Université catholique de Louvain

Integrated Project

Text Mining with LSA & LDA

Gabriel Carestia
Timothée Clément
Nicolas Deffense
François Duquesne
Thomas Feron
Antoine Rummens
Brieuc Ryelandt

Professors:
Emmanuel Hanert
Patrick Bogaert

December 19, 2018

UCLouvain
Faculté des bioingénieurs

Contents

1 Introduction

The purpose of this project is to build a search engine using the powerful tools of text mining. This search engine will aim to help pharmaceutical researchers searching for relevant information in the thousands of articles on the PubMed data base. Before explaining the technical aspects of our project, we will briefly explain what text mining is all about (2).

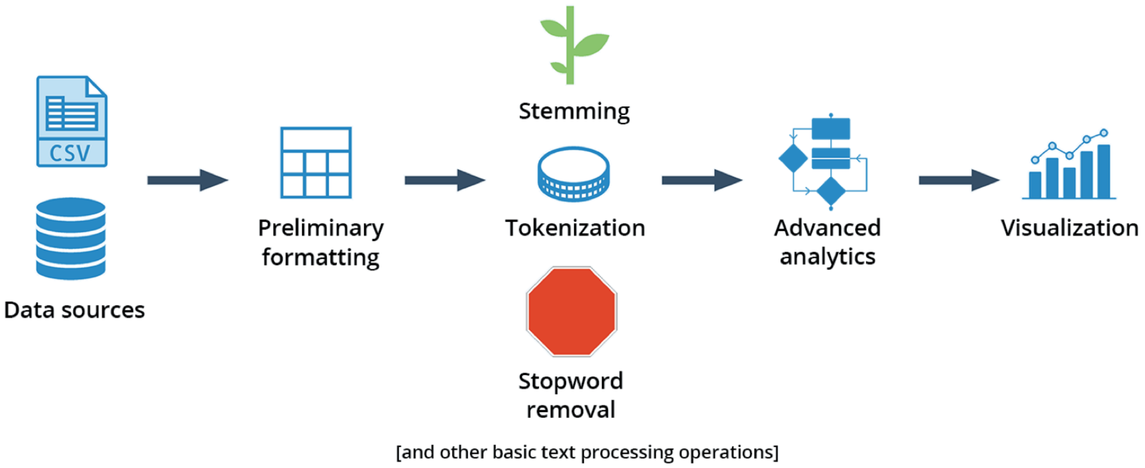


Figure 1: Text Mining Pipeline

The figure 1 shows the different steps of text mining. It can be seen as a pipeline, running from raw data to output.
According to Gartner's *Market Guide for Text Analytic* "At a high level, a text analytic system is a pipeline process comprising four stages:

- Text acquisition and preparation *(chap. 2)*
- Text processing *(chap. 3)*
- Text analysis *(chap. 4)*
- Output - User interface *(chap. 5)*."

We will end this report by discussing the difference between the two algorithms used. Also, we will discuss what could be done to improve our work.

All our codes are in an open source folder on GitHub to allow others to improve or use this project. Please, go to the following address : *https://github.com/Option10/Text-Mining-with-R.*

2 Text acquisition and preparation

For this project, we're using one of the PubMed datasets provided on the web ("pubmed18n0924.xml"). This document is a XML file which contains information such as the title, authors, abstracts, etc. for each article and for approximately 28,600 documents.

To be able to begin our text analysis, we need to extract the mentioned elements into a new dataframe that will be much easier to manipulate. At this point we are only taking into account the ID, Title, Abstract, Publishing date, Main authors, and Keywords of each article. The most important object in the data frame is without any doubt the Abstract since our text mining relies entirely on it.

It is important to mention that better results could be obtained if we had access to the entirety of the documents since we would have more information to process and thus a better precision. Also, involving keywords in text mining analysis could be a quick and easy way to extract relevant information about articles.

3 Text processing

- **Tokenisation** : The first step to process our text is tokenisation. The objective is to divide the text into tokens that can be words (unigram) or groups of words. For instance *This is a sentence* will be divided in [this], [is], [a], [sentence]. It is also possible to divide into groups of words (n-grams) to keep the information that is given by the words position e.g [not happy] $\Longleftrightarrow$ [not] [happy].
- **Stopword removal** : The stopwords are words that don't give any information for the analysis. For instance the verbs *to be* or *to have* or the determiners are present in every text and will tilt the analysis.
- **Stemming** : The stemming process is used to gather words that share the same grammatical structure. For instance, *linked*, *links*, *link*, *linkage* will all be change into *link*. Stemming reduces words to their basic form to avoid the repetition of words with the same meaning.
  In our case, we did not include this step in the data preprocessing because we were not convinced of the results we obtained.

Example:

This sentence is helping you to understand the tokenisation.

Tokens: [this] [sentence] [is] [helping] [you] [to] [understand] [tokenisation]

Stopwords: this, is, you, to

Tokens remaining: [sentence] [helping] [understand] [tokenisation]

Stemming: [sentence] [help] [understand] [token]

4 Text analysis

4.1 Words frequency

Once the text is divided into tokens, it is possible to analyse the words frequency to have an idea of a text's topic. The tool used for that purpose is the *TF-IDF* (Term frequency - Inverse document frequency).

TF: the term frequency is simply the normalised count of each token in a text.

IDF: the inverse document frequency is the count of each token in the whole corpus of texts analysed. The idea is to give a low impact to a token that is present in many texts.

TF-IDF: is the product between TF and IDF. It gives for each token its importance in a text lowered by its importance in the corpus. If it has a high value, it means that the word appears a lot in the text but not much in the corpus, so it will be very significant for this text.

4.2 Latent Semantic Analysis

Steyvers & Griffiths say that : "The Latent Semantic Analysis approach makes three claims :

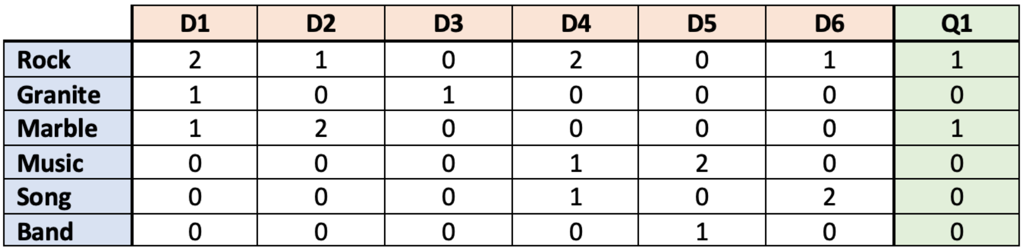
1. semantic information can be derived from a term-document co-occurrence matrix;
2. dimensionality reduction is an essential part of this derivation;
3. terms and documents can be represented as points in Euclidean space."

4.2.1 Term-document matrix

The LSA makes it possible to establish relationships between a set of documents and the terms they contain, by constructing "concepts" or "topics" related to documents and terms. A topic can be seen as the subject matter of a text, the meaning of the text.

For various purposes, we can choose to represent a text by a "Bag of words". In this way it's possible to see the most frequent words but how can we determine the topics of the text ?

The main idea of LSA is to build a "Term-Document matrix" that describes the occurrence of certain terms in documents. It's a spare matrix whose rows correspond to "terms" and whose columns correspond to "documents".



	D1	D2	D3	D4	D5	D6	Q1
Rock	2	1	0	2	0	1	1
Granite	1	0	1	0	0	0	0
Marble	1	2	0	0	0	0	1
Music	0	0	0	1	2	0	0
Song	0	0	0	1	0	2	0
Band	0	0	0	0	1	0	0

Figure 2: Term-Document Matrix

The figure 2 shows an example with a corpus of 6 documents (D1,..., D6) containing 6 terms. Here it is very simple to notice the appearance of 2 topics, one concerning mineralogy and the other rock music. The last column is an example of query. Someone is looking relevant information about "rock" and "marble". (Bao)

4.2.2 Dimensionality reduction

Secondly, we use the Singular Value Decomposition (SVD) to find topics in text and reduce dimensionality of the term-document matrix. In our project, we chose to use the “Irlba” package which is a fast and memory-efficient way to compute a partial SVD.

The dimensionnality reduction is a factorization technique that separates any matrix A into a product of 3 separates matrices $A = T * S * D^T$:

- A : Term - Document matrix
- T : Orthogonal matrix of the Left Singular Vectors
- S : Diagonal matrix of the Singular Values of A
- D^T : Orthogonal matrix of the Right Singular Vectors

SVD reduces dimensionality by selecting only the k largest singular values, and only keeping the first k columns of T and D^T : $\hat{A} = T_k * S_k * D_k^T$

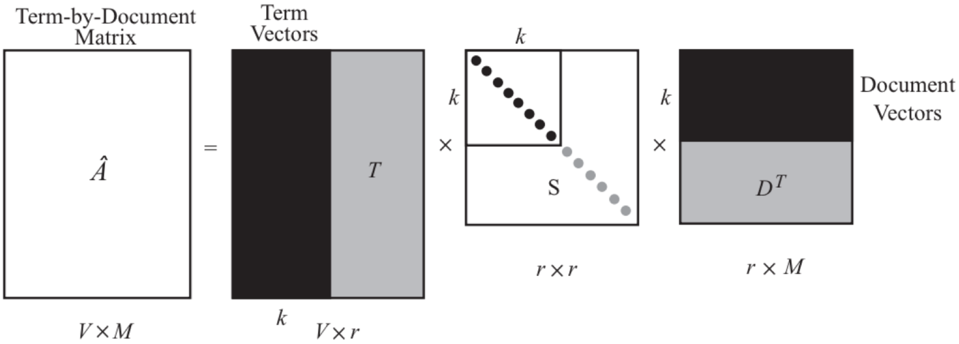


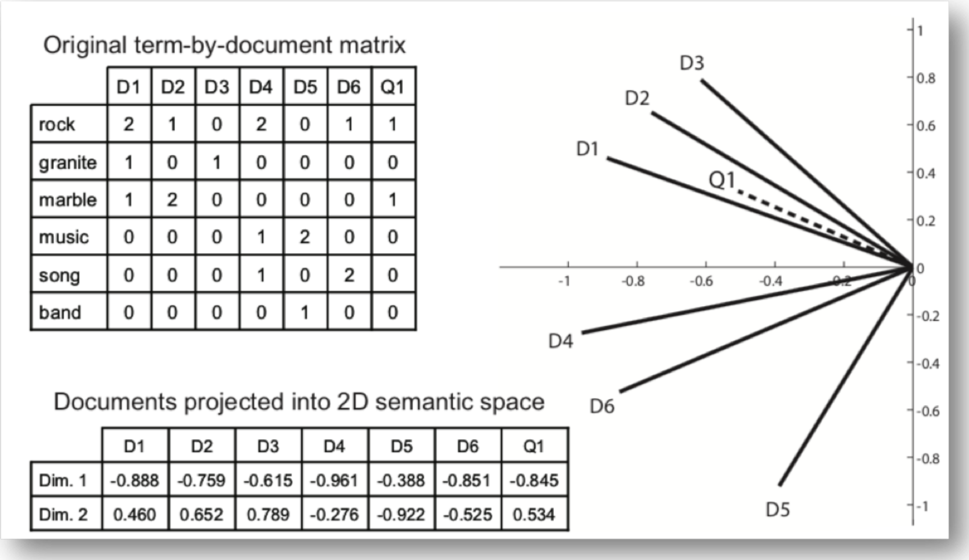
Figure 3: Singular Value Decomposition

The matrix S is called the *"topic importance matrix"* where each singular value represent the "strength" of each topic. With this matrix, it's possible to remove the least important topics (associated with the lowest singular values) and thus reduce the dimensionality of the matrix.

The number of topics is a hyperparameter that can be modified by an experienced user. In our project, we have chosen to set it arbitrarily at 100.

4.2.3 Representation in space

Once we have the matrices resulting from the SVD, we can project all the documents - in our case the abstracts from PubMed - in spaces. The axes of the space correspond to the topics. The closer the documents are spatially, the more they are supposed to talk about the same topics. When the user submits a request, it is projected in the same space. It is therefore possible to calculate Euclidean distances in order to see the abstracts that are most relevant for this query.



Original term-by-document matrix

	D1	D2	D3	D4	D5	D6	Q1
rock	2	1	0	2	0	1	1
granite	1	0	1	0	0	0	0
marble	1	2	0	0	0	0	1
music	0	0	0	1	2	0	0
song	0	0	0	1	0	2	0
band	0	0	0	0	1	0	0

Documents projected into 2D semantic space

	D1	D2	D3	D4	D5	D6	Q1
Dim. 1	-0.888	-0.759	-0.615	-0.961	-0.388	-0.851	-0.845
Dim. 2	0.460	0.652	0.789	-0.276	-0.922	-0.525	0.534

Figure 4: Representation in Euclidean space

4.2.4 Output and query system for LSA

As we have seen before, the LSA gives us three output matrix (T_k,S_k and D_k^T). Those matrix will allow us to compare words with the documents we have analysed with our algorithm. Because we want to compare terms with documents (a query which is a combination of words and the corpus), we have to calculate the coordinates of that combination of words in the Document-topic matrix (eq 1).

$$\hat{q} = S_k^{-1} T_k^T q \tag{1}$$

The $\hat{q}$ vector that can be compared with document matrix (D_k^T) is the cross-product between the topic importance matrix (S_k), the term-topic matrix (T_k^T) and a query vector (q) as shown in fig. 4.
Finally the euclidean distance between the $\hat{q}$ vector and all the documents of the corpus (D_k^T) can be calculated and the documents sorted by distance, the shortest distances between the query and documents are thus the most relevant documents.

4.3 Latent Direchlet Allocation

An other way to extract text which we're interested in is to use a statistical approach. Rather than using a Euclidean Space like in LSA and measuring distances between texts, the Latent Direchlet Allocation allows us to generate **probabilities** that a certain subject of interest (topic) is included in a text and a certain mixture of words is included within this topic.

Indeed, every text in our corpus is composed of several topics and each topic is composed of a certain number of words. What we want is to create an algorithm that will generate all the probabilities of a certain topic to be in a certain text and the probabilities that a numerous words will appear in a topic.

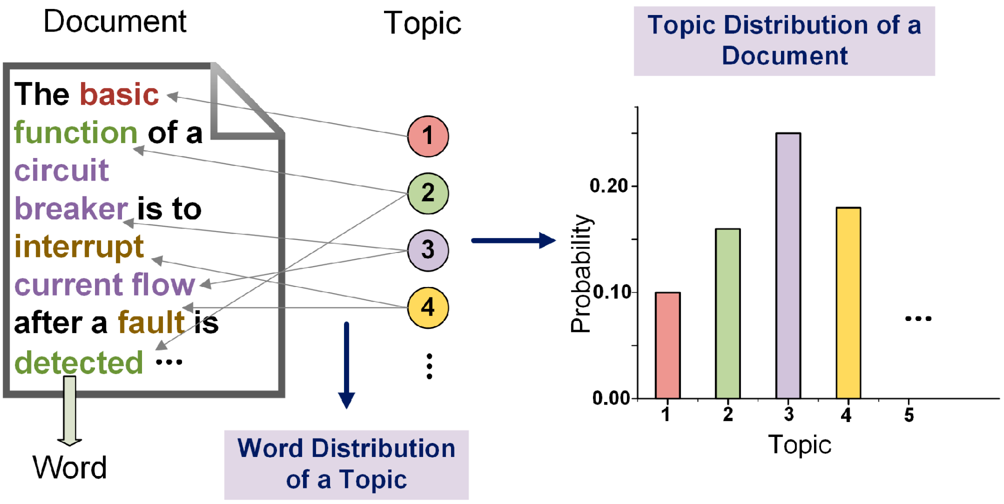

Document
Topic
Topic Distribution of a Document
The basic function of a circuit breaker is to interrupt current flow after a fault is detected ...
1
2
3
4
⋮
Word
Word Distribution of a Topic
Probability
0.00
0.10
0.20
1
2
3
4
5
...
Topic

Figure 5: LDA Summary

The first thing to do before using the LDA algorithm is to set the number of topics we want. This is represented by the value of k, our hyperparameter. We read in the literature that this hyperparameter is usually set between 20 or 30. We will set k equal to 20 here for convenience reasons but it can be modified regarding the will of the user.

As explained above, outputs of the LDA are probabilities. The first probability we will be talking about is the Beta probability. This probability generated by the LDA function represent the per-topic-per-word probability. For example, if we run the function we get:

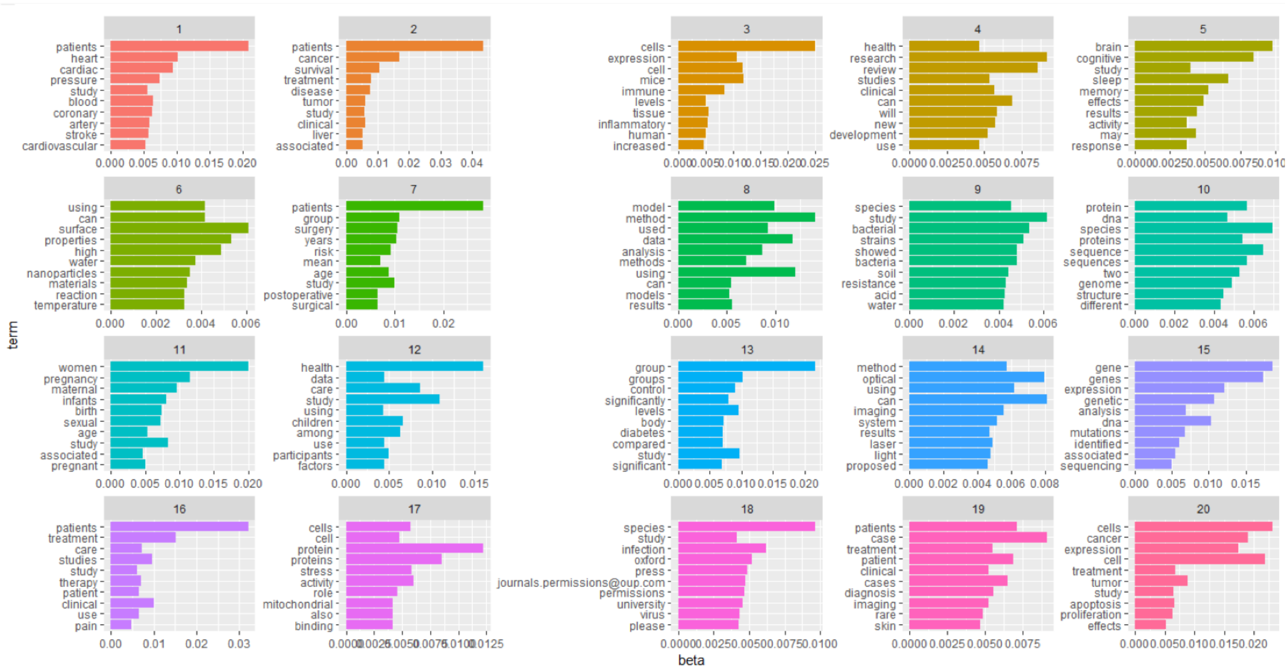
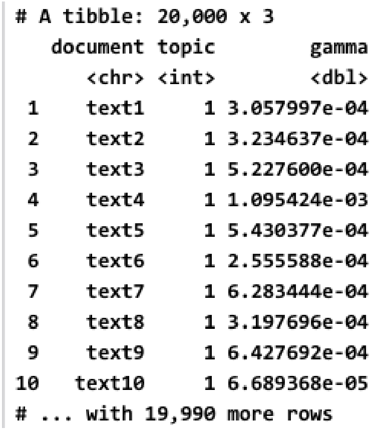


Figure 6: Beta probability for each topic

So, for each combination of word and topics, the model generates the probability of a term to be generated from a certain topic. In our example, the word "patient" has 0.02 percent chance to be present in the topic 1 but 0.04 probability to be in topic 2.

We can now extract for each word, the probability of being present in a topic. This also means that every topic is composed of a certain number of words that are more or less representative of a topic depending on the value of Beta.

Furthermore, LDA allows us to consider each document as a mixture of topics. To quantify this, we will use the Gamma probability. Indeed, we can generate the probability of a topic to be present in a certain document. So for our previous example we will get :



```
# A tibble: 20,000 x 3
   document topic        gamma
      <chr> <int>        <dbl>
 1    text1     1 3.057997e-04
 2    text2     1 3.234637e-04
 3    text3     1 5.227600e-04
 4    text4     1 1.095424e-03
 5    text5     1 5.430377e-04
 6    text6     1 2.555588e-04
 7    text7     1 6.283444e-04
 8    text8     1 3.197696e-04
 9    text9     1 6.427692e-04
10   text10     1 6.689368e-05
# ... with 19,990 more rows
```

Figure 7: Code insight for Gamma values

Here we can see that the model estimate that about 0.0003% of the words in document 1 were generated from topic 1.

The main idea of the project is to submit a "positive request" for a word we are looking for in the PubMed data base and a negative one for the opposite reason. To find the documents we are interested in, let's remind that each document is a mixture of topics and each topic is composed of a mixture of words.

Indeed, the algorithm will find the documents that have a higher proportion of topics having themselves a higher probability to contain the positive query word. For example, if the positive query is “cancer”, we want the algorithm to sort the document having a big proportion of topics talking about cancer i.e the Beta value for the word cancer in a topic will be high enough.

For the negative query, the algorithm will omit all documents composed of topics for which the negative query word is representative of the topic. Here, the representativeness of a word in a topic is set considering that the first 200 words with the biggest Beta values are representative for the chosen topic.

For example, if the negative query is “breast”, the algorithm will delete all documents with topics that have the word “breast” in their 200 first highest Beta values.

Next thing to consider is that we want to select all texts with the highest values for the parameters Beta ad Gamma. For that, we will first normalize Gamma because it is generally way bigger than Beta. Then, we will sum Beta and Gamma for all the selected texts and keep the ones with the highest values.

Finally, we have sorted all the corpus to select only documents having a high probability dealing with a positive request which the user can edit regarding its need.

Please note that the representativeness of a topic discussing above and setting to the 200 first highest Beta values can be discussed as well as the weighting between the Beta and Gamma values. For this code, these assumptions work pretty well but can be further investigated.

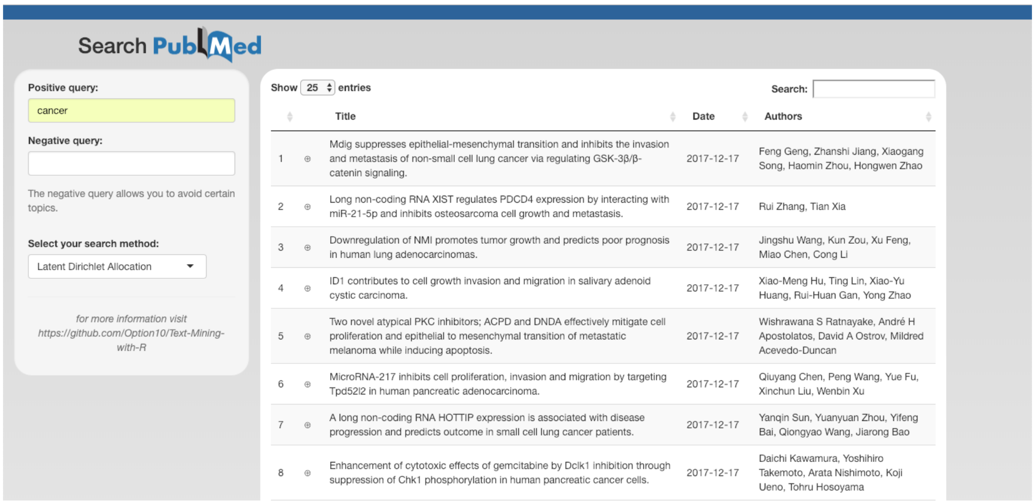
4.4 LSA or LDA

In practice, LSA is much faster to train than LDA. For this project, we have implemented both methods and leave the choice of method to the user. It'll therefore be possible to launch a quick but "coarse" search or a slower but accurate search. It is interesting to note that LSA focuses in the raw frequency of the queries words in the articles, while LDA try to spot the meaning/the idea represented by the queries words.

5 Output - User interface

The idea is to create a Google search-like application that will allow the user to find the Articles that correspond to his or her query. At this point of our project we are using the "Shiny" library from R-Cran to implement our query system in a visually appealing app.

Here is an example of how our user interface works : In the left part, you can choose your positive and negative query as well as the method you want to use. The right-hand side returns the list of articles in order of relevance. You can also see the title of the article, its authors and the date of publication. Here is a example with 'cancer' as positive query and LDA as method :



Positive query: cancer

Negative query:

The negative query allows you to avoid certain topics.

Select your search method: Latent Dirichlet Allocation

for more information visit https://github.com/Option10/Text-Mining-with-R

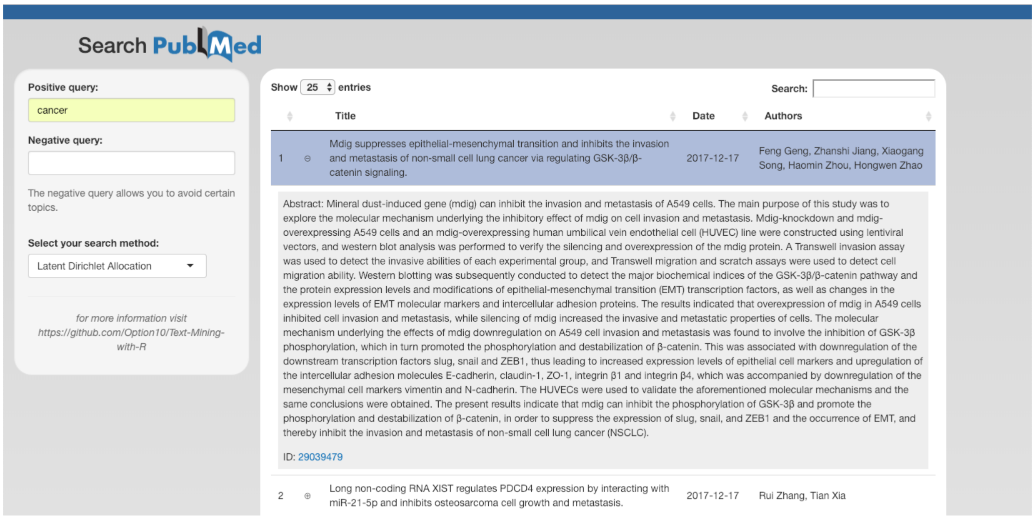
Show 25 entries

Search:

		Title	Date	Authors
1	⊕	Mdig suppresses epithelial-mesenchymal transition and inhibits the invasion and metastasis of non-small cell lung cancer via regulating GSK-3β/β-catenin signaling.	2017-12-17	Feng Geng, Zhanshi Jiang, Xiaogang Song, Haomin Zhou, Hongwen Zhao
2	⊕	Long non-coding RNA XIST regulates PDCD4 expression by interacting with miR-21-5p and inhibits osteosarcoma cell growth and metastasis.	2017-12-17	Rui Zhang, Tian Xia
3	⊕	Downregulation of NMI promotes tumor growth and predicts poor prognosis in human lung adenocarcinomas.	2017-12-17	Jingshu Wang, Kun Zou, Xu Feng, Miao Chen, Cong Li
4	⊕	ID1 contributes to cell growth invasion and migration in salivary adenoid cystic carcinoma.	2017-12-17	Xiao-Meng Hu, Ting Lin, Xiao-Yu Huang, Rui-Huan Gan, Yong Zhao
5	⊕	Two novel atypical PKC inhibitors; ACPD and DNDA effectively mitigate cell proliferation and epithelial to mesenchymal transition of metastatic melanoma while inducing apoptosis.	2017-12-17	Wishrawana S Ratnayake, André H Apostolatos, David A Ostrov, Mildred Acevedo-Duncan
6	⊕	MicroRNA-217 inhibits cell proliferation, invasion and migration by targeting Tpd52l2 in human pancreatic adenocarcinoma.	2017-12-17	Qiuyang Chen, Peng Wang, Yue Fu, Xinchun Liu, Wenbin Xu
7	⊕	A long non-coding RNA HOTTIP expression is associated with disease progression and predicts outcome in small cell lung cancer patients.	2017-12-17	Yanqin Sun, Yuanyuan Zhou, Yifeng Bai, Qiongyao Wang, Jiarong Bao
8	⊕	Enhancement of cytotoxic effects of gemcitabine by Dclk1 inhibition through suppression of Chk1 phosphorylation in human pancreatic cancer cells.	2017-12-17	Daichi Kawamura, Yoshihiro Takemoto, Arata Nishimoto, Koji Ueno, Tohru Hosoyama

Figure 8: Search with LDA

By clicking on the '+' button on the left of the title you can also read the entire abstract if you wish or also click on the ID of the article, which will take you to the corresponding pubmed page.



Positive query: cancer

Negative query:

The negative query allows you to avoid certain topics.

Select your search method: Latent Dirichlet Allocation

for more information visit https://github.com/Option10/Text-Mining-with-R

Show 25 entries

Search:

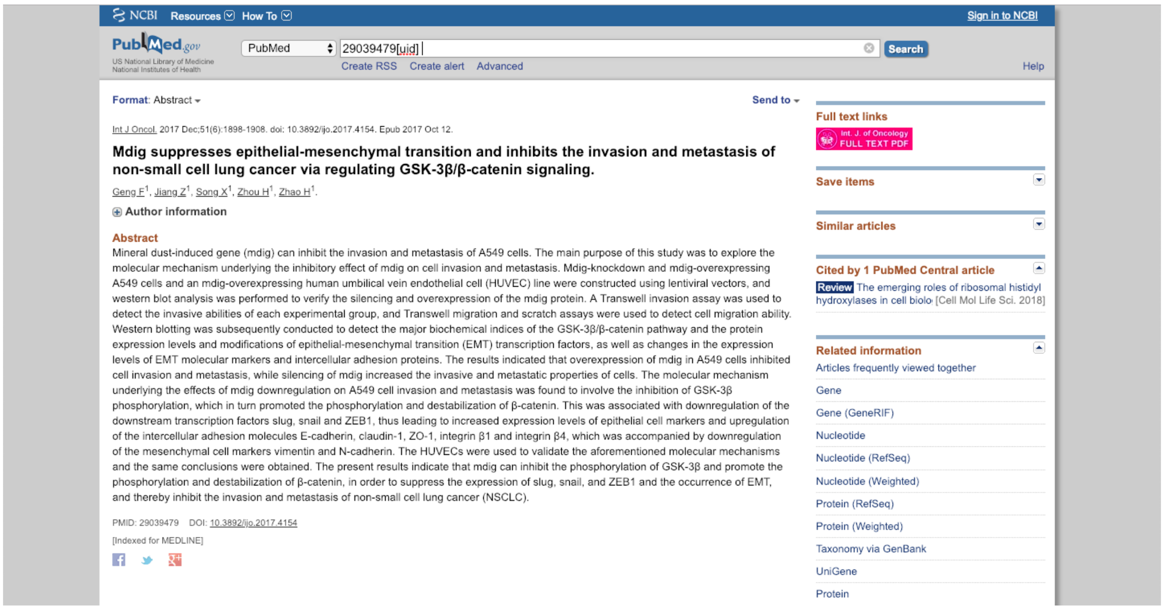
		Title	Date	Authors
1	⊖	Mdig suppresses epithelial-mesenchymal transition and inhibits the invasion and metastasis of non-small cell lung cancer via regulating GSK-3β/β-catenin signaling.	2017-12-17	Feng Geng, Zhanshi Jiang, Xiaogang Song, Haomin Zhou, Hongwen Zhao

Abstract: Mineral dust-induced gene (mdig) can inhibit the invasion and metastasis of A549 cells. The main purpose of this study was to explore the molecular mechanism underlying the inhibitory effect of mdig on cell invasion and metastasis. Mdig-knockdown and mdig-overexpressing A549 cells and an mdig-overexpressing human umbilical vein endothelial cell (HUVEC) line were constructed using lentiviral vectors, and western blot analysis was performed to verify the silencing and overexpression of the mdig protein. A Transwell invasion assay was used to detect the invasive abilities of each experimental group, and Transwell migration and scratch assays were used to detect cell migration ability. Western blotting was subsequently conducted to detect the major biochemical indices of the GSK-3β/β-catenin pathway and the protein expression levels and modifications of epithelial-mesenchymal transition (EMT) transcription factors, as well as changes in the expression levels of EMT molecular markers and intercellular adhesion proteins. The results indicated that overexpression of mdig in A549 cells inhibited cell invasion and metastasis, while silencing of mdig increased the invasive and metastatic properties of cells. The molecular mechanism underlying the effects of mdig downregulation on A549 cell invasion and metastasis was found to involve the inhibition of GSK-3β phosphorylation, which in turn promoted the phosphorylation and destabilization of β-catenin. This was associated with downregulation of the downstream transcription factors slug, snail and ZEB1, thus leading to increased expression levels of epithelial cell markers and upregulation of the intercellular adhesion molecules E-cadherin, claudin-1, ZO-1, integrin β1 and integrin β4, which was accompanied by downregulation of the mesenchymal cell markers vimentin and N-cadherin. The HUVECs were used to validate the aforementioned molecular mechanisms and the same conclusions were obtained. The present results indicate that mdig can inhibit the phosphorylation of GSK-3β and promote the phosphorylation and destabilization of β-catenin, in order to suppress the expression of slug, snail, and ZEB1 and the occurrence of EMT, and thereby inhibit the invasion and metastasis of non-small cell lung cancer (NSCLC).

ID: 29039479

		Title	Date	Authors
2	⊕	Long non-coding RNA XIST regulates PDCD4 expression by interacting with miR-21-5p and inhibits osteosarcoma cell growth and metastasis.	2017-12-17	Rui Zhang, Tian Xia

Figure 9: Advance Search



29039479[uid]

Int J Oncol. 2017 Dec;51(6):1898-1908. doi: 10.3892/ijo.2017.4154. Epub 2017 Oct 12.

Mdig suppresses epithelial-mesenchymal transition and inhibits the invasion and metastasis of non-small cell lung cancer via regulating GSK-3β/β-catenin signaling.

Geng F[1], Jiang Z[1], Song X[1], Zhou H[1], Zhao H[1].

⊕ Author information

Abstract

Mineral dust-induced gene (mdig) can inhibit the invasion and metastasis of A549 cells. The main purpose of this study was to explore the molecular mechanism underlying the inhibitory effect of mdig on cell invasion and metastasis. Mdig-knockdown and mdig-overexpressing A549 cells and an mdig-overexpressing human umbilical vein endothelial cell (HUVEC) line were constructed using lentiviral vectors, and western blot analysis was performed to verify the silencing and overexpression of the mdig protein. A Transwell invasion assay was used to detect the invasive abilities of each experimental group, and Transwell migration and scratch assays were used to detect cell migration ability. Western blotting was subsequently conducted to detect the major biochemical indices of the GSK-3β/β-catenin pathway and the protein expression levels and modifications of epithelial-mesenchymal transition (EMT) transcription factors, as well as changes in the expression levels of EMT molecular markers and intercellular adhesion proteins. The results indicated that overexpression of mdig in A549 cells inhibited cell invasion and metastasis, while silencing of mdig increased the invasive and metastatic properties of cells. The molecular mechanism underlying the effects of mdig downregulation on A549 cell invasion and metastasis was found to involve the inhibition of GSK-3β phosphorylation, which in turn promoted the phosphorylation and destabilization of β-catenin. This was associated with downregulation of the downstream transcription factors slug, snail and ZEB1, thus leading to increased expression levels of epithelial cell markers and upregulation of the intercellular adhesion molecules E-cadherin, claudin-1, ZO-1, integrin β1 and integrin β4, which was accompanied by downregulation of the mesenchymal cell markers vimentin and N-cadherin. The HUVECs were used to validate the aforementioned molecular mechanisms and the same conclusions were obtained. The present results indicate that mdig can inhibit the phosphorylation of GSK-3β and promote the phosphorylation and destabilization of β-catenin, in order to suppress the expression of slug, snail, and ZEB1 and the occurrence of EMT, and thereby inhibit the invasion and metastasis of non-small cell lung cancer (NSCLC).

PMID: 29039479 DOI: 10.3892/ijo.2017.4154

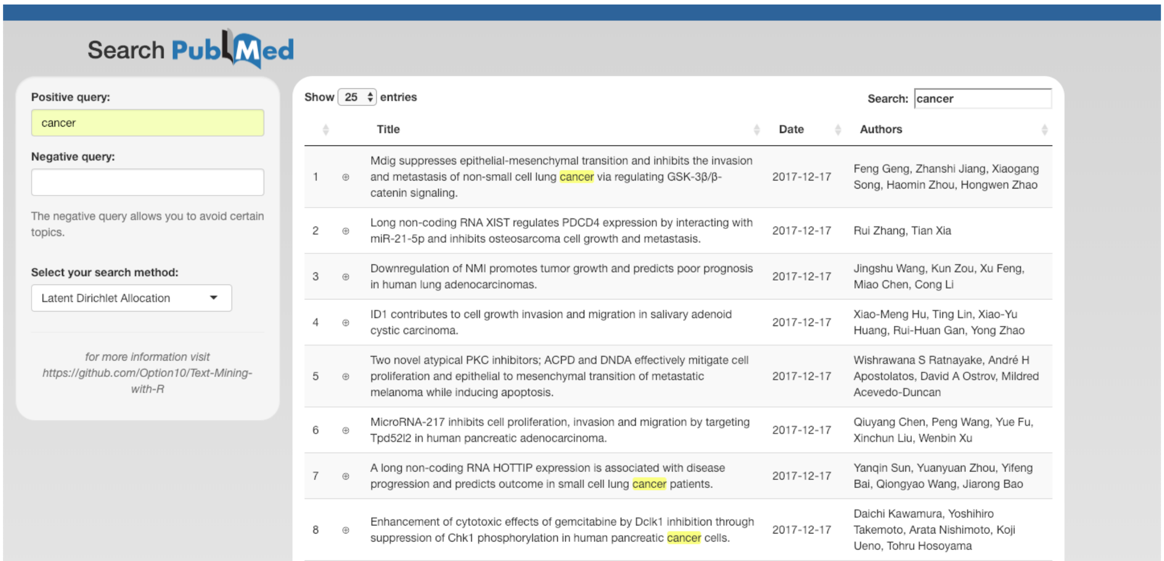
[Indexed for MEDLINE]

Cited by 1 PubMed Central article

Review The emerging roles of ribosomal histidyl hydroxylases in cell biolo [Cell Mol Life Sci. 2018]

Figure 10: Advance Search

In this interface it is also possible to search in the titles where a certain word appears by entering it in the box located at the top right. This highlights the words and allows you to quickly spot their occurrence.



Positive query: cancer

Negative query:

The negative query allows you to avoid certain topics.

Select your search method: Latent Dirichlet Allocation

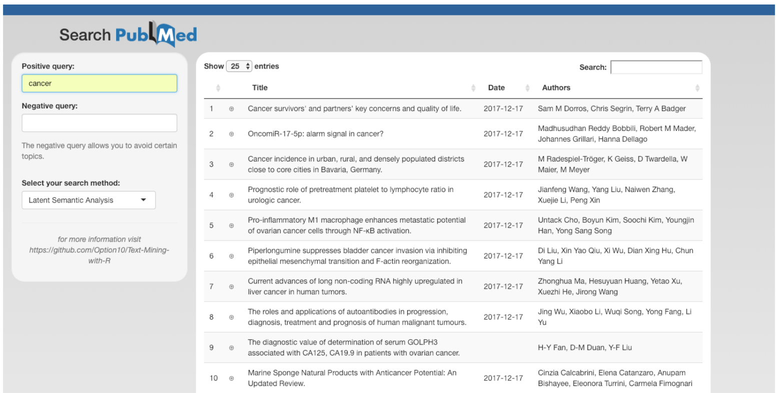
for more information visit https://github.com/Option10/Text-Mining-with-R

Search: cancer

		Title	Date	Authors
1	⊕	Mdig suppresses epithelial-mesenchymal transition and inhibits the invasion and metastasis of non-small cell lung cancer via regulating GSK-3β/β-catenin signaling.	2017-12-17	Feng Geng, Zhanshi Jiang, Xiaogang Song, Haomin Zhou, Hongwen Zhao
2	⊕	Long non-coding RNA XIST regulates PDCD4 expression by interacting with miR-21-5p and inhibits osteosarcoma cell growth and metastasis.	2017-12-17	Rui Zhang, Tian Xia
3	⊕	Downregulation of NMI promotes tumor growth and predicts poor prognosis in human lung adenocarcinomas.	2017-12-17	Jingshu Wang, Kun Zou, Xu Feng, Miao Chen, Cong Li
4	⊕	ID1 contributes to cell growth invasion and migration in salivary adenoid cystic carcinoma.	2017-12-17	Xiao-Meng Hu, Ting Lin, Xiao-Yu Huang, Rui-Huan Gan, Yong Zhao
5	⊕	Two novel atypical PKC inhibitors; ACPD and DNDA effectively mitigate cell proliferation and epithelial to mesenchymal transition of metastatic melanoma while inducing apoptosis.	2017-12-17	Wishrawana S Ratnayake, André H Apostolatos, David A Ostrov, Mildred Acevedo-Duncan
6	⊕	MicroRNA-217 inhibits cell proliferation, invasion and migration by targeting Tpd52l2 in human pancreatic adenocarcinoma.	2017-12-17	Qiuyang Chen, Peng Wang, Yue Fu, Xinchun Liu, Wenbin Xu
7	⊕	A long non-coding RNA HOTTIP expression is associated with disease progression and predicts outcome in small cell lung cancer patients.	2017-12-17	Yanqin Sun, Yuanyuan Zhou, Yifeng Bai, Qiongyao Wang, Jiarong Bao
8	⊕	Enhancement of cytotoxic effects of gemcitabine by Dclk1 inhibition through suppression of Chk1 phosphorylation in human pancreatic cancer cells.	2017-12-17	Daichi Kawamura, Yoshihiro Takemoto, Arata Nishimoto, Koji Ueno, Tohru Hosoyama

Figure 11: LDA Search with CTRL-F

It works exactly the same way with LSA as method. It is interesting to note that the results are not exactly identical.



Search PubMed

Positive query: cancer

Negative query:

The negative query allows you to avoid certain topics.

Select your search method: Latent Semantic Analysis

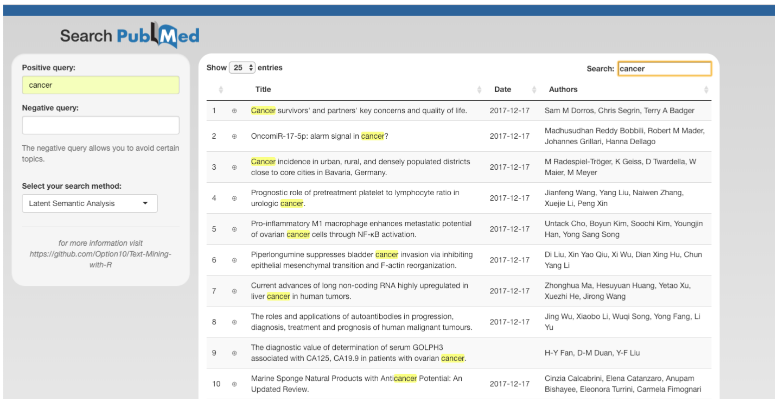
for more information visit https://github.com/Option10/Text-Mining-with-R

Show 25 entries

Search:

		Title	Date	Authors
1	⊕	Cancer survivors' and partners' key concerns and quality of life.	2017-12-17	Sam M Dorros, Chris Segrin, Terry A Badger
2	⊕	OncomiR-17-5p: alarm signal in cancer?	2017-12-17	Madhusudhan Reddy Bobbili, Robert M Mader, Johannes Grillari, Hanna Dellago
3	⊕	Cancer incidence in urban, rural, and densely populated districts close to core cities in Bavaria, Germany.	2017-12-17	M Radespiel-Tröger, K Geiss, D Twardella, W Maier, M Meyer
4	⊕	Prognostic role of pretreatment platelet to lymphocyte ratio in urologic cancer.	2017-12-17	Jianfeng Wang, Yang Liu, Naiwen Zhang, Xuejie Li, Peng Xin
5	⊕	Pro-inflammatory M1 macrophage enhances metastatic potential of ovarian cancer cells through NF-κB activation.	2017-12-17	Untack Cho, Boyun Kim, Soochi Kim, Youngjin Han, Yong Sang Song
6	⊕	Piperlongumine suppresses bladder cancer invasion via inhibiting epithelial mesenchymal transition and F-actin reorganization.	2017-12-17	Di Liu, Xin Yao Qiu, Xi Wu, Dian Xing Hu, Chun Yang Li
7	⊕	Current advances of long non-coding RNA highly upregulated in liver cancer in human tumors.	2017-12-17	Zhonghua Ma, Hesuyuan Huang, Yetao Xu, Xuezhi He, Jirong Wang
8	⊕	The roles and applications of autoantibodies in progression, diagnosis, treatment and prognosis of human malignant tumours.	2017-12-17	Jing Wu, Xiaobo Li, Wuqi Song, Yong Fang, Li Yu
9	⊕	The diagnostic value of determination of serum GOLPH3 associated with CA125, CA19.9 in patients with ovarian cancer.		H-Y Fan, D-M Duan, Y-F Liu
10	⊕	Marine Sponge Natural Products with Anticancer Potential: An Updated Review.	2017-12-17	Cinzia Calcabrini, Elena Catanzaro, Anupam Bishayee, Eleonora Turrini, Carmela Fimognari

Figure 12: Search with LSA



Search PubMed

Positive query: cancer

Negative query:

The negative query allows you to avoid certain topics.

Select your search method: Latent Semantic Analysis

for more information visit https://github.com/Option10/Text-Mining-with-R

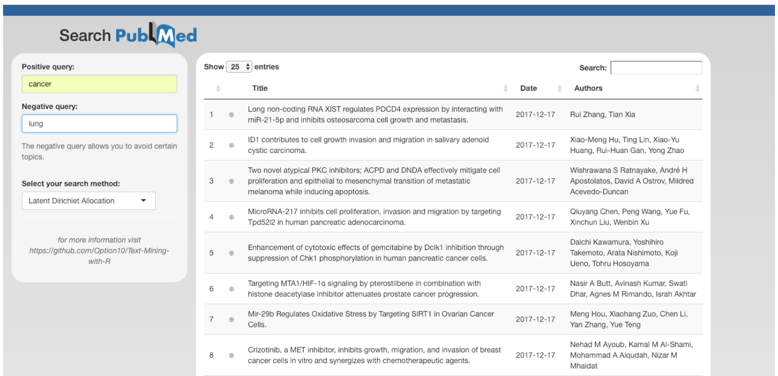
Show 25 entries

Search: cancer

		Title	Date	Authors
1	⊕	Cancer survivors' and partners' key concerns and quality of life.	2017-12-17	Sam M Dorros, Chris Segrin, Terry A Badger
2	⊕	OncomiR-17-5p: alarm signal in cancer?	2017-12-17	Madhusudhan Reddy Bobbili, Robert M Mader, Johannes Grillari, Hanna Dellago
3	⊕	Cancer incidence in urban, rural, and densely populated districts close to core cities in Bavaria, Germany.	2017-12-17	M Radespiel-Tröger, K Geiss, D Twardella, W Maier, M Meyer
4	⊕	Prognostic role of pretreatment platelet to lymphocyte ratio in urologic cancer.	2017-12-17	Jianfeng Wang, Yang Liu, Naiwen Zhang, Xuejie Li, Peng Xin
5	⊕	Pro-inflammatory M1 macrophage enhances metastatic potential of ovarian cancer cells through NF-κB activation.	2017-12-17	Untack Cho, Boyun Kim, Soochi Kim, Youngjin Han, Yong Sang Song
6	⊕	Piperlongumine suppresses bladder cancer invasion via inhibiting epithelial mesenchymal transition and F-actin reorganization.	2017-12-17	Di Liu, Xin Yao Qiu, Xi Wu, Dian Xing Hu, Chun Yang Li
7	⊕	Current advances of long non-coding RNA highly upregulated in liver cancer in human tumors.	2017-12-17	Zhonghua Ma, Hesuyuan Huang, Yetao Xu, Xuezhi He, Jirong Wang
8	⊕	The roles and applications of autoantibodies in progression, diagnosis, treatment and prognosis of human malignant tumours.	2017-12-17	Jing Wu, Xiaobo Li, Wuqi Song, Yong Fang, Li Yu
9	⊕	The diagnostic value of determination of serum GOLPH3 associated with CA125, CA19.9 in patients with ovarian cancer.		H-Y Fan, D-M Duan, Y-F Liu
10	⊕	Marine Sponge Natural Products with Anticancer Potential: An Updated Review.	2017-12-17	Cinzia Calcabrini, Elena Catanzaro, Anupam Bishayee, Eleonora Turrini, Carmela Fimognari

Figure 13: LSA Search with CTRL-F

We have implemented the negative query as described earlier in this report. Here are the results using the word "lung" as a negative query. The first article is indeed deleted from our list because it explicitly talks about lung cancer.



Search PubMed

Positive query: cancer

Negative query: lung

The negative query allows you to avoid certain topics.

Select your search method: Latent Dirichlet Allocation

for more information visit https://github.com/Option10/Text-Mining-with-R

Show 25 entries

Search:

		Title	Date	Authors
1	⊕	Long non-coding RNA XIST regulates PDCD4 expression by interacting with miR-21-5p and inhibits osteosarcoma cell growth and metastasis.	2017-12-17	Rui Zhang, Tian Xia
2	⊕	ID1 contributes to cell growth invasion and migration in salivary adenoid cystic carcinoma.	2017-12-17	Xiao-Meng Hu, Ting Lin, Xiao-Yu Huang, Rui-Huan Gan, Yong Zhao
3	⊕	Two novel atypical PKC inhibitors; ACPD and DNDA effectively mitigate cell proliferation and epithelial to mesenchymal transition of metastatic melanoma while inducing apoptosis.	2017-12-17	Wishrawana S Ratnayake, André H Apostolatos, David A Ostrov, Mildred Acevedo-Duncan
4	⊕	MicroRNA-217 inhibits cell proliferation, invasion and migration by targeting Tpd52l2 in human pancreatic adenocarcinoma.	2017-12-17	Qiuyang Chen, Peng Wang, Yue Fu, Xinchun Liu, Wenbin Xu
5	⊕	Enhancement of cytotoxic effects of gemcitabine by Dclk1 inhibition through suppression of Chk1 phosphorylation in human pancreatic cancer cells.	2017-12-17	Daichi Kawamura, Yoshihiro Takemoto, Arata Nishimoto, Koji Ueno, Tohru Hosoyama
6	⊕	Targeting MTA1/HIF-1α signaling by pterostilbene in combination with histone deacetylase inhibitor attenuates prostate cancer progression.	2017-12-17	Nasir A Butt, Avinash Kumar, Swati Dhar, Agnes M Rimando, Israh Akhtar
7	⊕	Mir-29b Regulates Oxidative Stress by Targeting SIRT1 in Ovarian Cancer Cells.	2017-12-17	Meng Hou, Xiaohang Zuo, Chen Li, Yan Zhang, Yue Teng
8	⊕	Crizotinib, a MET inhibitor, inhibits growth, migration, and invasion of breast cancer cells in vitro and synergizes with chemotherapeutic agents.	2017-12-17	Nehad M Ayoub, Kamal M Al-Shami, Mohammad A Alqudah, Nizar M Mhaidat

Figure 14: Negative Search

6 Conclusion

This project is a result of a four-month period of work, and we are aware of its limits.

The user interface is very intuitive, and the user can select the algorithm he wants for its research. Both LSA and LDA work pretty well, they both have their pros and cons.

LSA will be preferentially used when the user knows what kind of documents he wants. The user can also put several words in his query's, something LDA can't do.

On the other hand, LDA will be very useful when users want a more prospective approach. Furthermore, we have observed that LSA algorithm deals more with the frequencies of words and select documents with a very high frequency for the researched word. LDA in contrast will select documents dealing mainly with all the "concepts" behind the query. If the positive query is "cancer", LDA will select documents dealing with cancer but also tumor, metastasis etc.

Finally, one thing we can discuss is the weighting of the beta and gamma values for the LDA and the Euclidean distance calculations in LSA. Indeed, the weighting we used is a basic normalization of gamma values and a sum between those new gammas with beta values, but others weighting methods may be used.

For LSA, we've chose to measure distance between our query's and document using an Euclidean distance. We justify this choice because of its simplicity, but other methods may be used.

References

[Bao] Bao, H. T. Latent Semantic Analysis and Topic Modeling: Roads to Text Meaning. page 16.

[2] Harris, D. (2016). What Is Text Analytics? We Analyze the Jargon.