

Text Mining of Shwachman and Thalassemia disease papers

Anton Heijs ^{*}
Liesbeth Siderius [†]

December 24, 2020

1 Introduction

In this paper we describe the analysis of rare disease papers using natural language processing, machine learning and visualization techniques to determine the topic areas described in the papers about Shwachman disease and Thalassemia disease. We also describe how one can use these techniques to determine the most important papers for each topic area which results in a ranking of papers and helps a doctor or medical researcher to select these papers first. Additionally we investigate how one can link LOINC codes to the relevant papers. The papers we use for this study are selected from pubmed using a broad query to obtain all relevant papers and those that appear not relevant are removed from the data sets later in the process.

The query "(Shwachman) OR (Shwachman syndrome) OR (Shwachman disease)" resulted in 231 papers from 1975 to 2020 and "(Thalassemia) OR (Thalassemia syndrome) OR (Thalassemia disease)" 3385 papers from 1951 to 2020

From the data which pubmed provides we needed to extract the text from the title and the abstract for our analysis. Software was written to extract and reformat the pubmed data to perform the text mining analysis. All the other fields are not relevant for text mining.

^{*}Commissioned by Stichting Shwachman Syndroom Support Holland, The Netherlands

[†]Stichting Shwachman Syndroom Support Holland, The Netherlands

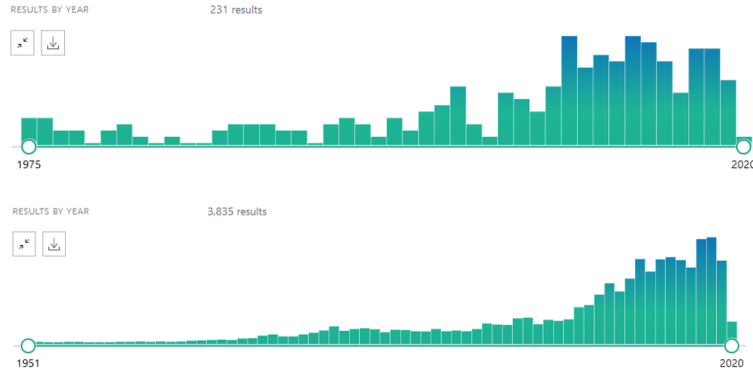


Figure 1: Distribution of published papers for Shwachman disease and Thalassemia disease from pubmed over time.

2 Text Mining

The goal of text mining is *“to discover or derive new information from text data, finding patterns across data-sets and/or separating signal from noise”* by means of unsupervised or supervised algorithms. This is significantly different from information retrieval and search where information which is already known is retrieved from a large collection of data. Although text collections represent vast and rich amounts of information, the actual automatic interpretation of text collections is a difficult process.

The first step in the text mining process consists of retrieving a collection of text documents. This can be a data-set from a single source or it can be composed of data from multiple sources. In order to make text documents open for computation, we need to transform them into numbers. This can be done in many ways. The process illustrated in figure 11 is known as a vector space model and is the current state-of-the-art text processing method for text analytics.

- Tokenization: A text is broken up into component words, called tokens.
- Stop-word removal: Stop-words are words that occur frequently but hold no real information value. Some examples for stop-words are ‘the’, ‘a’ and ‘and’. As these words contain no real information they are removed.
- Stemming: Stemming is the process of reducing words to their stem, base or root form. This means that words are reduced to their most simple form. Multiple words will generally reduce to a single stem. Terms that are reduced to the same stem are combined.
- Vectorization : The vector space model is a numerical representation for text documents. Documents are represented as vectors of term occurrences. Each

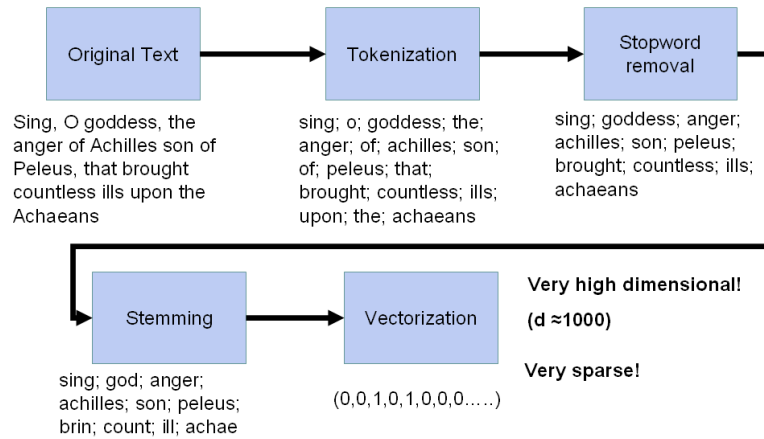


Figure 2: Building a vector space model on an example text to illustrate the process.

dimension in the vector corresponds to a unique term.

Additional steps include some form of vector normalization, as that enables the comparison between longer and shorter documents. For some techniques, term weighting is performed. Term weighting has the effect of making words that are specific to a subset of documents more important. This is also important for the pubmed papers since the text length can vary between the papers.

3 Unsupervised Machine Learning for Text

In machine learning there are two main approaches to use an algorithmic approach to learn from data, supervised machine learning and unsupervised machine learning. Supervised machine learning needs examples, training data, to train an algorithm to perform a ranking of the data, such a collection of documents. Unsupervised machine learning differs from supervised machine learning in that no manual input or labels are used. Instead of using a human-defined set of pre-labelled training data, the data is classified automatically without human interference based on some common trait which is present in the data. Unsupervised machine learning can thus be described as a descriptive approach because it attempts to find new sets of categories as opposed to supervised machine learning which takes a predictive approach.

3.1 Clustering

Clustering is the assignment of documents, into groups, called clusters, so that the documents from the same cluster are more similar to each other than objects from

different clusters. Clustering is a common technique for statistical data analysis.

It is important to realize that although clustering is used to find natural groupings by means of common traits in the data, by using a technique which dissects a dataset into groupings, structure is implicitly imposed upon the dataset. After clustering the groupings can be evaluated, named and their properties summarized. The results from clustering can be used to identify natural groupings present in the data, data reduction or to generate hypotheses for the dataset that was analyzed.

3.2 Spatial Placement

Spatial placement is the name for techniques that generate a graphical representation of the documents arranged in such a way that the distance relations (similarities) are preserved since these distances are calculated in a high dimensional space and projected onto a two dimensional plane. Since documents contain many words one has to represent them as vectors and these vectors can represent for instance 1000 terms and thus the distance between two documents is the distance between 2 vectors in a 1000-dimensional space. When all the distances between all the documents are calculated the projection algorithm can determine the position of all documents on a plane and we can calculate also contour lines to generate a "landscape view". Objects that are similar will then be placed close together, and objects that are different will be placed further apart.

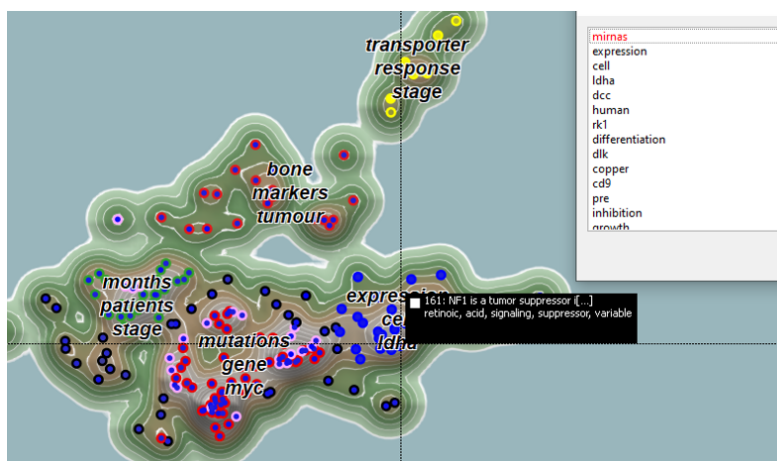


Figure 3: Example of spatial placement with density estimation. This enables us to clearly see the document distribution and highlights the subgrouping of the right-hand cluster. Relevant terms related to one Shwachman paper are indicated to support selection of documents.

In the figure we see the results of a technique known as *iterative point placement* on a sample collection of pubmed data. Every point in the visualization represents a pubmed document. The spatial placement algorithm attempts to preserve in the 2D plot the

distance between each pair of points as computed in the high-dimensional vector space model. The distance is a measure for document similarity, so similar documents will be grouped together. The user can interact with the visualization and select documents and document groups for further inspection or processing.

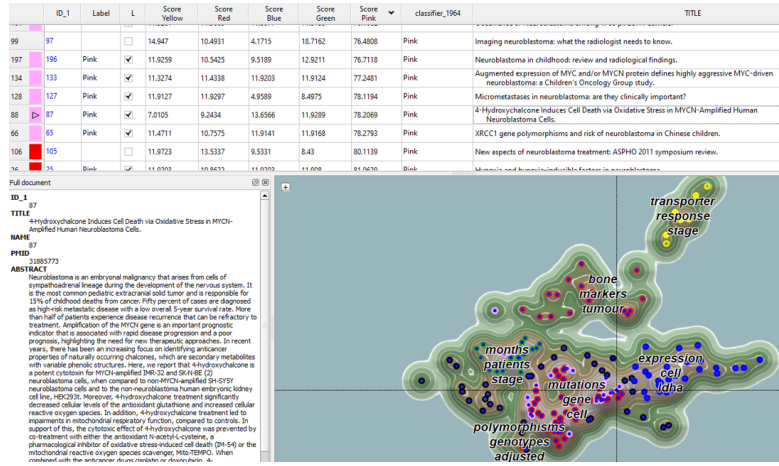


Figure 4: View of the software showing the visualization of the Shwachmann pubmed papers and also the color labeled documents in the topic clusters. The documents are also ranked for the different topics which is shown above the landscaping view. The text of one selected pubmed document (see cross hair in the visualization) is also shown.

The strength of the cluster (landscape) visualization is that it makes it easy to select documents that are examples (training documents) for a class and they can be used to train a classification algorithm. This was done for all topics and the classification scores are shown in the GUI. By selecting a topic area of interest and then selecting the top ranking documents one can quickly identify the most relevant papers for a disease aspect (a topic cluster).

4 Automated Text Classification

Classification or *categorization* is a task focused on assigning an document to one or more categories. This decision is taken based on the values of properties of this item, known as *features*.

In figure 5 we show an example of a classification problem. For the purposes of the illustration we show a two dimensional problem. Real-world usage often involves over 1000 independent dimensions. Based on the training examples (black and white circles) we construct a separation hyper-plane. This hyper-plane is then used to assign classification scores to an unknown example (red).

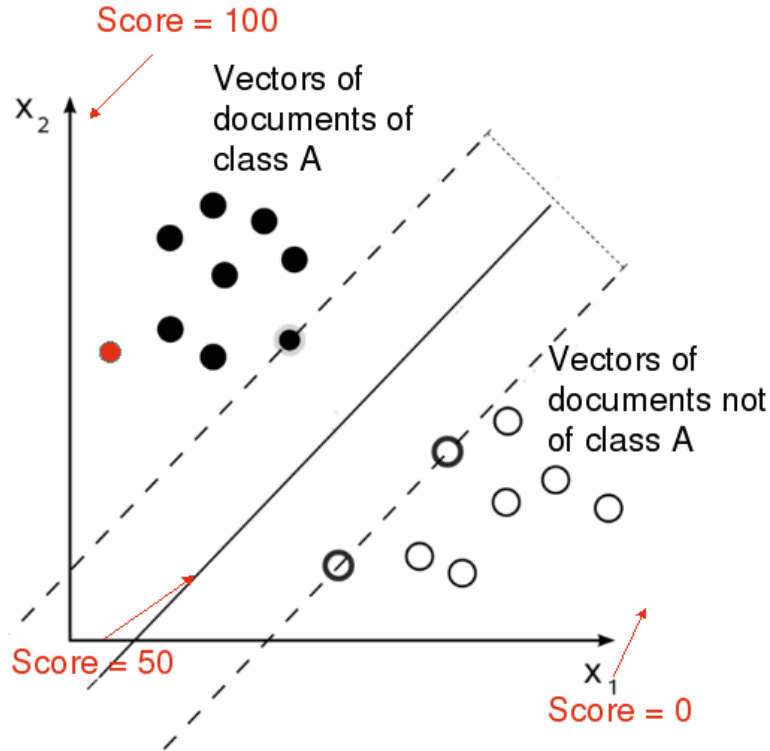


Figure 5: Low-dimensional example of text classification shows training examples, the separation (hyper-)plane and one unknown item.

The main task of the user of an automated classification system is finding relevant training examples and determining their classification labels. Often it is up to the user to select and label training examples from a large, unknown, set. The unsupervised techniques mentioned in the previous section, especially interactive spatial placement visualizations, can help in uncovering new categories of documents in the data set and finding training data to train an automated text classifier. This is done also for the pubmed documents.

There are many different algorithms that can be used to determine the separation hyperplane. For text classification, support vector machines (SVM) is a proven, robust algorithm that performs close to the theoretical maximum for the given data in the wide majority of cases. Because of the structure of the algorithm it is especially well-suited to computations using sparse vectors that span thousands of dimensions.

In ranking problems the goal is to learn to order a new set of objects as accurately as possible. Such ranking problems naturally occur in applications like search engines and recommendation systems. In our application we use it to identify relevant example documents using clustering, selecting and using them as training documents to train

the SVM classification algorithm and then run the classifiers on all the documents to calculate the classification scores. When these score are sorted we obtain a ranking of all documents belonging to a topic. In the case of these rare diseases one can use a broad search on pubmed, since one often does not know how to search for some specific documents and after the described analysis approach one obtains the most relevant documents.

5 Using LOINC codes

Medical laboratory observations are identified by LOINC codes which stands for the **Logical Observation Identifiers Names and Codes** and is an a universal standard and an electronic database for clinical care and management. The database includes not just medical laboratory code names but also nursing diagnosis, nursing interventions, outcomes classification, and patient care data sets. LOINC provides thus universal code names and identifiers to medical terminology related to electronic health records The purpose is to assist in the electronic exchange and gathering of clinical results (such as laboratory tests, clinical observations, outcomes management and research).

LOINC terminology has two main parts ; laboratory LOINC and clinical LOINC.

- Laboratory LOINC covers laboratory tests, microbiology tests (including antibiotic susceptibilities)
- Clinical LOINC covers a variety of non-lab concepts (ECG, cardiac echo, ultra-sound)

Each term for test or observation is described by a unique 6-part name in the LOINC database. This database currently has over 71,000 observation terms that can be accessed and used universally.

Each database record includes six fields for the unique specification of each identified single test, observation, or measurement:

1. Component- so what is measured, evaluated, or observed (example: urea,...)
2. Kind of property- so the characteristics of what is measured, such as length, mass, volume, time stamp and so on
3. Time aspect- which is the interval of time over which the observation or measurement was made
4. System- which is the context or specimen type within which the observation was made (example: blood, urine,...)
5. Type of scale- which is the scale of measure. The scale may be quantitative, ordinal, nominal or narrative

6. Type of method- which determines the procedure used to make the measurement or observation

Other database fields include status and mapping information for database change management, synonyms, related terms, substance information (e.g. molar mass, CAS registry number), choices of answers for nominal scales, translations.

When we look-up the LOINC codes for Shwachmann disease we will find the 6-digit number **41764-2**¹ and references to a set of terms.

1. Component : SBDS gene targeted mutation analysis
2. Property : Find
3. Time : Pt
4. System : Bld/Tiss
5. Scale : Doc
6. Method : Molgen

There are "Additional Names" and "Basic Attributes" mentioned and a set of "Related Names" which are the terms we want to search for in the analysed papers. For Shwachman these "Related Names" are:

Blood
CGI-97
Document
Finding
Findings
FLJ10917
Molecular genetics
Molecular pathology
MOLPATH
MOLPATH.MUTATIONS
Mut
Mut Anl
Mutations
PCR
Point in time
Random
SDS
Shwachman-Bodian-Diamond syndrome
SWDS

¹A search on Shwachman give the LOINC code **41764-2** and details are on webpage: <https://loinc.org/41764-2/>.

Tissue
 Tissue, unspecified
 WB
 Whole blood
 Whole blood or Tissue

When we search over all documents for the set of LOINC terms for Shwachman² we expect to find those papers that relevant to medical test procedures as defined by the LOINC code for Shwachman.

When we then use the classifiers created in the text mining process we can also determine the most relevant papers by sorting the on the classification score.

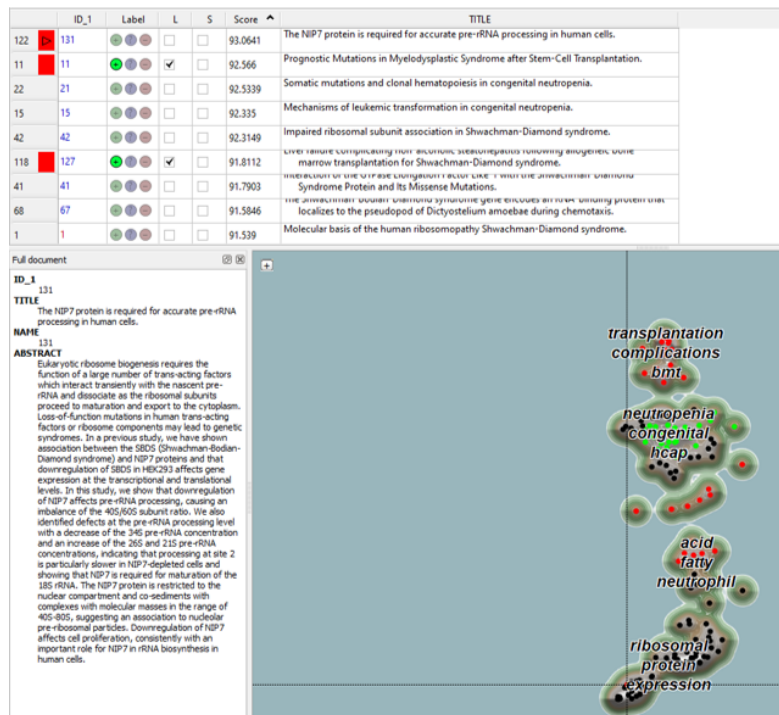


Figure 6: Shwachman disease papers classified to determine the most relevant pubmed papers related to Neutropenia, which is an abnormally low concentration of neutrophils (a type of white blood cell) in the blood.

After the classification, which provides the most relevant documents, we performed a search for the terms used to describe the LOINC code for Shwachman disease. When

²Search terms for Shwachman to find papers relevant for medical test related to Shwachman : Blood CGI-97 Document Finding Findings FLJ10917 Molecular genetics Molecular pathology MOLPATH MOLPATH.MUTATIONS Mut Mut Anl Mutations PCR Point in time Random SDS Shwachman-Bodian-Diamond syndrome SWDS Tissue Tissue, unspecified WB Whole blood Whole blood or Tissue

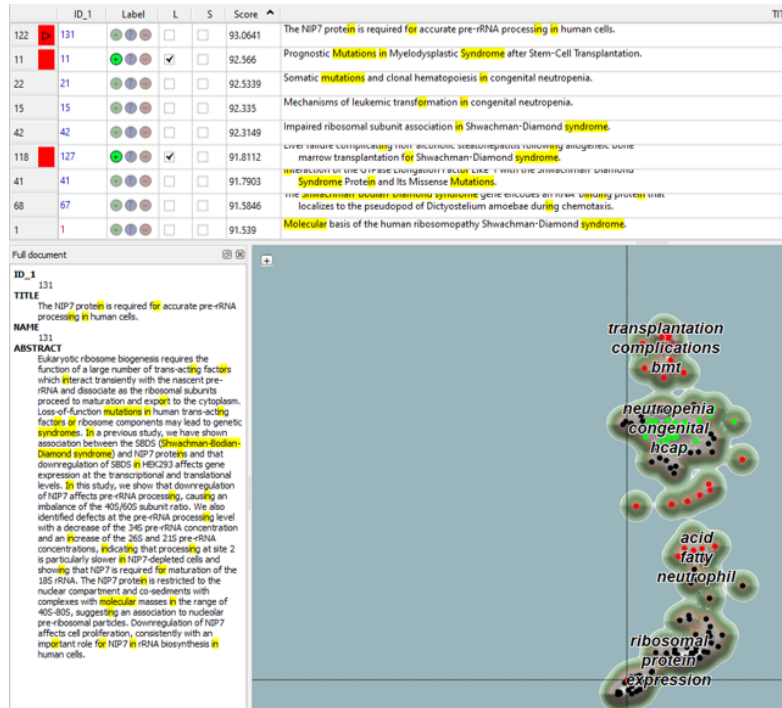


Figure 7: Using a search for the terms describing the LOINC code for Shwachman disease we can see where they all appear in the papers.

they most relevant terms are used in a filtering operation we obtain only those relevant documents that are also mentioning the terms that are captured by the LOINC code so the pubmed papers that mention a medical test related to Shwachman disease. In the visualization one can see a smaller set of documents (85 out of 143 papers).

We have in detail described the text mining process using the 143 pubmed papers on Shwachmann disease. For Thalassaemia disease we perform the same processing steps except we use multi-class classification which means that for each topic cluster in the document set we make a classifier. After training the classifiers we run them on all documents which provides a multiple probability scores for each paper.

Again a smaller set of papers mention terms that belong to the LOINC code 55234-9 for Alpha thalassemia gene panel - Blood by Molecular genetics method (see <https://loinc.org/55234-9/>). These papers are also ranking by importance (in this example to "transfusion" since the results of this classifier is shown).

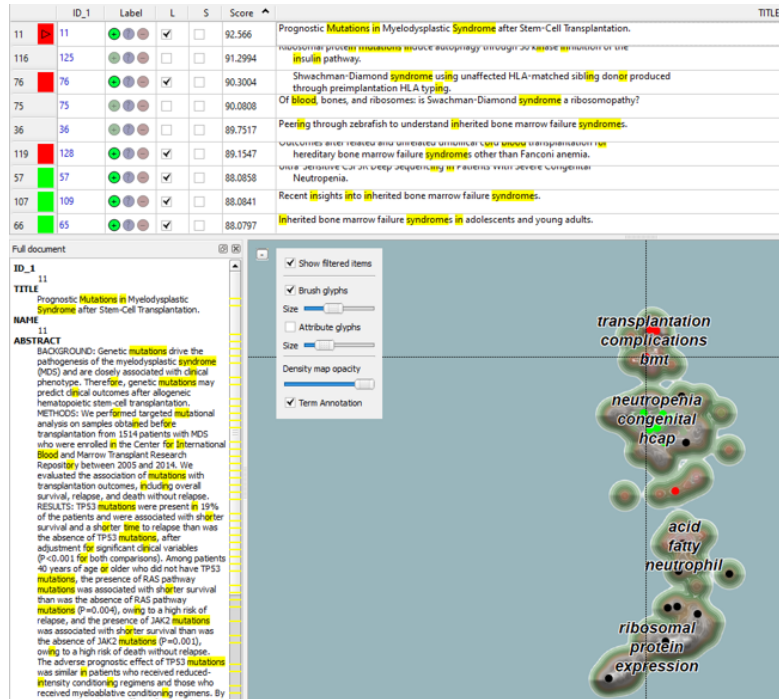


Figure 8: Using a filtering for the most relevant terms of the LOINC code for Shwachman disease we can identify the pubmed papers on Shwachman that are also related to a medical test as captured by the LOINC code for Shwachman disease.

6 Conclusion

Text mining is a powerful tool for processing pubmed document sets. Using unsupervised techniques such as clustering and spatial placement, one can quickly gain insight into the contents of the documents, discover hidden properties and determine how to further explore and label the data.

Using classification techniques, the user can create high-performance text classifiers that can sort through the pubmed documents in seconds rather than days. By first using the visualizations for document selection and labelling, using these labelled documents to create and run classifiers one can quickly find relevant information in rare disease data such as shown for Shwachmann and Thalassemia disease. By using the terms that belong to LOINC codes in filtering the whole (ranked) document set one can identify also the important rare disease papers that most likely are relevant for medical test.

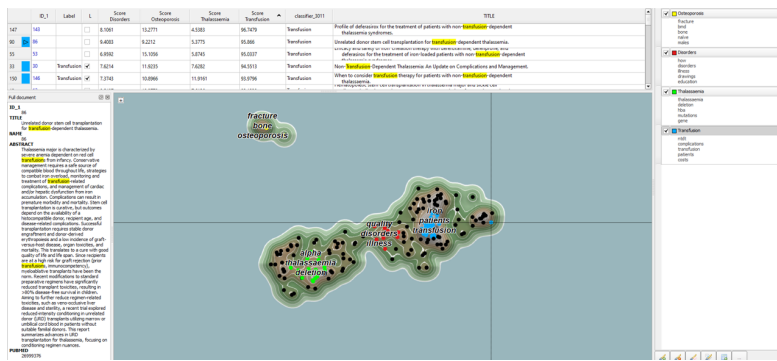


Figure 9: Multi class classification to show the relevance of each document for the 4 topics.

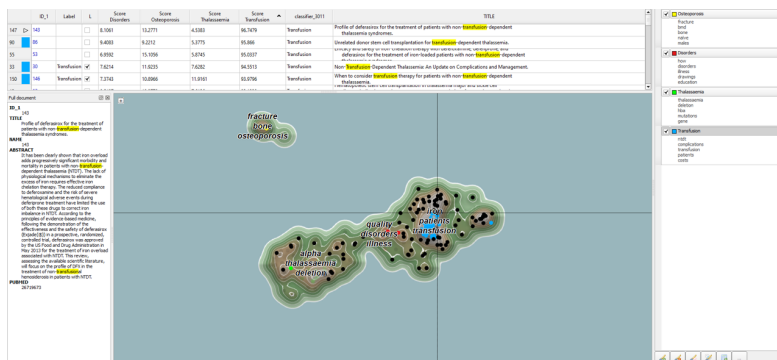


Figure 10: Thalassaemia papers which are relevant to "transfusion".

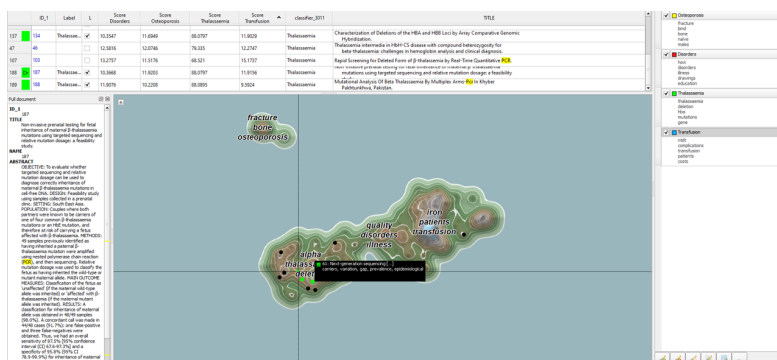


Figure 11: A subset of the Thalassaemia papers after filtering on LOINC terms belonging to LOINC code 55234-9.