International Journal of Social Science and Economic Research

ISSN: 2455-8834

Volume:07, Issue:09 "September 2022"

USING SOFTWARE TO EXTRACT KEY FINDINGS FROM SCIENTIFIC RESEARCH PAPERS: A CASE STUDY USING RESEARCH ON DIETS

Jai Agarwal, Taruna Agarwal, Hrithik Jain, Yilun Wu, John Leddo, Saanvi Lamba, Rohan Penmetsa, Helena Gabrial, Satvik Matta, Rohan Matta, John Wu, AmulyaGottipati, JatinPalvai, Rohit Rajakumar, Kinnari Chaubal, Nayan Shrestha, Deepika Ravi, SathvikRedrouthu, Dillon Michlena, Dev Doshi, Reva Jasty, Deeya Sharma, Mitra Manikandan, Julian Burke, Aneesh Sreedhara, ShriyanBachigari, SmaranPasupuleti, Riya Pasupuleti, Vihaan Cherukuri, Samarth Jain, Jishnu Patel, Aryan Dotiwalla, Manav Sabharwal, Joy Thumma, Rishon Nimmakayala, Eshwar Dokku, Annie Li, SudhitSangela, Nanda Pailla, Owen Yeung, SharanyaChilukuri, Prathima Prakash, Rohan Suri, Surya Vallamkonda, Tanush Bhardwaj, Arin Rahman, Charu Mehta, Anish Somu, Pranav Velleleth, Sumit Kamath, SudarsanNallabola, AdvaiitIyer, Trisha Nittala, HashmitaNittala, Kabilan Prasanna, Pranav Narravula, Nihar Xavier, Nevin Philip

Dr. John Leddo is the director at research at MyEdMaster

MyEdMaster, LLC., Herndon, Virginia, USA

DOI: 10.46609/IJSSER.2022.v07i09.026 URL: https://doi.org/10.46609/IJSSER.2022.v07i09.026

Received: 25 September 2022 / Accepted: 5 October 2022 / Published: 10 October 2022

ABSTRACT

With the explosion of information on the Internet, search engine users still find themselves having to weed through a myriad of websites to ensure that they find the relevant information. This is even more cumbersome in dynamic subject areas, such as scientific research, where research findings may not be stable and even contradictory. Laypeople are especially burdened since they may lack the knowledge to evaluate what scientific papers are actually concluding. The present paper describes software that reads scientific papers and distills their principal findings in a format that laypeople can understand. This software is evaluated in the topic area of research on diet. A separate paper evaluates this software in the topic area of research on Covid-19.

Introduction

Over time, there has been an explosion of information available to people over the Internet. While search engines have become more sophisticated at retrieving information, virtually any search term a person enters into a search engine yields millions of websites. This creates

ISSN: 2455-8834

Volume:07, Issue:09 "September 2022"

inherent problems. No search engine user will have the time and inclination to go through the retrieved websites to determine which websites contain the best and most relevant information. Even within a given website, it is not always immediately apparent where the relevant information is. Of course, once the information is found, it may be difficult for the user to understand, particularly if the topic in question is technical in nature. Perhaps no subject area exemplifies this point more than medicine.

In our previous work (Boina et al, 2021a, b, c, d), we argued that search engines and personal assistants of the future need to do more than retrieve websites but alsoread through them and extract and learn information necessary to answer users' questions. In these studies, we developed software that accepted medical conditions from users, went on the Internet, retrieved information from the sites, learned it and then diagnosed medical conditions in the users with greater than 90% accuracy. These medical conditions included, among others, a variety of skin ailments and stress.

One of the characteristics that the medical conditions described in our previous work had in common is that the knowledge about these conditions was fairly stable, and that there were agreed upon criteria for diagnosing them. However, this is not true for all medical topics. For example, while professionals have studied the effects of aging on people and the effects of factors like diet and exercise on health for years, recently the two fields have converged and researchers are now looking at whether lifestyle choices regarding such things as diet and exercise can actually slow down or even reverse the aging process.

As with any new field of inquiry, such investigation still involves tracking a moving target. The optimal diet to slow down aging has not been fully articulated and it may, in fact, depend on the particulars of the person eating that diet (e.g., age, gender, lifestyle, presence of co-morbidities).In the area of diet, research has centered on three main categories: what to eat, when to eat and how much to eat.

For example, when it comes to what to eat, many diets have become popular. One example of this is the ketogenic diet, a diet high in fat and protein, but low in carbohydrates. A study with mice on a ketogenic diet has shown that this diet extended their longevity. Motor functions, memory, and muscle mass were preserved in ketogenic mice. Also, protein acetylation was increased in the liver and skeletal muscle of the ketogenic mice (Roberts et al., 2017).

Multiple studies have also been done on the Mediterranean diet, a diet low in protein and higher in carbohydrates and have concluded that a diet that adheres to the principles of the traditional Mediterranean one is associated with longer survival (Trichopulou and Vasilopoulou, 2000). The

ISSN: 2455-8834

Volume:07, Issue:09 "September 2022"

Greek version of the Mediterranean diet is dominated by the consumption of olive oil and by the high consumption of vegetables and fruits.

When it comes to when to eat, the concept of fasting has received great currency in the literature. Intermittent and periodic fasting (IF and PF, respectively) are emerging as safe strategies to affect longevity and healthspan by acting on cellular aging and disease risk factors while causing no or minor side effects. IF lasting from 12 to 48 hours and repeated every 1 to 7 days and PF lasting 2 to 7 days and repeated once per month or less have the potential to prevent and treat disease, but their effect on cellular aging and the molecular mechanisms involved are only beginning to be unraveled. There are also the therapeutic potential and side effects of IF and PF with a focus on cancer, autoimmunity, neurodegeneration, and metabolic and cardiovascular disease (Longo et al., 2021).

As for how much to eat, there is a lot of evidence that calorie restriction, limitation of available nutrients without malnutrition, can reduce the incidence of and slow the progression of many age-related pathologies. A broad review of meta-analyses of studies conducted on rodents showed that a low-calorie diet played a vital role in increasing the lifespan of an individual (Ekmekcioglu, 2020).

It is challenging enough for professionals to keep up with an evolving research base, but what if you are a layperson who just wants to live longer and be healthier? Is protein good and carbohydrates bad as suggested by the keto diet or vice versa, as suggested by the Mediterranean diet? Is it better to fast a certain amount each day or periodical skip one or more days of eating?

To address a problem such as this, it would be beneficial to have software that can read scientific journals and then summarize the key findings such that laypeople (and even professionals) have a tool that helps them weed through the complexities of virtually unlimited data and weed out the relevant findings. The purpose of the present project is to create such a tool.

The software

The code performs automatic information extraction from academic journal articles to the literature review template with certain specifications. This information extraction script is written in Python and utilizes natural language processing and text analysis. Applying this script saves significant time compared to manually reading journal articles and filling out information based on these articles. Note that this script only fills out partial columns of the template since the other information required to extract from the journal articles and papers earns high variance in context and needs subjective judgment from readers. The columns that will be filled out using this script: Title, Paper source (website), Year of Publication, Journal Type (review), Link, Human trial or not, Data (Interventions), Positive/negative findings, Conclusion Summary.

ISSN: 2455-8834

Volume:07, Issue:09 "September 2022"

We use the script by keeping the code, the academic journal article pdf files, and the literature review template into the same directory. We then record all the file names and store them into a Python list under the "if__name__= = '__main__':". The first thing that happens after this is the extraction of DOI (Digital Object Identifier), which is a string of numbers, letters and symbols used to permanently identify an article or document and link to it on the web.

It is extracted by finding the start and the end index of the string which represents the doi and returns the string, this string can sometime contain a period in the end. So, that period is removed by removing the last character, also the doi might start with an "org/" so we ignore the first 4 characters. The last character in the doi should be a number, so we remove the last non-numeric characters unless a number is encountered and then return the string. This returned string is the DOI. Now, the DOI is converted to json using the doi2json function so that the query function can form the URL and then use the GET method to retrieve information from the given server using that URL. From this retrieved information, we can extract the title, publisher, year and journal type. Next, we find whether the given research is a human trial or not. This is done by checking the frequency of the word 'human'. After this we find out the interventions, which are basically the health habits/diet/sleep/stress that the study analyses. The approach taken here is to analyze the sentences between the "keywords" paragraph and "introduction" paragraph in the paper. Each word in the sentence between these paragraphs is matched with the values of the dictionary intervention Word Bank. Wherever a match is found, its corresponding key of the value is returned. If no match is found, then 'others' is returned. The next thing to do is to find the effects of the experimental study (beneficial, detrimental or neutral). Sentiment analyses is performed on the paragraph that lies between the Abstract and the Keywords, if the label of the analyses is positive, then beneficial is returned. If negative then detrimental is returned, else neutral is returned. This is done by getting the start index of the Abstract and the end index of the Abstract. If the difference between the start and the end index is more than 512 words, then only 510 words starting from the start index is taken. Finally, we summarize the paper by analyzing the text between the heading Conclusion and Acknowledgements (if Acknowledgements are not there in the paper, then we analyze until References). This is done by finding the differences of first instance of the end index and the first instance of the words "Conclusion", "Conclusion ", "Conclusions ", "Summary", "Discussion", "CONCLUSIONS", "CONCLUSION". The smallest of all these differences is taken as input to generate the summary. A dictionary of word-frequency pair is created where words that are not part of STOP_WORDS and punctuations are included with their value being the frequency of those words. After this, a dictionary of sentence-score pair is created, the score of the sentences is calculated by checking the number of words that the sentence and the word-frequency dictionary have in common, and then the frequency of that common word in the word-frequency pair dictionary is taken/added as score in the sentence-score pair dictionary. Then, the summary is calculated by finding the nlargest scores of the

ISSN: 2455-8834

Volume:07, Issue:09 "September 2022"

sentence where n here is the number of sentence times the percentage of paragraph we want as summary. After getting all these values, the columns that can be filled in the template will be filled.

Testing the Software

The goal of the present software is to read scientific articles and write summaries of them that are understandable by laypeople, thus enabling them to keep up with the latest in scientific research. Accordingly, we tested the accuracy of the software by having it read scientific articles and write summaries. The summaries focused on the main relationship between the independent variables and the dependent variables. These summaries were compared with actual text taken from the scientific articles themselves. A total of 11 articles, each focusing on research involving interventions related to health and longevity, were used. In a separate paper, we evaluate the software's effectiveness in reading and summarizing papers involving Covid-19.

Results

Each of the research papers' own text-based summaries and the software's summaries are shown below to demonstrate the effectiveness of the software.

Paper 1 Title: The longitudinal effects of induction on beginning teachers' stress

Link:https://pubmed.ncbi.nlm.nih.gov/29998489/

Conclusion from paper:

Perceived stress causes and stress responses can change over time. Specific induction arrangement elements appear to be powerful elements to reduce the level, and the change over time, of specific perceived stress causes and stress responses.

Conclusion from model:

Workload reduction increases the quality of life and decreases all the negative social aspects and stress responses

Paper 2 Title: Mobile phone use and stress, sleep disturbances, and symptoms of depression among young adults--a prospective cohort study

Link:https://pubmed.ncbi.nlm.nih.gov/21281471/

Conclusion from paper:

ISSN: 2455-8834

Volume:07, Issue:09 "September 2022"

High frequency of mobile phone use at baseline was a risk factor for mental health outcomes at 1-year follow-up among the young adults. The risk for reporting mental health symptoms at follow-up was greatest among those who had perceived accessibility via mobile phones to be stressful. Public health prevention strategies focusing on attitudes could include information and advice, helping young adults to set limits for their own and others' accessibility.

Conclusion from model:

Excessive phone use at any age causes more mental health issues in the future.

Paper 3 Title: Extending healthy life span--from yeast to humans

Link:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3607354/

Conclusion from paper:

Because extreme DR can lead to several detrimental health effects it will be crucial to determine whether chronic severe DR in humans increases susceptibility to infections and wound-related pathologies and mortality. Additional studies are warranted to evaluate the calorie intake and relative macro- and micronutrient composition needed for optimal health and successful aging.

Conclusion from model:

Dietary Restriction is very beneficial, but severe DR can be detrimental, and more studies must be conducted to determine the negative effects of severe DR.

Paper 4 Title: Effects of Calorie Restriction on Health Span and Insulin Resistance: Classic Calorie Restriction Diet vs. Ketosis-Inducing Diet

Link:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8071299/

Conclusion from paper:

Calorie restriction is the most robust intervention known until the moment to increase maximal lifespan and health span in large studies, with high evidence, biological plausibility and explained biochemical mechanisms

Conclusion from model:

Caloric Restriction is an intervention that can greatly improve both longevity and quality of life.

Paper 5 Title: Physical activity and life expectancy with and without diabetes: life table analysis of the Framingham Heart Study

ISSN: 2455-8834

Volume:07, Issue:09 "September 2022"

Link:https://pubmed.ncbi.nlm.nih.gov/16373893/

Conclusion from paper:

Moderately and highly active people have a longer total life expectancy and live more years free of diabetes than their sedentary counterparts but do not spend more years with diabetes.

Conclusion from model:

Type two diabetes (one of the leading causes of premature death in America) can be prevented through exercise and not having diabetes increases your lifespan. Even a low physical activity will create a longer lifespan without diabetes which means a better quality of life.

Paper 6 Title: Healthy lifestyle and life expectancy in people with multimorbidity in the UK Biobank: A longitudinal cohort study

Link:https://pubmed.ncbi.nlm.nih.gov/32960883/

Conclusion from paper:

Regardless of the presence of multimorbidity, engaging in a healthier lifestyle was associated with up to 6.3 years longer life for men and 7.6 years for women; however, not all lifestyle risk factors equally correlated with life expectancy, with smoking being significantly worse than others.

Conclusion from model:

Living a healthy lifestyle improves length of life

Paper 7 Title: A Periodic Diet that Mimics Fasting Promotes Multi-System Regeneration, Enhanced Cognitive Performance, and Healthspan

Link:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4509734/

Conclusion from paper:

Prolonged fasting (PF) promotes stress resistance but its effects on longevity are poorly understood. We show that alternating PF and nutrient-rich medium extended yeast lifespan independently of established pro-longevity genes.

Conclusion from model:

Periodic diet that mimics fasting improves quality of life.

ISSN: 2455-8834

Volume:07, Issue:09 "September 2022"

Paper 8 Title: Goldenrod Extract has a Senolytic Effect

Link:https://www.lifespan.io/news/goldenrod-extract-has-a-mild-senolytic-effect-in-cell-culture-study/

Conclusion from paper:

The results show that the extract was effective at slowing down the cells' journey to senescence and could suppress a number of proinflammatory pathways. However, what happens in the culture does not always happen in the human body. Therefore, don't rush out and buy goldenrod supplements based on this initial data

Conclusion from model:

The plant-compound known as goldenrod extract has the capability of ridding the body of senescent and inflamed cells. Natural remedies could be used in the future for less aggressive methods of medicines.

Paper 9 Title: Alpha-Ketoglutarate Decreases Biological Age in Human Study

Link:https://www.lifespan.io/news/alpha-ketoglutarate-decreases-biological-age-in-human-study/

Conclusion from paper:

AKG levels generally decrease with aging. AKG supplementation has been shown to improve lifespan and healthspan in various model animals, but human studies are scarce.

Conclusion from model:

Certain supplements based on "longevity drugs" could possibly slow the concept of aging.

Paper 10 Title: Plasmalogens Alleviate Age-Related Cognitive Decline in Mice

Link:https://www.lifespan.io/news/plasmalogens-alleviate-age-related-cognitive-decline-in-mice/

Conclusion from paper:

Plasmalogen deficiency looks increasingly interesting as a target for treating Alzheimer's disease and other types of age-related cognitive decline. It would also be interesting to know whether plasmalogen supplementation has any effect on lifespan.

ISSN: 2455-8834

Volume:07, Issue:09 "September 2022"

Conclusion from model:

Synaptic structures that enable individuals to process information were kept intact for longer after surveying the mice introduced to the drug. For human individuals, this drug can promote brain functionality.

Paper 11 Title: The Sleep-Immune Crosstalk in Health and Disease

Link:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6689741/

Conclusion from paper:

Clinicians should be aware that many diseases are comorbid with sleep disturbances and encourage the patients to improve their sleep behavior and educate them about sleep hygiene (i.e., good sleep habits). This may have a beneficial effect on the severity and progression of the disease.

Conclusion from model:

Sleep deprivation is unhealthy for a body and create severe immune system and blood-count related aging problems.

Discussion

Generally speaking, the summaries produced by the software represented good abstractions of the main independent-dependent variable relationships described in the articles. As such software such as this can serve as a powerful tool to help people sift through large quantities of information to distill out the key takeaways. This can prove a valuable resource for both laypeople and professionals.

In spite of the success of the present software, there are enhancements that could make the software even more useful. As is often the case, research and other papers present different or even conflicting findings. In the field of scientific publication, these differences and also similarities are aggregated and summarized in review papers. An obvious enhancement to the present software is to enable it to perform similar meta-analyses. In this enhancement, the software could separate studies by their findings and report studies that report consistent findings and those that disagree with those findings. The next step would be to look at differences between the studies to identify potential variables that can account for differences. This could be a valuable source of potential research questions to investigate.

The present software also requires its user to retrieve the scientific articles and then feed them into the software. This can be a time-intensive process. Another enhancement is to integrate the

ISSN: 2455-8834

Volume:07, Issue:09 "September 2022"

current technology with our automated search and retrieval technology that we reported in our previous papers (Boina et al., 2021a, b, c, d). This would make the tool even more powerful as it would then accept a topic from a user, perform the Internet search and retrieval of the technical articles and then summarize their main points. By combining these technologies and enhancing them with the capabilities described above, we can create a next generation of search engines that do more than just retrieve information but also provide its users with the key takeaways from that information.

References

Boina, N., Agarwal, J., Agarwal, T., Leddo, J. et al**.** (2021a). A Novel Meta-Machine Learning Approach to Diagnose Stress from Environmental Factors Using Automated Knowledge Graphs, International Journal of Social Science and Economic Research, *6(12), 4961-4970.*

Boina, N., Agarwal, J., Agarwal, T., Leddo, J. et al. (2021b). A Novel Meta-Machine Learning Approach to Diagnose Stress from Individual Factors Using a Self-retrieved Dataset and then Provide Directed Treatment, International Journal of Social Science and Economic Research, *6(12), 4933-4944.*

Boina, N. et al. (2021c). A Novel Meta-Machine Learning Platform Able to Autonomously Learn How to Diagnose Acne and Jaundice. International Journal of Social Science and Economic Research, 6(10), 4151-4158.

Boina, N. et al. (2021d). A Novel Meta-Machine Learning Platform Able to Autonomously Learn How to Diagnose Autism, Breast Cancer, Melanoma Mole Cancer and Pink Eye. International Journal of Social Science and Economic Research, 6(10), 4159-4171.

Ecmekcioglu, C. (2020). Nutrition and Longevity – From mechanisms to uncertainties. Critical Reviews in Food and Science Nutrition, *60(18), 3063-3082.*

Longo, V.D., Di Tano, M., Mattson, M.P. &Guidi, M. P. (2021). Intermittent and periodic fasting, longevity and disease. Nature Aging, *1, 47-59.*

Roberts, M.N., Wallace, M.A., Tomilov, A.A., Cortopassi, G.A., Ramsey, J.J.& Lopez-Dominguz, J.A. (2017). A Ketogenic Diet Extends Longevity and Healthspan in Adult Mice.Cell Metabolism, *26, 539-546.*

Trichopoulou, A. &Vasilopoulou, E. (2000). Mediterranean Diet and Longevity. British Journal of Nutrition, Dec;84 Suppl 2:S205-9. doi: 10.1079/096582197388554. PMID: 11242471.

