MDPI

*Article*

# Forecasting Emerging Topics in Research and Development in Pharmaceutical Industry

**MahParsa** [1,†,‡],

[1] mahboobeh.parsapoor@mail.mcgill.ca
* Correspondence: mahboobeh.parsapoor@mail.mcgill.ca
† Current address: McGill University

Academic Editor: name
Version March 23, 2019 submitted to Journal Not Specified

**Keywords:** Bibliometrics Tools, Clustering Algorithms, Machine Learning Tools, Text Analysis, Topic Modeling, Topic Trends

## 1. Abstract

Predicting the future trends in topics can impact on research projects in the academy by helping grant agencies recognize emerging topics and assign future grants to scientists that are working on those topics. Identifying future trends in a topic can also help companies and start ups invest in emerging themes to change the game in their fields. This paper suggests a new strategy to identify developing drifts.

## 2. Introduction

Discovering research trends in pharmaceutical industry can directly influence governmental policies, healthcare services, educational organizations, research institutes. Thus, it is entirely necessary to predict what is the future trends in pharmaceutical industries. For example, discovering the future trends in biosensors [1] not only assist pharmaceutical and biomedical researchers in grasping information of emerging drifts in this field of bio electronics but also help institutions and grant agencies to assign future grants to scientists that are working on emerging topics [2]. Online tools such as "Google Trends" can be utilised to forecast the latest trending topics or determine (Figure 1) the popularity of research themes (Figure 5) in the different fields of science and technology, e.g., biosensors. They can provide a comparison interest to different terms such as "Dendritic cells" (blue line), "Biosensors" (yellow line) and "Cancer Treatment" (red line) Cell" see Figure 3. These tools cannot detect emerging topics by just analysing scholarly articles. Thus, it is necessary to use a method to analyze documents, discover topics and predict the future topics.

Text analysis methods includes techniques that are useful for processing text that are obtained from equipment logs, blogs, articles and and social media such as Facebook and twitter. Using text analysis we can extract useful information from text, remove unnecessary information from text and represent text using numerical values. In other words, text analysis tools help us to develop statistical tools. One of main class of text analysis is machine learning methods such as clustering algorithms,

[1] The term "biosensors", which firstly introduced by Griffin and Nelson, refers an analytical device which converts a biological response into an electrical signal [1]. A typical biosensor has two components, biological elements such as enzyme, antibody, cell, tissue and a physical transducer such as voltammetric, amperometric, conductometric, spectrophotometric


**Figure 1.** The number of interest of biosensors over time

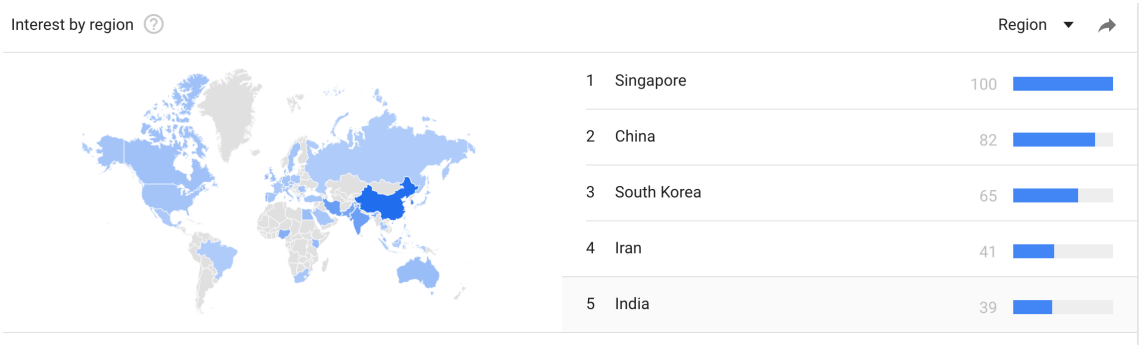


**Figure 2.** Different locations that the terms "biosensors" have been most noted during the specified time frame

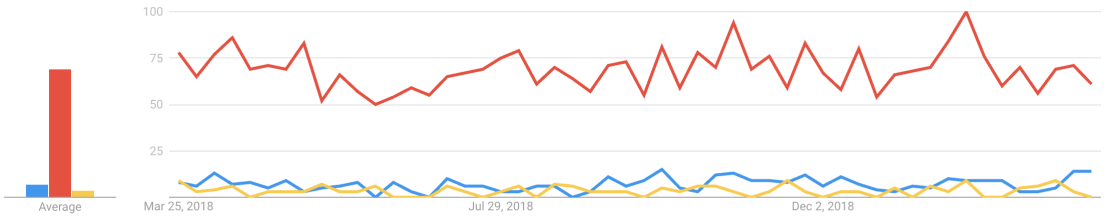


**Figure 3.** Different locations that the terms "Dendritic cells", "Biosensors" and "Cancer Treatment" Cell" have been most noted during the specified time frame

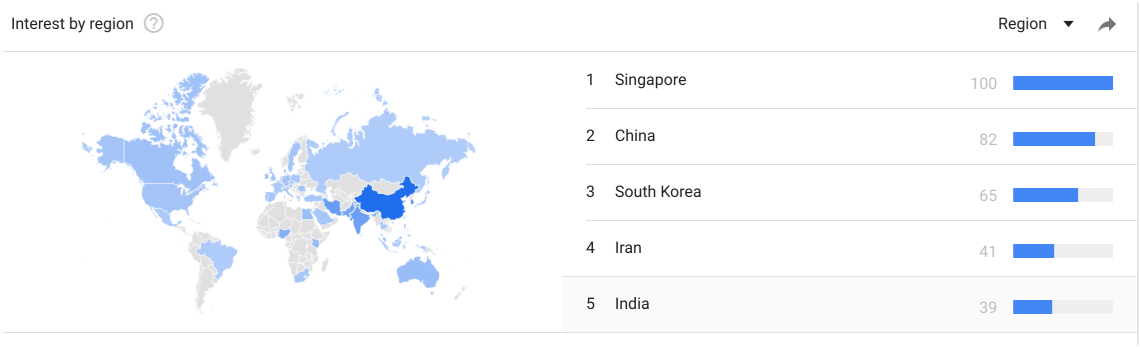


**Figure 4.** Different locations that the terms "biosensors" have been most noted during the specified time frame

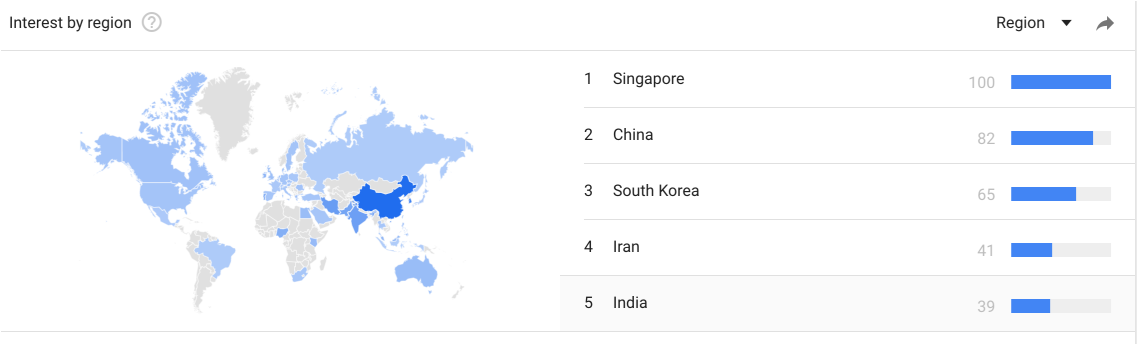


**Figure 5.** Different locations that the terms "biosensors" have been most noted during the specified time frame

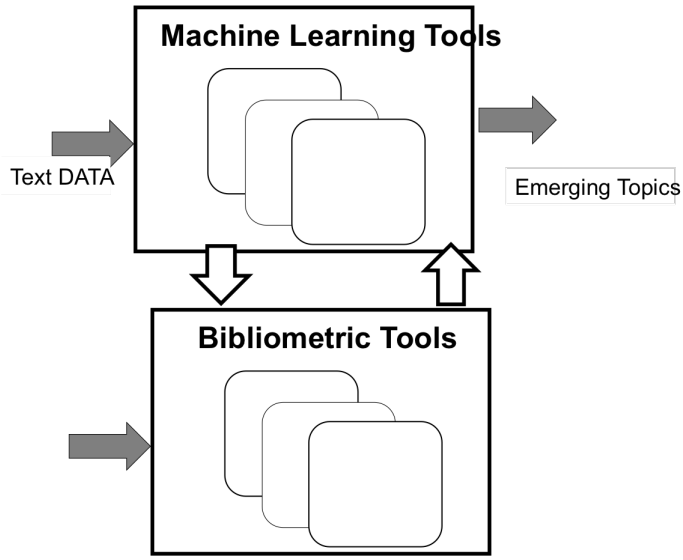


**Figure 6.** The general structure of framework to forecast emerging topics; The framework receives text documents (an abstract of publication) and discover emerging topics. Note that, MLT and BT modules utilize a bidirectional connections to exchange data and useful information

LDA and LSA. It should be noted that outputs of text analytical tools must be used with other machine learning tools to provide meaning full results. This paper introduces a framework that merges machine learning algorithms and bibliometrics to discover emerging topics and forecasts the latest trending of emerging topics. The framework automatically detects emerging topics from a set of text documents and estimate evolving topics. Figure 6 describes the framework with its two main modules that are named "Machine Learning Tools" ("MLT") and "Bibliometrics Tools"("BT").

## 3. Framework

The MLT module in the framework integrates various statistical-based and bio-inspired machine learning tools to do the following tasks.

1) Select useful terms
2) Find meaningful topics and assign text documents to highest frequency terms of topics
3) Categorise text documents
5) Predict bibilmetric indicators
6) Integrate data
7) Identify emerging topics
8) Predict emerging topic trend

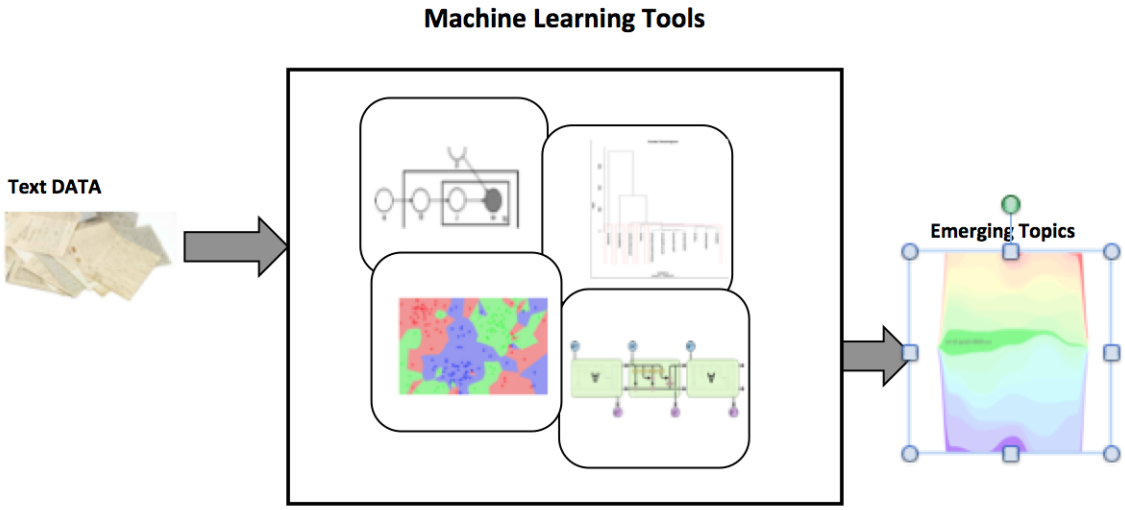


**Figure 7.** An example of the MLT module

This paper is focusing on the three first steps of the above tasks. The first phase is to select useful terms to gather documents from Blogs, Papers, websites. For example, we selected some keywords such as "biosensors", "Continuous glucose monitoring", "glucose measurement", "Hypoglycemia", "Insulin", "medical devices", "metabolic", "pharmacological target", "Protein", "real-time applications", "wearable sensors" to gather documents related to biosensors. At the second phase, we can run natural processing language tools and topic modeling tools (e.g., LDA) to find terms with highest frequency and topics that are hidden in documents. The output of LDA is topics and terms that are related to each topic. Thus, for each topic, we can determine one term with the highest probability of belonging to topics. The third step categorises text documents to most important topics using a clustering algorithm. The MLT module is an ensemble method [2], that integrates multiple algorithms (such as topic modelling, ensemble classifiers, clustering algorithms, multivariate time-series prediction models such as adaptive neuro fuzzy inference system (ANFIS) and Brain Emotional Learning inspired Model (BELiMs) [3] and [4], multivariate multiple regression) to analyse texts, detect emerging topics and determine the evolution of emerging topics.

For example, an MLT module (see Figure 7) can encompasses latent dirichlet allocation (LDA) as a topic modelling algorithm, hierarchical clustering clgorithm (HCA) as a clustering algorithm, K nearest neighbour (KNN) as a classifier, and long short-term memory (LSTM) [3] as a time-series prediction model) [5].

Another part of this framework is the BT (Bibliometrics Tools) module to analyse metadata of academic literature related to most important topics. For example,"bibliometrix" (an R package for bibliometric study) can interpret meta-data of scholarly articles related to topics to indicators (e.g., the number of publications) that would be sent to the MLT module to construct multivariate time series data sets.

Bibliometric tools (see Figure 8) implement quantitative analysis of academic literature to discover the impact of research fields, the impression of a set of researchers, or the impact of particular articles.

---

[2] Ensemble learning is the process to combine multiple learning models such as classifiers, prediction to solve a complex problem. The main advantage of an ensemble approach is to increase the accuracy

[3] LSTM has been inspired by human memory and learning systems and can learn about long-term dependencies of values of a variable in the past to predict future values of variable

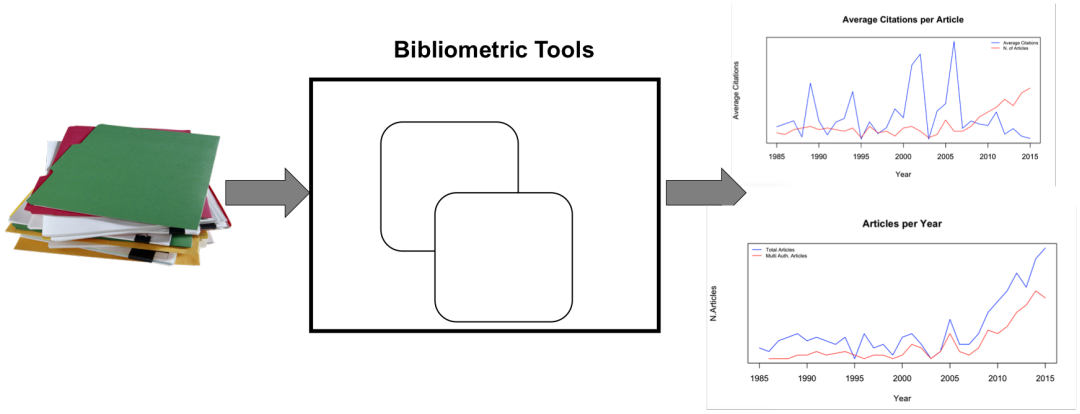


**Figure 8.** The BT module that encompasses bibliometrix

The performance and effectiveness of proposed framework would be evaluated on a variety of text datasets (e.g., Scopus); the obtained results would be compared with other methodologies that most are based on bibliometric tools.

## 4. Latent Dirichlet Allocation (LDA) of MLTs

This section explains Latent Dirichlet Allocation (LDA) as a well-known topic modelling. Topic modelling algorithms mostly are statistical methods and belongs to a group of algorithms that are named generative probabilistic modeling methods. Generative probabilistic modelling methods are are based on a probabilistic generative procedure to find find hidden variables underlying the observed data. Then they utilize a joint probability distribution over both the observed and hidden variables and compute the posterior distribution (conditional distribution) of the hidden variables given the observed variables. aim to determine a probabilistic model of a corpus that can assign high probability to members of the corpus (i.e unseen documents).

One well-known generative probabilistic topic modeling algorithm is Latent Dirichlet Allocation (LDA) which have been excellent performance in finding topics in a collection of documremts. In LDA, Latent variables are topics, while apparent variables are words. LDA receives words as an input vector and generates topics which are probability distribution over words based on a generative process. LDA uses a joint probability distribution over both the observed and hidden random variables and compute the posterior distribution (conditional distribution) of the hidden variables given the observed variables. The fundamental assumption of LDA is that documents can be assigned to multiple topics. Another assumption is that topics are hidden variables and words in documents are apparent variables. Thus, LDA performs a generative process by receiving words as an input vector to provide topics which are probability distribution over words. In addition to the above assumption, the following assumption should be considered:

1. The number of topics,$k$, is known.
2. The size of vocabulary,$V$, is known.
3. Each topic can be defined as a Dirichlet distribution over words.
4. Each document has two probabilistic distributions, the first one is a Dirichlet distribution over topics to choose a topic proportion. The second one is mutlinomial disterbution that be used to choose a word for the document.
5. The number of words, $N$ is choosen using a poisson distribution as $\sim (\xi)$.
6. Choose $\theta \sim (\alpha)$.
7. For each of the $N$ words $w_n$:

   (a) Choose a topic $z_n \sim (\theta)$.

(b) Choose a word $w_n$ from $(w_n z_n, \beta)$, a multinomial probability conditioned on the topic $z_n$.

It should be noted that, LDA considers the dimensional of the topic variable $z$ and the dimensional of the Dirichlet distribution $k$ are known; moreover, LDA uses a matrix as $\beta$ with $k \times V$ where $\beta_{ij} = p(w^j = 1 z^i = 1)$, the matrix is named as word probability.

A $k$-dimensional Dirichlet random variable $\theta$ can take values in the $(k-1)$-simplex (a $k$-vector $\theta$ lies in the $(k-1)$-simplex if $\theta_i \geq 0$, $\sum_{i=1}^{k} \theta_i = 1$), and has the following probability density on this simplex:

$$(\theta\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}, \tag{1}$$

where the parameter $\alpha$ is a $k$-vector with components $\alpha_i > 0$, and where $\Gamma(x)$ is the Gamma function. The Dirichlet is a convenient distribution on the simplex — it is in the exponential family, has finite dimensional sufficient statistics, and is conjugate to the multinomial distribution. In inference-and-estimation, these properties will facilitate the development of inference and parameter estimation algorithms for LDA.

Given the parameters $\alpha$ and $\beta$, the joint distribution of a topic mixture $\theta$, a set of $N$ topics , and a set of $N$ words is given by:

$$(\theta, , \alpha, \beta) = (\theta\alpha) \prod_{n=1}^{N} (z_n\theta)(w_n z_n, \beta), \tag{2}$$

where $(z_n\theta)$ is simply $\theta_i$ for the unique $i$ such that $z_n^i = 1$. Integrating over $\theta$ and summing over $z$, we obtain the marginal distribution of a document:

$$p(\alpha, \beta) = \int p(\theta\alpha) \left( \prod_{n=1}^{N} \sum_{z_n} p(z_n\theta) p(w_n z_n, \beta) \right) d\theta. \tag{3}$$

Finally, taking the product of the marginal probabilities of single documents, we obtain the probability of a corpus:

$$p(\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d\alpha) \left( \prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}\theta_d) p(w_{dn} z_{dn}, \beta) \right) d\theta_d. \tag{4}$$

## 5. Clustering Algorithms of MLTs

The main aim of clustering methods is to categorize unstructured documents into a list of meaningful groups [6]. Clustering algorithms can divide data object into separated groups [6]. Each group encompasses similar objects; while objects between two groups must be different. In other words, similar objects could be partitioned to a cluster, while patterns that are not alike to each other should be distributed in separated clusters. The similarity measurement is defined using a clustering algorithm A clustering algorithm needs a clear, implementable similarity rule. Defining different rules lead to developing a different variation of clustering algorithms. Two main categories of clustering algorithms are partitional clustering methods and hierarchical clustering methods. Partitional clustering algorithm partitions data between separated groups which have not any common members, so each object only belongs to one group. In contrast, hierarchical clustering provides a nested tree-like framework to partition data. It should be noted that a member in lower clusters could be considered as a member of upper clusters in the hierarchical structure. Clustering algorithms [7] as a group of unsupervised learning algorithm have applications in text analysis and organizing documents.

### 5.0.1. **K-means Clustering Algorithm**

K-means clustering is an iterative unsupervised machine learning and has the goal to partition objects into K separate groups. In the following, I have explained the steps of a K-means algorithm.

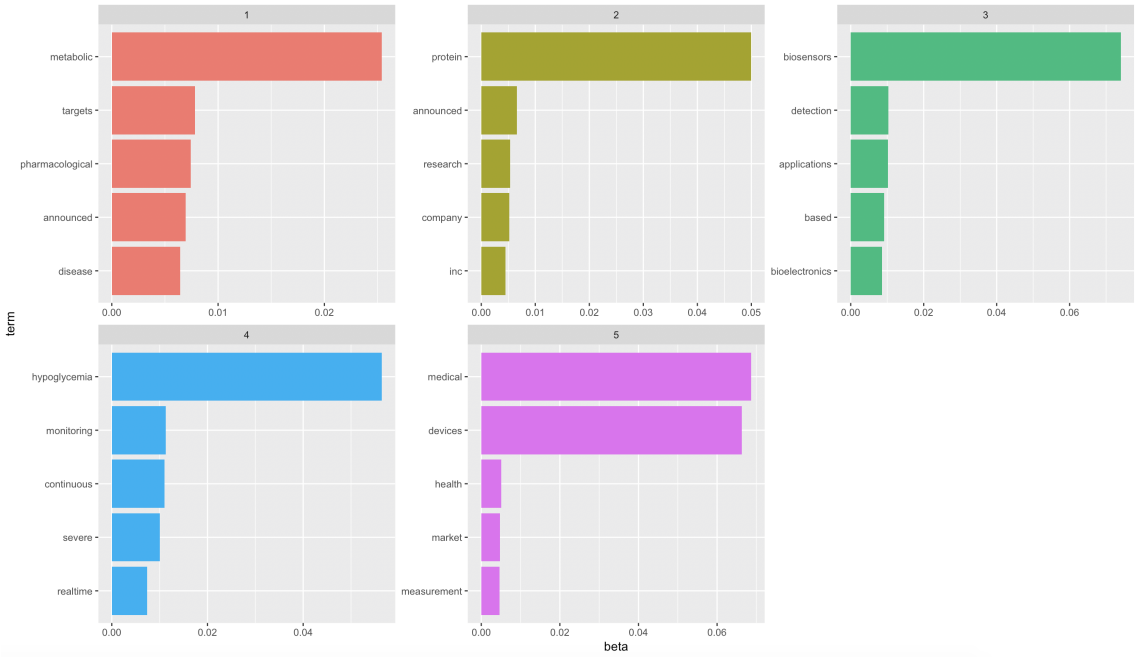


**Figure 9.** Generated five topics for a corpus with documents related to bio-sensors. The following terms, "Metabolic", "Protein", "Medical", "biosensors" and "Hypoglycemia" have highest probability for topics.

1. Determination Centroids which means with K cluster centres randomly selected.
2. Partitioning which means for each pattern cluster centre is found, and that pattern is considered from that cluster.
3. Replacement: Each centroid is replaced by average of data points in its partition
4. Iteration: The following step is repeated until convergence criteria is satisfied.

Step1: $x_i = (x_{i1}, ...x_{ip})$:
Step2: If centroids are $m_1, m_2, ...m_k$, and partitions are
$c_1, c_2, ...c_k$, then one can show that K-means converges to a *local* minimum of

$$\sum_{k=1}^{K} \sum_{i \in c_k} ||x_i - m_k||^2 \qquad \text{Euclidean distance}$$

## 6. Experimental Results

The experimental results of this section are focusing on text analysis of two types of text data sets. The first data set is realted to 6 documents that are gathered using the following keywords: "biosensors", "Continuous glucose monitoring", "glucose measurement", "Hypoglycemia", "Insulin", "Medical devices", "Metabolic", "Pharmacological target", "Protein", "Real-time Applications", "Wearable sensors". The second data set is documents that have been extracted by using "Dendritic cells" as keywords.

### *6.1. Topic Modeling Results*

### *6.2. First Data Set*

Figures 9, 10 and 11 show different 5, 10 and 15 topics have been generated by using LDA models.
As it can be observed from the above figures, topic modeling algorithms provide us information about words with highest probability. The information can be used to predict the future topics.

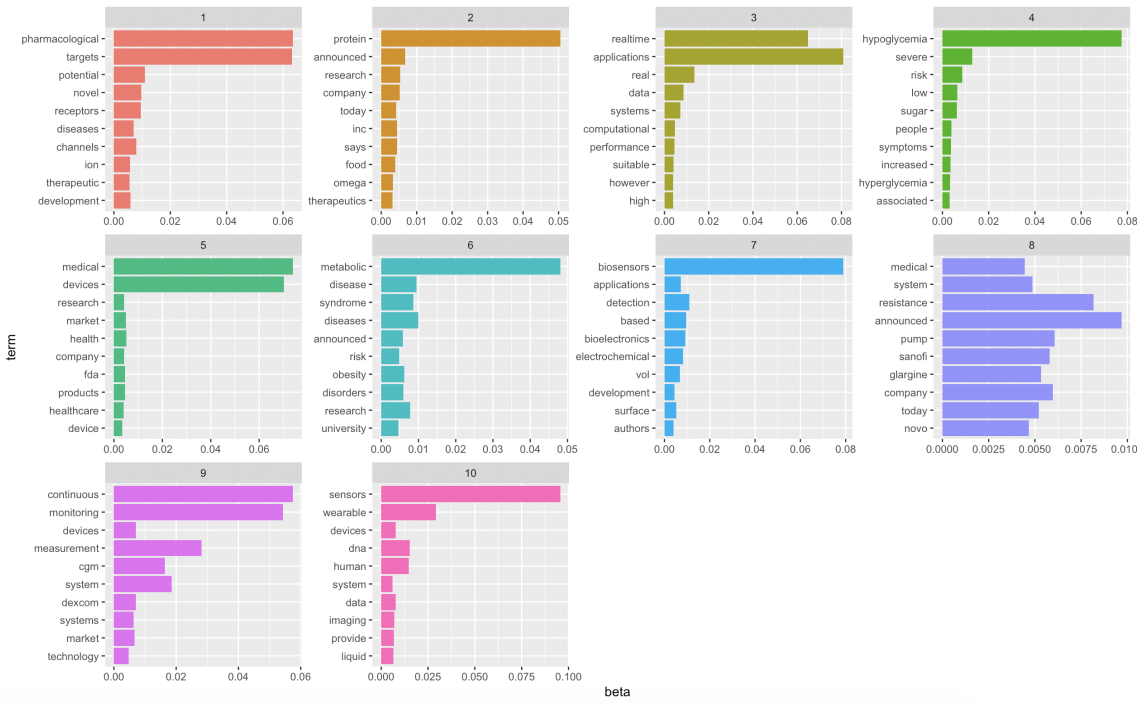


**Figure 10.** Generated ten topics for a corpus with documents related to bio-sensors

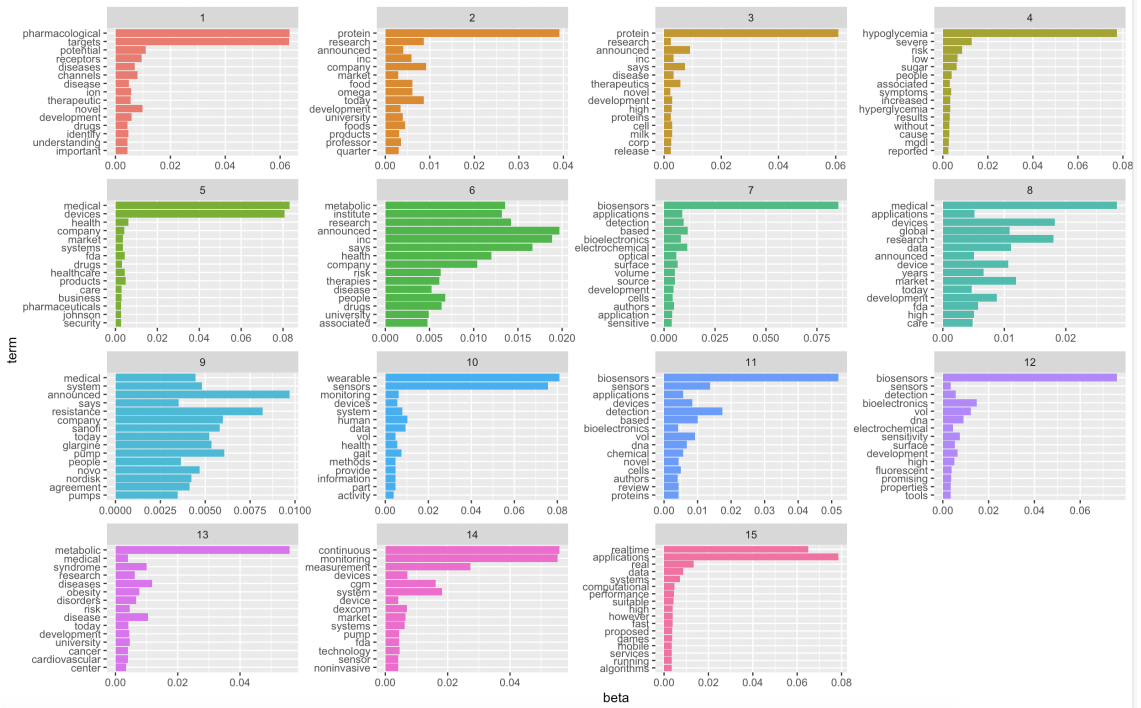


**Figure 11.** Generated fifteen topics for a corpus with documents related to bio-sensors

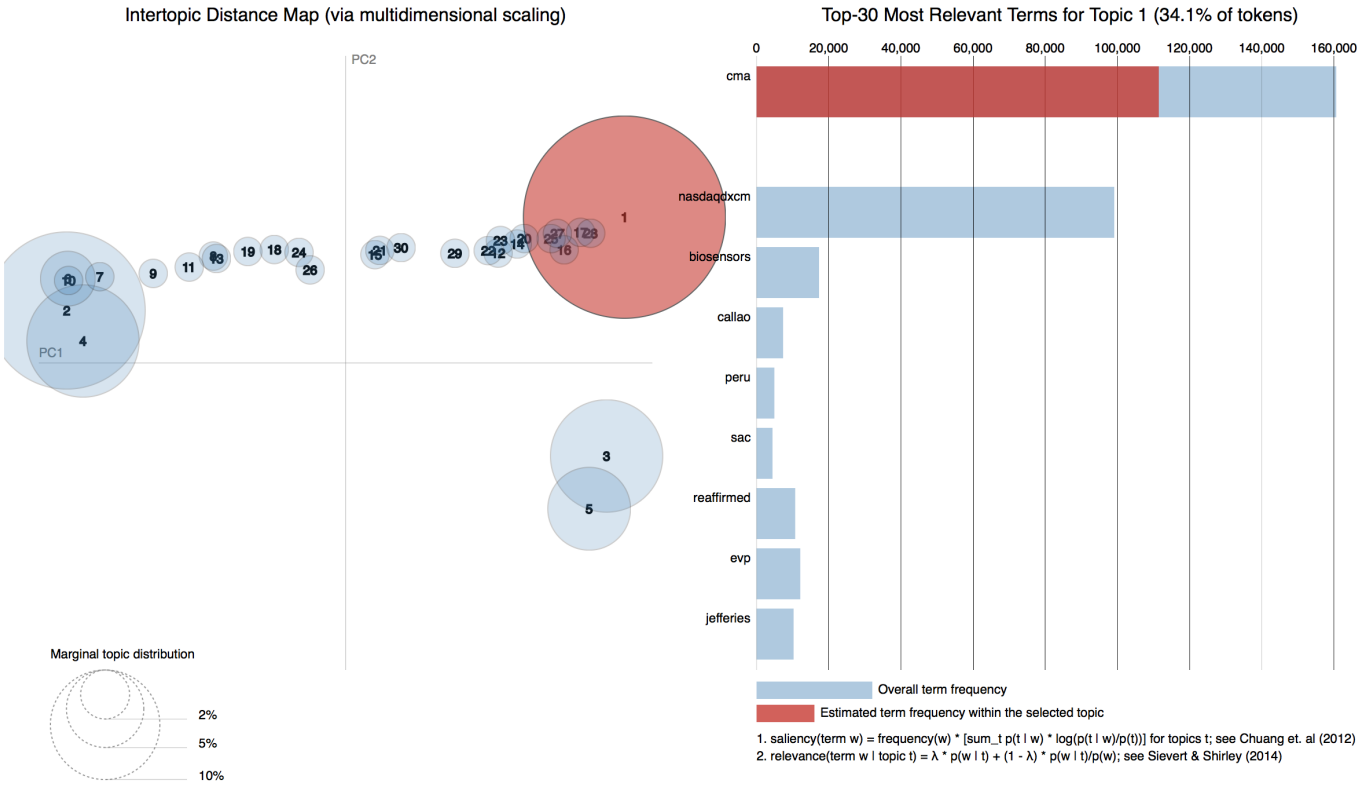


**Figure 12.** A visualization of 30 topics selected by LDA

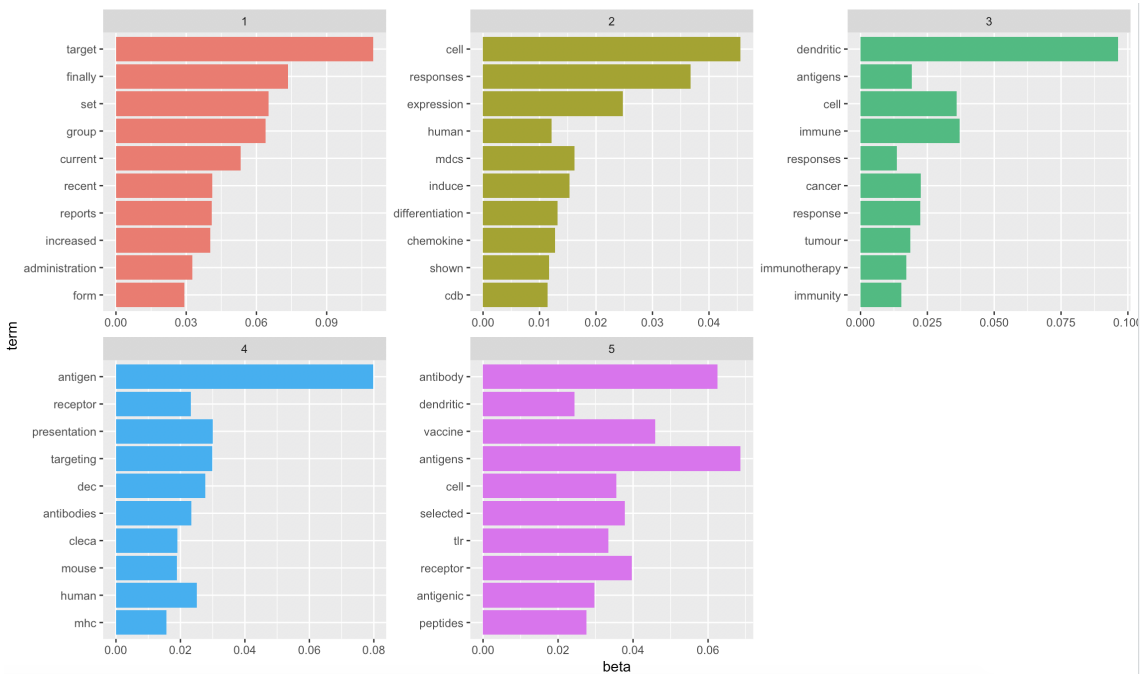


**Figure 13.** Generated five topics for a corpus with documents related to Dentritic Cells

Figure 12presents a nice visualization using "LDAvis" [8]of 30 topics that obtained from our documents.

### *6.3. second Data set*

The second example of this part is related to documents that are focusing on Dendritic cells [4].
Figures **??**, **??** and **??** show different 5, 10 and 15 topics have been generated by using LDA models.

### *6.4. Clustering Results*

This section show how clustering algorithms can partition documents to clusters.It should be mentioned we apply clustering algorithms on the documents and also the obtained results of topic modeling.

#### 6.4.1. The First Data Set

Figures 16, 18 and **??** show the hierarchical clusters obtained from applying hirachical clustering algorithms on the documents that are related to biosensors.
Figures 19, 20,21, 22 and 23 show the hierarchical clusters obtained from applying K-means clustering algorithms on the documents that are related to biosensors.

#### 6.4.2. The Second Data Set

This section show how clustering algorithms can partition documents related to Dendritic cells. Figures 24, 25 and 26 show the hierarchical clusters obtained from applying hierarchical clustering algorithms on those documents.

---

[4] Dendritic cells (DCs) can be classified as antigen-presenting cells (accessory cells) of the mammalian immune system. These cells are responsible for processing antigen material and offering it on the cell surface to the T cells of the immune system. In other words, DSc can play a role to exchange message be


**Figure 14.** Generated ten topics for a corpus with documents related to Dentritic Cells

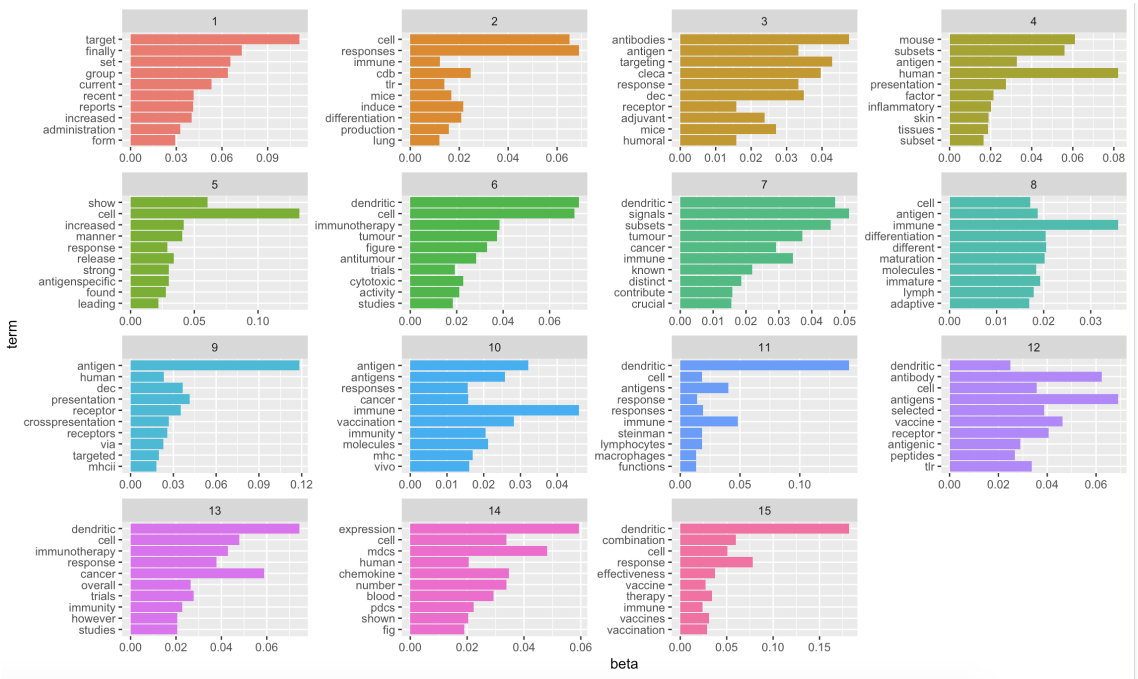


**Figure 15.** Generated 15 topics for a corpus with documents related to Dentritic Cells

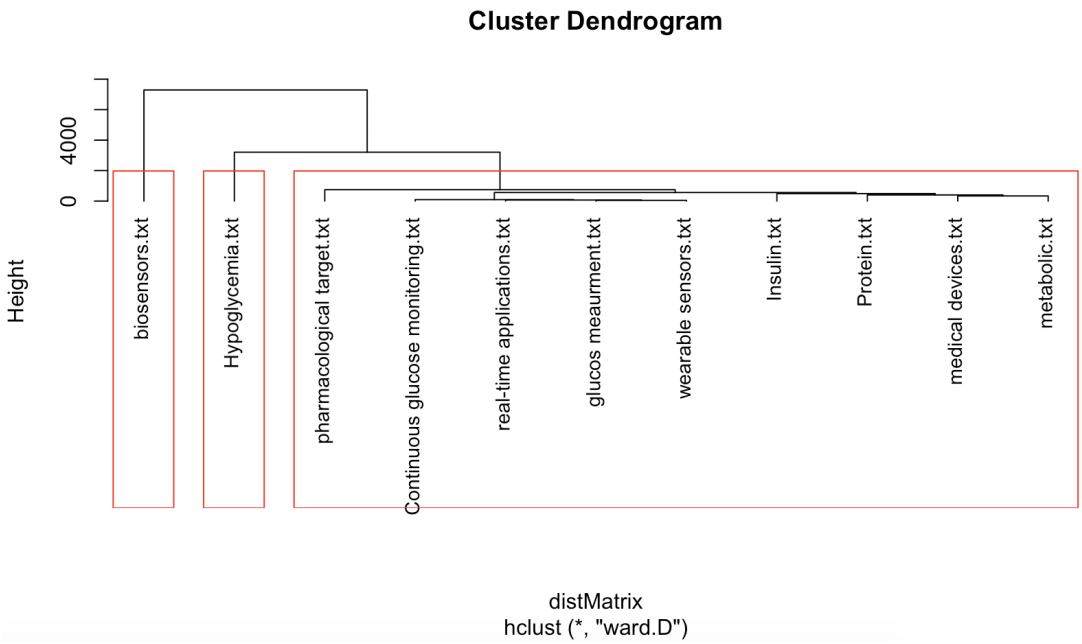


**Figure 16.** Using hierarchical clustering algorithm to have 3 cluster for 18 documents

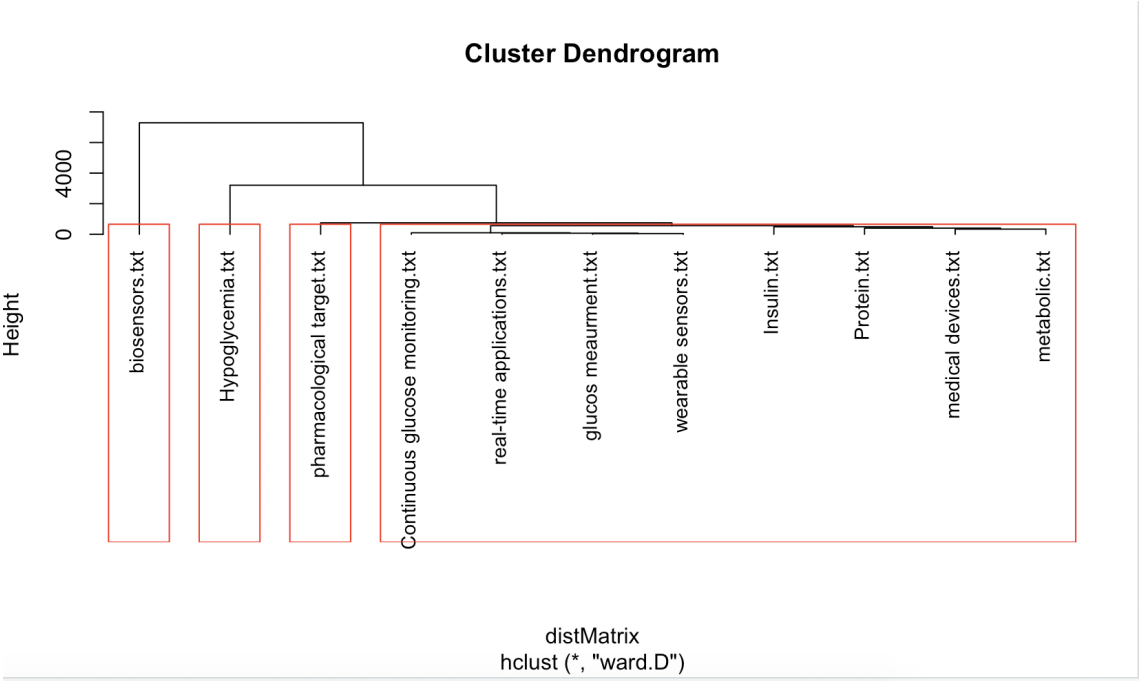


**Figure 17.** Using hierarchical clustering algorithm to have 4 clusters for 18 documents

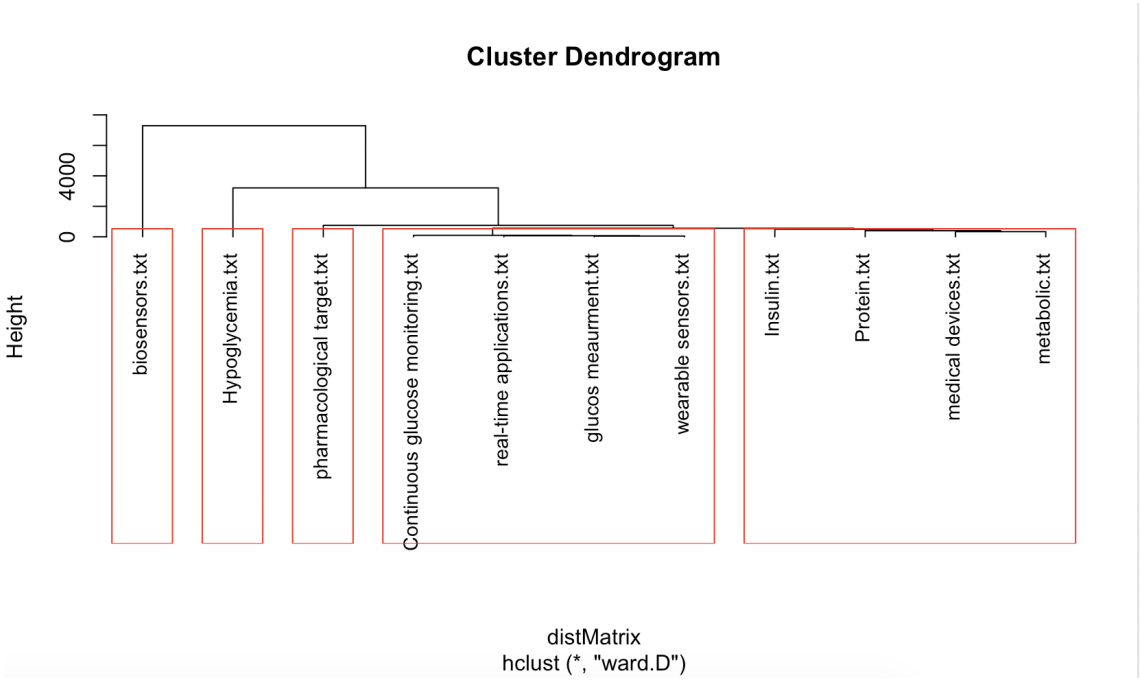


**Figure 18.** Using hierarchical clustering algorithm to have 4 clusters for 18 documents

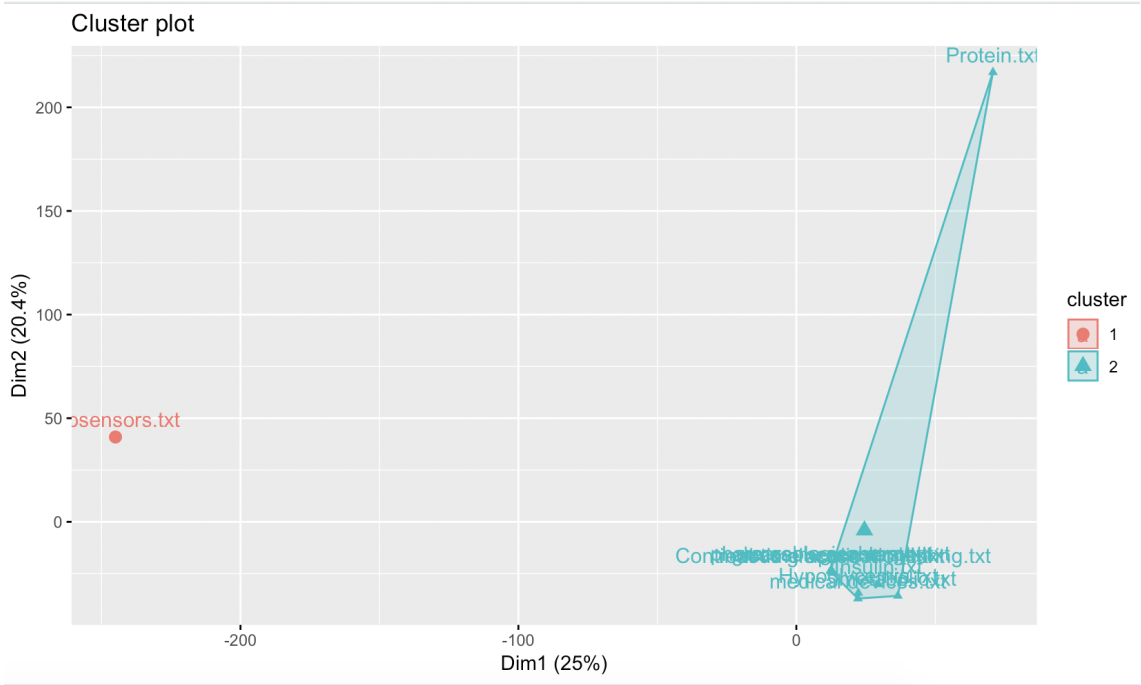


**Figure 19.** Using hierarchical clustering algorithm to have 2 cluster for 18 documents

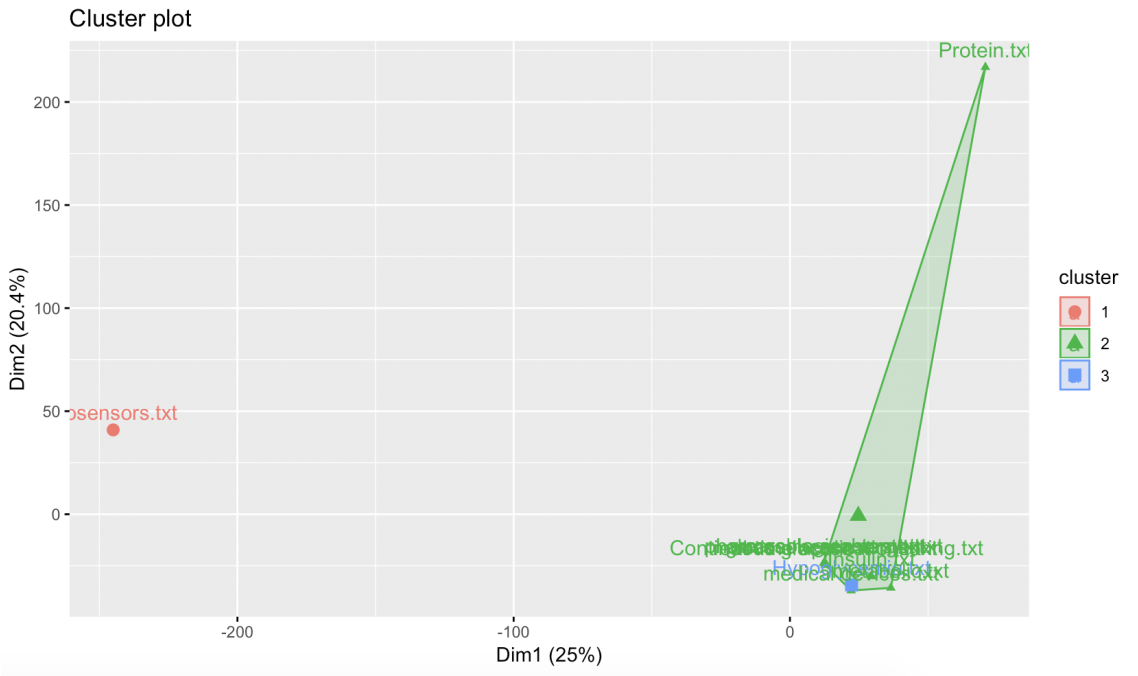


**Figure 20.** Using K-means clustering algorithm to have 3 clusters for 18 documents

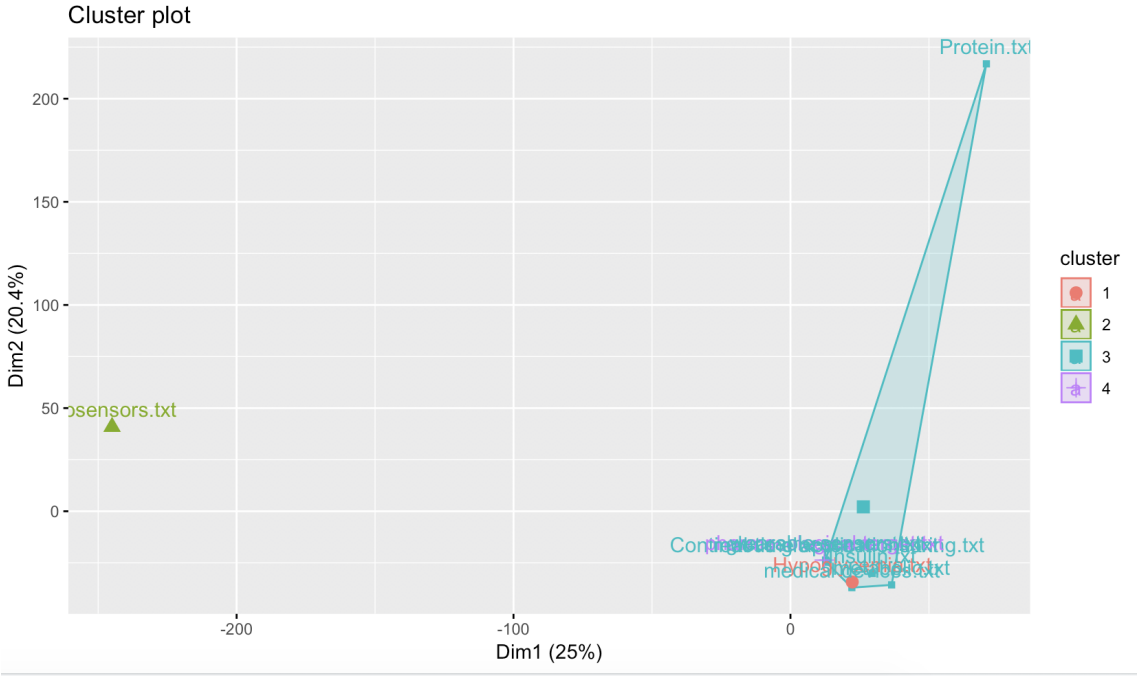


**Figure 21.** Using K-means clustering algorithm to have 4 clusters for 18 documents

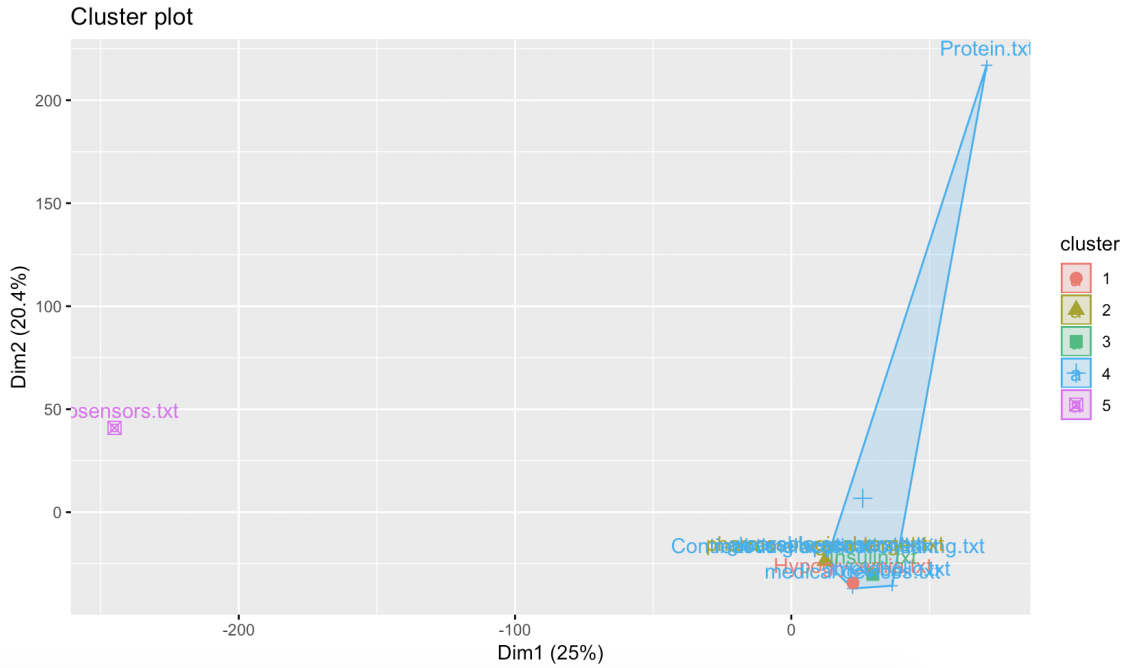


**Figure 22.** Using K-means clustering algorithm to have 5clusters for 18 documents

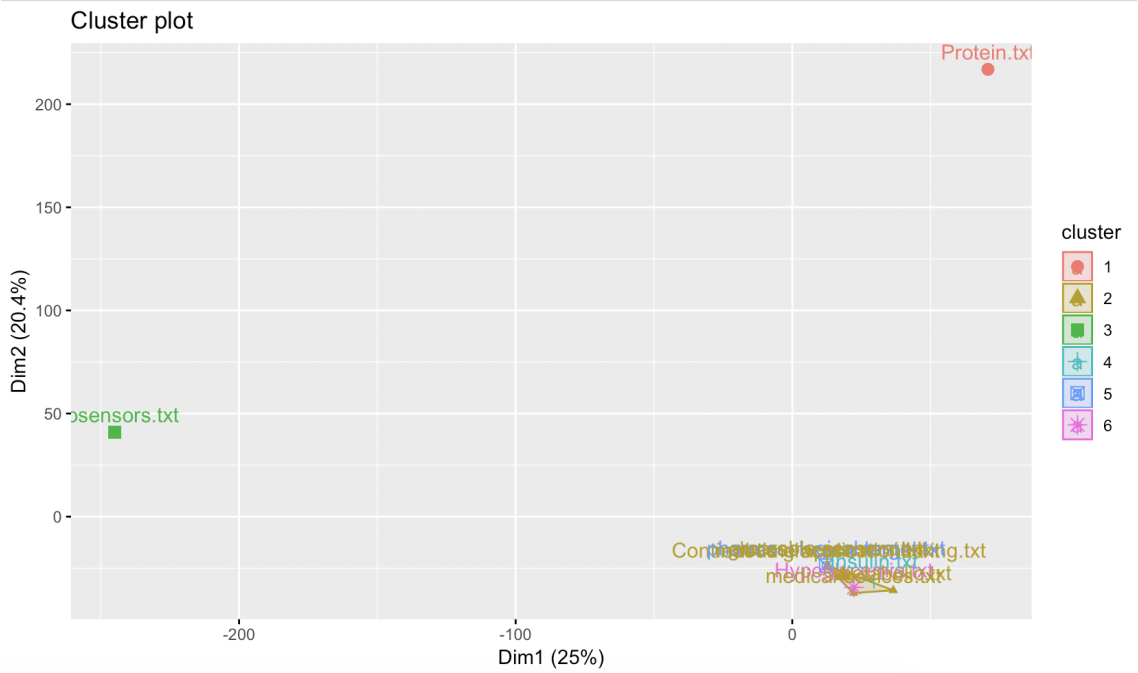


**Figure 23.** Using K-means clustering algorithm to have 6 clusters for 18 documents

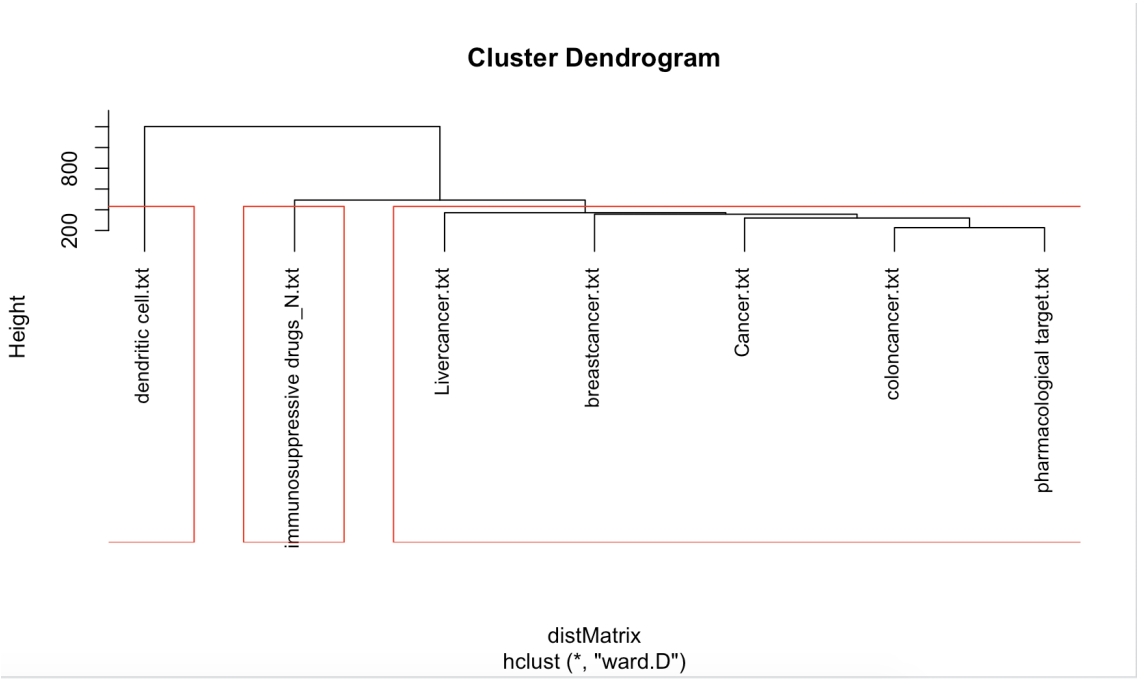


**Figure 24.** Using hierarchical clustering algorithm to have 2 clusters for documents related to DCs

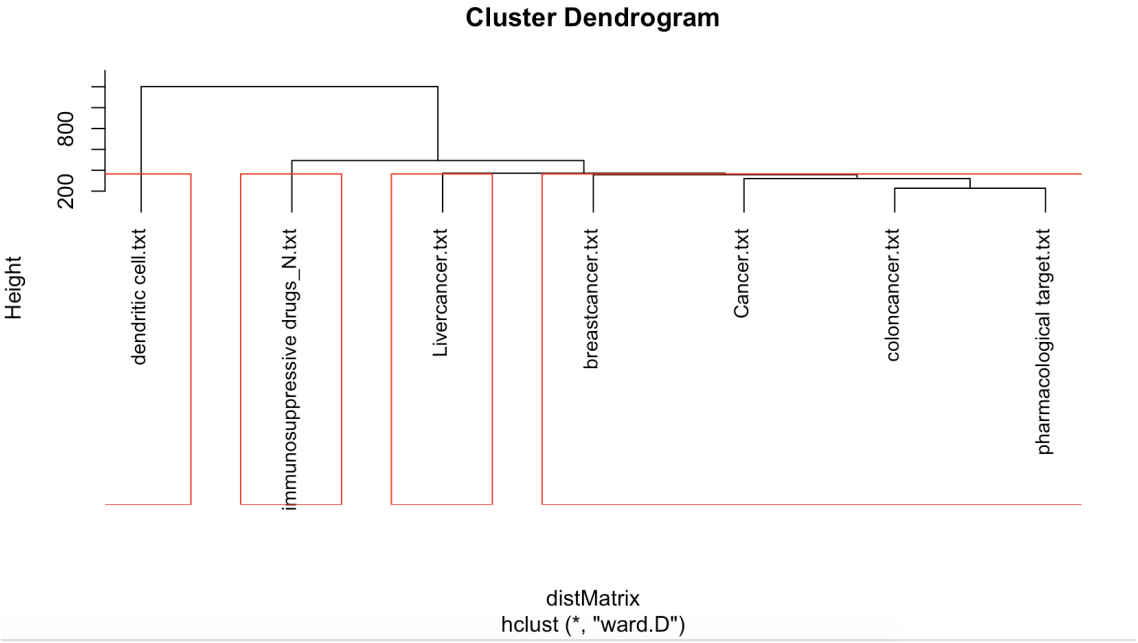


**Figure 25.** Using hierarchical clustering algorithm to have 3 clusters for documents related to DCs

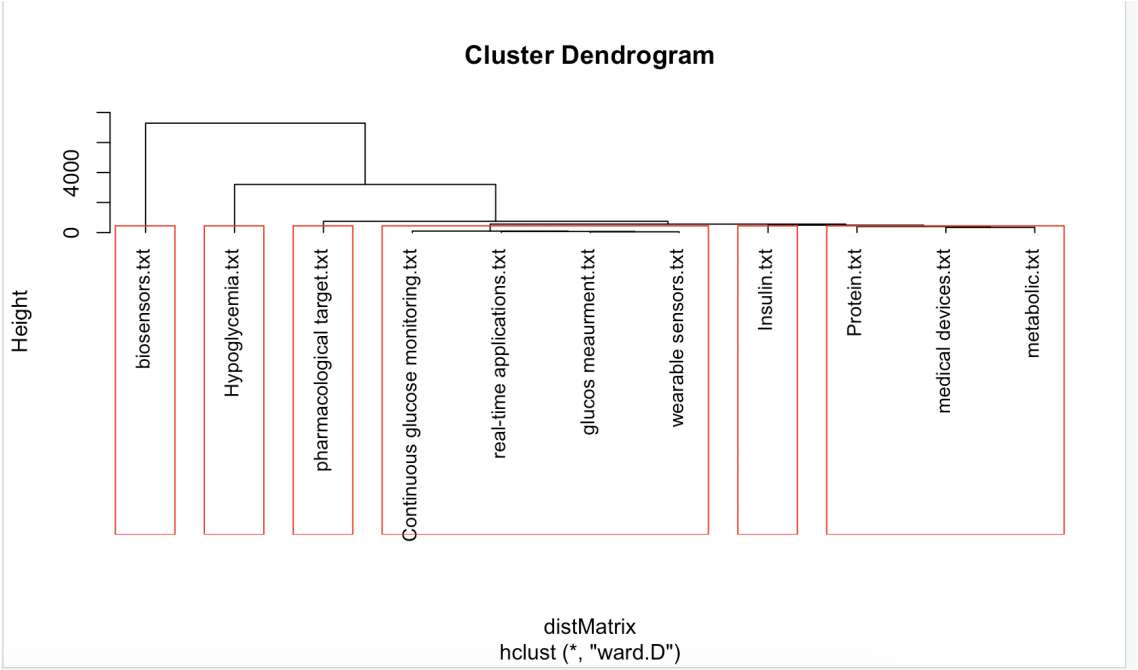


**Figure 26.** Using hierarchical clustering algorithm to have 4 clusters for documents related to DCs

Figures 27, 28 and 29 show clusters obtained from applying K-means clustering algorithm on those documents.

## 7. Conclusion

This paper explained how Multiple machine learning tools (MLT) and bibliometric tools can be used to identify emerging topics and forecast trends of emerging topics. The MLT module takes advantage of fusion algorithm to select important topics and determine relevant indices of those topics. The BT module is responsible for constructing time-series data set of the indices. The motivation behind presenting fusion algorithm is to develop a scalable data-driven framework with high performance. The purpose of utilising bibliometrics is to provide multivariate time series data set. In this paper, we showed how topic modeling and clustering algorithms are working and provided some results for our next step The performance and effectiveness of proposed framework would be evaluated on a variety of text data sets (e.g., Scopus, ); the obtained results would be compared with other methodologies that most are based on bibliometric tools. The fact that algorithm fusion of the MLT module integrate time series prediction and identification system to elaborate topic evolution and evolving emerging topic is one of the significant benefits of the framework.

## References

1. Mehrotra, P. Biosensors and their applications–A review. *Journal of oral biology and craniofacial research* **2016**, *6*, 153–159.
2. Salatino, A. Early Detection and Forecasting of Research Trends **2015**.
3. Parsapoor, M. Brain Emotional Learning-Inspired Models **2014**.
4. Parsapoor, M. Towards Emotion-inspired Computational Intelligence (EiCI). PhD thesis, Halmstad University, 2015.
5. Gers, F.A.; Eck, D.; Schmidhuber, J. Applying LSTM to time series predictable through time-window approaches. In *Neural Nets WIRN Vietri-01*; Springer, 2002; pp. 193–200.
6. Xu, R.; Donald Wunsch, I. Survey of Clustering Algorithms. *IEEE TRANSACTIONS ON NEURAL NETWORKS* **2005**, *16*, 645.

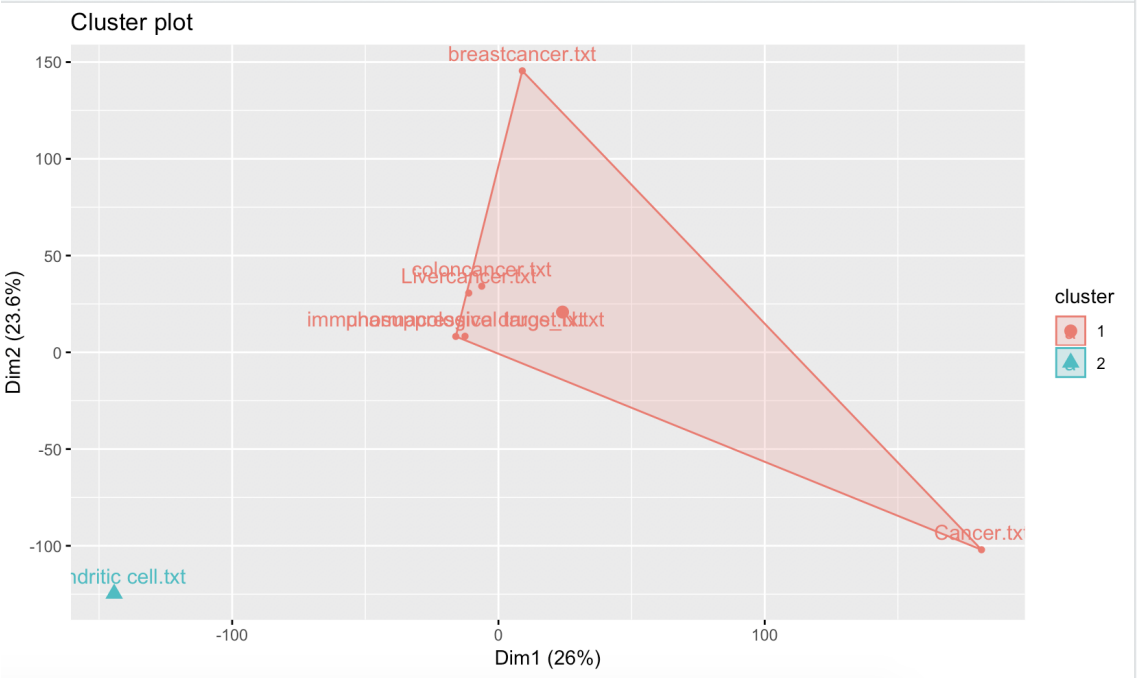


**Figure 27.** Using K-meeans clustering algorithm to have 2 clusters for documents related to DCs

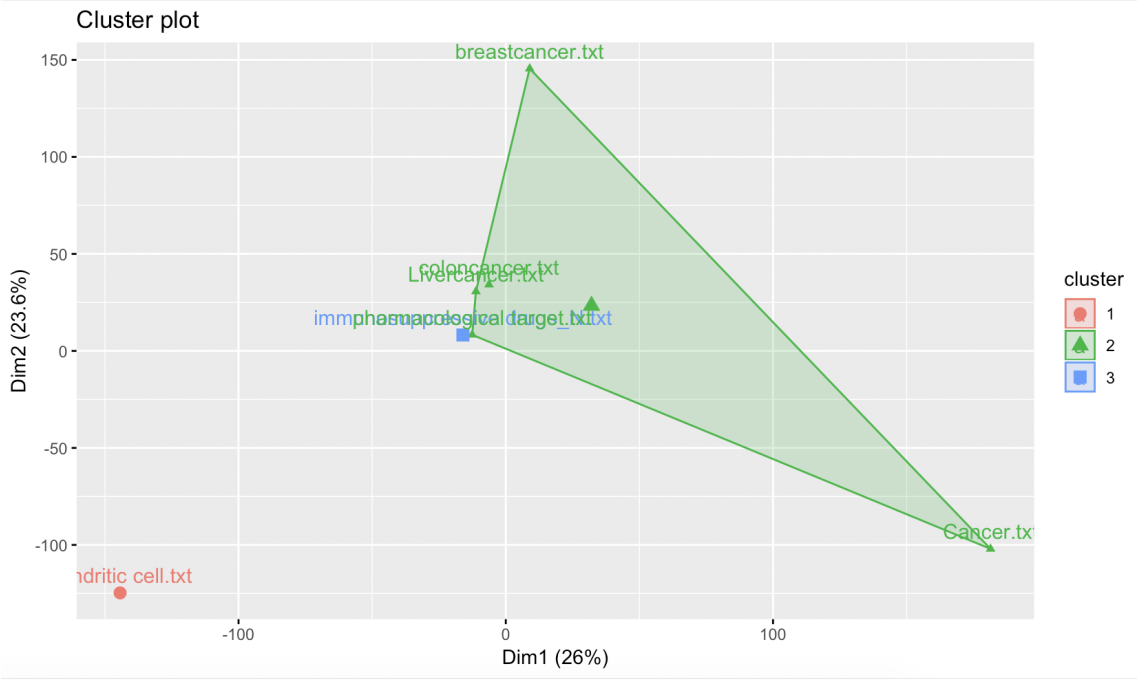


**Figure 28.** Using K-means clustering algorithm to have 3 clusters for documents related to DCs

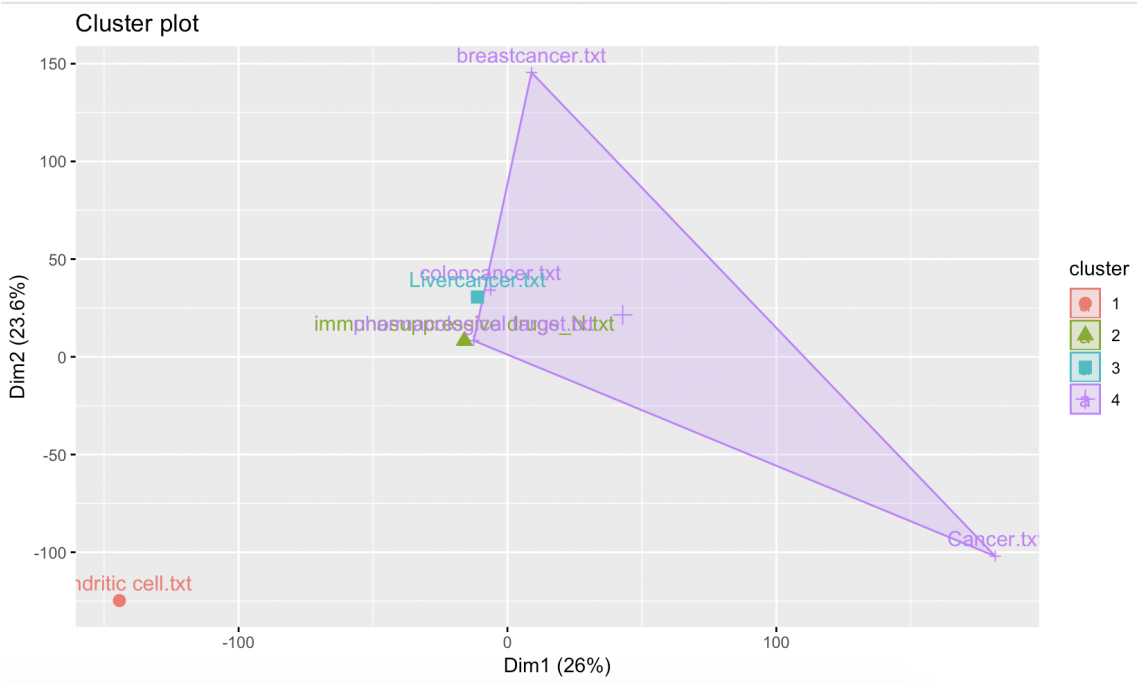


**Figure 29.** Using K-means clustering algorithm to have 4 clusters for documents related to DCs


7. Allahyari, M.; Pouriyeh, S.; Assefi, M.; Safaei, S.; Trippe, E.D.; Gutierrez, J.B.; Kochut, K. A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919* **2017**.
8. Sievert, C.; Shirley, K. LDAvis: A method for visualizing and interpreting topics. Proceedings of the workshop on interactive language learning, visualization, and interfaces, 2014, pp. 63–70.
9. Blei, D.M. Probabilistic topic models. *Communications of the ACM* **2012**, *55*, 77–84.
10. Hurtado, J.L.; Agarwal, A.; Zhu, X. Topic discovery and future trend forecasting for texts. *Journal of Big Data* **2016**, *3*, 7.