

# Web Application for the study of the relationship between Obesity and intestinal microbiome based on PubMed text mining

**Neus Torrent Ample**

Máster Universitario en Bioinformática y Bioestadística. UOC-UB  
Inteligencia artificial y desarrollo web

**Romina Astrid Rebrij**

**Antoni Pérez Navarro**

2<sup>nd</sup> of June of 2022



Esta obra está sujeta a una licencia de  
Reconocimiento-NoComercial-  
SinObraDerivada3.0 España de  
Creative Commons

## FICHA DEL TRABAJO FINAL

<b>Título del trabajo:</b>	<i>Web Application for the study of the relationship between Obesity and intestinal microbiome based on PubMed text mining</i>
<b>Nombre del autor:</b>	<i>Neus Torrent Ample</i>
<b>Nombre del consultor/a:</b>	<i>Romina Astrid Rebrij</i>
<b>Nombre del PRA:</b>	<i>Antoni Pérez Navarro</i>
<b>Fecha de entrega (mm/aaaa):</b>	<i>06/2022</i>
<b>Titulación:</b>	<i>Máster Universitario en Bioinformática y Bioestadística UOC-UB</i>
<b>Área del Trabajo Final:</b>	<i>Inteligencia artificial y desarrollo web</i>
<b>Idioma del trabajo:</b>	<i>Inglés</i>
<b>Número de créditos:</b>	<i>15</i>
<b>Palabrasclave</b>	<i>Data Mining, PubMed, Metabolic</i>
<b>Resumen del Trabajo:</b>	
<p>La prevalencia de obesidad ha alcanzado proporciones epidémicas y actualmente afecta a más de 2000 millones de personas.</p> <p>Durante la última década, la obesidad y otras enfermedades metabólicas se han asociado con 227 variantes genéticas involucradas en diferentes vías metabólicas así como al desequilibrio del microbioma intestinal.</p> <p>El crecimiento exponencial de la literatura biomédica supone un reto para mantenerse al día con los últimos hallazgos. Además, encontrar relaciones entre diferentes disciplinas se ha vuelto aún más difícil para la comunidad científica. Últimamente, la minería de datos se propone como una herramienta útil para descubrir nuevos patrones y tendencias contenidas en los documentos científicos.</p> <p>Se aplican técnicas de minería de textos de la literatura biomédica, concretamente resúmenes de PubMed, con el fin de desarrollar una aplicación web interactiva, METAVOLIKOS, para estudiar la relación de genes y microbioma intestinal en pacientes con obesidad y otras enfermedades</p>	

metabólicas.

Concretamente, se estudia la relación entre términos a través del análisis semántico latente y posterior similitud por coseno.

El pipeline del trabajo se integra al desarrollo de la aplicación web utilizando el paquete Shiny de R.

Se infiere una relación a partir de publicaciones sobre enfermedades metabólicas donde no necesariamente hay artículos que incluyen el término “Obesidad” o el gen o población bacteriana relacionada.

Este trabajo muestra cómo el análisis semántico latente puede arrojar luz para crear nuevas hipótesis de investigación.

METAVOLIKOS es una aplicación de minería de texto fácilmente utilizable por científicos no especialistas, que facilita el desarrollo de nuevas líneas de investigación.

#### **Abstract:**

Worldwide, the prevalence of obesity has reached epidemic proportions. The global obesity epidemic continues its relentless advance, currently affecting >2 billion people.

Over the last decade, 227 genetic variants involved in different biological pathways and distortion of the normal microbial balance have been associated with obesity and other metabolic diseases.

Exponential growth of biomedical literature has supposed a challenge for professionals to keep up with the latest findings. Moreover, to find connections between different topics or disciplines has become even more difficult for the scientific community. In recent decades, the use of data mining has been proposed to glean useful insights, discovering novel patterns, relationships and trends contained within the documents.

Text mining techniques were applied from the biomedical literature, specifically PubMed abstracts, in order to develop an interactive web application, METAVOLIKOS, to study the relationship of genes and intestinal microbiome in patients with obesity and other metabolic diseases.

Concretely, the relationship between terms was studied through latent semantic

analysis and subsequent cosine similarity.

The work pipeline was integrated into the development of the web application using the Shiny package of R.

A relationship has been inferred from publications on metabolic diseases where articles that include the term “Obesity” or explicitly the related gene or bacteria are not necessarily included. So, this work shows how latent semantic analysis can shed light to create new research hypotheses.

METAVOLIKOS is a text mining application that is easily adaptable and usable by bench scientists, which inspires to develop new research lines.

## Index

1	Summary .....	1
2	Introduction .....	2
	2.1 Context and justification of the work .....	2
	2.2 Objectives.....	3
	2.3 Approach and method to follow .....	4
	2.4 Planning with goals and timing .....	7
	2.5 Expected results .....	8
	2.6 Brief description of the other chapters of the report.....	9
3	State of the art .....	10
4	Methodology .....	16
5	Results and discussion .....	29
6	Conclusions .....	41
	6.1 Conclusions .....	41
	6.2 Future lines.....	42
	6.3 Planning follow-up .....	43
7	Glossary .....	44
8	Bibliography .....	45

## List of figures

Figure 1. Work scheme for the extraction of genes and changes in the intestinal microbiota associated with obesity. Some text-mining facilities of pubmed.minerR package are displayed [modified from 10]. .....	6
Figure 2. Gantt Diagram of the project. ....	8
Figure 3. This figure shows the exploding number of articles available from Medline over 1940-2005 (data retrieved from the SRS server at the European Bioinformatics Institute; <a href="http://www.ebi.ac.uk/">www.ebi.ac.uk/</a> ). In 2003, about 560,000 articles were added to Medline, and from 2000 to 2003, 2 million articles [13]......	10
Figure 4. Stages of knowledge discovery process [14]. ....	11
Figure 5. Data mining techniques from 2000-2011. It shows the important data mining techniques trends for association rules, genetic algorithms, clustering, artificial neural networks, a priori algorithms, support vector machines, feature selection, customer relationship management, classification, neural networks and decision trees [16]. ....	13
Figure 6. General schematic of the methods used in text and data mining: (A) information recovery, (B) information extraction, (C) interpretation of the information (the integration of various previously corroborated hypotheses can produce a new combined hypothesis). The three steps required for extracting information are shown in panel B: (1) breakdown of the information into basic units (e.g. words), (2) identification of biological entities, and (3) interpretation of the relationship between biological entities [17]. ....	14
Figure 7. Search field descriptions and tags of PubMed. ....	16
Figure 8. Search in PubMed using [MH] and [DP] tags. ....	17
Figure 9. Inspection of S4 class object. ....	18
Figure 10. Genes cited into the publications of the corpus. ....	20
Figure 11. Words used into the publications of the corpus. ....	20

Figure 12. Bacteria population and related concepts extracted from the secondary corpus.....	21
Figure 13. LSA uses bag of word model, which results in a term-document matrix (occurrence of terms in a document). Rows represent terms and columns represent documents. LSA learns latent topics by performing a singular matrix decomposition on the document-term matrix using singular value decomposition [19]. .....	23
Figure 14. Two key components of every Shiny app [21].....	27
Figure 15. The shortest viable Shiny app [21]. .....	27
Figure 16. Reactivity programming components of Shiny apps [21].....	28
Figure 17. Home of METAVOLIKOS. ....	29
Figure 18. Search publications menu. ....	30
Figure 19. Query visualization in PubMed from Publications Search tab. ....	30
Figure 20. Results of the PubMed search form “Cushing’s syndrome” and publication date between 2019 and 2020.....	31
Figure 21. Detail of the bottom of the table where the title and abstract of the selected publication are displayed.....	31
Figure 22. Page to which the link of each article redirects. ....	32
Figure 23. Word and gene cloud for the search of “Cushing’s syndrome” and publication date between 2019 and 2020.....	32
Figure 24. Result of frequency cited words and genes of “Cushing’s syndrome” and publication date between 2019 and 2020 in table format. ....	33
Figure 25. Boxes to modify the minimum and maximum frequency of words and genes to be displayed into the cloud graphs and tables.....	34
Figure 26. Genes and metabolic diseases relation tab.....	35
Figure 27. Related gene table (first tab). ....	35
Figure 28. Related gene table (second tab). ....	36
Figure 29. Obesity-gene relation 3D graph.....	37



Figure 30. Microbiota and metabolic diseases relation tab.....	38
Figure 31. Related microbiota table (first tab).....	38
Figure 32. Related microbiota table (second tab).....	39
Figure 33. Obesity-microbiota relation 3D graph (default output). ....	40
Figure 34. Obesity-microbiota relation 3D graph (rotated output).....	40

## **List of tables**

Table 1. Tasks and goals of the project.....	7
Table 2. Top mining algorithms identified by the IEEE International Conference on Data Mining in December 2006. These 10 algorithms cover classification, clustering, statistical learning, association analysis, and link mining, which are all among the most important topics in data mining research and development [15]. .....	12

# 1 Summary

Worldwide, the prevalence of obesity has reached epidemic proportions. The global obesity epidemic continues its relentless advance, currently affecting >2 billion people.

Over the last decade, 227 genetic variants involved in different biological pathways and distortion of the normal microbial balance have been associated with obesity and other metabolic diseases.

Exponential growth of biomedical literature has supposed a challenge for professionals to keep up with the latest findings. Moreover, to find connections between different topics or disciplines has become even more difficult for the scientific community. In recent decades, the use of data mining has been proposed to glean useful insights, discovering novel patterns, relationships and trends contained within the documents.

Text mining techniques were applied from the biomedical literature, specifically PubMed abstracts, in order to develop an interactive web application, METAVOLIKOS, to study the relationship of genes and intestinal microbiome in patients with obesity and other metabolic diseases.

Concretely, the relationship between terms was studied through latent semantic analysis and subsequent cosine similarity.

The work pipeline was integrated into the development of the web application using the Shiny package of R.

A relationship has been inferred from publications on metabolic diseases where articles that include the term “Obesity” or explicitly the related gene or bacteria are not necessarily included. So, this work shows how latent semantic analysis can shed light to create new research hypotheses.

METAVOLIKOS is a text mining application that is easily adaptable and usable by bench scientists, which inspires to develop new research lines.

## 2 Introduction

### 2.1 Context and justification of the work

Obesity is a multifactorial disorder that results in excessive accumulation of adipose tissue [1]. Contrary to just being a medical condition or risk factor for other diseases, the rise in obesity has fuelled the current debate over its classification as a disease [2].

Worldwide, the prevalence of obesity has reached epidemic proportions [3]. The global obesity epidemic continues its relentless advance, currently affecting >2 billion people according to [4]. The prevalence of obesity in established market economies (Europe, USA, Canada, Australia, etc.) varies greatly, but a weighed estimate suggests an average prevalence in the order of 15-20% of the population [5].

Excess adipose tissue increases the work of the heart and leads to anatomical changes in this organ. It alters pulmonary, endocrine and immunological functions, all with adverse effects on health. Some of the complications of obesity include cardiovascular disease, non-insulin-dependent diabetes mellitus, obstructive pulmonary disease, arthritis and cancer [2].

Obesity is a multifactorial condition in which environmental, biological and genetic factors all play essential roles [3].

With the emergence of genome-wide association studies over the last decade, 227 genetic variants involved in different biological pathways (central nervous system, food sensing and digestion, adipocyte differentiation, insulin signaling, lipid metabolism, muscle and liver biology, gut microbiota) have been associated with obesity [6].

On the other hand, animal and human studies have implicated distortion of the normal microbial balance in obesity and metabolic diseases associated. The gut microbiota affects host metabolism and obesity through several pathways involving gut barrier integrity, production of metabolites affecting satiety and insulin resistance, epigenetic factors (to induce the expression of genes related to lipid and carbohydrate metabolism thereby leading to greater energy harvest

from the diet), and metabolism of bile acids and subsequent changes in metabolic signaling [7-8].

Exponential growth of biomedical literature has supposed a challenge for professionals to keep up with the latest findings of the specific topic. Moreover, to find connections between different topics or disciplines has become even more difficult for the scientific community.

In recent decades, the use of data mining has been proposed to glean useful insights, in the biomedical literature [9], and specially to study gene-disease relationship [10].

Identifying all gene-disease relationship by wet methods is highly costly and time-consuming. Nevertheless, a bioinformatics-based approach may provide certain gene or gut microbiome candidates for specific diseases before large-scale population studies are carried out.

The development of an interactive web application would allow the user to enjoy a friendly interface, avoiding the requirement of programming knowledge or special software. It could allow explore the content of PubMed publications obtaining key information (such as the gene-disease or microbiome-disease relationship) presented in the form of tables or graphs.

## 2.2 Objectives

Text mining techniques were applied from the biomedical literature, specifically PubMed abstracts, in order to search for publications on a disease of interest and summarize its content word clouds and genes. Specially, an interactive web application was developed to study the relationship of genes and intestinal microbiome in patients with obesity and other metabolic diseases.

### Main objective:

Develop an interactive web application that allows discovering information contained in the summaries of PubMed publications from data mining.

### Specific objectives:

- Obtain a list of genes potentially related to different metabolic diseases.

- Explore the relationship between the intestinal bacterial populations in patients with obesity and other metabolic diseases.
- Identify relationships between genes, obesity and other metabolic diseases.

### 2.3 Approach and method to follow

The application is based on text mining. For text mining the dataset usually comprises the documents themselves and the features are extracted from the documents automatically based on their content [11]. The scheme consists of the following steps taken according to [10].

#### 1. Create the corpus

To identify the texts to be used to solve the problems, the abstracts of scientific publications were used. This allows access to a very good summary of the publication and in turn has the advantage that this information is available to the public.

To download the abstracts from PubMed, libraries that allow information to be downloaded from R were used. In addition, the forms of searching in PubMed using MeSH tags and publication dates were taken into account.

#### 2. Document Preparation

The first step is to prepare the texts even before convert raw text documents to the multidimensional format. It was done taken into account the following steps:

- Stop word removal: Stop words are frequently occurring words in a language that are not very discriminative for mining applications.
- Stemming: Variations of the same word will be consolidated.
- Punctuations marks: After stemming has been performed, punctuation marks are removed.

After the aforementioned steps, the resulting document may contain only semantically relevant words [9].

#### 3. Extract the features

There are several ways in which the conversion of documents from plain text to instances with a fixed number of attributes in a training set can be carried

out. For the purposes of this work was carry out a simple word-based conversion, known as a bag-of-words representation. A document is just a collection of words placed in an arbitrary order, together with a count of how many times each one occurs (that is, there is no semantic analysis of the text, but rather one works with words) [11]. Moreover, the genes and bacterial population and microbiome related concepts were extracted from the text.

#### 4. Terms election to latent semantic analysis (LSA)

Processing of the abstracts with essential libraries includes the possibility of extraction words lists, sentence lists, disease, chemical agents, genes and PID numbers, among others.

This was used to building the semantic space. These come from the list of genes, intestinal microbioma symptoms, the names of metabolic diseases and other terms in the abstracts, such as the keywords of the study topic.

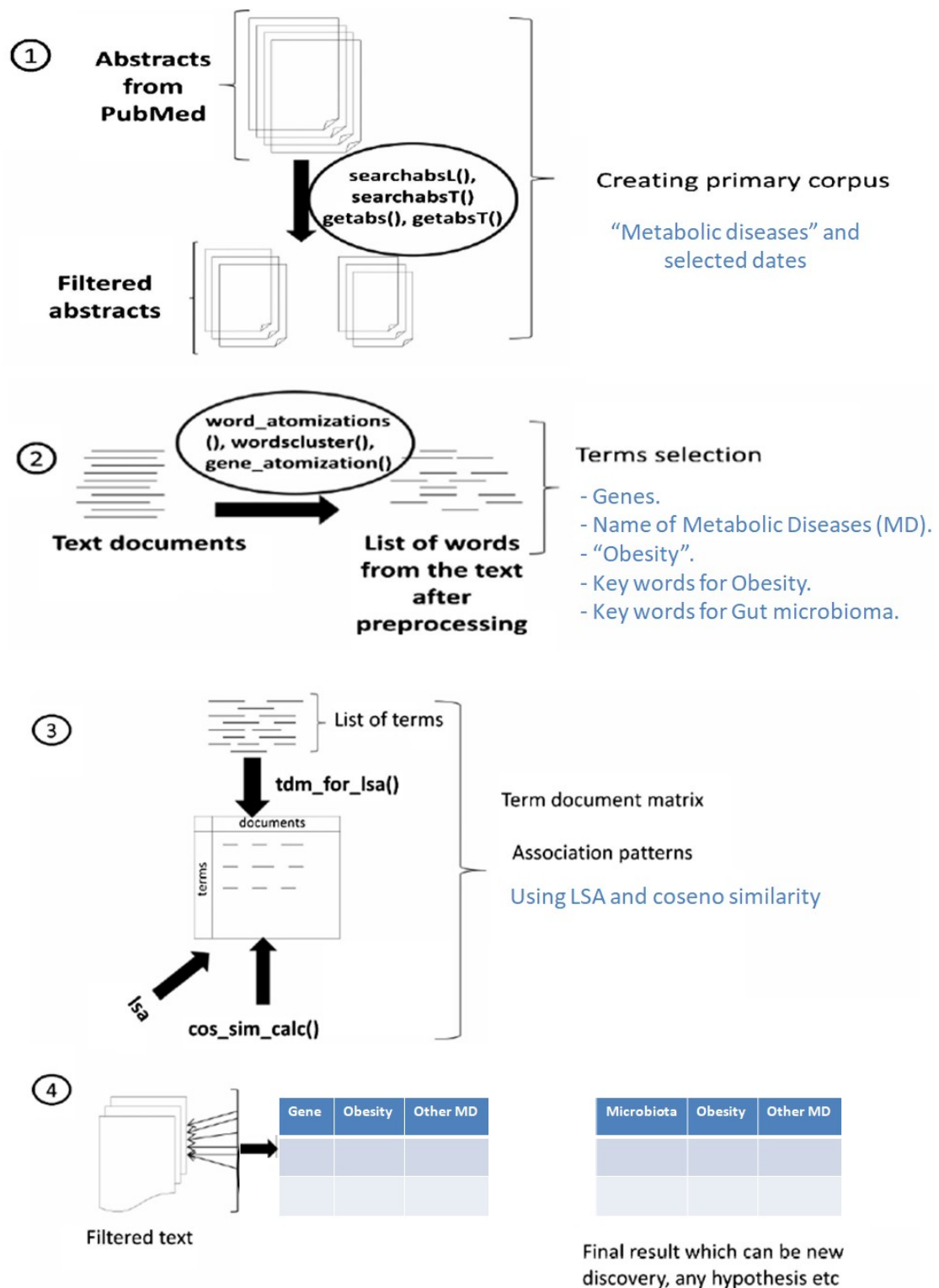
#### 5. LSA and coseno similarity

The relationship between terms was studied through LSA and subsequent cosine similarity.

Probabilistic latent semantic analysis is an expectation maximization-based mixture modeling algorithm. This is because the underlying generative process is optimized to discovering the correlation structure of the words rather than the clustering structure of the documents [11].

#### 6. Basic visualizations

In this case, graphs (clouds of words or genes, neighbor graphs) and tables were used to visualization the results.



**Figure 1. Work scheme for the extraction of genes and changes in the intestinal microbiota associated with obesity. Some text-mining facilities of pubmed.minerR package are displayed [modified from 10].**



## 7. Interactive web application development

Finally, the works pipeline was integrated into the development of the web application using the Shiny package of R. The server used to upload the web application was Shinyapps.io.

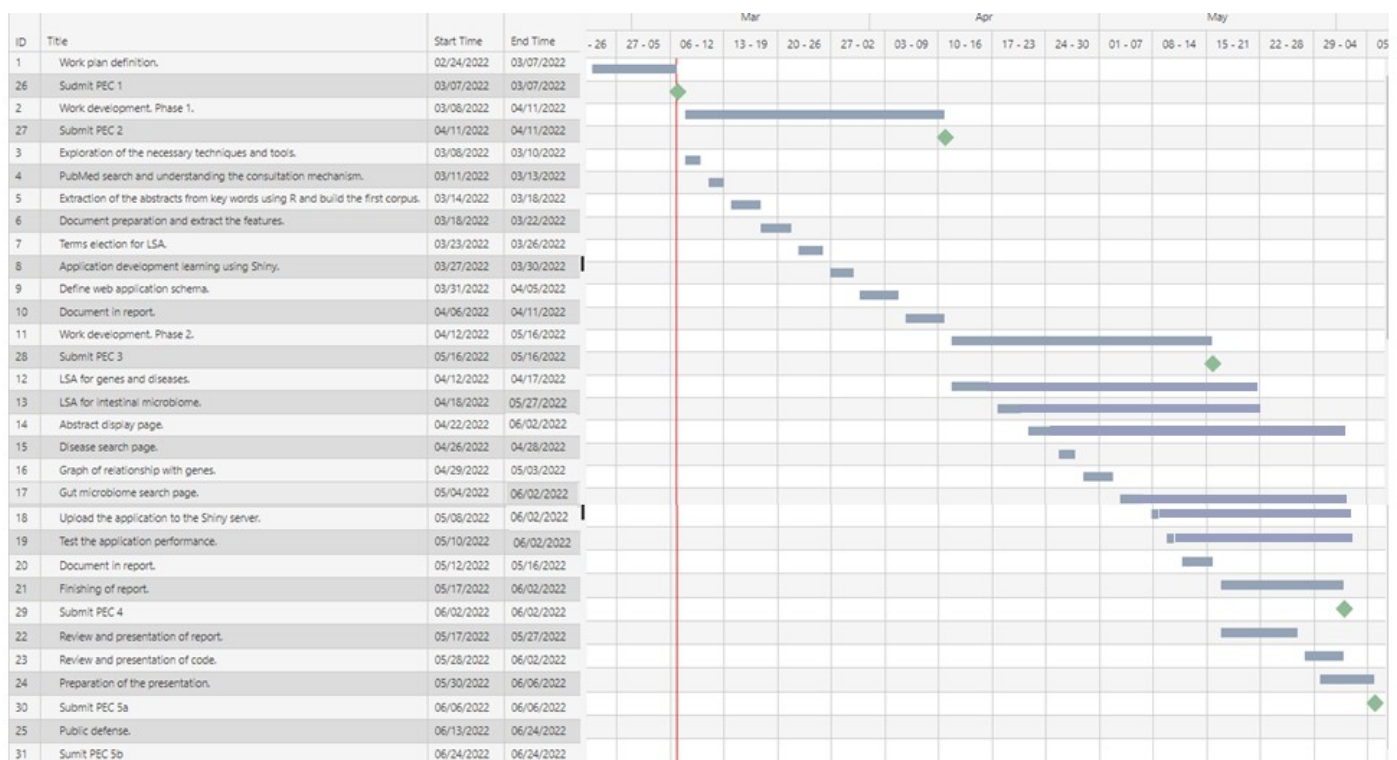
### 2.4 Planning with goals and timing

**Table 1. Tasks and goals of the project.**

<b>Description</b>	<b>Start date</b>	<b>Finish date</b>	<b>Goal</b>
Work plan definition.	24/02/2022	7/03/2022	Submit PEC 1
Work development. Phase 1.	08/03/2022	11/04/2022	Submit PEC 2
Exploration of the necessary techniques and tools.	08/03/2022	10/03/2022	
PubMed search and understanding the consultation mechanism.	11/03/2022	13/03/2022	
Extraction of the abstracts from key words using R and build the first corpus.	14/03/2022	18/03/2022	
Document preparation and extract the features.	18/03/2022	22/03/2022	
Terms election for LSA.	23/03/2022	26/03/2022	
Application development learning using Shiny.	27/03/2022	30/03/2022	
Define web application schema.	31/03/2022	05/04/2022	
Document in report.	06/04/2022	12/04/2022	
Work development. Phase 2.	12/04/2022	16/05/2022	Submit PEC 3
LSA for genes and diseases.	12/04/2022	17/04/2022	
LSA for intestinal microbiome.	18/04/2022	21/04/2022*	
Abstract display page.	22/04/2022	25/04/2022**	
Disease search page.	26/04/2022	28/04/2022	
Graph of relationship with genes.	29/04/2022	03/05/2022	
Gut microbiome search page.	04/05/2022	07/05/2022**	
Upload the application to the Shiny server.	08/05/2022	09/05/2022**	

Test the application performance.	10/05/2022	11/05/2022**	
Document in report.	12/05/2022	16/05/2022	
Finishing of report.	17/05/2022	2/06/2022	Submit PEC 4
Review and presentation of report.	17/05/2022	27/05/2022	
Review and presentation of code.	28/05/2022	02/06/2022	
Preparation of the presentation.	30/05/2020	06/06/2022	Submit PEC 5a
Public defense.	13/06/2022	24/06/2022	Submit PEC 5b

\*Extended to 27/05/2022. \*\*Extended to 02/06/2022.



**Figure 2. Gantt Diagram of the project.**

## 2.5 Expected results

### 1. Work plan definition

PDF document containing the work description as well as the general and specific objectives. In addition, it describes the method to be followed and the

specific scheduled tasks to achieve the objectives, a risk analysis and the expected results.

## 2. Report

Document in PDF format that details all research development, results, discussion and conclusion obtained throughout the master's thesis, as well as the implemented code.

## 3. Product

- METAVOLIK: Web Application for the study of the relationship between Obesity and intestinal microbiome based on PubMed text mining.
- Access to the web application: <https://neustorrentample.shinyapps.io/METAVOLIKOS/>
- Access to a public Github repository that allows access to the developed code: <https://github.com/NeusTorrent/METAVOLIKOS>

## 4. Presentation

Presentation in PPT format to present and defend the work.

### 2.6 Brief description of the other chapters of the report

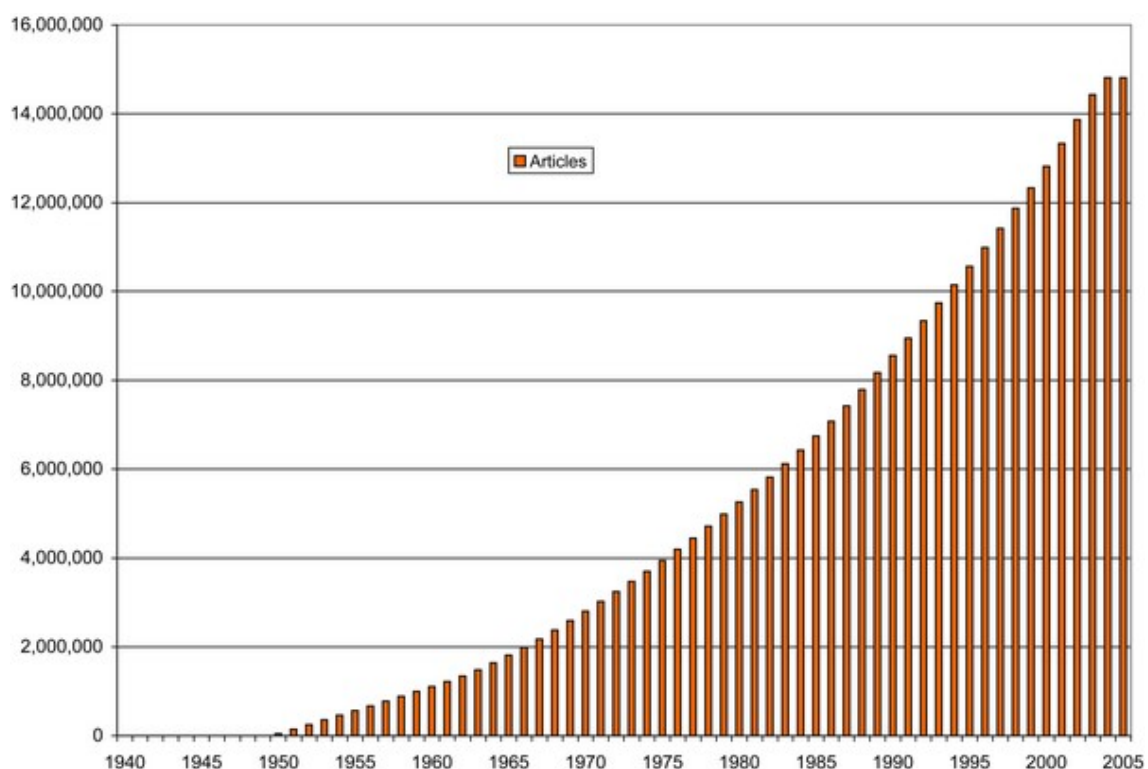
Below are the chapters of State of the Art, Methodology, Results and Discussion, and Conclusions of the works.

Through them it is detailed the evolution and current situation of text mining analysis, the details of the analytical techniques used to reach the specific objectives, the performance of METAVOLIK tool. Finally, the conclusions, the impact, and limitations of the developed web application are exposed.

### 3 State of the art

The biological sciences are generating enormous quantities of data, ushering in the era of “big data”. Sequencing data alone constitutes 35 petabases/year and will grow to 1 zettabase/year by 2025. Today, the data being generated is massive, complex and increasingly diverse owing to recent technological innovations. However, the impact of this data revolution on our lives is hampered by the limited amount of data that has been analyzed [12].

With biomedical literature increasing at a rate of several thousand papers per week, it is impossible to keep abreast of all developments. This is becoming an increasingly important practical problem as the volume of material in many fields keeps increasing and even in specialist fields it can be very difficult to locate relevant documents [11]. Therefore, automated means to manage the information overload are required.

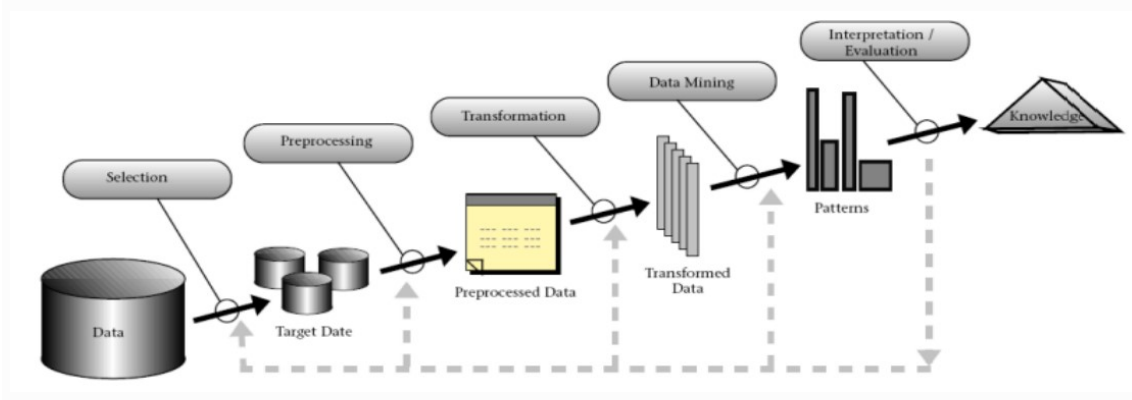


**Figure 3.** This figure shows the exploding number of articles available from Medline over 1940-2005 (data retrieved from the SRS server at the European Bioinformatics Institute; [www.ebi.ac.uk/](http://www.ebi.ac.uk/)). In 2003, about 560,000 articles were added to Medline, and from 2000 to 2003, 2 million articles [13].

Data mining is the act of computationally extracting new information from large amounts of data. It is the study of collecting, cleaning, processing, analyzing, and gaining useful insights from data. Therefore, “data mining” is a broad umbrella term that is used to describe these different aspects of data processing. From an analytical perspective, data mining is challenging because of the wide disparity in the problems and data types that are encountered [9].

Databases and data mining tools are nevertheless indispensable in the era of data abundance and excess, which contrasts the not-so-ancient era when the problem was the access to the scarce data.

Data Mining came into existence in the middle of 1990’s and appeared as a powerful tool that is suitable for fetching previously unknown pattern and useful information from huge dataset. Various studies highlighted that data mining techniques help the data holder to analyze and discover unsuspected relationship among their data which in turn helpful for making decision. In general, data mining and Knowledge Discovery in Databases (KDD) are related terms and are used interchangeably but many researchers assume that both terms are different as data mining is one of the most important stages of the KDD process. The knowledge discovery process are structured in various stages: Whereas the first stage is data selection where data is collected from various sources; the second stage is pre-processing of the selected data; the third stage is the transformation of the data into appropriate format for further processing; the fourth stage is data mining where suitable data mining technique is applied on the data for extracting valuable information; and evaluation is the last stage as shown in **Figure 4** [14].



**Figure 4. Stages of knowledge discovery process [14].**

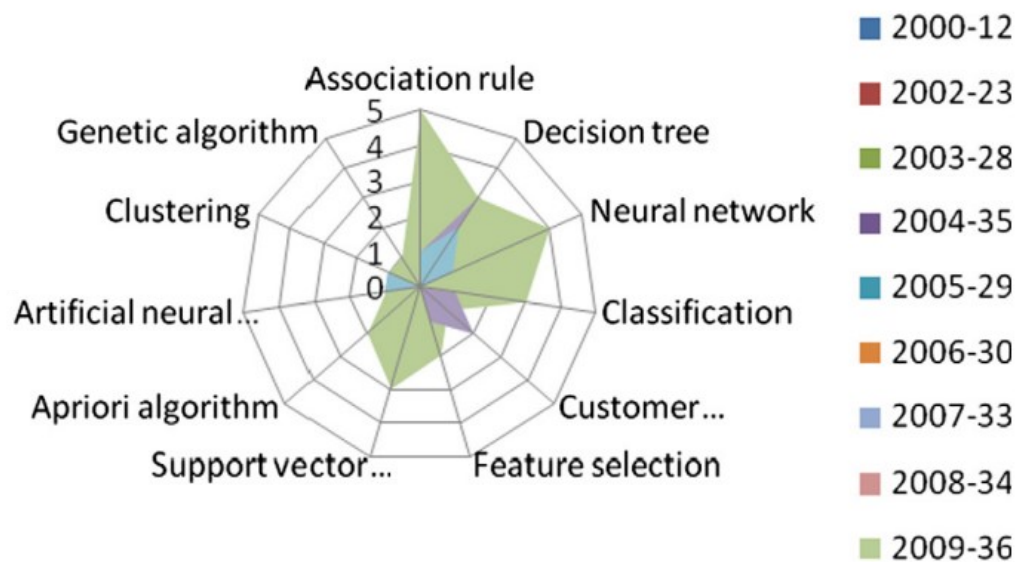
A wide variety of methods for extracting value from different types and models of data fall under the umbrella of “data mining”.

**Table 2. Top mining algorithms identified by the IEEE International Conference on Data Mining in December 2006. These 10 algorithms cover classification, clustering, statistical learning, association analysis, and link mining, which are all among the most important topics in data mining research and development [15].**

<b>Algorithm</b>	<b>Type</b>	<b>Method</b>	<b>Authors</b>
<b>C4.5</b>	Classification algorithms	Decision trees	Australian Research Council
<b>The k-means algorithm</b>	Clustering algorithms		Lloyd (1957, 1982), Forgey (1965), Friedman and Rubin (1967), and McQueen (1967)
<b>Support vector machines</b>	Classification algorithms		
<b>The Apriori algorithm</b>	Frequent pattern algorithms	Association rule mining	
<b>The EM (Expectation–Maximization) algorithm</b>	Clustering algorithms	Maximum likelihood estimation of normal mixtures	Arthur Dempster, Nan Laird, and Donald Rubin (1977)
<b>PageRank</b>	Link-based algorithms.	Search ranking algorithm using hyperlinks on the Web	Sergey Brin and Larry Page (1998)
<b>AdaBoost</b>	Classification algorithms	Ensemble methods	Yoav Freund and Robert Schapire (1995).
<b>kNN: k-nearest neighbor classification</b>	Classification algorithms		Evelyn Fix and Joseph Hodges in 1951
<b>Naive Bayes</b>	Classification algorithms		
<b>CART: Classification and Regression Trees</b>	Classification algorithms	Decision trees	Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone (1984)

Data mining integrates techniques from several fields including machine learning, statistics, patterns recognition, artificial intelligence, and database systems, for the analysis of large volumes of data.

Data mining techniques developed recently include generalization, characterization, classification, clustering, association, evolution, pattern matching, data visualization and meta-rule guided mining. The techniques for mining knowledge from different kinds of databases include relational, transactional, object oriented, spatial and active databases, as well as global information systems [16].

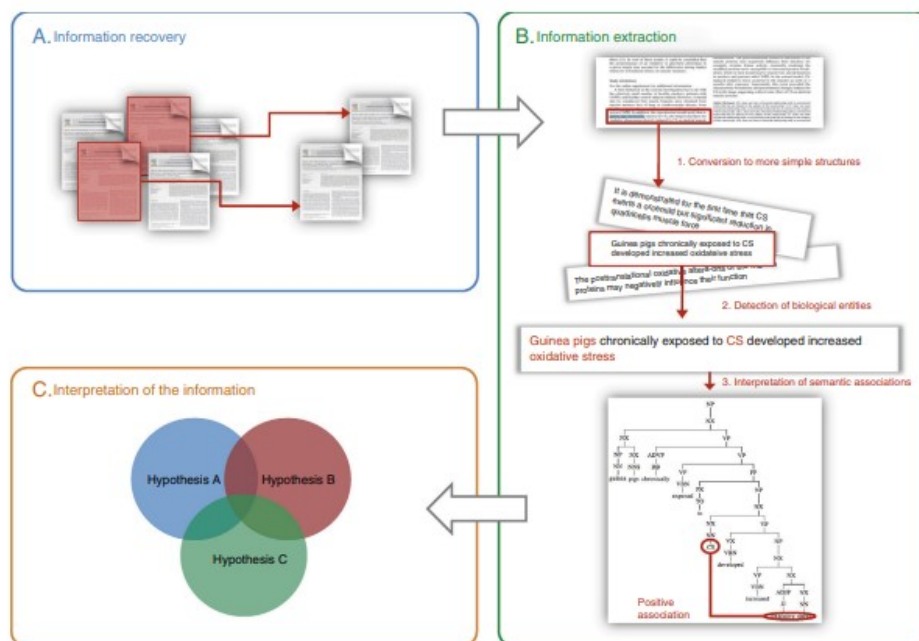


**Figure 5. Data mining techniques from 2000-2011. It shows the important data mining techniques trends for association rules, genetic algorithms, clustering, artificial neural networks, a priori algorithms, support vector machines, feature selection, customer relationship management, classification, neural networks and decision trees [16].**

Most researchers consider some other areas, including text mining, as being under the data mining umbrella.

Text mining is a subfield of data mining that seeks to extract valuable new information from unstructured (or semi structured) sources. Text mining extracts information from within those documents and aggregates the extracted pieces over the entire collection of source documents to uncover or derive new information. Thus, given as input a set of documents, text mining methods seek

to discover novel patterns, relationships and trends contained within the documents [12].



**Figure 6. General schematic of the methods used in text and data mining: (A) information recovery, (B) information extraction, (C) interpretation of the information (the integration of various previously corroborated hypotheses can produce a new combined hypothesis). The three steps required for extracting information are shown in panel B: (1) breakdown of the information into basic units (e.g. words), (2) identification of biological entities, and (3) interpretation of the relationship between biological entities [17].**

In the end, a set of the different subtask solutions are used in a pipeline that allows information to be integrated and analyzed toward knowledge discovery. However, this multiplies the effects of errors down the pipeline, leaving systems highly vulnerable. An overarching challenge for biomedical text mining is to incorporate the many knowledge resources that are available to us into the pipeline. In the biomedical domain, unlike the general text mining domain, we have access to large numbers of extensive, well-curated ontologies and knowledge bases. Biomedical ontologies provide an explicit characterization of a given domain of interest. The quality of data mining efforts would likely increase if existing ontologies (e.g. UMLS and BioPortal) were used as sources



of terms in building lexicons, for figuring out what concept subsumes another, and as a way of normalizing alternative names to one identifier [12].

Because of its potential power for solving complex problems, data mining has been successfully applied to diverse areas such as business, engineering, social media, and biological science.

The biomedical domain is one of the most interesting application areas for text mining, given both the potential impact of the information that can be discovered and the specific characteristics and volume of information available [12].

Computational methods contribute to this field by bringing knowledge from literature, together with high-throughput data sets to identify both known and new relationships between genes, pathways, drugs, environmental contaminants and diseases. Systems that can extract relationships from both literature and data simultaneously present the opportunity to identify meaningful patterns from data, identify literature support for those patterns, and where warranted, identify relationships that are highly consistent in large-scale throughput data sets but absent from literature [12].

Finally, an area in which the field has fallen short is that of making text mining applications that are easily adaptable by end users. Many researchers have developed systems that can be adapted by other text mining specialists, but applications that can be tuned by bench scientists are mostly lacking [12].

## 4 Methodology

### Investigate about the database of PubMed

To identify the texts to be used to solve the problems, the abstracts of scientific publications were used. This allows access to a very good summary of the publication and in turn has the advantage that this information is available to the public.

PubMed contains more than 34 million citations from the biomedical literature. It gives access to different databases such as MEDLINE, GeneBank, Complete Genome, etc.

MEDLINE is the database of the National Library of Medicine of United States and includes citations and abstracts from medicine and biological sciences fields [22]. This library has open access from PubMed and is the largest database of this platform.

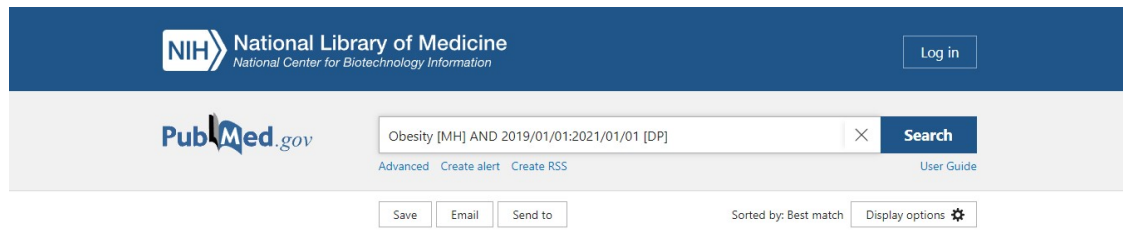
PubMed uses a controlled vocabulary: the medical subject headings (MeSH tags). These terms have hierarchy and are used to index articles in MEDLINE [22].

Usually, PubMed performs as automatic term mapping search, searching among the MeSH terms, journals and authors, consecutively [22]. However, it is possible to indicate, through using tags, a more specified search.

Affiliation [ad]	Full Investigator Name [fir]	Pagination [pg]
All Fields [all]	Grant Number [gr]	Personal Name as Subject [ps]
Article Identifier [aid]	Investigator [ir]	Pharmacological Action [pa]
Author [au]	ISBN [isbn]	Place of Publication [pl]
Author Identifier [auid]	Issue [ip]	PMCID and MID
Book [book]	Journal [ta]	PMID [pmid]
Comment Correction Type	Language [la]	Publication Date [dp]
Completion Date [dcom]	Last Author Name [lastau]	Publication Type [pt]
Conflict of Interest Statement [cois]	Location ID [lid]	Publisher [pubn]
Corporate Author [cn]	MeSH Date [mhda]	Secondary Source ID [si]
Create Date [crdt]	MeSH Major Topic [majr]	Subset [sb]
EC/RN Number [rn]	MeSH Subheadings [sh]	Supplementary Concept [nm]
Editor [ed]	MeSH Terms [mh]	Text Words [tw]
Entry Date [edat]	Modification Date [lr]	Title [ti]
Filter [filter] [sb]	NLM Unique ID [jid]	Title/Abstract [tiab]
First Author Name [1au]	Other Term [ot]	Transliterated Title [tt]
Full Author Name [fau]	Owner	Volume [vi]

**Figure 7. Search field descriptions and tags of PubMed.**

Some of these tags were used in the present work in order to search in a more precise way.



**Figure 8. Search in PubMed using [MH] and [DP] tags.**

MeSH term specification into the search (tag [MH]) guarantees that the abstracts which compose the corpus, effectively are treating the desired subject. Moreover, the tag [DP] was used to specify the publication date.

### Create the corpus

The current project was carried out using R and RStudio, version R version 4.1.2 (2021-11-01). R is a programming language used for statistical computing while RStudio uses the R language to develop statistical programs.

To download the abstracts from PubMed, the function `batch_pubmed_download()` of the package `easyPubMed` was used. That function allows download the information from PubMed to R, restricting the search by keywords and years of publication. Abstracts of articles were saved locally in form of text files and it formed the starting point, the primary corpus.

```
> output_obesity = batch_pubmed_download(pubmed_query_string = 'obesity [MH] AND "2019/01/01"[PDAT] : "2021/01/01"[PDAT]', format="abstract", batch_size= 1500)
[1] "PubMed data batch 1 / 20 downloaded..."
[1] "PubMed data batch 2 / 20 downloaded..."
[1] "PubMed data batch 3 / 20 downloaded..."
[1] "PubMed data batch 4 / 20 downloaded..."
[1] "PubMed data batch 5 / 20 downloaded..."
```

In our case, it was selected the abstracts from PubMed, which contains the MeSH term "Obesity", and are published in 2019 and 2020.

Using the following command, we can verify, that we obtained 20 text files, which are allocated into the working directory.

```
> list.files(path = "C:/Users/Neus/Desktop/MÀSTER BIOINFORMÀTICA I BIOESTADÍSTICA/TFM/Corpus")
[1] "easyPubMed_data_001.txt" "easyPubMed_data_002.txt" "easyPubMed_data_003.txt"
[4] "easyPubMed_data_004.txt" "easyPubMed_data_005.txt" "easyPubMed_data_006.txt"
[7] "easyPubMed_data_007.txt" "easyPubMed_data_008.txt" "easyPubMed_data_009.txt"
[10] "easyPubMed_data_010.txt" "easyPubMed_data_011.txt" "easyPubMed_data_012.txt"
[13] "easyPubMed_data_013.txt" "easyPubMed_data_014.txt" "easyPubMed_data_015.txt"
[16] "easyPubMed_data_016.txt" "easyPubMed_data_017.txt" "easyPubMed_data_018.txt"
[19] "easyPubMed_data_019.txt" "easyPubMed_data_020.txt"
```

Following, we joint all these 20 files in 1 to work properly into the next steps.

```
> file.create("pubmed_result_obesity.txt")
[1] TRUE
> for (i in 1:length(output_obesity)){
+   + file.append("pubmed_result_obesity.txt", output_obesity[i])
+ }
```

The package `pubmed.mineR` contains the function `readabs()`, which read the abstracts downloaded and saved locally and transform them into a S4 class object.

```
> corpus_obesity = readabs("pubmed_result_obesity.txt")
```

At this point, we can inspect the created object. It is class S4 and contains the information of Journal, Abstract and PMID of 29511 articles selected.

Name	Type	Value
corpus_obesity	S4 (pubmed.mineR::Abstracts)	S4 object of class Abstracts
Journal	character [29511]	'1. Front Endocrinol (Lausanne). 2022 Feb 7;12:803545. doi: "EPOC...
Abstract	character [29511]	'10.3389/fendo.2021.803545. eCollection 2021. The Complex Que...
PMID	double [29511]	35197927 35186858 35180034 35173682 35154005 35153648 ...

**Figure 9. Inspection of S4 class object.**

Moreover, we can explore the corpus in detail using the `@`sign. In the following output we show the journal, abstract and PMID number of the first element of the S4 object.

```
> corpus_obesity@Journal[1]
[1] "1. Front Endocrinol (Lausanne). 2022 Feb 7;12:803545. doi:"
> corpus_obesity@Abstract[1]
[1] "10.3389/fendo.2021.803545. eCollection 2021. The Complex Quest of Preventing Obesity in Early Childhood: Describing Challenges and Solutions Through Collaboration and Innovation. Seidler AL(1)(2), Johnson BJ(2)(3)(4), Golley RK(2)(3)(4), Hunter KE(1)(2). Author information: (1)National Health and Medical Research Council Clinical Trials Centre, University of Sydney, Camperdown, NSW, Australia. (2)Transforming obesity Prevention in Children (TOPCHILD) Collaboration, Sydney, NSW, Australia. (3)Caring Futures Institute, Flinders University, Bedford Park, SA, Australia. (4)College of Nursing and Health Sciences, Flinders University, Bedford Park, SA, Australia. Childhood obesity remains a major public health issue and priority area for action. Promisingly, obesity prevention interventions in the first 2000 days of life have shown modest effectiveness in improving health behaviours and healthy weight status in children. Yet, researchers in this field face several challenges. This can lead to research waste and impede progress towards delivering effective, scalable solutions. In this perspective article, we describe some of the key challenges in early childhood obesity prevention and outline innovative and collaborative solutions to overcome these. Combining these solutions will accelerate the generation of high-quality evidence that can be implemented into policy and practice. Copyright © 2022 Seidler, Johnson, Golley and Hunter. DOI: 10.3389/fendo.2021.803545 PMID: PMC8859836"
> corpus_obesity@PMID[1]
[1] 35197927
```

Obesity is a multifactorial pathology, related to many other health disorders, with genetic component, and influenced by lifestyle and diet. Due to that the scientific publications related to this concept come from very diverse areas.

In order to work with a corpus more related to our topic, the relation of obesity with other metabolic diseases and changes in intestinal microbiome, a secondary corpus was created. It was used when the relation of microbiota and obesity was established.

The function `SearchabsL()` from the `pubmed.minerR` package searched the abstracts for the given term or combinations of several terms. In this method the argument "include" uses the boolean operator 'OR' and is liberal.

```
> secondary_corpus_obesity = searchabsL(corpus_obesity, include = "microbiome | microbiota")
[1] "28 abstracts microbiome | microbiota"
```

The abstracts containing the term "microbiome" or "microbiota" are selected from primary corpus to perform the secondary corpus. It is composed by 28 abstracts.

### Extract the features

The features extracted from the abstracts of the corpus were the genes and the words. We used the functions `word_atomizations()` and `gene_atomizations()` of the `pubmed.minerR` package.

```
> genes_EA_obesity = gene_atomization(corpus_obesity)
> genes_EA_list_obesity = genes_EA_obesity[,1]
```

`gene_atomization()` automatically fetch the genes (HGNC approved Symbol) from the text and report their frequencies.

	Gene_symbol	X...Genes	Freq
1	TG	thyroglobulin	491
2	CRP	C-reactive protein	422
3	HR	HR lysine demethylase and nuclear receptor corepressor	402
4	FTO	FTO alpha-ketoglutarate dependent dioxygenase	381
5	SSB	small RNA binding exonuclease protection factor La	273
6	FGF21	fibroblast growth factor 21	257
7	MC4R	melanocortin 4 receptor	252
8	NHS	NHS actin remodeling regulator	225
9	SDS	serine dehydratase	225
10	CS	citrate synthase	211
11	SIRT1	sirtuin 1	200

**Figure 10. Genes cited into the publications of the corpus.**

```
> word_obesity = word_atomizations(corpus_obesity)
```

The function `word_atomizations()` tokenizes text into words and reports them in descending order of their occurrence frequencies. Common English words, extra white space, and punctuation marks are automatically removed. That's why an extra document preparation is not carried out. The top ranking words are high-occurrence frequency terms.

	words	Freq
13905	obesity	1305
14466	patients	1008
18560	weight	813
17089	study	697
16015	results	583
16105	risk	573
13893	obese	567
6517	associated	514
7067	body	514
7020	bmi	463
17252	surgery	442

**Figure 11. Words used into the publications of the corpus.**

Moreover, the microbiota and related concepts were extracted from the secondary corpus. For that purpose, a research of all intestinal bacteria population, and related concepts was carried out based on [23] and [24].



```
> Microbiota_one_word = c("dysbiosis", "acids", "serratia", "enterobacter", "morganelle", "skunlikevirus", "phl", "fllikevirus", "roseburia", "blautia", "clostridium", "akkermansia", "ruminococcus", "lactobacillus", "microbial", "diversity", "metabolites", "metagenomic", "butyrate", "firmicutes", "bacteroidetes", "bacteroides", "prevotella", "xylanibacter", "proteobacteria", "microbiota", "propionate", "acetate", "christensenellaceae", "tenericutes", "prevotella", "microbes", "microbiome", "verrucomicrobia", "lachnospiraceae", "bifidobacterium", "fusobacterium", "faecalibacterium", "roseburia", "eubacterium", "bilophila", "desulfovibrio", "blautia", "turicibacter", "bilophila", "adlercreutzia", "actinobacteria", "streptococcus", "lactic acid bacteria", "lachnospiraceae", "rikenellaceae", "parasutterella", "sutterella", "lachnospiraceae", "veillonellaceae", "alistipes", "actinobacteria", "enterobacteriaceae", "staphylococcus", "escherichia", "methabacteriodes", "betaproteobacteria", "firmicutes", "clostridia", "proteobacteria", "tenericutes", "actinobacteria", "mollicutes", "negativicutes", "bacteroidia", "erysipelotrichales", "selenomonadales", "bacteroidales", "coriobacteriales", "clostridiales", "prevotellaceae", "lachnospiraceae", "peptostreptococcaceae", "ruminococcaceae", "coriobacteriaceae", "collinsella")
```

Then, the words of the secondary corpus were extracted as described before using the `word_atomization()` function.

```
> micro_MD_comb = data.frame(word_atomizations(secondary_corpus_obesity))
```

Using the `str_subset()` and `str_extract()` the terms of the "Microbiota\_one\_word" vector were extracted from the abstracts of the secondary corpus.

```
> colour_match <- str_c(Microbiota_one_word, collapse = "|")
> has_colour <- str_subset(micro_MD_comb$words, colour_match)
> matches <- str_extract(has_colour, colour_match)
> micro_MD_list_comb <- unique(matches)
```

Finally, the microbiota terms included into the "Microbiota\_one\_word" vector and present into the secondary corpus were extracted.

```
[1] "microbiota"      "microbiome"      "diversity"        "firmicutes"      "dysbiosis"
[6] "bacteroidetes"   "akkermansia"     "lactobacillus"    "acids"            "bifidobacterium"
[11] "clostridia"      "metabolites"     "microbes"         "proteobacteria"   "actinobacteria"
[16] "bacteroides"     "butyrate"        "faecalibacterium" "prevotella"       "streptococcus"
```

**Figure 12. Bacteria population and related concepts extracted from the secondary corpus.**

### Terms election to latent semantic analysis (LSA)

Processing of the abstracts with essential libraries includes the possibility of extraction words lists, sentence lists, disease, chemical agents, genes and PID numbers, among others.

It was used to building the semantic space. These come from the list of genes (extracted from the corpus), intestinal microbiome population (detailed previously), the names of metabolic diseases (according to [25]) and other terms in the abstracts, such as the keywords of the study topic.

```
> Disease_one_word = c("Diabetes", "Obesity", "Dyslipidemia", "Hypolipidemia", "Hyperlipidemia", "Hyperthyroidism", "Hypothyroidism", "Hypoparathyroidism", "Hyperparathyroidism", "Cushing", "Hyperuricemia", "Hemochromatosis", "Metabolic", "Fatty liver", "Hypercholesterolemia")
> terms_obesity = c("bmi", "metabolic", "diabetes", "glucose", "liver", "adipose")
```

## LSA and cosine similarity

Before go deep into the statistical analysis, it is necessary to understand a number of specific characteristics that are unique to text data:

1. Number of “zero” attributes: Although the base dimensionality of text data may be of the order of several hundred thousand words, a single document may contain only a few hundred words. If each word in the lexicon is viewed as an attribute, and the document word frequency is viewed as the attribute value, most attribute values are 0. This phenomenon is referred to as high-dimensional sparsity. There may also be a wide variation in the number of nonzero values across different documents. This has numerous implications for many fundamental aspects of text mining, such as distance computation [9].

2. Non-negativity: The frequencies of words take on nonnegative values. When combined with high-dimensional sparsity, the non-negativity property enables the use of specialized methods for document analysis. In general, all data mining algorithms must be cognizant of the fact that the presence of a word in a document is statistically more significant than its absence. Unlike traditional multidimensional techniques, incorporating the global statistical characteristics of the data set in pair wise distance computation is crucial for good distance function design [9].

Although the vector-space representation of text can be considered a sparse numeric data set with very high dimensionality, this special numeric representation is not very amenable to conventional data mining algorithms. The first step is to use latent semantic analysis to transform the text collection to a non-sparse representation with lower dimensionality. Rather, traditional text mining algorithms are directly applied to the reduced representation obtained from LSA [9].

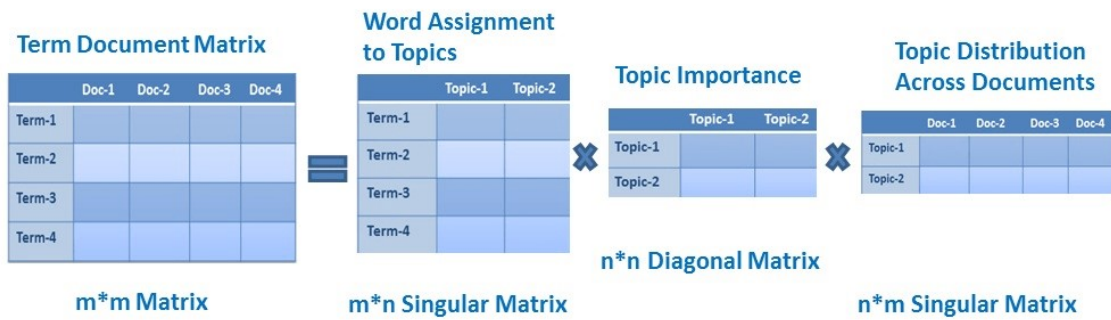
At the philosophical core underlying LSA is that text embeds knowledge not only by conveying information explicitly through sentences but also implicitly through how words co-occur with each other. Sometimes, the co-occurrence of words may also reveal ideas, because ideas are embedded into language through word co-occurrences. Terms can be related to one another in LSA



even if they do not co-occur in the same document as long as both terms co-occur with shared other terms [18].

Mathematically, this is done by running a singular value decomposition (SVD) on the term-document (frequency) matrix (TDM). This matrix contains the number of times any term of interest (or to be exact, any term not excluded) appears in any of the documents being analyzed. SVD is a two mode data reduction analysis that transforms the TDM into three matrices: 1, terms; 2, documents; and 3, a matrix that multiplied by the other two reconstruct the original TDM matrix. Running SVD on the TDM is what defines LSA and makes it more than mere word co-occurrences analysis [18].

$$M = U \times \Sigma \times V^T$$



**Figure 13. LSA uses bag of word model, which results in a term-document matrix (occurrence of terms in a document). Rows represent terms and columns represent documents. LSA learns latent topics by performing a singular matrix decomposition on the document-term matrix using singular value decomposition [19].**

The original  $M$  matrix could, therefore, be reconstructed by multiplying the  $U$ ,  $\Sigma$ , and  $V$  matrices. However, in LSA, a truncated SVD is used wherein only a portion of the  $\Sigma$  matrix is calculated or retained and the remaining singular values are replaced with zeroes. If the matrices were multiplied back together, it would create an approximation of the original matrix where the number of singular values used determines how close the approximation is. The reconstructed matrix is known as the rank  $k$  approximation, where  $k$  is the number of singular values used. That multiplying the reduced rank matrices only creates an approximation of the original matrix may seem to be

a problem but is actually one of the most powerful features of LSA. Because SVD seeks to minimize error, it combines vectors that are closest to each another, thus preserving as much of the original information as possible in fewer dimensions. As a result, selecting an appropriate rank is critically important in LSA.  $k$  is often set at 100, 200, or 300. There is no rule on how best to select  $k$  a priori. The SVD transformation creates a semantic space out of the TDM [18].

$$M \approx A_k = U_k \times \Sigma_k \times V_k^T$$

Once the semantic space has been created, much can be done with the term and document matrices created within that space. One common analysis is to compare vectors of terms by applying cosine similarity. This kind of analysis can be applied to find which terms are related to one another by calculating the cosine similarity between vectors in the  $\Sigma \times U$  matrix. Likewise, such an analysis can be applied to determine which documents are related to one another by calculating cosine similarity between vectors in the  $\Sigma \times V^T$  matrix [18].

$$\cos(\vec{u}, \vec{v}) = \frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\|_2 \cdot \|\vec{v}\|_2} = \frac{\sqrt{\sum_i (\vec{u}_i \times \vec{v}_i)}}{\sqrt{\sum_i (\vec{u}_i^2)} \times \sqrt{\sum_i (\vec{v}_i^2)}}$$

The closer the cosine of two vectors is to 1.0, the more similar they are. However, caution is advised when interpreting low cosine similarities. A similarity near 0.0 may indicate that terms have opposite meanings, but it may also indicate that they are unrelated [18].

In the case of the present work, the matrix of terms and documents was build using the corpus, the keywords defined previously and the genes that have a frequency higher than 10.

```
> tdm_main = tdm_for_lsa(corpus_query_comb, c(genes_MD_list_comb, "obesity", terms_obesity))
```

The function `lsa()` was used to carry out the LSA. The dimensions were calculated by the function instead to be pre determined.

```
> lsa_corpus = lsa(tdm_main, dims = dimcalc_share())
```

It contains the tk matrix: the matrix of terms of dimension terms x dimension of semantic space. It corresponds to the “m\*n matrix” (see **Figure 13**).

```
> lsa_corpus
$tk
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]
TG	-1.257396e-02	2.964872e-02	-9.388284e-03	2.643229e-02	-3.742126e-03	6.014150e-03	-2.499784e-02
CRP	-8.418265e-03	1.272986e-02	-4.235522e-03	8.746745e-03	1.342459e-02	5.874651e-03	3.234913e-03
HR	-4.481128e-02	5.630278e-02	5.208072e-02	9.170090e-02	1.304472e-01	-5.808899e-02	1.873377e-01
FTO	-7.115344e-03	-4.625841e-05	-1.294405e-03	2.194980e-03	7.075037e-03	3.421991e-03	-2.485907e-03

The element dk is the matrix of documents of dimension documents x dimension of semantic space. It corresponds to the “n\*m matrix” of the **Figure 13**.

```
$dk
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]
[1,]	-6.157684e-03	8.037422e-03	-1.150633e-02	-6.385761e-03	5.667844e-04	6.441809e-04	-5.377751e-03
[2,]	-4.957673e-05	1.721535e-04	-1.666999e-06	5.491743e-04	3.383513e-04	3.318941e-04	1.162188e-04
[3,]	-1.499948e-02	-7.642762e-03	1.395408e-03	-7.566974e-05	1.370081e-03	-1.588679e-03	6.862571e-05
[4,]	-3.885815e-03	9.644958e-04	-2.115122e-04	7.686501e-03	1.013455e-02	3.438525e-02	1.251108e-02

The sk vector contains de coefficients, which are the elements of the diagonal matrix (“n\*n matrix” of the **Figure 13**).

```
$sk
```

[1]	592.26982	274.28241	233.28309	206.57590	190.52449	176.14992	168.29227	156.52149	152.91351	143.95059
[11]	134.96586	117.59497	111.07298	89.08725	84.07727	79.27931	77.57601	75.30502	72.97089	69.95516
[21]	66.88328	66.25989	63.46951	62.89202	62.08050	59.60480	57.94799	57.40429	56.96912	52.43354
[31]	52.29790	50.66784	49.52806	47.85780	47.44381	46.66590	44.50926			

The function as.textmatrix() was used to construct the estimated tdm matrix from LSA.

```
> matrix = as.textmatrix(lsa_corpus)
> matrix[1:4,1:4]
```

	[,1]	[,2]	[,3]	[,4]
TG	0.9976024345	5.423148e-05	-0.0003419574	0.018788963
CRP	1.9930414597	2.148416e-05	-0.0008675667	0.015961310
HR	-0.0001554622	1.081804e-05	0.0001428707	0.008116661
FTO	-0.0011569662	-7.366126e-04	-0.0001334554	0.036793167

Finally, the correlation between the terms and one selected term (“Obesity” in the current work) was determined. A cutoff value to define which relationships were displayed was established.

```
> aa = associate(matrix, "obesity", measure = "cosine", threshold = 0.1)
> aa
```

	CNP	S100A16	SFN	PAH	FAAH	KRAS	NES	MYT1L	MPO	GLS
	0.8607444	0.8503768	0.8084020	0.7823686	0.7649736	0.7485364	0.7315080	0.7014775	0.6507965	0.6062416
	BID	KY	GK	SSTR2	CCR2	LEAP2	PSD	CAT	OXT	APP
	0.5885117	0.5598003	0.5503918	0.5494050	0.5450880	0.5381766	0.5286841	0.5201827	0.5123753	0.4776944
	bmi	APOE	SH2B1	ACR	LBP	MET	ADA	OSM	FGL1	AIP
	0.4707412	0.4536297	0.4535384	0.4520901	0.4456180	0.4372056	0.4370593	0.4366659	0.4365070	0.4269469
	STS	STAT3	EMD	ANGPTL3	TH	SCD	IVD	RBP4	EPO	SI
	0.4212412	0.4200825	0.4163839	0.4013022	0.3992198	0.3962637	0.3920488	0.3879004	0.3856479	0.3832225
	TLR9	EDA	VDR	CD36	FES	TMEM18	IMPACT	KLF4	NOS3	LPA
	0.3708001	0.3706992	0.3490318	0.3476166	0.3461392	0.3459114	0.3382303	0.3354622	0.3340333	0.3287345

It is possible to observe that the most related genes with the term “Obesity” were CNP and S100A16.

In order to explore the relationship between the intestinal bacterial populations in patients with obesity and other metabolic diseases, the same analyses described above for the gene relation was carried out.

### Development of the application in Shiny

Shiny is an open source R package that provides an elegant and powerful web framework for building web applications using R. Shiny helps you turn your analyses into interactive web applications without requiring HTML, CSS, or JavaScript knowledge.

Shiny allows you to take your work in R and expose it via a web browser so that anyone can use it.

In the past, creating web apps was hard for most R users because:

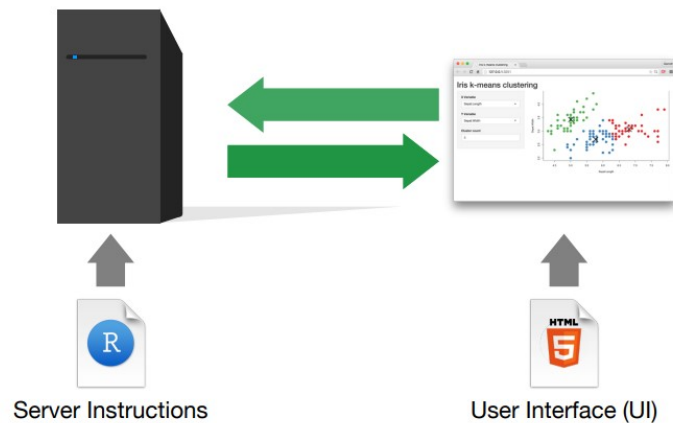
- You need a deep knowledge of web technologies like HTML, CSS and JavaScript.
- Making complex interactive apps requires careful analysis of interaction flows to make sure that when an input changes, only the related outputs are updated.

Shiny makes it significantly easier for the R programmer to create web apps by [20]:

- Providing a carefully curated set of user interface (UI for short) functions that generate the HTML, CSS, and JavaScript needed for common tasks.
- Introducing a new style of programming called reactive programming which automatically tracks the dependencies of pieces of code. This means that whenever an input changes, Shiny can automatically figure out how to do the smallest amount of work to update all the related outputs.

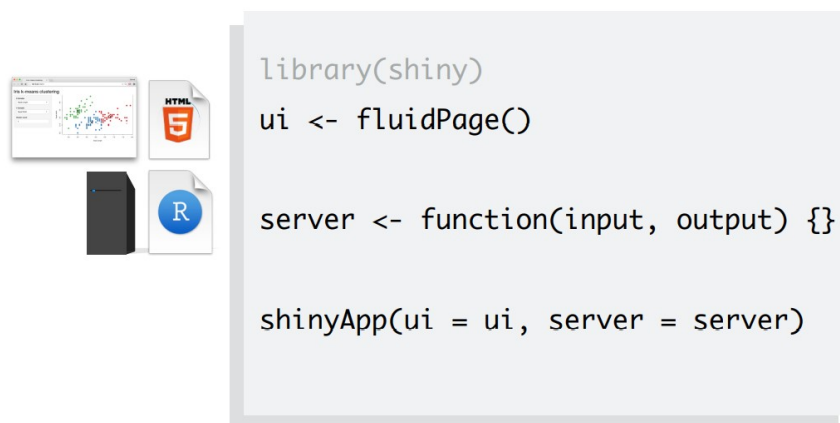
Today, Shiny is used in almost as many niches and industries as R itself is. It's used in academia as a teaching tool for statistical concepts, a way to get undergrads excited about learning to write code, a splashy medium for showing off novel statistical methods or models. It's used by big pharma companies to speed collaboration between scientists and analysts during drug development. It's used by Silicon Valley tech companies to set up real-time metrics dashboards that incorporate advanced analytics [20].

As we introduced before, the two key components of every Shiny app are: the UI (short for user interface) which defines how your app looks, and the server function which defines how your app works. Moreover, Shiny uses reactive programming to automatically update outputs when inputs change [20].



**Figure 14. Two key components of every Shiny app [21].**

There are several ways to create a Shiny app. The simplest is to create a new directory for the app, and put a single file called app.R in it. This app.R file is used to tell Shiny both how the app should look (UI), and how it should behave (Server).

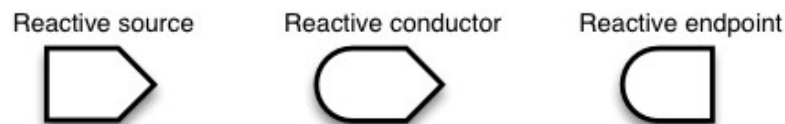


**Figure 15. The shortest viable Shiny app [21].**

The final expression executes `shinyApp(ui, server)` to construct and start a Shiny application from UI and server.

It's easy to build interactive applications with Shiny, but to get the most out of it, it is necessary to understand the reactive programming model used by Shiny.

Regarding to reactivity, there are three kinds of objects in reactive programming: reactive sources, reactive conductors, and reactive endpoints, which are represented with these symbols:



**Figure 16. Reactivity programming components of Shiny apps [21].**

In a Shiny application, the source typically is user input through a browser interface. For example, when the user selects an item, types input, or clicks on a button, these actions set values that are reactive sources. A reactive endpoint is usually something that appears in the user's browser window, such as a plot or a table of values [21].

It's also possible to put reactive components in between the sources and endpoints. These components are called reactive conductors. A conductor can both be a dependent and have dependents. In other words, it can be both a parent and child in a graph of the reactive structure. Sources can only be parents (they can have dependents), and endpoints can only be children (they can be dependents) in the reactive graph. Reactive conductors can be useful for encapsulating slow or computationally expensive operations [21].

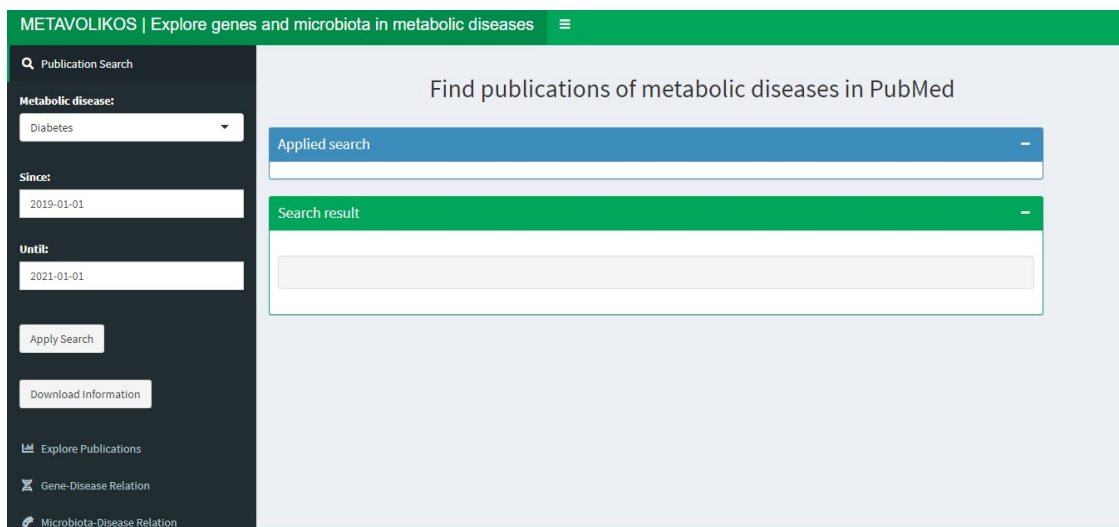
## 5 Results and discussion

As a result of the techniques described above, the METAVOLIK web application was obtained. It allows discovering information contained in the summaries of PubMed publications from data mining. Specifically, the web application can:

- Obtain a list of genes potentially related to different metabolic diseases.
- Explore the relationship between the intestinal bacterial populations in patients with obesity and other metabolic diseases.
- Identify relationships between genes, obesity and other metabolic diseases.

### Home

The main page of the tool contains a side menu which gives access to all the tabs.



**Figure 17. Home of METAVOLIKOS.**

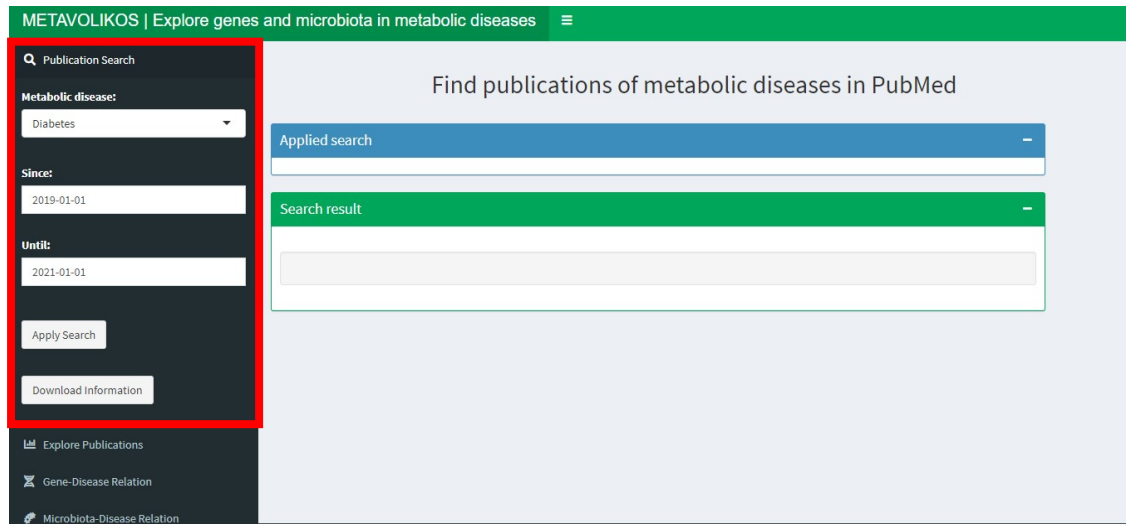
The app contains four main tabs:

- Publication search.
- Explore publications.
- Gene-Disease relation.
- Microbiota-Disease relation.

## Publication Search

This tab allows to search articles from PubMed, which contain the MeSH tag associated with one of the diseases in a range of dates chosen by the user.

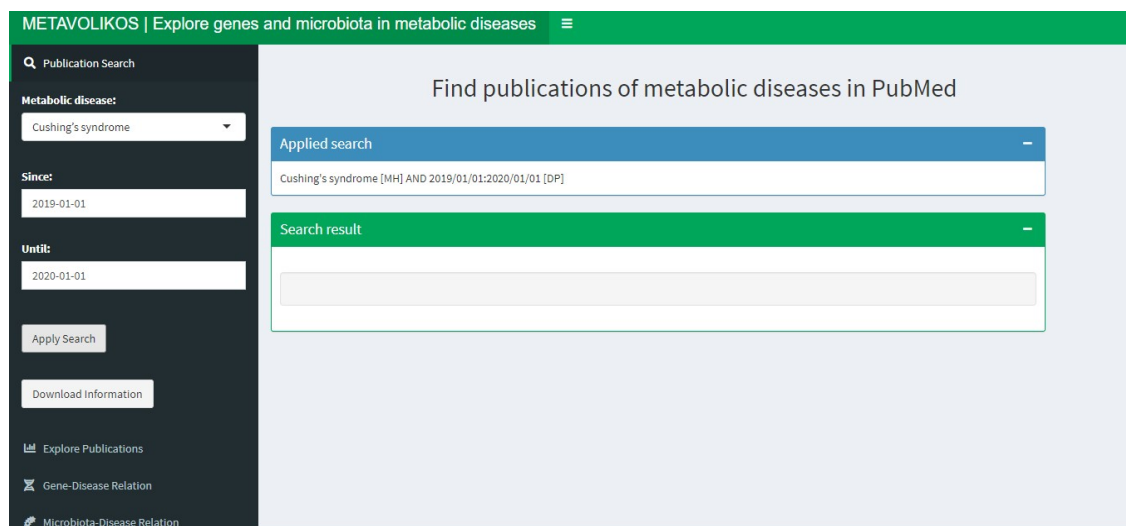
From the side menu, the disease and the date range can be selected for the search in PubMed.



The screenshot shows the METAVOLIKOS interface with a green header bar containing the text "METAVOLIKOS | Explore genes and microbiota in metabolic diseases". On the left, a dark sidebar menu is highlighted with a red border. It includes a "Publication Search" section with a search icon, a "Metabolic disease:" dropdown menu set to "Diabetes", "Since:" and "Until:" date input fields (set to "2019-01-01" and "2021-01-01" respectively), and "Apply Search" and "Download Information" buttons. Below this are three menu items: "Explore Publications", "Gene-Disease Relation", and "Microbiota-Disease Relation". The main content area has a light blue background and the heading "Find publications of metabolic diseases in PubMed". It contains two expandable sections: "Applied search" (blue header) and "Search result" (green header), both currently collapsed.

**Figure 18. Search publications menu.**

From “Search” button the query is displayed as it is sent to PubMed. In the example shown it is possible to observe the search for the disease “Cushing’s syndrome”. In the upper box of results, the MeSH concept associated to this disease can be seen, as well as the publication dates selected with the label [DP].



This screenshot shows the same METAVOLIKOS interface as Figure 18, but with the search parameters changed. The "Metabolic disease:" dropdown is now set to "Cushing's syndrome". The "Since:" date is "2019-01-01" and the "Until:" date is "2020-01-01". The "Applied search" section is now expanded, displaying the query: "Cushing's syndrome [MH] AND 2019/01/01:2020/01/01 [DP]". The "Search result" section remains collapsed.

**Figure 19. Query visualization in PubMed from Publications Search tab.**



“Download Information” button trigger the articles downloading from PubMed related to the applied search.

Results are shown in a table in the “Search result” box. The table shows the PMID number and the title of each downloaded article. In addition, a new button is enabled in the side menu to perform a new search.

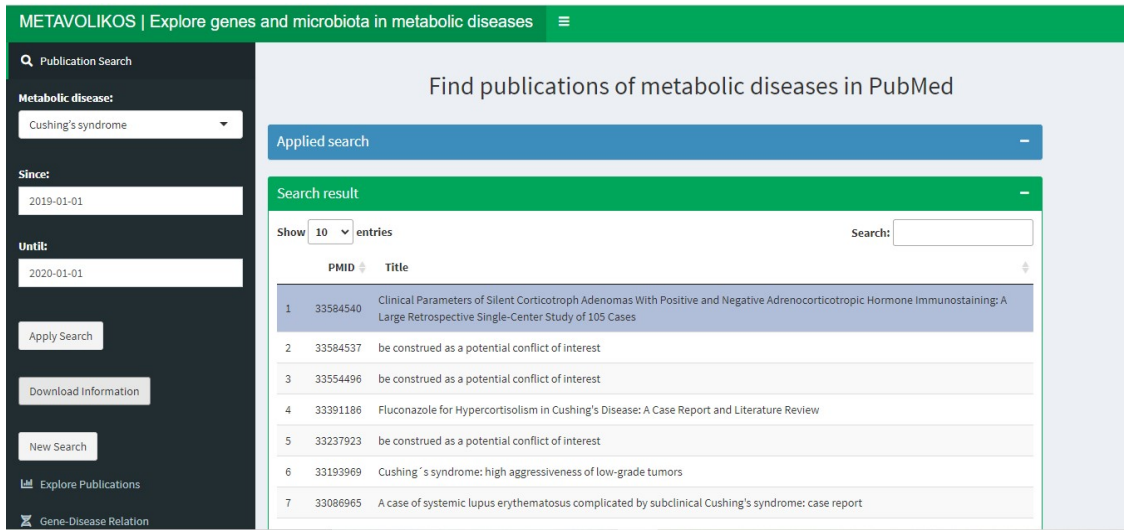


Figure 20. Results of the PubMed search form “Cushing’s syndrome” and publication date between 2019 and 2020.

At the bottom of the table, the title and abstract of the selected publication is displayed, allowing a quickly reading of the content of the article.

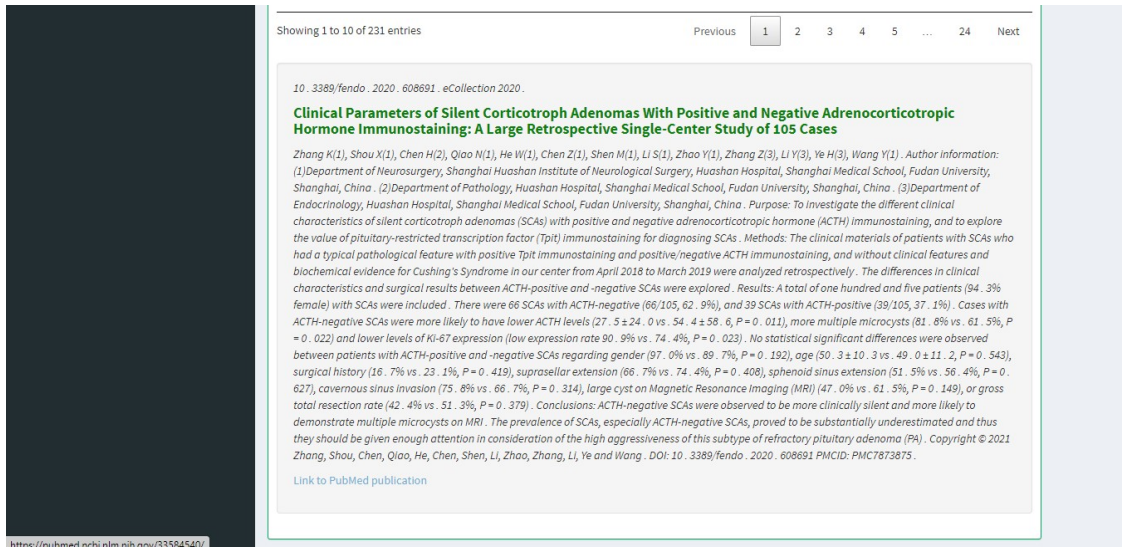


Figure 21. Detail of the bottom of the table where the title and abstract of the selected publication are displayed.

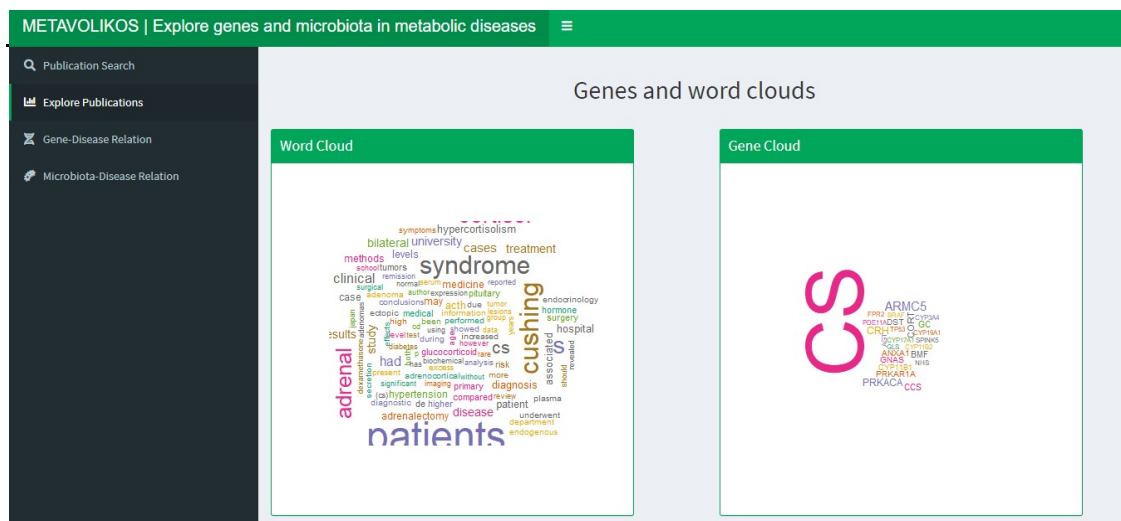
In addition, a link that leads to the article in PubMed is facilitated.



**Figure 22. Page to which the link of each article redirects.**

## Explore Publications

The second tab allows exploring previously downloaded abstracts. Genes and words are extracted from them and are displayed in cloud graph and table format.



**Figure 23. Word and gene cloud for the search of “Cushing’s syndrome” and publication date between 2019 and 2020.**

In this graph bigger is the size of the term higher is its frequency into the corpus.

Word Table	
Word	Frequency
patients	520
cushing	328
syndrome	304
s	280
cortisol	258
adrenal	253
cs	184
clinical	136
had	122
study	112

Gene Table		
Symbol	Name	Frequency
CS	citrate synthase	184
ARMC5	armadillo repeat containing 5	16
CRH	corticotropin releasing hormone	10
PRKACA	protein kinase cAMP-activated catalytic subunit alpha	10
BMF	Bcl2 modifying factor	8
CCS	copper chaperone for superoxide dismutase	8
GC	GC vitamin D binding protein	8
CORT	cortistatin	7
DST	dystonin	6
GNAS	GNAS complex locus	6

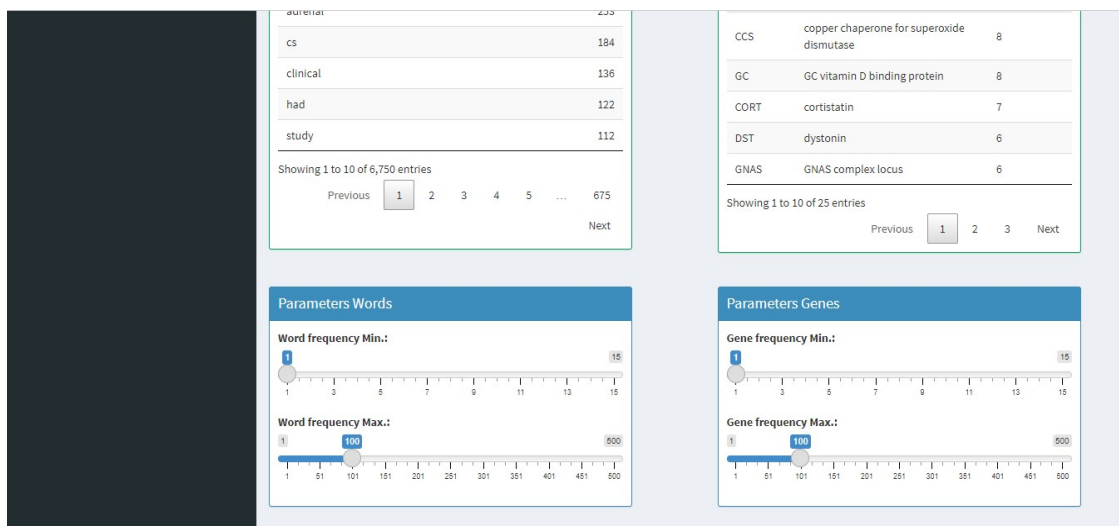
**Figure 24. Result of frequency cited words and genes of “Cushing’s syndrome” and publication date between 2019 and 2020 in table format.**

Regarding to the words result for the current search, it is possible to see that the most frequent used terms are nonspecific words like patients, cushing, syndrome, cortisol or adrenal.

In relation with the genes, clearly, the gene most cited into the corpus was the gene “CS” (citrate synthase).

Taking into account these results it is possible to hypothesize that the cortisol, the adrenal glands and the gene “CS” are involved into the physiopathology of the Cushing’s syndrome.

Maximum and minimum frequency of words and genes can be selected. When this selection is modified, the cloud graphs and the tables are updated automatically.



**Figure 25. Boxes to modify the minimum and maximum frequency of words and genes to be displayed into the cloud graphs and tables.**

### Gene-Disease Relation

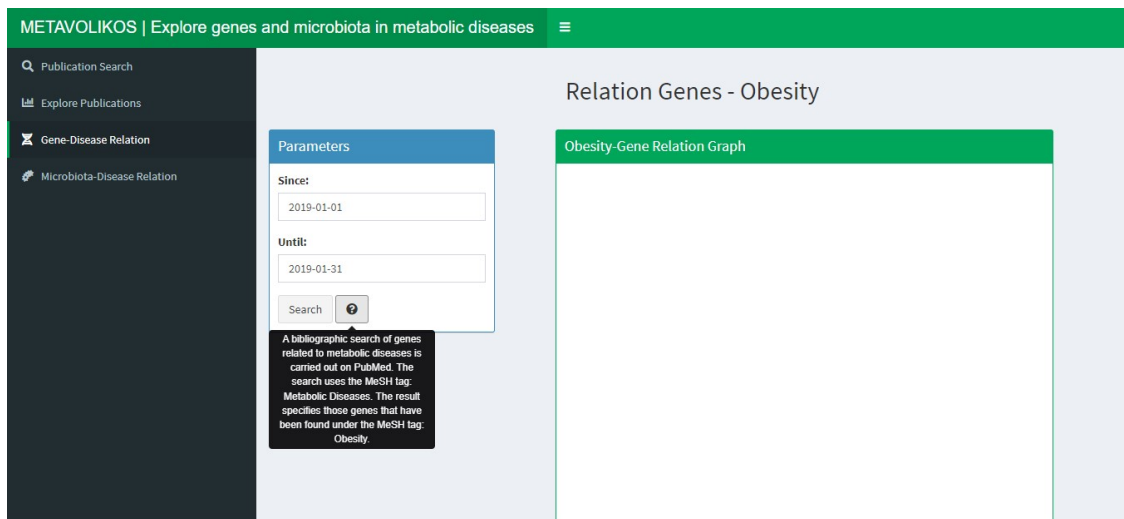
In this tab the relation between genes, obesity and other metabolic diseases is carried out. That relation is established based on the genes mentioned in publications on metabolic diseases to infer their relation with the term “Obesity” through latent semantic analysis.

The result is a table indicating if the gene was already found in MeSH search-based obesity publications.

By obtaining the document terms matrix, and consequently, the terms terms matrix, it is possible to find for those genes related to obesity with which other metabolic diseases have high similarity (measured by cosine).

Results are displayed in table and graph form.

Moreover, it is possible to select a range of dates between which to search the publications. The instructions for the user and showed when the question sign button is clicked.



**Figure 26. Genes and metabolic diseases relation tab.**

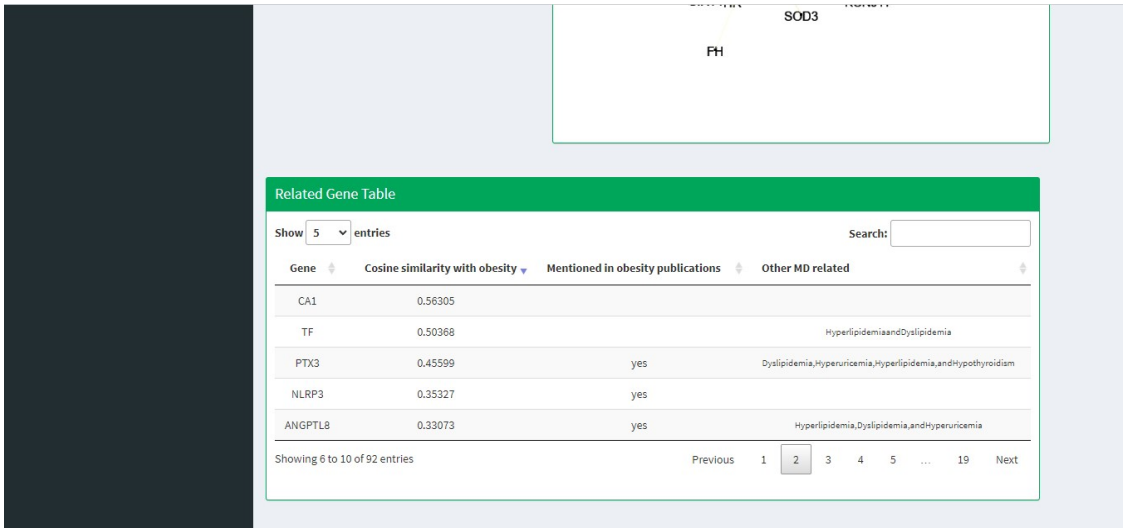
The result of the search is shown in a table, which detail the genes mentioned in the publications, their relation with the term “Obesity” (this relation is evaluated by cosine), and a note about if the gene is mentioned in the publication with MeSH term “Obesity” in the selected period. In addition, for each gene a list of other related metabolic diseases is included.

Gene	Cosine similarity with obesity	Mentioned in obesity publications	Other MD related
RBP4	0.90923	yes	
SDS	0.73361	yes	HyperuricemiaandDyslipidemia
SPX	0.68052	yes	Hypothyroidism
FTO	0.66035	yes	
GLS	0.65423	yes	HyperuricemiaandDyslipidemia

**Figure 27. Related gene table (first tab).**

The table is ordered by cosine value, so those genes that are mostly related to the term “Obesity” appear in the first’s rows of the table.

In the current example, it is possible to see that the gene RBP4 has high relation with the term “Obesity” (0.90 cosine similarity). Due to that, this gene is already present into the abstracts of publications with the MeSH term “Obesity” associated. It doesn’t have any other metabolic disease related.



**Figure 28. Related gene table (second tab).**

In the second tab of the related gene table, appears the gene TF, which has a medium cosine similarity (0.50) with the term “Obesity” and it doesn’t appear into the publications about obesity of the selected period of publications. Moreover, this gene has relation with other metabolic diseases like hyperlipidemia and dyslipidemia.

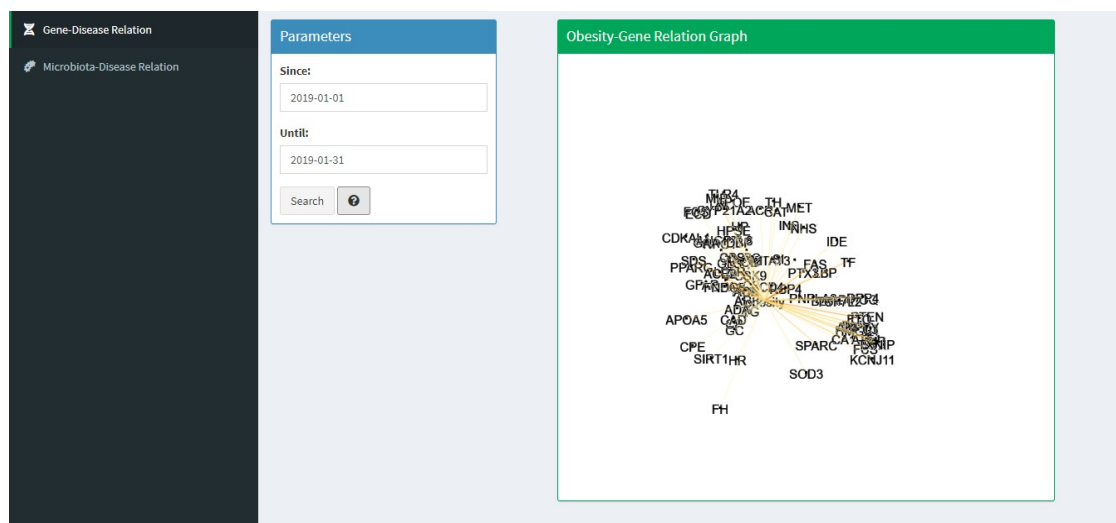
This case represents how latent semantic analysis, and specifically METAVOLIKOS tool, can shed light to create new research hypotheses.

A relationship has been inferred from publications on metabolic diseases where articles that include the term “Obesity” or explicitly this gene are not necessarily included.

It is not unreasonable to think that the gene TF is related with the obesity and it is not related directly in scientific publications. It opens a new research line.

Results are also displayed in “3D graph” form. It is a three dimensions graph of the neighbors of the term “Obesity”. The connections show a different color depending on the cosine value of each pair of terms. The colors follow a scale

of yellow-red: from pale yellow (cosine of 0) to dark red (cosine of 1). The graph can be rotated to make it easier to explore all pair of terms.



**Figure 29. Obesity-gene relation 3D graph.**

### Microbiota-Disease Relation

This section follows the same structure than the previous one but relating intestinal bacteria population, obesity and other metabolic diseases. That relation is established based on the bacteria population and related concepts mentioned in publications on metabolic diseases (which include the term “Microbiota” or “Microbiome”) to infer their relation with the term “Obesity” through latent semantic analysis.

The result is a table indicating if a specific bacteria population or related concept was already found in MeSH search-based obesity publications (which contains the term “Microbiota” or “Microbiome”).

By obtaining the document terms matrix, and consequently, the terms terms matrix, it is possible to find for those bacteria related to obesity with which other metabolic diseases have high similarity (measured by cosine).

Results are displayed in table and graph form.

Moreover, it is possible to select a range of dates between which to search the publications. The instructions for the user and showed when the question sign button is clicked.



**Figure 30. Microbiota and metabolic diseases relation tab.**

The result of the search is shown in a table, which detail the bacteria population mentioned in the publications, their relation with the term “Obesity” (this relation is evaluated by cosine), and a note about if the bacteria is mentioned in the publication with MeSH term “Obesity” in the selected period (articles must include the term “Microbiota” or “Microbiome”). In addition, for each bacteria population or related concept a list of other related metabolic diseases is included.

Microbiota	Cosine similarity with obesity	Mentioned in obesity publications	Other MD related
dysbiosis	0.9204	yes	Metabolic, Diabetes, and Dyslipidemia
acids	0.81662	yes	Hyperlipidemia, Metabolic, and Diabetes
metabolites	0.8149	yes	Hyperlipidemia, Diabetes, Dyslipidemia, and Metabolic
butyrate	0.76918	yes	Hyperlipidemia, Diabetes, Metabolic, and Dyslipidemia
metagenomic	0.75095	yes	Dyslipidemia, Hyperlipidemia, and Diabetes

**Figure 31. Related microbiota table (first tab).**



The table is ordered by cosine value, so those bacteria population or related concept that are mostly related to the term “Obesity” appear in the firsts rows of the table.

In the current example, it is possible to see that the concept “dysbiosis” has high relation with the term “Obesity” (0.92 cosine similarity). Due to that, this concept is already present into the abstracts of publications with the MeSH term “Obesity” associated. Moreover, this concept is also related with metabolic diseases, diabetes and dyslipidemia.

Microbiota	Cosine similarity with obesity	Mentioned in obesity publications	Other MD related
microbes	0.73645	yes	Dyslipidemia,Hyperlipidemia,andDiabetes
diversity	0.71278	yes	Diabetes,Dyslipidemia,andHyperlipidemia
microbiome	0.64457	yes	Dyslipidemia
microbiota	0.62877	yes	HyperlipidemiaandDiabetes
lactobacillus	0.58011	yes	DiabetesandHyperlipidemia

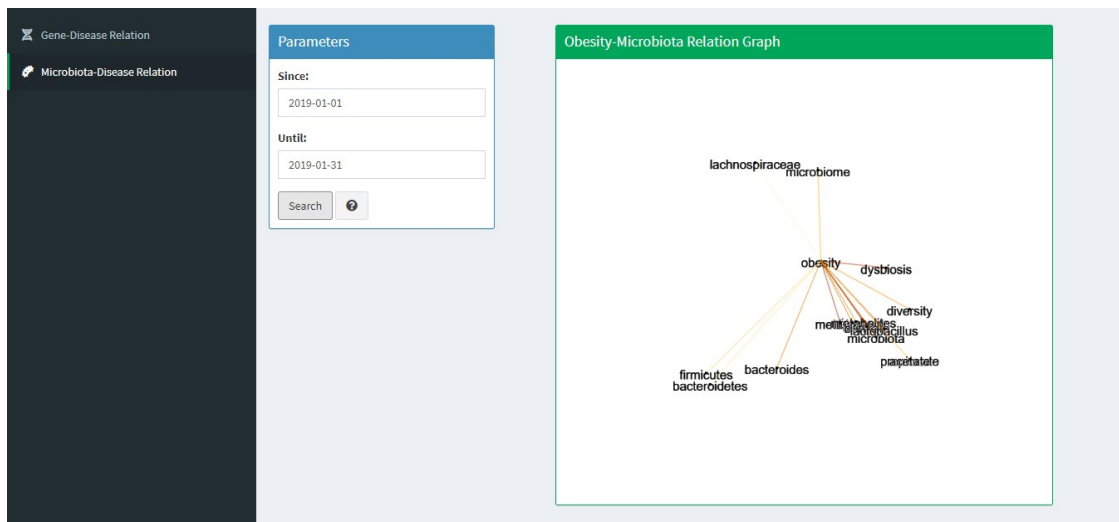
Showing 6 to 10 of 16 entries

Previous 1 2 3 4 Next

**Figure 32. Related microbiota table (second tab).**

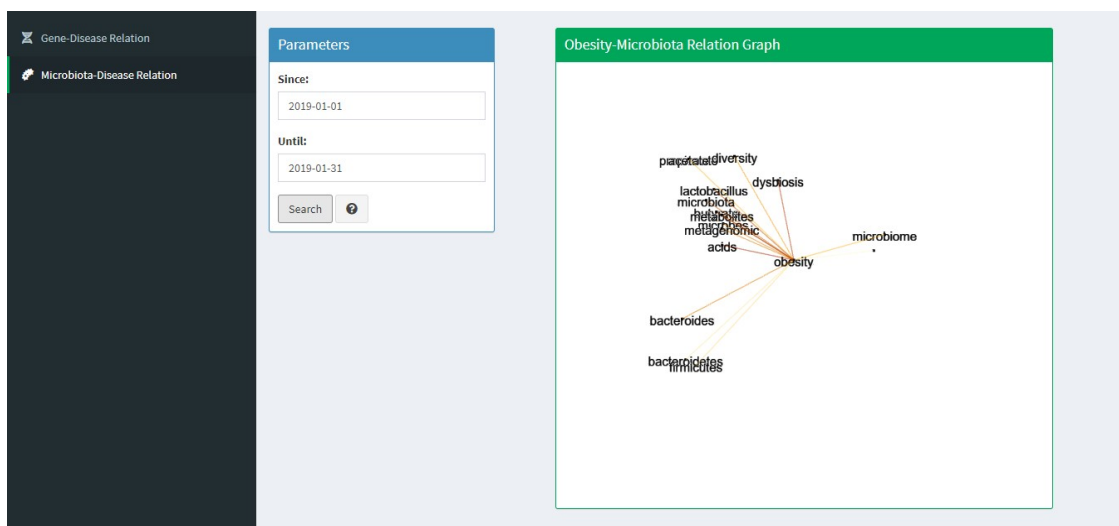
Is in the second tab of the related microbiota table is possible to find the first bacterial population related to the term “Obesity” (cosine of 0.58). The most related concepts to term “Obesity” are concepts related to intestinal microbiome (like dysbiosis or acids), but not bacterial population. This bacteria is lactobacillus, which is already present into the obesity publications and also is related with other metabolic diseases like diabetes and hyperlipidemia.

Results are also displayed in “3D graph” form. It is a three dimensions graph of the neighbors of the term “Obesity”. The connections show a different color depending on the cosine value of each pair of terms. The colors follow a scale of yellow-red: from pale yellow (cosine of 0) to dark red (cosine of 1). The graph can be rotated to make it easier to explore all pair of terms.



**Figure 33. Obesity-microbiota relation 3D graph (default output).**

Observing the three dimensions graph it is possible to see the relation between the term “Obesity” and the bacterial population of firmicutes and bacterioidetes/bacterioides. They can be defined as quite related to the term “Obesity” too.



**Figure 34. Obesity-microbiota relation 3D graph (rotated output).**

Using the functionality of graph rotation can modify the output to properly show the relation of the concept dysbiosis and obesity. The connection is represented in dark red color; due this pair has the highest similarity by cosine.

## 6 Conclusions

### 6.1 Conclusions

The development of METAVOLIKOS, a web application that allows discovering information contained in the abstracts of scientific publications on metabolic diseases, has been achieved.

In relation with the objectives proposed, it has been fully reached. Concretely the application allows:

- Obtain a list of genes potentially related to different metabolic diseases.
- Obtain a list of terms potentially related to different metabolic diseases.
- Explore the relationship between the intestinal bacterial populations in patients with obesity and other metabolic diseases.
- Identify relationships between genes, obesity and other metabolic diseases.

The web application developed is interactive and would allow to the user to enjoy of a friendly interface, avoiding the requirement of programming knowledge or special software. It could allow explore the content of PubMed publications obtaining key information (such as the gene-disease or microbiome-disease relationship) presented in the form of tables and graphs.

This application provides solution to the current problem since it is a text mining tool that is easily adaptable and usable by end users. Many researchers have developed systems that can be adapted by other text mining specialists, but applications that can be tuned by bench scientists are mostly lacking.

The application is available at the following link and is open to the public: <https://neustorrentample.shinyapps.io/METAVOLIKOS/>

During the discussion, it was detected cases in which the original corpus of abstracts didn't contain the term "Obesity" and the term of a specific gene in the same document. However, the LSA predicted a relationship between both terms. The latter evidences the power of the LSA technique and its potential to infer relationships between two terms in a context, shedding light to the biomedical literature and facilitating the creation of new research hypotheses.

The Shiny tool was useful for the development of the interactive application since programming knowledge is not required to use it.

## 6.2 Future lines

Regarding to the methods used during the development of this application, the detection of the abstract titles was challenging due to the function used does not always obtain satisfactory results, detecting erroneous phrases as the title of the publication.

The atomization of words shows many words that are not specific to the study area (such as patient, study, etc.), which makes the tool less useful.

All this makes interesting to think of new lines of work that allow text mining to be combined with a manual information curation process or to debug functions for these purposes.

Regarding to the relationship between the intestinal microbiota and the metabolic diseases, there was difficulty in finding a function to extract the terms related to this topic from the abstracts of the corpus. The available functions (as `pubtator()`) are specialized in the study of genes, diseases, chemical agents, mutations and species, making it difficult to apply them to other type of terms, such as the intestinal microbiome. It would be interesting to develop a new tool for this purpose.

Another aspect to consider is the computation time, especially in the strategy of searching for new related genes starting from the corpus of metabolic diseases and using a high number of terms (such as the genes found and keywords). Increasing the size of the corpus and, consequently, the terms (genes and keywords are dependent on the size of the corpus), the matrix of documents terms grows quickly, limiting the response of the application at the point of the LSA. It would be interesting to optimize the functions already used to optimize the computation time.

As source of information, although the PubMed abstracts are a good starting point since they contain a good summary of the publication and have a public access, sometimes detailed information or terms of interest to be related don't appear on them. It would be more accurate (and interesting to compare) to work with a corpus created with the whole scientific publications.

Thanks to be developed using Shiny, METAVOLIKOS, is a very easily adaptable application. Due to that, in the future could be interesting to add more features on it to improve the usability or increase the functionalities. Moreover, it is also possible to adapt it to other medical area.

### 6.3 Planning follow-up

The applied methodology was the proper one. However, the planning was not fully in line with the deadlines of the evaluation tests.

Main sections of the platform could not be programmed before the delivery date of the evaluation test number 3 as it was planned.

It would have been key in the planning to begin programming in Shiny during the phase 1 of the project development, instead of postponing everything for phase 2.

It was unexpected to have to incorporate new knowledge about the development of the applications using Shiny to through the entire development phase of the tool.

## 7 Glossary

**Metabolic disease:** any of the diseases or disorders that disrupt normal metabolism, the process of converting food to energy on a cellular level.

**Obesity:** abnormal or excessive fat accumulation that presents a risk to health.

**MD:** abbreviation of metabolic disease.

**Microbiota/microbiome:** the range of microorganisms that may be commensal, symbiotic, or pathogenic found in and on all multicellular organisms, including plants.

**LSA:** abbreviation of latent semantic analysis. Technique that allow process texts and to infer relations between terms and documents.

**Data mining:** the process of sorting through large data sets to identify patterns and relationships that can help to find new information and knowledge through data analysis.

**Text mining:** the process of exploring and analyzing large amounts of unstructured text data that can help to find new information and knowledge through data analysis.

**Corpus:** set of abstracts of scientific publications on which text mining techniques are applied.

**Shiny:** Shiny is an R package that makes it easy to build interactive web apps straight from R.

**R:** open source programming language with a huge variety of statistics libraries.

## 8 Bibliography

1. Sanmiguel, C. Gut Microbiome and Obesity: A Plausible Explanation for Obesity. *Current Obesity Reports*, **2015**, 4(2):250-61.
2. Conway, B.; Rene, A. Obesity as a disease: no lightweight matter. *Obesity reviews*, **2004**, 5(3):145-51.
3. Tanvig M. Offspring body size and metabolic profile - effects of lifestyle intervention in obese pregnant women. *Danish medical Journal*, **2014**, 61(7):B4893.
4. Caballero, B. Humans against Obesity: Who Will Win?. *Advances in Nutrition*, **2019**, 10(suppl\_1):S4-S9.
5. Seidell, J.C. Obesity, insulin resistance and diabetes: a worldwide epidemic. *The British Journal of Nutrition*, **2000**, 83 Suppl 1:S5-8.
6. Pigeyre, M.; Yazdi, F.T.; Kaur, Y.; Meyre, D. Recent progress in genetics, epigenetics and metagenomics unveils the pathophysiology of human obesity. *Clinical Science*, **2016**, 130(12):943-86.
7. George Kunnackal John, G.K.; Mullin, G.E. The Gut Microbiome and Obesity. *Current Oncology Reports*, **2016**, 18(7):45.
8. Lee, C.J.; Sears, C.L.; Maruthur, N. Gut microbiome and its role in obesity and insulin resistance. *Annals of the New York Academy of Sciences*, **2020**, 1461(1):37-52.
9. Charu C. Aggarwal. *Data Mining*. (Springer, 2015).
10. Rani; Jyoti; Shah, A.R.; Ramachandran, S. PubmedminerR: An R package with text-mining algorithms to analyse PubMed abstracts. *Journal of Biosciences*, **2015**, 40(4), 671-682.
11. Max Bramer. *Principles of Data Mining*. (Springer, 2013).
12. Gonzalez, G.H.; Tahsin, T.; Goodale, B.C.; Greene, A.C.; Greene, C.S. Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery. *Briefings in Bioinformatics*, **2016**, 17(1), 33–42.

13. Rebholz-Schuhmann, D.; Kirsch, H.; Couto, F. Facts from Text—Is Text Mining Ready to Deliver?. *PLOS Biology*, **2005**, 3(2): e65.
14. Tomar, D.; Agarwal, S. A survey on Data Mining approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology*, **2013**, 5(5), 241-266.
15. Wu, X.; Kumar, V.; Quinlan, J.R.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G.J.; Ng, A.; Liu, B.; Yu, P.S.; Zhou, Z.H.; Steinbach, M.; Hand, D.J.; Steinberg, D. Top 10 algorithms in data mining. *Knowl Inf Syst*, **2008**, 14, 1–37.
16. Liao, S.H.; Chu, P.H.; Hsiao, P.Y. Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications*, **2012**, 39(12), 11303-11311.
17. Piedra, D.; Ferrer, A.; Joaquim, G. Text Mining and Medicine: Usefulness in Respiratory Diseases. *Archivos de Bronconeumología*, **2013**, 50(3), 113-119.
18. Gefen, D.; Endicott, J.E.; Fresneda, J.E.; Miller, J.; Larsen, K.R. A Guide to Text Analysis with Latent Semantic Analysis in R with Annotated Code: Studying Online Reviews and the Stack Exchange Community. *Communications of the Association for Information Systems*, **2017**, 41(21), pp-pp.
19. *Datacamp* [on line] [consult: 16th of May of 2022]. Available in: <https://www.datacamp.com/tutorial/discovering-hidden-topics-python>
20. Hadley Wickham. *Mastering Shiny*. (O'Reilly Media, 2020).
21. Shiny from R Studio [on line] [consult: 15th of May of 2022]. Available in: <https://shiny.rstudio.com/articles/reactivity-overview.html>
22. National Library of Medicine: Learning Resources Database [on line] [consult: 27th of May of 2022]. Available in : <https://learn.nlm.nih.gov/documentation/training-packets/T0042010P/>
23. Almeida, A.; Mitchell, A.L.; Boland, M.; Forster, S.C.; Tarkowska, A.; Lawley, T.D.; Finn, R.D. A new genomic blueprint of the human gut microbiota. *Nature*, **2019**, 568, 499-504.
24. Shanahan, F.; Sheehan, D. Microbial contributions to chronic inflammation and metabolic disease. *Curr Opin Clin Nutr Metab Care*, **2016**, 19(4), 257-262.



25. Hernández-Granados, M.J.; Ramírez-Emiliano, J.; Franco-Robles, E. *Experimental Animal Models of Human Diseases - An Effective Therapeutic Strategy*. (Ibeh Bartholomew, 2018).