis/NN from/IN black/JJ cumin/NN (/-LRB- nigella/NN sativa/NN)/-RRB- oil/NN after/IN topical/JJ use /NN ./.", and the filtered nouns were "contact;dermatitis;cumin;nigella;sativa;oil;use". Finally, 3,628 nouns were extracted.
- In general, the nouns were expressed in italics, and these italicized nouns were extracted using nouns between "" and "" tags in the article, because the word within "" and "" were italicized in XML files. Finally, 6,770 nouns were extracted.

These nouns were used to identify the species whether in abstracts related to essential oils or not, with 46,009 names being taken from the names of the species table. These species names from the table were again identified as to whether they appeared in abstracts related to essential oil or not. Finally, 25,771 abstracts were determined as containing 10,045 species names.

III. RESULT

The result of these identifications in the abstracts can be downloaded from https://www.baiforu.tw/essential_oil_species.txt. Totally, 156,371 records contain three columns: the first is "Taxonomy ID", the second is "Scientific Name", and the third column is "PMID". The first-ranked detected species name from abstracts related to essential oils was "Embryophyta", with a total of 3,805 abstracts containing this species name. We used the keywords "essential AND oil AND anti AND fungi" to quarry articles from PubMed on 2024/03/19 and obtained 1,847 results, which shows that essential oils are related to fungi; consequently, it was found that the number of abstracts related to fungi outnumbered those for plants. After the ranking of "Thymus vulgaris", most of them were common essential oil plants. Statistical results can be downloaded from https://www.baiforu.tw/essential_oil_species_identification_count.txt. The first column is "Taxonomy ID", the second column is "Scientific Name", and the third column is the counting number of abstracts related to specific species.

IV. CONCLUSION AND DISCUSSION

To our best knowledge, this is the first article using text mining and natural language processing to detect species names mentioned in abstracts related to essential oils. Although many species of fungi have been identified, many of the species used in producing essential oils have also been identified. In future research, we will further remove the names of various species of fungi and only select plant species for producing essential oils.

REFERENCES

[1]. B. Cooke and E. Ernst, "Aromatherapy: a systematic review," Br J Gen Pract, 50(455): pp. 493-498, 2000.
[2]. E.W. Sayers, et al., "Database resources of the national center for biotechnology information," Nucleic Acids Res, 50(D1): pp. D20-D26, 2022.
[3]. D. Bi, Ju-E Guo, E. Zhao, S. Sun and S. Wang, "Identifying environmental and health threats in unconventional oil and gas violations: evidence from Pennsylvania compliance reports," Environ Sci Pollut Res Int, 29(15): pp. 22742-22755, 2022.
[4]. K. Domingues, N.H. Franco, I. Rodrigues, G. Stilwel and M.M.-S. Ana, "Bibliometric trend analysis of non-conventional (alternative) therapies in veterinary research," Vet Q, 42(1): pp. 192-198, 2022.
[5]. Dos Santos, N.S.S., et al., "Biotransformation of 1-nitro-2-phenylethane [Formula: see text] 2-phenylethanol from fungi species of the Amazon biome: an experimental and theoretical analysis," J Mol Model, 29(8): pp. 223, 2023.
[6]. Sayers E., "The E-utilities In-Depth: Parameters, Syntax and More, " 2009 2022/11/30 [cited 2024 04/18]; Available from: https://www.ncbi.nlm.nih.gov/books/NBK25499/.
[7]. Schoch C.L., et al., "NCBI Taxonomy: a comprehensive update on curation, resources and tools, Database (Oxford), 2020, 2020.
[8]. "The 9 E-utilities and Associated Parameters," [cited 2024 4/18]; Available from: https://www.nlm.nih.gov/dataguide/eutilities/utilities.html.
[9]. Manning C., Surdeanu M., Bauer J., Finkel J., Bethard S., and McClosky, D., "The Stanford CoreNLP natural language processing toolkit," in Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, 2014.
[10]. Steinmann A., Schätzle M., Agathos M. and Breit R., "Allergic contact dermatitis from black cumin (Nigella sativa) oil after topical use," Contact Dermatitis, 36(5): pp. 268-276, 1997.