12-2018

A Generative Statistical Approach for Data Classification in a Biologically Inspired Design Tool

Marvin Manuel Arroyo Rujano
University of Arkansas, Fayetteville

A Generative Statistical Approach for Data Classification in a Biologically Inspired Design Tool

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Mechanical Engineering

by

Marvin Arroyo Rujano
University of Arkansas
Bachelor of Science in Mechanical Engineering, 2015

December 2018
University of Arkansas

This thesis is approved for recommendation to the Graduate Council.

David C. Jensen, Ph.D.
Thesis Director

Zhenghui Sha, Ph.D.
Committee Member

Rick J. Couvillion, Ph.D.
Committee Member

ABSTRACT

The objective of the research this thesis describes is to find a way to classify text-based descriptions of biological adaption to support Biologically Inspired design. Biologically inspired design is a fairly new field with ongoing research. There are different tools to assist designers and biologists in bio-inspired design. Some of the most common are BioTRIZ and AskNature. In recent years, more tools have been proposed to aid and make research in the field easier, for example, the Biologically Inspired Adaptive System Design (BIASD) tool. This tool was designed with the goal of helping designers in early design stages generate more robust and innovative designs. Even though this tool offers a vast database of biological examples, many limitations have been encountered in the tool. The most noticeable is the order in which the biological examples are distributed within the tool. The process used to classify them was very subjective and does not follow a pattern. Another challenge is the way in which the user of the tool reaches the biological examples. By addressing these issues, we provide a more objective way to classify the biological adaptive strategies. To do this, we needed a meta classification in order for the questions to be rationally organized within the tool. Then, approaches such as k-means and Latent Dirichlet Allocation (LDA) techniques from machine learning were employed to minimize the randomness and increase the objectivity of the tool. Out of the two, the LDA model provided a more useful classification. A validation of the LDA model was needed, so we used perplexity, which is used in statistical models to measure the accuracy of a language model and better understand datasets. At the end, a rational classification for the analogues of the BIASD tool was generated.

ACKNOWLEDGMENTS

David C. Jensen, Ph.D. Associate Professor of Mechanical Engineering, who was the adviser and assisted in the research and funding of this work.

Darin W. Nutter, Ph.D. Professor and Department Head and the Twenty-First Century Leadership Chair in Engineering, who also collaborated in the funding of this research.

TABLE OF CONTENTS

LIST OF PUBLISHED PAPERS

Arroyo, Marvin, Nicholas Huisman, and David C. Jensen. "Exploring Natural Strategies for Bio-Inspired Fault Adaptive Systems Design." *Journal of Mechanical Design* 140.9 (2018): 091101.

1. INTRODUCTION

Biologically-inspired design is used as a tool to help designers generate more robust design ideas. In this work, we discuss a biologically inspired design tool that uses biological examples to help designers generate more robust ideas in early design stages. The name of the tool is Biologically Inspired Adaptive Systems Design (BIASD), and it is a set of four binary trees; it has a collection of 161 analogues from nature that were gathered and distributed among the four binary trees. In this work, we addressed many challenges and issues with the existing tool. Specifically, we modified and deleted some of the original questions on the binary trees to better navigate them. Additionally, we introduce a classification for the questions to improve the organization of the tool. Finally, we generated a novel classification for the analogues in the tool using a text-based machine learning technique.

Analogues in biologically inspired design are ideas that designers use to generate creative solutions in products and systems. In their work, Malshe et. al exemplify this concept by showing different ways in which organisms in nature repel water, change color, attach to certain surfaces, etc. and how scientists have managed to mimic their structure in tools that we use on day-to-day basis [1]. This concept is also known more as biomimicry; the BIASD tool is more bio-inspired rather than biomimicry. The difference in the two is that biomimicry copies the structure (sometimes to a micro or nano level) to replicate their functions. On the other hand, bio-inspired design takes ideas from the way organisms achieve something in nature and try to generate robust designs based on those examples. Appendix B shows all examples of the analogues in the BIASD tool.

The structure of the BIASD tool has some drawbacks. Since the classification of the analogues was done in an ad hoc and subjective way, a more impartial method to classify them was applied. Furthermore, the way in which the questions were established throughout the tool was confusing in terms of order and terminology; additionally, the questions lacked a meta-structure. This brings us to the main objective of this work: we are trying to find a better way to classify and structure the data set in the BIASD tool, especially the analogues. Machine learning provides different methods that can be used to classify the analogues in a less subjective way. From machine learning, methods like K-means clustering and Latent Dirichlet Allocation (LDA) were used. A third method that we called dictionary search was used to try to classify the analogues. This method was inspired from different strategies used in reliability engineering to avoid failures.

2. BACKGROUND

In this work, we experimented with different approaches to classify our dataset. In this section, we discuss the BIASD tool itself and how it works to help the reader understand the structure and why we want to ultimately change it. Furthermore, we talk about machine learning and the methods used and how they were applied to the existing biologically inspired tool for design.

2.1 BIASD Tool

Biologically inspired design is an emerging area in computer and design engineering as well as in biology; its main objective is to methodically extract biological knowledge to assist in solving design problems. Existing bio-inspired design tools include BioTRIZ and AskNature; these record fundamental biology principles and propose analogies to inspire designers solve problems. [2].

In this work, we focus our attention on a bio-inspired tool that is structured as a binary tree. This collection of binary trees is called BIASD tool, and in order to understand this tool better, we need to familiarize ourselves with three key definitions. First, we have adaptation, which is the response to an internal or external event, allowing for a modification of the system's goals. Secondly, there is strategy, which is the group of possible actions that move a system from one state to another. Lastly, we have analogues that are a particular example of the implementation of a strategy found in in nature [2].

The Bio-Inspired Adaptive Systems Design (BIASD) tool is a collection of 161 biological examples of fault adaptation [3]. The tool takes the form of four binary trees and represent a group of paths that lead to the biological examples; each path represents a different strategy. The user starts answering the firs binary question at the top of the tree and works their way down. The questions are answered in relation to the problem the user is trying to solve. Answering each question in the binary tree will lead the user to the next question until a leaf of the tree is reached. This leaf represents the analogue (biological example) that satisfies the path or group of questions followed. The four different binary trees are as follows: repair, reprogram, replace and reconfigure.

1. Replace: An adaptation where a faulty component is completely or partially replaced by a new one.
2. Repair: An adaptation where a faulty component is brought back to full functionality.
3. Reconfigure: Adaptations which uses other components and characteristics of the system to mimic the function of the faulty component in that system.

   Reprogram: Adaptations which adjust the programming or behavior of the system to work around the fault without fixing it. That is, changing its goals.

2.1.1 Understanding the Bio-Inspired Adaptive Systems Design tool.

Decision tree analysis is a way in which one can make decision making a little easier. The goal of this binary tree structure is to facilitate the different strategies available for users of the tool and also in identifying analogues that provide flexibility and more robustness. At the same time, designers can compare different strategies, so they can pick the best fit for their design problem [2]. Some important characteristics about binary trees used in this tool are worth mentioning. Understanding these concepts and rules will help the reader understand how to navigate and use the BIASD tool.

First, we have the size of the tree overall:

1. Depth: the depth of a binary tree is the number of edges that need to be traversed when traveling from the root of the tree to node n.
2. Height: the height of a binary tree is one more than the depth of the deepest node.
3. Level: a node with depth d is said to be at level d of the tree.
4. Complete Binary Tree: a binary tree is said to be complete if all levels except the last have two child nodes [2].

The BIASD Tool is a complete binary tree because a node is only created if there are differences in the adaptation strategies selected by the user. Moreover, for this tree, Huffman coding [4] is used to denote paths. The path from top to any leaf is designated by the answers to each binary question. In other words, using Hoffman’s code, the user has an easier way to identify the different paths that have been followed. The code for a given analogue consists of the sequence of zeros and ones encountered on the path from the root of the tree to the leaf nodes; the zeros will represent questions answered with “no,” and ones will represent questions answered with “yes.” For instance, a code written as [1.1.0.1] means that the questions to arrive to that analogue were answered as [Yes, Yes, No, Yes] [2]. Furthermore, Equation 1 quantifies the number of paths from root to every leaf of a decision tree. Let us say that $Path_G^{(n)}$ represents a path followed from head to tail of the tree. The number of leaves nodes represents the number of paths in the decision tree. That is $i = 1,2, \ldots, N$ and at the same time the number of leaves in a tree [2].

$$p = \sum_{i=0}^{N} Path_G^{(n)}$$

Once the user knows how to travel down the binary tree, the order in which the paths are traversed can be, in a way, played with depending on what the user is looking for. Some of the orders in which a user can go through the decision tree are as follows:

1. Inorder Traversal: the user visits a node's left child (and consequently, the left child's subtree), then the node itself, and lastly, that node's right child (and the right child's subtree).
2. Preorder Traversal: in this order, the user visits a given node before visiting either of the node's children and their subtrees.
3. Postorder traversal: in this order, one visits a given node's two subtrees before visiting the node itself.

Pruning a binary tree is the action of purposely ignoring part of the tree to save time and effort when analyzing it. The more you can prune the binary tree, the more time you save in the decision-making process. If a designer were to focus on a subsection of the tree based on the context of the question on a node, then he or she could prune the right side of that subtree saving effort on analyzing a few nodes and focus on the rest of the tree. In the BIASD tool, this occurs when one of the questions eliminates a large set of strategies. For example, selecting the No path to the question “Are there external recourses available to support the repair?” eliminates searching many branches of the repair strategy tree.

A lot of editing went on after the BIASD Tool was created. A new structure was built in GraphViz to help with visualization and organization. In addition, new analogues were added to database and some of them were taken away. The questions in all four decision trees went through some changes as well to make the decision process smoother and less confusing. Furthermore, a meta-classification for the questions was also necessary. The inspiration came from BioTRIZ. We started looking for underlying principles which differentiate strategies; this does not mean trying to find principles in the analogues but a cross-cutting relationship in the classification itself.

2.2 Machine Learning and K-means Clustering

In this work, we are trying to classify the database of 161 analogues in a different way than the four Rs that we know as Repair, Replace, Reprogram and Reconfigure. Different methods were tried to find a relationship among all examples. In this section, we are going to introduce machine learning and the model that was chosen to ultimately classify the analogues of the BIASD Tool. Nevertheless, we will mention the other methods used to try to classify the database as well and explain why they were not a good fit.

Machine learning enables computers to make successful predictions based on past experiences. Over the past few years, this field has shown a great improvement with the assistance of the increase of storage capacity and processing power of computers. Machine learning’s main goal is to model the relationship between a set of observable quantities, which are usually called inputs, and another set of variables that are related to these quantities and that are usually called outputs [5]. In this work, we are interested in looking in the relationship between the analogues as mentioned before. However, some real-life problems are too complicated to model directly as a closed form input-output relationship. Here is where some of the methods in machine learning come into play. Some of these methods can provide techniques that automatically create a computational model of these complex relationships by processing the data and maximizing a problem dependent performance criterion [5]. This process is known as "training," and the date used for training is called "training data." This trained model generates new insights into how input variables are mapped to the output; at the same time, this data can be used to predict new inputs that did not belong to the training data [5].

A learning algorithm searches a space S of hypotheses to find the best hypothesis in the space. A problem can arise when the existing training data is too small compared to the size of the hypothesis space. If there is not enough data, the learning algorithm finds different hypotheses in S that all give the same accuracy on the training data [6]. Furthermore, other issues may arise when working with training datasets. Many learning algorithms function only when the user performs a local search, which makes it hard for learning algorithms to find the best hypothesis [6].

For the clustering approach, we built a MATLAB code to find a relationship in the analogues. With this method, we are trying to find clusters in a set of existing data points. One can achieve this by using a deterministic technique called the K-means algorithm. To begin with, we consider the problem of identifying groups, or clusters, of data points in a multidimensional space. Suppose we have a dataset $\{x_1 \dots x_N\}$ consisting of N observations of a random D-dimensional variable x. The objective is to separate the data set into some number K of clusters. Instinctively, one might consider a cluster as comprising a group of data points whose inter-points distances are small compared with the distances to points outside the cluster. First, we introduce a set of D-dimensional vectors μ_k, where $k = 1,2,3, \dots, K$, in which μ_k is a prototype related to the k^{th} cluster [7]. Furthermore, we can say that μ_k is, in essence, the centers of the clusters. With all of this defined, we can now say that we are looking for an array of data points to clusters, as well as a set of vectors $\{\mu_k\}$, so that the sum of the squares of the distances of each data point to its closest vector μ_k, is a minimum.

Examples of this work are shown in [8-15]. Now let us consider a multivariate data $x_i = \left(x_{i1}, x_{i2}, \dots, x_{ip}\right)', i = 1, \dots, n$, containing n independent objects measured on p variables. For any separation of the n objects into k clusters (P_k), denote by C_m the set of objects allocated to the *m*th cluster and by *n*th the number of objects in C_m. Furthermore, in their paper, Yan and Ye denote $d_{i,i'}$ the distance between objects i and i'. They then look for the mth number of clusters, and with a defined value of k, they define W_k. This value is a characteristic quantity for the within-clusters homogeneity associated with P_k. Therefore, they conclude that the value of W_k drops quickly as k increases, which will generate an “elbow” in the W_k curve, and it indicates an optimal estimate of the number of clusters in the data [15]. We used the elbow curve approach to determine the number of clusters for our data set, and it is shown in the Analysis and Methodology section.

2.3 Latent Dirichlet Allocation

Machine learning helps sort, adjust, assemble and classify information that exist in the form of vectors. The machine learning method we found most useful was the Latent Dirichlet Allocation model. There are different tools out there that help with this process, and those tools or codes can be classified in two: supervised and unsupervised learning. The difference between the two is that in supervised learning, the machine assumes a function from a set of training examples. On the other hand, unsupervised learning means that the machine attempts to find a hidden structure in unlabeled data. In this work, our machine learning model is of the unsupervised type, and it is called Latent Dirichlet Allocation (LDA).

First, we give a general definition of LDA and what it does before analyzing it in depth. In natural language processing, LDA is a generative topic bag of words model that automatically discovers topics in text documents. This model assesses each document as a mixture of various topics, and then each word in the document belong to one of the document's topics the model generates. For instance, when classifying engineering classes, Group A comprehends a topic with the words "heat," "energy," "thermal," and "power." It is reasonable to assume that Group A is about classes related to Thermodynamics, Heat Transfer, Energy Systems, etc., i.e. mechanical engineering classes, whereas Group B may generate topics with words such as "circuit," "current," "power," "systems." One can assume these are words about electrical engineering classes. LDA is useful when analyzing a set of documents and trying to find patterns within those documents. We try to use LDA to generate topics and understand better the set of documents that we already have.

LDA is an unsupervised generative statistical technique for modeling a corpus (collection of documents), and it is the most commonly used topic modeling method [16]; it is also a Bayesian probabilistic model of text documents. This text modeling approach is applicable to general collections of discrete data; in LDA, the user assumes there is an underlying number of topics according to which documents are generated. Additionally, each of these topics is represented as multinomial distribution over the number of $|V|$words in the vocabulary [18]. Moreover, each topic is assumed to have been drawn from a Dirichlet distribution - $\beta_k \sim Dirichlet(\eta)$. Given these topics, LDA assumes for each word i in the document. After that, it draws a topic index $z_{di} \in [1, \dots, K\}$ from the topic weights $z_{di} \sim \theta_d$and draws the observed word w_{di} from the selected topic, $w_{di} \sim \beta_{z_{di}}$ [19]. Moreover, we can think of LDA as a probabilistic factorization of the matrix of word counts n, where n_{dw} is the number of times the word w appears in document d, into a matrix of topic weights θ and a dictionary of topics β [20]. This model is parameterized by the k-dimensional Dirichlet parameters $\alpha = \langle \alpha_1, \ldots, \alpha_k \rangle$ and a $k \times |V|$ matrix β, which are parameters controlling the k multinomial distributions over words. In the analysis section, we discuss the code used in Python to calculate the number of topics and words in a corpus. The code in Python can be found in Appendix A, figure 10.

To compare the performance of LDA, one can use what is called perplexity. In language modeling, perplexity is used as a quality measure for language models [21]. In information-theoretic approaches, perplexity is a widely-used measure. When we have a language model and a corpus, perplexity is the measure of the size of the set of words from which the next word is chosen given that we observe a group of words [22]. Furthermore, perplexity is usually used to measure how good a language modeling strategy is. In order for a model to be defined as "good," perplexity should be low – when low perplexity is achieved, the language model is said to have a high accuracy [23]. In a way, perplexity can be defined as the measure of entropy of the model, which means that it measures how much information is lost in a language modeling strategy, and it is defined by equation 2.

$$2^{H(p)} = 2^{-\sum_x p(x) log_2 p(x)}$$

In equation 2, the exponent $-\sum_x p(x) log_2 p(x)$ represents the entropy of the distribution [23]. The lowest the value of perplexity, the better. Perplexity can also be defined as a measure of how loosely a distribution fits to some samples. A small value indicates that the distribution fits tightly to the samples. On the other hand, a big value indicates that it fits loosely to the samples. Usually, one’s objective is to have low perplexity; however, these values should not be too low to avoid overfitting.

2.4 Reliability Engineering Strategies to Prevent Failure

Another of the methods utilized to try and classify the analogues in the BIASD tool was a dictionary search. This dictionary search was inspired in engineering strategies that handle adverse events. In fault adaptive design systems, designers implement reliability to assess the impact of adverse events. Some industries, quantitative analysis methods like fault tress are utilized to measure system performance [2]. Some of the engineering strategies used in the dictionary search were over-specification (design envelops, robustness, reliability) and redundancy (reconfigurability, parallelism, maintenance, scheduled and condition-based). These engineering strategies tend to be very broad, and sometimes, they overlap and are even interchangeable depending on the field where they are being used. Nevertheless, these strategies offer another perspective in which one can analyze other strategies, for instance, the ones in the BIASD tool. In the next section, we explain how we tried to use these strategies to classify the analogues and re-structure the tool at the same time.

3. ANALYSIS AND METHODOLOGY

Before deciding to work with LDA, different statistical methods were used to try to classify the database of the BIASD Tool. In order to better structure the data, we decided to start from bottom up – we are attempting to solve the problem by starting to sort out the analogues. The main idea is to create an algorithm that when the tool is fed any analogue (new or existing), the classification system that we are going to create allocates the analogue in certain category or group. Regarding the questions and their organization in the binary tree a meta-classification for these questions was proposed. Furthermore, two methods were used before LDA. The first one was a clustering approach in MATLAB; the code is included to show how the results did not have any relevant information for us to use. The second method tried was a dictionary search that dealt with existing engineering strategies used in Reliability Engineering; some of these strategies are over-specification, redundancy, maintenance, reconfigurability, etc.

3.1 Accessibility and Structure

First, we see at the meta-classification of the questions and accessibility of the BIASD tool. The tool started as a binary tree made in excel, and it is shown in figure 1. All four binary trees were constructed this way, and the Replace graph (which is the largest tree of them all) was composed of three different tiers because of its size. Having the tool in this format was not very accessible, so we decided to custom a different structure with the software GraphViz.

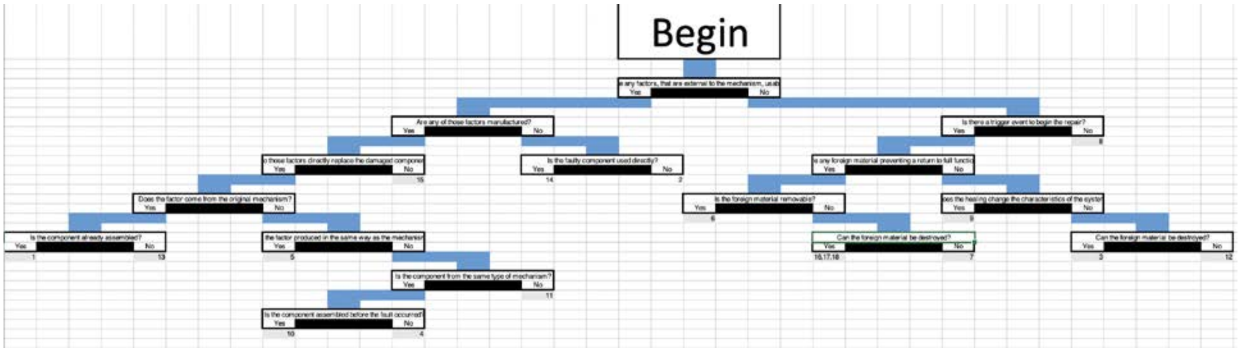

Begin
e any factors, that are external to the mechanism, usab
Yes
No
Are any of those factors manufactured?
Yes
No
Is there a trigger event to begin the repair?
Yes
No
8
o those factors directly replace the damaged componen
Yes
No
15
Is the faulty component used directly?
Yes
No
14
2
s any foreign material preventing a return to full functio
Yes
No
Does the factor come from the original mechanism?
Yes
No
Is the foreign material removable?
Yes
No
6
oes the healing change the characteristics of the syste
Yes
No
9
Is the component already assembled?
Yes
No
1
13
the factor produced in the same way as the mechanism
Yes
No
5
Can the foreign material be destroyed?
Yes
No
16,17,18
7
Can the foreign material be destroyed?
Yes
No
3
12
Is the component from the same type of mechanism?
Yes
No
11
Is the component assembled before the fault occurred?
Yes
No
10
4

Figure 1. First design of the Repair binary tree made in Excel.

Graphviz is a visualization software used represent structural information as diagrams or graphs. This way of graphing offers different benefits for applications in software engineering, database, visual interfaces, and web design. Furthermore, this software is open source for anyone to use. The software provides different layout programs and offers auxiliary tools, libraries and language bindings. The way the program works is by taking descriptions of graphs in a simple text language to make diagrams in several formats like PDF, JPEG, PNG, etc.

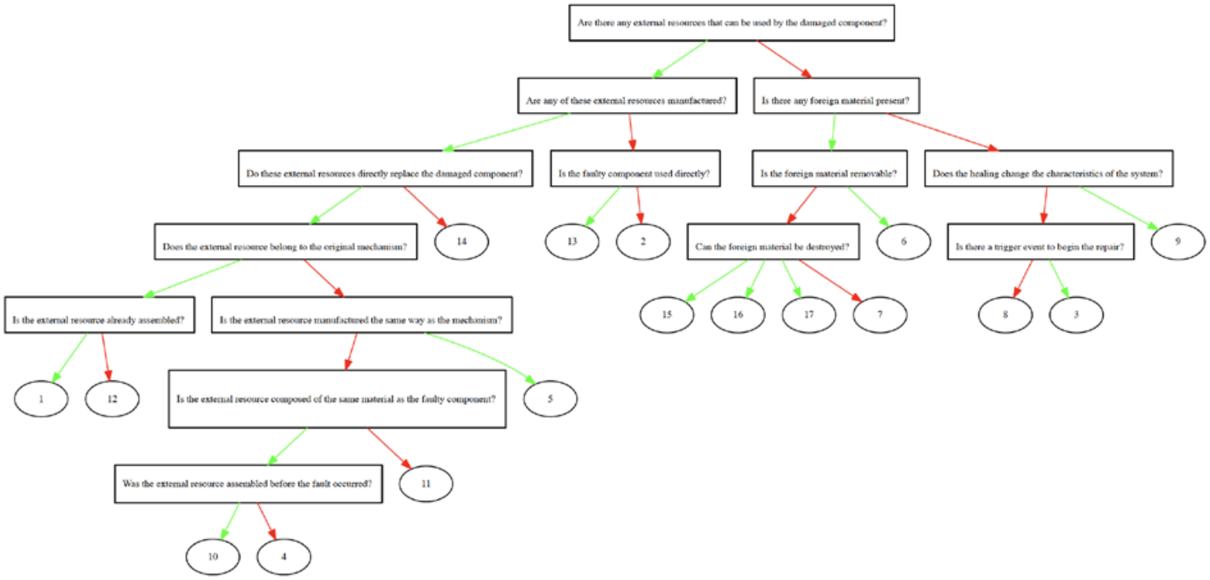

Are there any external resources that can be used by the damaged component?
Are any of these external resources manufactured?
Is there any foreign material present?
Do these external resources directly replace the damaged component?
Is the faulty component used directly?
Is the foreign material removable?
Does the healing change the characteristics of the system?
Does the external resource belong to the original mechanism?
14
13
2
Can the foreign material be destroyed?
6
Is there a trigger event to begin the repair?
9
Is the external resource already assembled?
Is the external resource manufactured the same way as the mechanism?
15
16
17
7
8
3
1
12
Is the external resource composed of the same material as the faulty component?
5
Was the external resource assembled before the fault occurred?
11
10
4

Figure 2. Last version of the Repair graph made with Graphviz.

Additionally, Graphviz has the ability to add hyperlinks to the nodes in the graphs, which we used to improve the functionality of the tool. Figure 2 shows the newest version of the same binary tree – Repair – in their Graphviz format.

Each leaf node of the tree in figure 2 has a hyperlink that takes the user to a website where the analogue is defined and the path that led to it is shown as well. Comparing graphs in figures 1 and 2, one can see which one is more user-friendly and offers more information about the analogues. In Appendix A, figure 8 shows the code to generate the Repair graph is shown. This code represents the Repair graph, which is fairly small; likewise, the Reprogram, Reconfigure and Replace have their own codes for their respective graphs. The Replace graph alone has more than 1900 lines in code.

Along with transforming the binary trees into Graphviz diagrams, a meta-classification for the questions was created to improve the navigation through the different trees. After analyzing the questions, six different categories were identified. These categories of strategy differentiation were based on energy, time, space, structure substance and information. The six different categories are as follows:

1. Energy Usage: all organisms require some type of energy to do work necessary for survival and reproduction. Within this category are sub-questions that further differentiate the observed strategies.
2. Downtime: downtime of a system is defined as the time in which the machine cannot perform any work. Systems experience downtime for different reasons such as maintenance, failure, machine modification or when the system is just not available. For this work, failure is going to be the main reason our system is going to experience downtime.
3. Material Removal: systems can dispose of materials that they do not longer use to either remove defective parts that have been already replaced or parts that will cause negative responses if not detached. In manufacturing, machining can result in the need to remove unwanted material from the system. There are special machines to dispose of these unwanted materials. Biological systems must use other mechanisms.
4. System Change: systems change by adding, removing or regrouping parts, or properties of parts or a material [24].
5. Foreign Assistance: autonomous engineering systems do not require any external agents or to assist them in their tasks. In this work, by foreign assistance, we mean that there is an agent helping the system through the process of returning to full functionality. Foreign material could refer to any material (alive or dead) that resides outside the original system. Some examples of these comprise the following: water (snow, rain, ice, etc.), air and any chemical reaction (such as oxidation) that the system utilizes to its benefit.

6. Resource Usage: resources are materials in the system's surroundings that are a result of environmental forces or manmade parts that are attached or stored in the system and designed to replace defective parts when necessary [2].

These six classifications assists differentiating the adaptation strategies observed in nature, and at the same time, re-structure the BIASD tool. Additionally, the classifications are meaningful in the sense that they help defining the significant characteristics of an adaption strategy. However, every category of this strategy is evenly useful for inspiring design solutions [2].

3.2 Dictionary Search Method

The second method that we tried to classify the analogues was a dictionary search that dealt with existing engineering strategies used in Reliability Engineering. The main objective of this approach was to generate a search of the existing analogues and putting them in groups. The groups would basically be the names of these strategies and their synonyms. For instance, if in analogue "x" appeared words such as margins, measures, etc., this analogue would be grouped into the "Over-Specification" group. The following list has all categories in this method:

1. Over-specification: it modifies dimensions and other parameters to decrease failure.
    a. Design Envelops: it uses margins and factors of safety to avoid failure.
    b. Robustness: this strategy uses techniques to reduce the system's performance fluctuations, for instance, the Taguchi approach.
    c. Reliability: it controls the probabilities of faults. This strategy helps the system function under designated conditions for a set period of time or number of cycles.
    d. Resiliency: it assesses a system's ability to recover after an adverse event.

2. Redundancy: it utilizes resources to increase the flexibility of a system while carrying more resources. Consequently, these systems are more complex which reduces the overall system's reliability.
    a. Reconfigurability: this strategy prepares the system for new objectives and customer requirements.
    b. Parallelism: a system has two or more elements executing the same function. If one is suddenly lost, this does not affect the overall performance of the system.
        i. Stand-by: with this strategy, a system has more components waiting to be used if a failure occurs. For instance, data-backups and spare tires.
        ii. Maintenance: it focuses on managing faults after they have occurred or before they happen.
    c. Scheduled: in this strategy, the time at which maintenance is needed is known in advance. The system's components are inspected at a given time to spot any safety problems.
    d. Condition-based: this strategy monitors, warns and provides plans of action in response to certain conditions [2].

After looking at the different strategies in reliability engineering, one can notice that most of the strategies share a similar definition. Furthermore, words such as reliability, resiliency, robustness and flexibility and adaptation share similar meanings that overlap and can be at times interchangeable depending on what field one is working on. These drawbacks made it difficult for us to generate a classification system for the analogues. Close to a third of the analogues did not contain some of these words at all, so they did not fall under any of the categories, and therefore, could not be classified. After this method, we moved to the next one: K-means clustering.

3.3 K-means Clustering Approach

Before explaining more of K-means clustering, we will describe the process that was followed in the MATLAB code. Our data set was comprised of 161 analogues sorted as follows: 18 in the Repair tree, 21 in the Reprogram tree, 16 in the Reconfigure tree, and 106 in the Replace tree. This list was modified in MATLAB, and we assigned a number to the name of each analogue for calculation purposes; also, a dictionary was created with all words that appeared in the data set. After that, the long list was separated into individual analogues and a counter was used to count the frequency of words in the whole data set. Then the code looked at the list of all words and populated a matrix counting the number of times the word in the dictionary appeared in each analogue. For instance, if word two in the dictionary did not appear in analogue number 14, then the value in the matrix would be zero. After this, one chooses how many clusters to divide the data into and plot the results. Figure 3 shows the results for a K-means algorithm of four clusters. The first interesting observation about this graph is how there is data on the negative *x* axis. This result is not desirable since K-means plots are based on measurements of distance. Also, every silhouette graph generated had a cluster that only contained one analogue. A silhouette is based on the comparison of the tightness of clusters and separation. Additionally, silhouettes show which objects lie well within their cluster, and which ones are in between clusters. Silhouettes also offer the advantage that they depend on the actual partition of the objects, and not on the clustering algorithm that was used to obtain it [25]. According to Rousseeuw, in his work, silhouettes best represent clusters when they are as wide, or as dark, as possible.

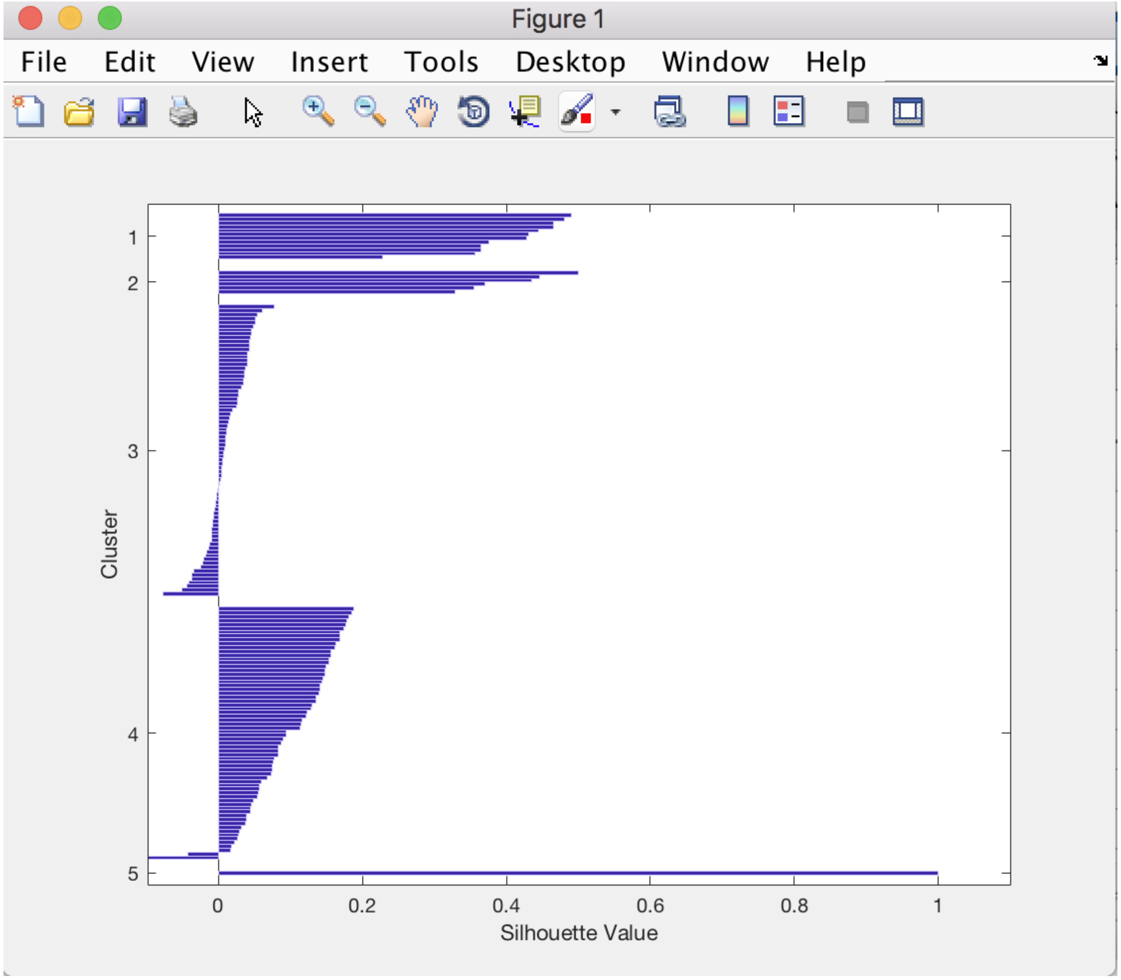


Figure 3. K-means algorithm clustering method (four clusters).

After trying with four clusters, the same procedure was done for five clusters. The same type of results was found. We concluded that as the clusters increased, all the clusters were grouped except for the last one that was always composed of only one analogue. Figure 4 depicts the result when *K* is five.

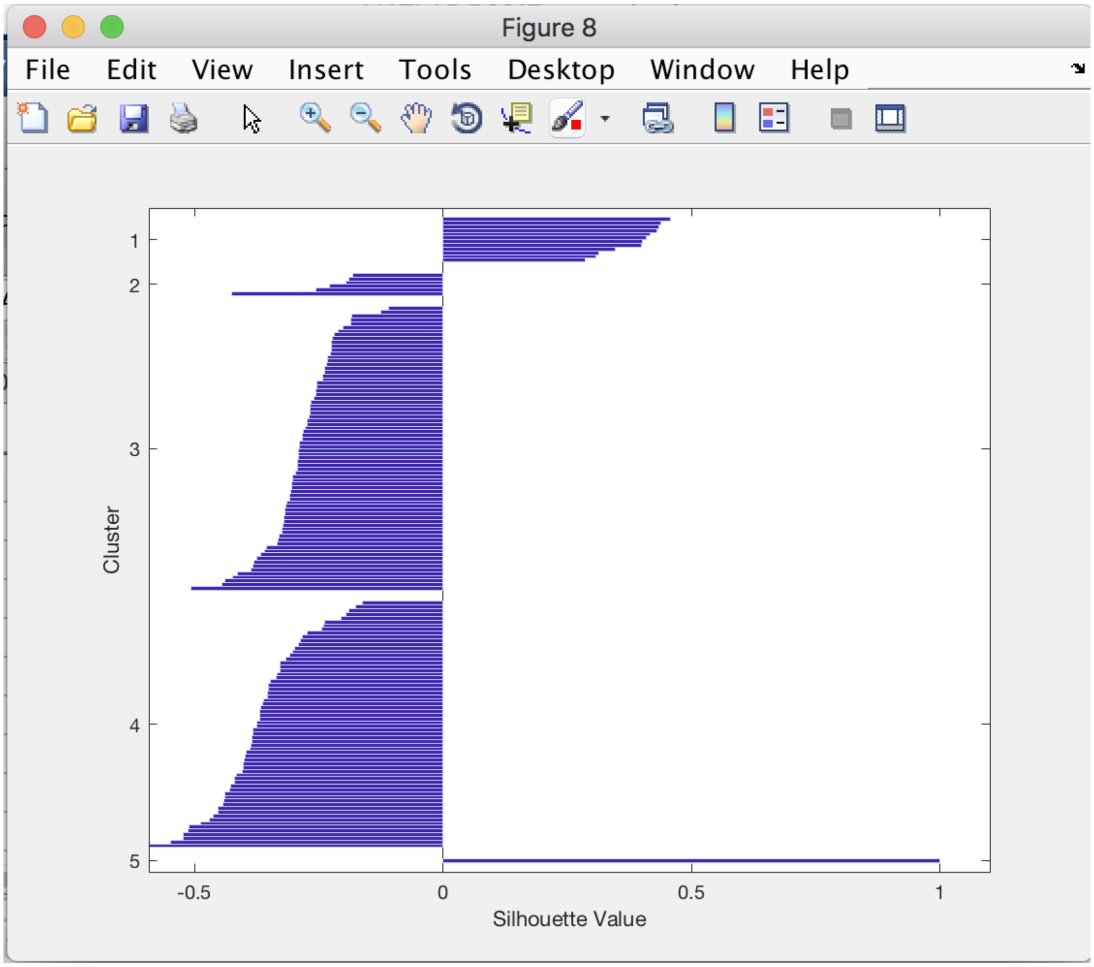


Figure 4. K-means algorithm clustering method (five clusters).

No valuable information was found from this experiment because some of them were showing in the negative *x* axis; if different points have low values or even negative values, then the clustering approach probably has too many or too few clusters [26]. In our case, if we look at figure 3 (four clusters) the silhouettes are not as wide as we want them, although we have only a few negative values. This means, we need more clusters. Once we move up one cluster (figure 4), most of the analogues move over to the negative axis, even when our silhouettes are wider. The same issue kept happening as we increased clusters. Moreover, there was always one cluster comprised of only one analogue, the one at the very bottom of the graph. In Appendix A, figure 9 shows the MATLAB code to help the reader understand the coding process.

One of the challenges we encountered when working on this method was determining the right number of clusters to obtain the best results. In cluster analysis, it is crucial to use efficient methods to determine the number of clusters. Similar work has been proposed by many researchers to deal with this number-of-clusters problem. Refer to background section. Our elbow curve is shown in Figure 5.

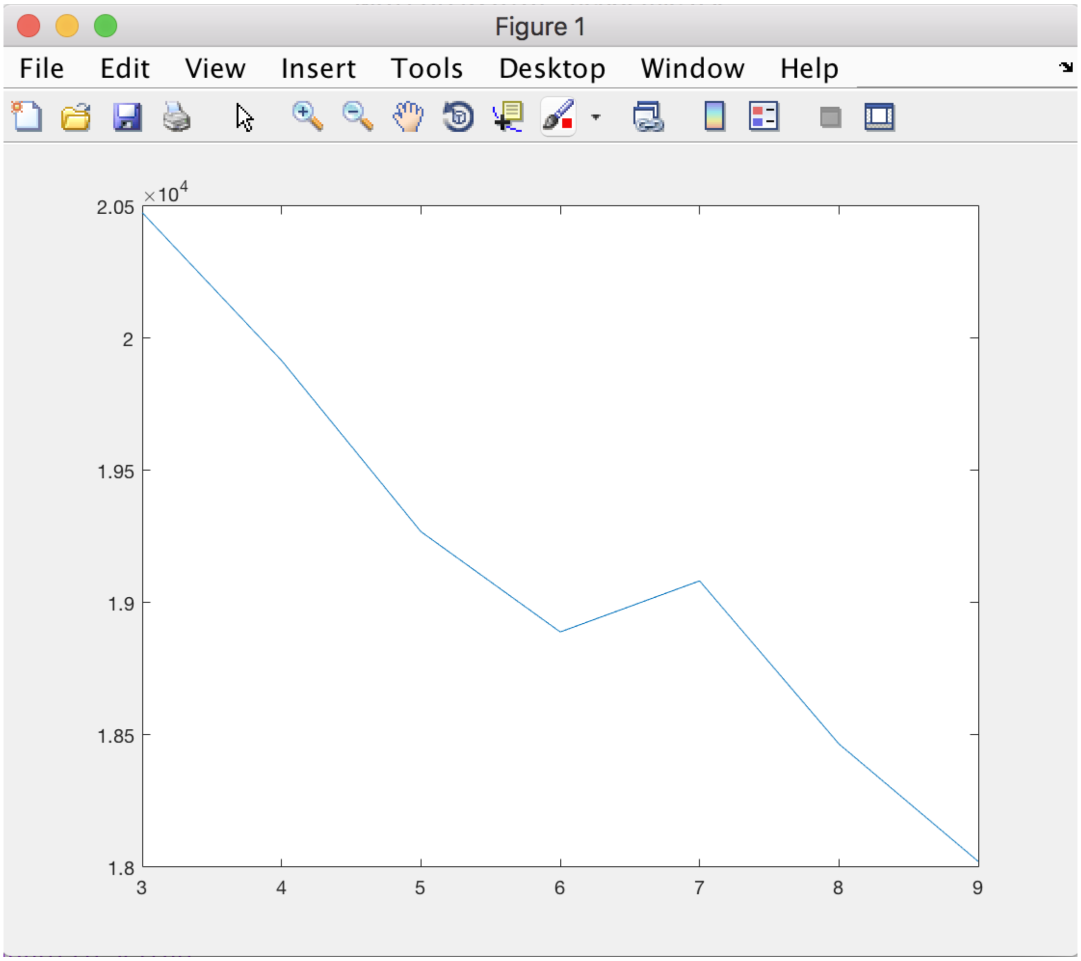


Figure 5. Elbow curve for the K-means clustering method with four clusters.

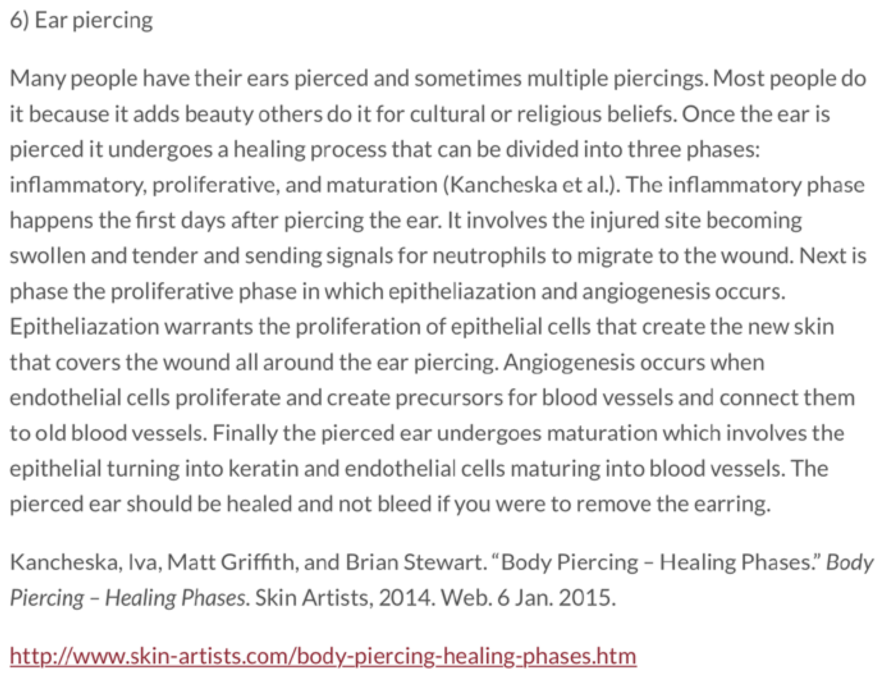
From all the elbow curves generated for different number of clusters, this was the one that fluctuated the least. However, from this graph one cannot say if six or seven are the most accurate number of clusters for this data since the graph does not converge on a value; instead it keeps decreasing as the numbers of clusters increases. From the elbow curve and the clusters, we obtained, we concluded that this method did not contribute any significant value to our main goal of classifying the analogues.

4. DATA COLLECTION

4.1 Extracting the text from BIASD tool

According to Algortithmia, in natural language, LDA generates topics comprised of words and automatically discovers topics in text documents. However, the topics are only the group of words; it does not assign a name to each topic. The LDA model we use for this work treats each document as a mixture of various topics and that each word in the document belongs to one of the topics in the document. The algorithm used was first present by David Blei, Andrew Ng and Michael Jordan. Furthermore, LDA works by calculating the probability that a word belongs to a topic. For instance, if we analyze the example given in section 2.3 about classifying engineering classes, in Group B, the word “circuit” would have a higher probability than the word “power” because, intuitively, the word “circuit” is more closely related to electrical engineering than the word “power.” The word “power” appears in both, Group A and Group B. In a sense, the word “power” is more general when compared to the rest of the words, which is why it will have a lower probability within the group.

In this work, the algorithm we use takes an object with an array of strings. The set of documents we use are those that describe every analogue in the BIASD Tool, which has a total of 161 analogues. Figure 6 shows an example from the BIASD Tool website. This is an example of what the analogues look like, and how long (on average) the text for each is. We copied and pasted each text, without the name and link, of each analogue into a word document, so they would not affect the results of the LDA model. The names of authors within each analogue was removed as well.



6) Ear piercing

Many people have their ears pierced and sometimes multiple piercings. Most people do it because it adds beauty others do it for cultural or religious beliefs. Once the ear is pierced it undergoes a healing process that can be divided into three phases: inflammatory, proliferative, and maturation (Kancheska et al.). The inflammatory phase happens the first days after piercing the ear. It involves the injured site becoming swollen and tender and sending signals for neutrophils to migrate to the wound. Next is phase the proliferative phase in which epitheliazation and angiogenesis occurs. Epitheliazation warrants the proliferation of epithelial cells that create the new skin that covers the wound all around the ear piercing. Angiogenesis occurs when endothelial cells proliferate and create precursors for blood vessels and connect them to old blood vessels. Finally the pierced ear undergoes maturation which involves the epithelial turning into keratin and endothelial cells maturing into blood vessels. The pierced ear should be healed and not bleed if you were to remove the earring.

Kancheska, Iva, Matt Griffith, and Brian Stewart. "Body Piercing – Healing Phases." *Body Piercing – Healing Phases*. Skin Artists, 2014. Web. 6 Jan. 2015.

http://www.skin-artists.com/body-piercing-healing-phases.htm

Figure 6. Analogue extracted from the Repair binary tree of the BIASD tool.

4.2 Preparing inputs in LDA model

After extracting all the text from the binary trees, we prepared the training and testing data sets (refer to section 2.2). We decided to make the training data bigger for accuracy purposes. Our training data was comprised of 80% of the analogues, while our testing data contained the remaining 20%. Since each binary tree has a different number of analogues (Replace 106, Reprogram: 21, Repair: 18, Reconfigure: 16), we decided to make the training and testing data a mixture of analogues from all four binary trees, so the results would not reflect a group of words from a single binary tree, but from all of them at the same time. Therefore, the training data contained a total of 129 analogues (80% of all analogues), 85 from the Replace tree, 17 from the Reprogram tree, 14 from the Repair tree and 13 from the Reprogram. Similarly, the testing data contained the remaining 32 analogues (20% of the complete sample), 21 from the Replace tree, 4 from the Reprogram tree, 4 from the Repair tree and 3 from the Reconfigure tree.

4.3 Running the LDA model

The next step in our experiment was to run the simulations to generate the topics and their respective words. In Appendix B, we have the code for one of the combinations to generate the topics and words in the experiment. A total of 15 combinations of words and topics were ran in the LDA model. Table 1 shows the different topic-word combinations. For each combination, the analogues in the training and testing data were randomized to see if the same topics were generated even when different analogues were used as testing data. After running the model 15 times, the perplexity for each combination was calculated. In the next section, we discuss the results of both topic-word combinations and perplexity of the model.

Table 1. Combinations of number of topics and words in the LDA model.

Combination	Number of topics	Number of words
1	1	1
2	2	5
3	2	2
4	2	10
5	4	4
6	5	4
7	5	6
8	4	6
9	4	8
10	6	4
11	4	10
12	8	4
13	6	8
14	6	6
15	8	8

5. RESULTS AND DISCUSSION

After running the LDA model 15 times, we ended with different topics and words for each topic. At the same time, the model generated each word's probability of belonging to each of the topics. Table 2 shows the topics and words generated by one of the combinations. The rest of the tables with topics and words generated are in Appendix B. Table 3 depicts the probabilities, of belonging to each topic, for each word. The tables for the remaining combinations can be found in Appendix B.

Table 2. Combination five containing four topics and four words.

Topics	Words			
1	Epithelial	Kidney	Size	Stage
2	Cells	Differentiate	Form	Proliferate
3	Body	Fish	Salt	Water
4	Arm	Blood	Skin	Tail

Table 3. Probabilities of documents belonging to each topic of combination five.

Probabilities			
Topic 1	Topic 2	Topic 3	Topic 4
0	0.3333	0.3333	0.3333
0	0.5	0.5	0
0	0.6667	0.1667	0.1667
0.25	0.5	0.25	0
0	0	0.3333	0.6667
0	0	0.4	0.6
0	0.5	0	0.5
0	0	0.8333	0.1667
0	0	0.5	0.5

Table 3 (Cont.) Probabilities of documents belonging to each topic of combination five.

Topic 1	Topic 2	Topic 3	Topic 4
0	0.1667	0.8333	0
0.3333	0.4444	0.2222	0
0.04	0.92	0	0.04
0.0909	0.9091	0	0
0.3333	0.3333	0.3333	0
0.4375	0.5	0.0625	0
0.2353	0.6471	0.0588	0.0588
0.3529	0.6471	0	0
0	0.6296	0.1111	0.2593
0	0.83333	0.1667	0
0	0.52	0.24	0.24
0.1071	0.5357	0.0357	0.3214
0	0.7727	0.0909	0.1364
0.0667	0.5333	0	0.4
0.0667	0.9333	0	0
0	0.9167	0.0833	0
0	1	0	0
0	1	0	0
0	1	0	0
0	0.8333	0.1667	0
0	1	0	0
0.125	0.75	0.125	0

In table 3, each column represents a topic generated by the LDA model. The rows represent the documents chosen for that combination, i.e. the analogues chosen for this particular testing data. As described in section 4.2, the testing data contains 32 analogues, therefore, the 32 rows in the table 32. Similarly, each table with the probabilities of each combination, which can be found in Appendix B, will have 32 rows. Except for combination one; since it is only one word and one topic, the probability was 1, which made the perplexity value equal to zero.

Following the calculations of all probabilities of all combinations, we found the perplexity for each combination. Table 4 shows the values found for perplexity.

Table 4. Perplexity calculations for each combination.

Words	Topics	Perplexity
4	4	238126683.4
4	5	72994566045
8	4	2.67497E+12
8	8	2.48053E+19
4	8	8.62043E+13
6	6	8.7481E+18
6	4	5.77899E+11
1	1	0
10	2	127315.505
10	4	1.0905E+13
8	6	1.75344E+17
4	6	6.24377E+12
2	2	1190.137314
5	2	685.3862756
6	5	3.28896E+11

After calculating the perplexity for each combination, we looked for the best combination of words and topics by using different contributing factors. The first one was the perplexity value. Then, we plotted the perplexity values to see how it behaves in our model. Figure 7 shows how perplexity increases for the combinations we tried.

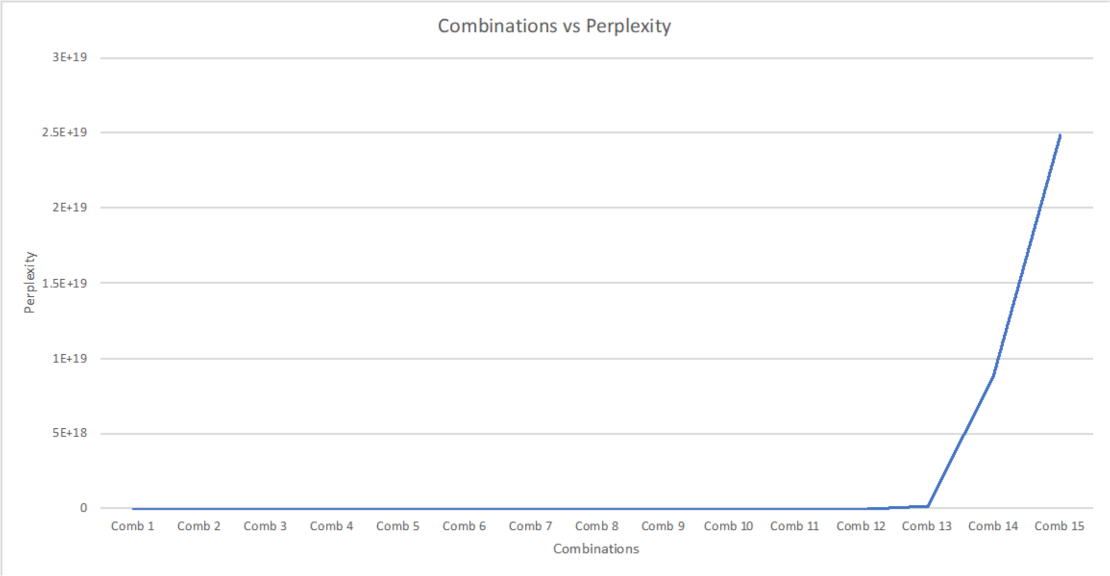


Figure 7. Increasing value of perplexity for our combinations.

We are interested in seeing how the perplexity increased. Right before the elbow, the value of perplexity stays low, and right after the elbow, perplexity increases by a lot for the points measured in this experiment specifically. According to Zhao et. all, the value of perplexity may generate meaningful results in some cases, but it is not always stable, and the results vary with the selected seeds even from the same data set. Although the best number of topics is unknown, different numbers of topics will likely result in different structuring of the corpus [16].

As mentioned before, since perplexity measures the value entropy of the data being analyzed, we want a low value to show that the model has a high accuracy. At the same time, we try to avoid a perplexity value that is too low. From the results, we can conclude that the more topics we have, the lower of perplexity, but too many topics can lead to complications as well. An insufficient number of topics could render an LDA model that is too uneven or hard to identify accurate classifiers for the analogues. On the other hand, an excessive number of topics could result in a model that is too complex, which makes its interpretation and validation very difficult [16].

From this analysis, we conclude that the best combination for topics and words is where the value of perplexity is equal to 8.62043E+13 – combination 12, which is right before the values of perplexity increase significantly. This value has a combination of eight topics and four words. Table 4 shows the topics and words generated for this combination; table 5 shows the probabilities for the testing data for this combination. Since a balance needs to be found, and it is too difficult to give topical names to those groups with eight words in each topic, some of the topics from the model chosen will be considered as subtopics. Our goal is to make a good model with the simplest topics. Furthermore, if only three analogues out of 161 belong to one topic, it is not a good thing. What we want are more explanatory of the model as a whole. In table 5, topics 3, 4, 5, 6, 7 and 8 are, to some extent, easier to name than topics 1 and 2. We can see how broad or specific the topics are by trying to name them. Topic 3 can be related to cell regeneration, whereas topics 4 and 5 are closely related to marine organisms. Some of the analogues related to topics 4 and 5 talk about fish being able to swap their reproductive organs to either have offspring or escape from predators. One of them have more relevance than the other; we mention which one it is when we talk about the probability tables. When coming up with names for the topics, one does not necessarily have to use every word. Moreover, topic 6 is related to limb loss or regrow. Lastly, topic 7 is considered a subtopic of 8 because topic 8 has to do with marine organisms' abilities to survive in different environment (salt and fresh water). Topic 7 is considered a subtopic because is more specific because we know from the analogues that species use their kidneys to regulate the amount water and salt in their bodies – osmoregulation. The ability to analyze each group deeper and know which of them is referring to osmoregulation, for instance, has to do with the user’s knowledge of this document and all its analogues; otherwise, it would be really hard to infer that information. As mentioned before, we explain one is considered a topic and why the other a subtopic in the following paragraphs.

Table 5. Combination fourteen containing eight topics and four words.

Topics	Words			
1	complex	food	lens	structure
2	exoskeleton	hair	smaller	system
3	cells	differentiate	form	proliferate
4	animal	change	sea	survive
5	ability	body	freshwater	waters
6	amputated	arm	grow	legs
7	cord	kidney	migrate	sea
8	environment	fish	salt	water

Table 6. Probabilities of documents belonging to each topic of combination fourteen.

Probabilities							
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
0	0.4	0.4	0	0	0	0	0.2
0	0	0.3333	0.3333	0	0.1667	0	0.1667
0	0	0.3333	0.3333	0	0.3333	0	0
0	0	0.3333	0.3333	0.1667	0	0	0.1667
0	0	0	0.1071	0	0	0.7143	0.1786
0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	1
0	0.0714	0	0.0714	0	0.1429	0.7143	0
0	0	0	0.2	0	0	0.8	0
0	0.75	0	0.25	0	0	0	0
0	0	0.3684	0.1053	0	0.0526	0.2632	0.2105
0	0	0.4	0.1	0	0.05	0.15	0.3
0	0.1176	0.1765	0.0588	0.0588	0.4118	0.1176	0.0588
0	0	0.3636	0.0606	0	0.0303	0.0909	0.4545
0.0455	0	0.3636	0.0455	0	0.0909	0	0.4545
0	0	0.4231	0.0769	0.0385	0.0769	0	0.3846
0	0	0.7143	0	0	0	0	0.2857
0.0385	0	0.3077	0.0769	0	0.0769	0.1923	0.3077

Table 6 (Cont.) Probabilities of documents belonging to each topic of combination fourteen.

Probabilities							
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
0	0	0.4667	0	0	0.2667	0	0.2667
0	0	0.5333	0.0667	0	0	0	0.4
0.0606	0.0606	0.6061	0.0606	0	0.0303	0	0.1818
0.2667	0	0.4	0	0	0.0667	0	0.2667
0	0	0.4211	0	0	0	0.1579	0.4211
0.027	0.027	0.5676	0.0541	0.027	0.027	0	0.2703
0	0.0909	0.4545	0.0909	0	0	0	0.3636
0.12	0	0.36	0.04	0	0.04	0.04	0.4
0	0	0.5667	0	0	0.0667	0	0.3667
0.0278	0.1111	0.5556	0	0	0.0556	0	0.25
0	0	0.5714	0.0714	0	0.0714	0	0.2857
0	0.0357	0.6071	0	0	0.0714	0	0.2857
0	0	0.5714	0.0286	0.0286	0	0	0.3714

Another contributing factor that helps deciding what topics of our result become subtopics are the zeros on the probability tables. As shown in table 6, most probability values for topic 1, 2 and 5 are zeros, which means that for the testing data, the analogues are not likely to be related to those topics. In a way, the zeros mean there are way too many topics. This is why topic 5 is a subtopic of topic 4 and topic 7 is a subtopic of topic 8. We can use perplexity to eliminate some choices, i.e. the groups of topics that have perplexity too high. At the same time, we could lower perplexity a little bit in other ones by increasing topics, but we do not gain a lot in doing so. Since we know the data very well and what words might be used more than others, we can tell which topics are not good for explaining the documents as a whole.

Table 6 as well as other probability tables in Appendix B have blank spaces. These blank spaces mean that nothing related to that specific document was found or related to the topics the LDA model generated. This is a result of the changes made to some of the analogues of the BIASD tool. After the first author of the tool completed the data, we took the initiative to add new analogues and eliminate some of them; this happened during the improvement of the accessibility and structure of the tool. Another confounding factor that explains this is the fact that the original analogues were written and researched by certain people, while the new analogues were written by a different person, so the pattern of speech used to write the analogues were different and noticed by the LDA model.

With the data we gathered we can structure our data into eight topics and four words; however, three of those topics are minor in importance and, therefore, considered subtopics. Five of the eight are more explanatory and much more useful. This model does not have a large perplexity and captures the data well. However, we focus on topics that are broadly applicable rather than a subtopic. Biology is a science highly structured with topics and subtopics, but LDA will not capture subtopics, but hierarchical modeling will.

Now, we have a new structure for the database, we ran the LDA model once again with all 161 analogues to generate 8 topics and 4 words. Table 7 shows the final topics and words used to classify our database.

Table 7. Final topics and words for new database classification.

Topics	Words			
1	kidney	membrane	survive	tubular
2	northern	skin	teeth	zebrafish
3	arm	body	layer	skin
4	fish	salt	sea	water
5	cells	differentiate	form	proliferate
6	cells	epithelial	kidney	signals
7	cord	grow	muscles	spinal
8	leg	lens	limb	lost

If we look at topic number 1 and 2, their words start becoming very specific, for instance, zebrafish. There are 161 examples in the whole model and most of them are animals, so having a specific kind of fish is too specific for our purposes. Topic 3 can be named “coating” or “organisms with protective coatings.” Topic 4 can be named “marine organisms.” We can give topic 5 the name “cellular recovery.” Then, topic 6 can be a subtopic of number 5, while topic 7 can be titled “organisms dealing with organ (limbs) losses.” Topic number 7 is not technically a subtopic but is the one topic with most zeros for all the data.

Table 8. Final classification for analogues organized by probability (topics 1 through 4).

Probabilities							
Number	Topic 1	Number	Topic 2	Number	Topic 3	Number	Topic 4
21	1	23	0.67	62	1	44	1
32	1	10	0.5	40	0.86	45	1
34	1	157	0.5	15	0.67	49	1
38	1	7	0.31	24	0.67	46	0.86
51	1	52	0.31	35	0.56	48	0.83
11	0.71	12	0.26	7	0.54	28	0.72
18	0.5	64	0.23	10	0.5	50	0.71
25	0.39	35	0.22	36	0.5	26	0.68
134	0.29	13	0.2	109	0.44	29	0.67
88	0.28	79	0.2	1	0.43	47	0.67
9	0.27	153	0.18	14	0.43	31	0.64
90	0.25	8	0.17	37	0.43	27	0.56
117	0.24	36	0.17	41	0.43	30	0.56
123	0.23	54	0.17	53	0.43	93	0.53
110	0.22	60	0.15	12	0.4	55	0.46
29	0.21	1	0.14	68	0.37	53	0.43
131	0.21	2	0.14	71	0.37	60	0.4
93	0.2	37	0.14	100	0.36	8	0.33
145	0.2	89	0.14	102	0.36	23	0.33
141	0.19	76	0.13	8	0.33	39	0.33
144	0.19	80	0.12	16	0.33	75	0.32
147	0.19	104	0.12	39	0.33	42	0.31
31	0.18	102	0.11	54	0.33	37	0.29
96	0.18	73	0.1	76	0.33	90	0.25
16	0.17	119	0.1	43	0.31	119	0.25
36	0.17	136	0.1	52	0.31	65	0.18
30	0.16	69	0.09	13	0.3	78	0.18
55	0.15	120	0.09	153	0.3	100	0.18
126	0.15	6	0.08	92	0.29	104	0.18
37	0.14	27	0.08	9	0.27	18	0.17
46	0.14	53	0.08	65	0.27	36	0.17
50	0.14	61	0.08	161	0.27	54	0.17

Table 8 (Cont.) Final classification for analogues organized by probability (topics 1 through 4).

Probabilities							
Number	Topic 1	Number	Topic 2	Number	Topic 3	Number	Topic 4
35	0.11	77	0.08	64	0.26	66	0.16
13	0.1	97	0.08	42	0.23	77	0.16
26	0.09	101	0.07	70	0.23	144	0.16
27	0.08	129	0.06	96	0.23	57	0.15
28	0.08	122	0.05	137	0.22	126	0.15
52	0.08	138	0.03	82	0.21	102	0.14
53	0.08	142	0.03	91	0.21	120	0.14
4	0.06	132	0.02	27	0.2	159	0.14
95	0.06	150	0.02	98	0.2	105	0.13
146	0.05	3	0	148	0.2	95	0.12
108	0.04	4	0	4	0.19	97	0.12
121	0.04	5	0	59	0.18	103	0.12
124	0.04	9	0	95	0.18	131	0.12
143	0.04	11	0	48	0.17	150	0.12
149	0.04	14	0	155	0.17	152	0.12
72	0.03	15	0	61	0.16	160	0.12
98	0.03	16	0	105	0.16	35	0.11
1	0	17	0	2	0.14	108	0.11
2	0	18	0	26	0.14	121	0.11
3	0	19	0	50	0.14	141	0.11
5	0	20	0	89	0.14	71	0.1
6	0	21	0	86	0.13	86	0.1
7	0	24	0	17	0.12	9	0.09
8	0	25	0	30	0.12	25	0.09
10	0	26	0	150	0.12	52	0.08
12	0	28	0	107	0.11	80	0.08
14	0	29	0	129	0.11	122	0.08
15	0	30	0	73	0.1	124	0.08
17	0	31	0	136	0.1	156	0.08
19	0	32	0	69	0.09	101	0.07
20	0	34	0	118	0.09	110	0.07
23	0	38	0	146	0.09	113	0.07

Table 8 (Cont.) Final classification for analogues organized by probability (topics 1 through 4).

Probabilities							
Number	Topic 1	Number	Topic 2	Number	Topic 3	Number	Topic 4
24	0	39	0	6	0.08	64	0.06
39	0	40	0	28	0.08	70	0.06
40	0	41	0	74	0.08	148	0.06
41	0	42	0	80	0.08	157	0.06
42	0	43	0	124	0.08	92	0.05
43	0	44	0	138	0.08	59	0.04
44	0	45	0	142	0.08	69	0.04
45	0	46	0	149	0.08	82	0.04
47	0	47	0	5	0.07	96	0.04
48	0	48	0	31	0.07	155	0.04
49	0	49	0	47	0.07	56	0.03
54	0	50	0	101	0.07	83	0.03
56	0	51	0	128	0.07	88	0.03
57	0	55	0	132	0.07	94	0.03
58	0	56	0	83	0.06	115	0.03
59	0	57	0	94	0.06	116	0.03
60	0	58	0	104	0.06	129	0.03
61	0	59	0	130	0.06	134	0.03
62	0	62	0	152	0.06	135	0.03
63	0	63	0	60	0.05	145	0.03
64	0	65	0	67	0.05	1	0
65	0	66	0	119	0.05	2	0
66	0	67	0	120	0.05	3	0
67	0	68	0	66	0.04	4	0
68	0	70	0	77	0.04	5	0
69	0	71	0	97	0.04	6	0
70	0	72	0	111	0.04	7	0
71	0	74	0	143	0.04	10	0
73	0	75	0	56	0.03	11	0
74	0	78	0	85	0.03	12	0
75	0	81	0	113	0.03	13	0
76	0	82	0	122	0.03	14	0

Table 8 (Cont.) Final classification for analogues organized by probability (topics 1 through 4).

Probabilities							
Number	Topic 1	Number	Topic 2	Number	Topic 3	Number	Topic 4
77	0	83	0	135	0.03	15	0
78	0	84	0	151	0.03	16	0
79	0	85	0	3	0	17	0
80	0	86	0	11	0	19	0
81	0	87	0	18	0	20	0
82	0	88	0	19	0	21	0
83	0	90	0	20	0	24	0
84	0	91	0	21	0	32	0
85	0	92	0	23	0	34	0
86	0	93	0	25	0	38	0
87	0	94	0	29	0	40	0
89	0	95	0	32	0	41	0
91	0	96	0	34	0	43	0
92	0	98	0	38	0	51	0
94	0	99	0	44	0	58	0
97	0	100	0	45	0	61	0
99	0	103	0	46	0	62	0
100	0	105	0	49	0	63	0
101	0	106	0	51	0	67	0
102	0	107	0	55	0	68	0
103	0	108	0	57	0	72	0
104	0	109	0	58	0	73	0
105	0	110	0	63	0	74	0
106	0	111	0	72	0	76	0
107	0	112	0	75	0	79	0
109	0	113	0	78	0	81	0
111	0	114	0	79	0	84	0
112	0	115	0	81	0	85	0
113	0	116	0	84	0	87	0
114	0	117	0	87	0	89	0
115	0	118	0	88	0	91	0
116	0	121	0	90	0	98	0

Table 8 (Cont.) Final classification for analogues organized by probability (topics 1 through 4).

Probabilities							
Number	Topic 1	Number	Topic 2	Number	Topic 3	Number	Topic 4
118	0	123	0	93	0	99	0
119	0	124	0	99	0	106	0
120	0	125	0	103	0	107	0
122	0	126	0	106	0	109	0
125	0	127	0	108	0	111	0
127	0	128	0	110	0	112	0
128	0	130	0	112	0	114	0
129	0	131	0	114	0	117	0
130	0	133	0	115	0	118	0
132	0	134	0	116	0	123	0
133	0	135	0	117	0	125	0
135	0	137	0	121	0	127	0
136	0	139	0	123	0	128	0
137	0	140	0	125	0	130	0
138	0	141	0	126	0	132	0
139	0	143	0	127	0	133	0
140	0	144	0	131	0	136	0
142	0	145	0	133	0	137	0
148	0	146	0	134	0	138	0
150	0	147	0	139	0	139	0
151	0	148	0	140	0	140	0
152	0	149	0	141	0	142	0
153	0	151	0	144	0	143	0
154	0	152	0	145	0	146	0
155	0	154	0	147	0	147	0
156	0	155	0	154	0	149	0
157	0	156	0	156	0	151	0
158	0	158	0	157	0	153	0
159	0	159	0	158	0	154	0
160	0	160	0	159	0	158	0
161	0	161	0	160	0	161	0
22		22		22		22	
33		33		33		33	

We organized the data in a way to show all documents with their probability belonging to each topic. The column called “number” is the number of the analogue. In Appendix B the list of 161 analogues is shown. Table 8 and table 36 are the final classification for the analogues. The whole table has been split into two tables for space reasons.

Table 8 shows the first 4 topics, and table 36 shows the probability distribution of the dataset from topics 5 to 8. Table 36 can be found in Appendix B. Table 37 (in Appendix B) shows the probability distribution with respect to the analogue number. One thing we can mention about these table are how analogues 22 and 33 do not appear at all in the probability table. As mentioned before one of the reasons might be because of the difference in writing styles between those who have contributed to add data to the BIASD tool. This new classification allows designers a different approach when looking for the analogues. Instead of using the binary trees, they can use the topics that were generated to find the analogues; at the same time, the can use the ones with highest probabilities since they contribute more to the overall document. Topic 7 was the one with most zeros, so it did not contribute a lot of information with respect to the topics generated. As expected, topic 5 and subtopic 6 are the ones with the least zeros in their probabilities. We can justify this by saying that 106 out of 161 of the analogues in the BIASD tool belonged to the Replace binary tree. Almost all of the analogues in this tree talked about cell regeneration.

6. CONCLUSION

In this work, we analyzed a database of 161 analogues of biological adaption from nature. These analogues were part of an existing tool called the Biologically Inspired Adaptive Systems Design (BIASD) tool. In this research, we generated a new classification for the analogues within the BIASD tool. We started by modifying the binary trees and classifying their questions since a lot of them did not follow a pattern or were set up in a random way. Our goal was to reduce the subjectivity of this method and classify the analogues in a more robust, objective way. We tried different methods such a search dictionary, K-means algorithm and finally used a combination LDA modelling from machine learning and perplexity from statistics to generate a new structure for the tool. Our LDA model helped us generate topics and words to classify the data, and the perplexity analysis helped us select the best combination for topics and words. The final result was a table divided into 8 different topics that were arranged in two different ways. One of them ordered by analogue number and the other one by the value of their probabilities. With this work, we significantly improved the BIASD tool. A way in which we envision the BIASD tool, instead of binary trees, is an ontology. Now that we have defined semantic topics, it is possible to create an ontology in a more objective way.

7. REFERENCES

[1] Malshe, A., Rajurkar, K., Samant, A., Hansen, H. N., Bapat, S., and Jiang, W., 2013. “Bioinspired functional surfaces for advanced applications”. CIRP Annals-Manufacturing Technology, 62(2), pp. 607–628.

[2] Arroyo, Marvin, Nicholas Huisman, and David C. Jensen. "Exploring Natural Strategies for Bio-Inspired Fault Adaptive Systems Design." *Journal of Mechanical Design* 140.9 (2018): 091101.

[3] Jensen, D. C., and Huisman, N., 2015. “Biologically inspired fault adaptive strategies for engineered systems”. In DS 80-2 Proceedings of the 20th International Conference on Engineering Design (ICED 15) Vol 2: Design Theory and Research Methodology Design Processes, Milan, Italy, 27-30.07. 15.

[4] Bird, R., and Wadler, P., 1988. Introduction to functional programming, Vol. 1. Prentice Hall New York.

[5] Baştanlar, Yalin, and Mustafa Özuysal. "Introduction to machine learning." *miRNomics: MicroRNA Biology and Computational Analysis*. Humana Press, Totowa, NJ, 2014. 105-128.

[6] Dietterich, Thomas G. "Ensemble methods in machine learning." *International workshop on multiple classifier systems*. Springer, Berlin, Heidelberg, 2000.

[7] Nasrabadi, Nasser M. "Pattern recognition and machine learning." *Journal of electronic imaging* 16.4 (2007): 049901.

[8] Caliński, Tadeusz, and Jerzy Harabasz. "A dendrite method for cluster analysis." *Communications in Statistics-theory and Methods* 3.1 (1974): 1-27.

[9] Hartigan, John A. "Clustering algorithms." (1975).

[10] Krzanowski, Wojtek J., and Y. T. Lai. "A criterion for determining the number of groups in a data set using sum-of-squares clustering." *Biometrics* (1988): 23-34.

[11] Kaufman, Leonard, and Peter J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons, 2009.

[12] Tibshirani, Robert, Guenther Walther, and Trevor Hastie. "Estimating the number of clusters in a data set via the gap statistic." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2 (2001): 411-423.

[13] Dudoit, Sandrine, and Jane Fridlyand. "A prediction-based resampling method for estimating the number of clusters in a dataset." *Genome biology* 3.7 (2002): research0036-1.

[14] Sugar, Catherine A., and Gareth M. James. "Finding the number of clusters in a dataset: An information-theoretic approach." *Journal of the American Statistical Association*98.463 (2003): 750-763.

[15] Yan, Mingjin, and Keying Ye. "Determining the number of clusters using the weighted gap statistic." *Biometrics* 63.4 (2007): 1031-1037.

[16] Zhao, Weizhong, et al. "A heuristic approach to determine an appropriate number of topics in topic modeling." *BMC bioinformatics*. Vol. 16. No. 13. BioMed Central, 2015.

[17] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Advances in neural information processing systems*. 2002.

[18] Hoffman, Matthew, Francis R. Bach, and David M. Blei. "Online learning for latent dirichlet allocation." *advances in neural information processing systems*. 2010.

[19] Buntine, Wray. "Variational extensions to EM and multinomial PCA." *European Conference on Machine Learning*. Springer, Berlin, Heidelberg, 2002.

[20] Goodman, Joshua. "An Empirical study of smoothing Techniques for Language Modeling stanley F. Chen and." (1998).

[21] Epstein, Mark, et al. "Statistical natural language understanding using hidden clumpings." *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*. Vol. 1. IEEE, 1996.

[22] Kilgarriff, Adam, and Tony Rose. "Measures for corpus similarity and homogeneity." *Proceedings of the Third Conference on Empirical Methods for Natural Language Processing*. 1998.

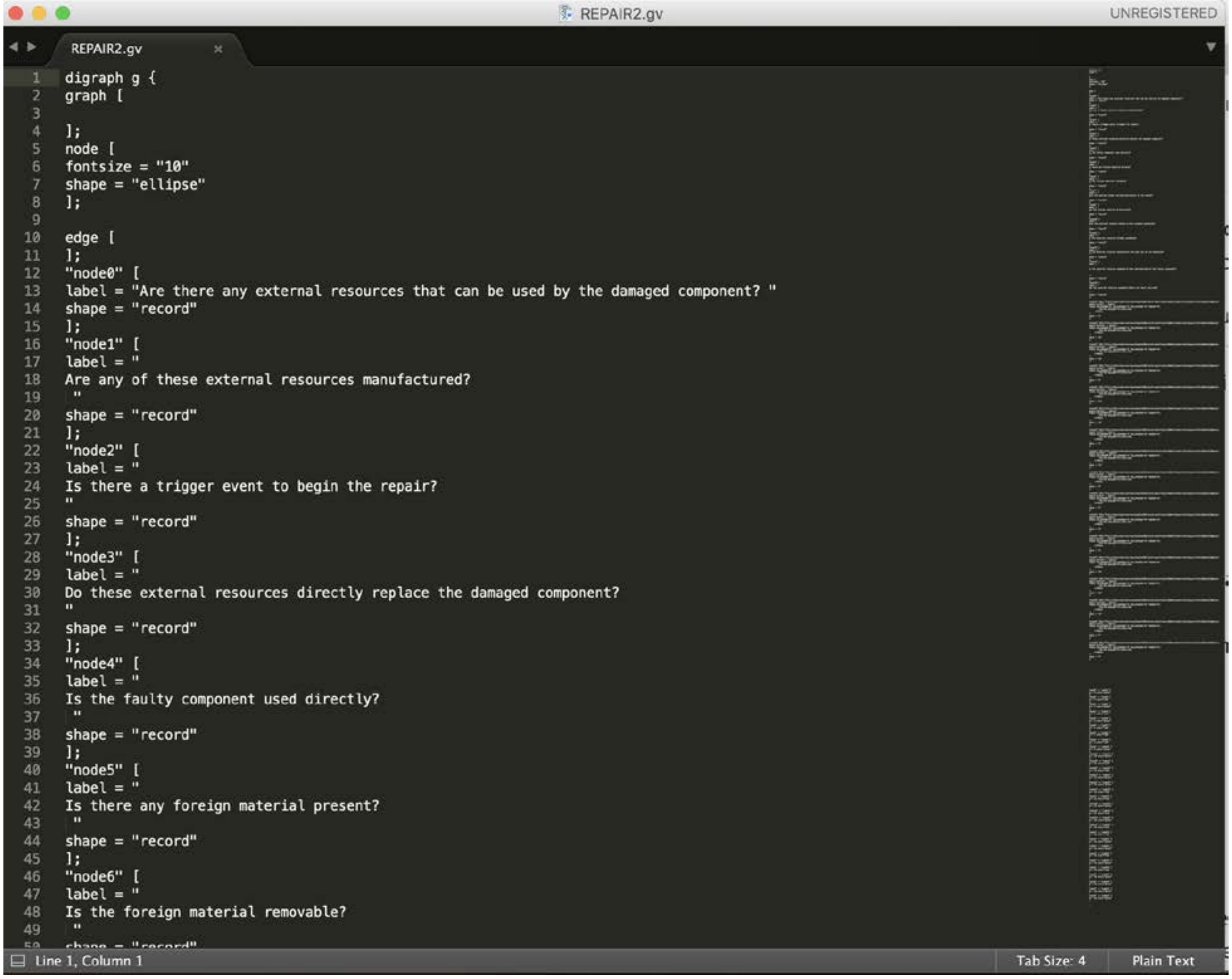
[23] Brown, Peter F., et al. "An estimate of an upper bound for the entropy of English." *Computational Linguistics* 18.1 (1992): 31-40.

[24] Vincent, J. F., Bogatyreva, O. A., Bogatyrev, N. R., Bowyer, A., and Pahl, A.-K., 2006. “Biomimetics: its practice and theory”. Journal of the Royal Society Interface, 3(9), pp. 471–482.

[25] Rousseeuw, Peter J. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of computational and applied mathematics* 20 (1987): 53-65.

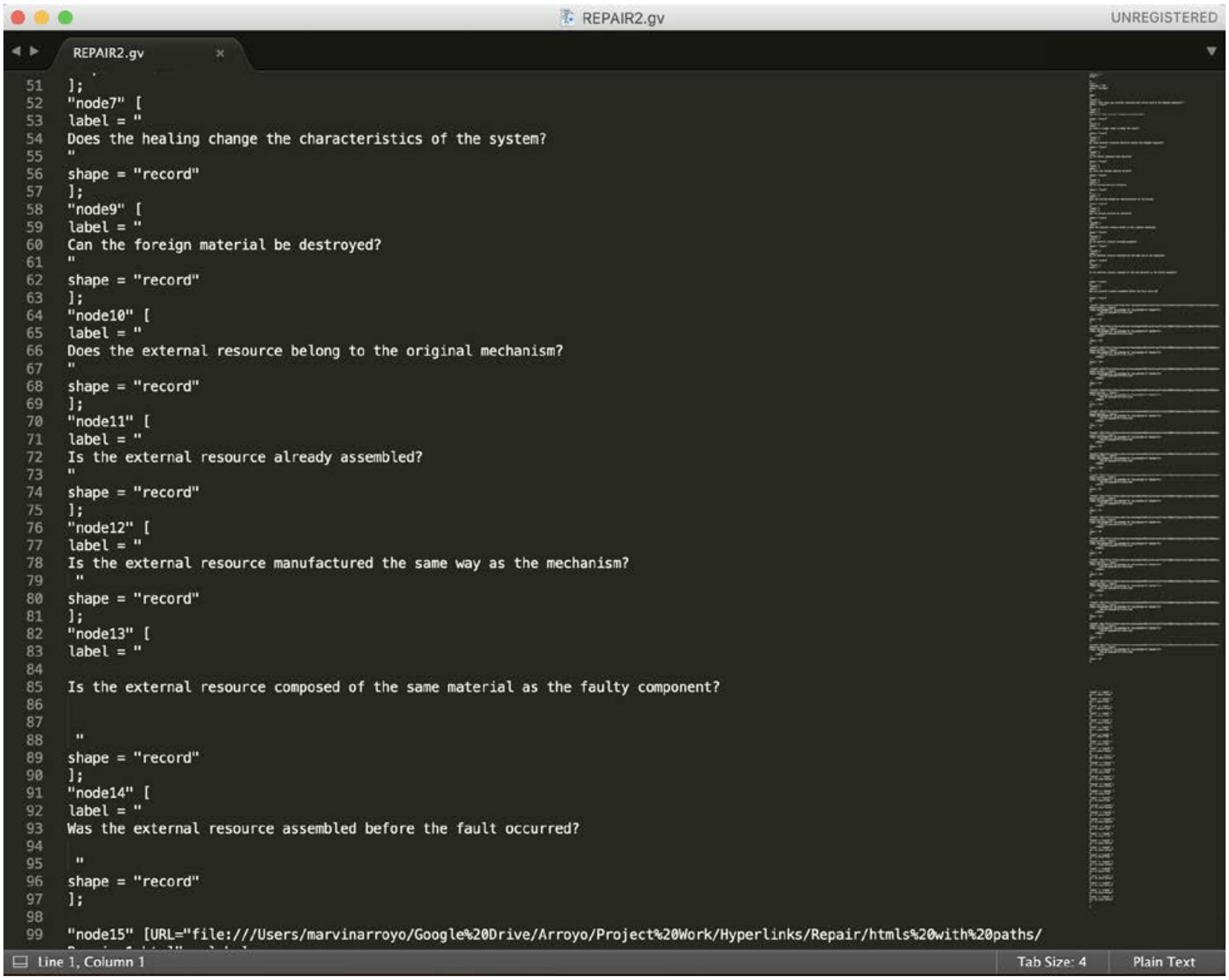
[26] de Amorim, Renato Cordeiro, and Christian Hennig. "Recovering the number of clusters in data sets with noise features using feature rescaling factors." *Information Sciences*324 (2015): 126-145.

APPENDIX A



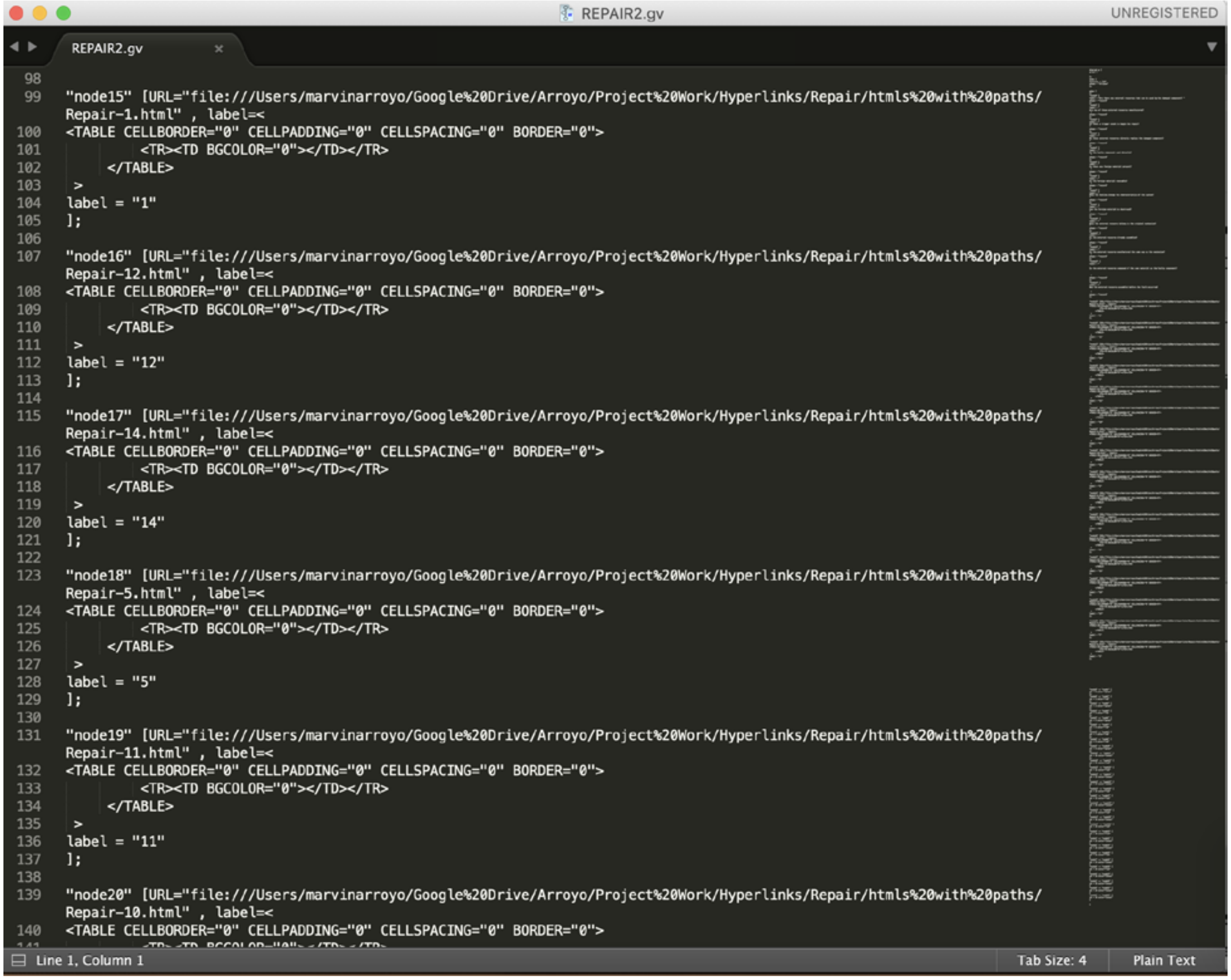
```
digraph g {
graph [

];
node [
fontsize = "10"
shape = "ellipse"
];

edge [
];
"node0" [
label = "Are there any external resources that can be used by the damaged component? "
shape = "record"
];
"node1" [
label = "
Are any of these external resources manufactured?
 "
shape = "record"
];
"node2" [
label = "
Is there a trigger event to begin the repair?
"
shape = "record"
];
"node3" [
label = "
Do these external resources directly replace the damaged component?
"
shape = "record"
];
"node4" [
label = "
Is the faulty component used directly?
 "
shape = "record"
];
"node5" [
label = "
Is there any foreign material present?
 "
shape = "record"
];
"node6" [
label = "
Is the foreign material removable?
 "
shape = "record"
```

Figure 8. Graphviz code to generate the Repair graph.



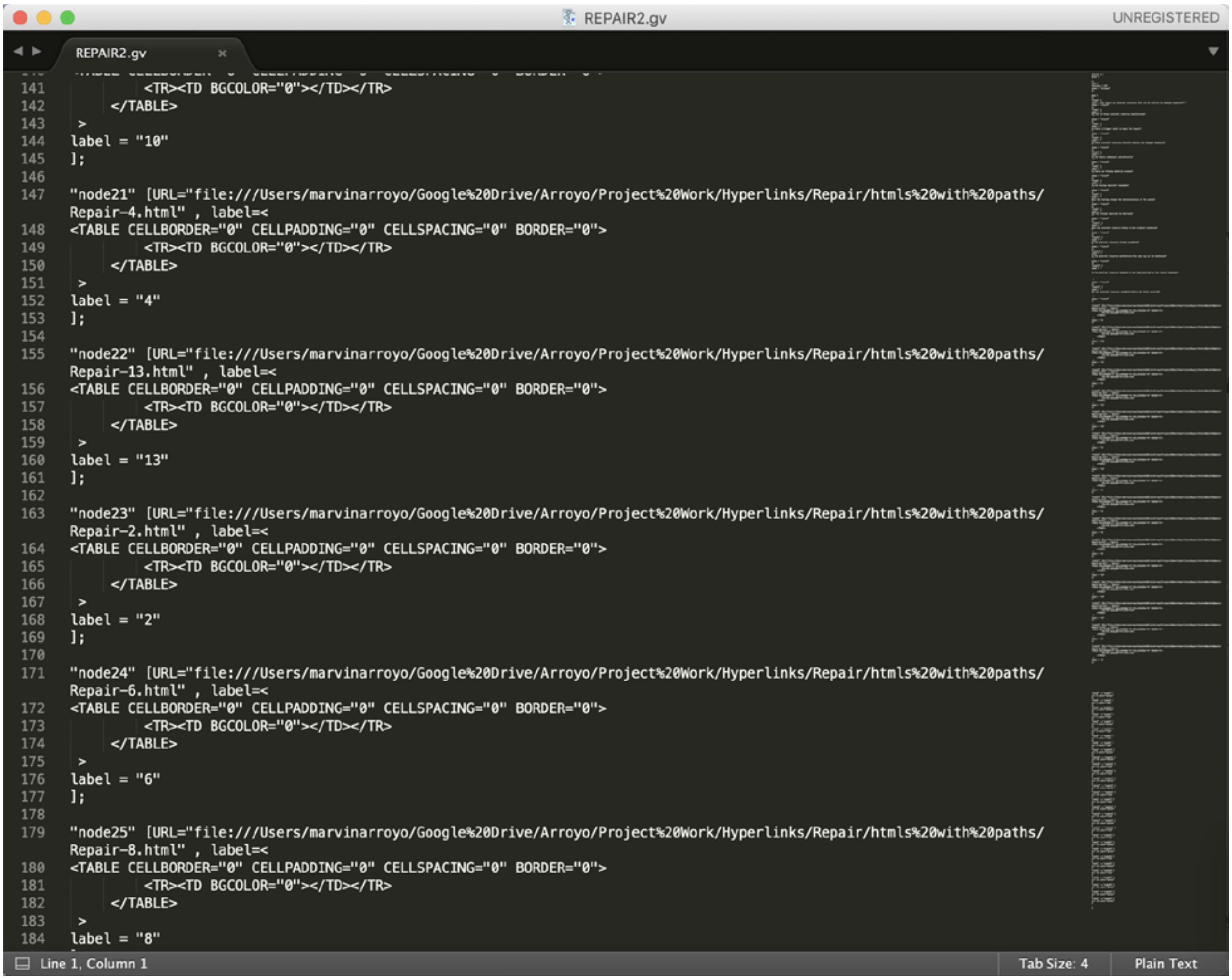
```
];
"node7" [
label = "
Does the healing change the characteristics of the system?
"
shape = "record"
];
"node9" [
label = "
Can the foreign material be destroyed?
"
shape = "record"
];
"node10" [
label = "
Does the external resource belong to the original mechanism?
"
shape = "record"
];
"node11" [
label = "
Is the external resource already assembled?
"
shape = "record"
];
"node12" [
label = "
Is the external resource manufactured the same way as the mechanism?
 "
shape = "record"
];
"node13" [
label = "

Is the external resource composed of the same material as the faulty component?


 "
shape = "record"
];
"node14" [
label = "
Was the external resource assembled before the fault occurred?

 "
shape = "record"
];

"node15" [URL="file:///Users/marvinarroyo/Google%20Drive/Arroyo/Project%20Work/Hyperlinks/Repair/htmls%20with%20paths/
```

Figure 8 (Cont.) Graphviz code to generate the Repair graph.



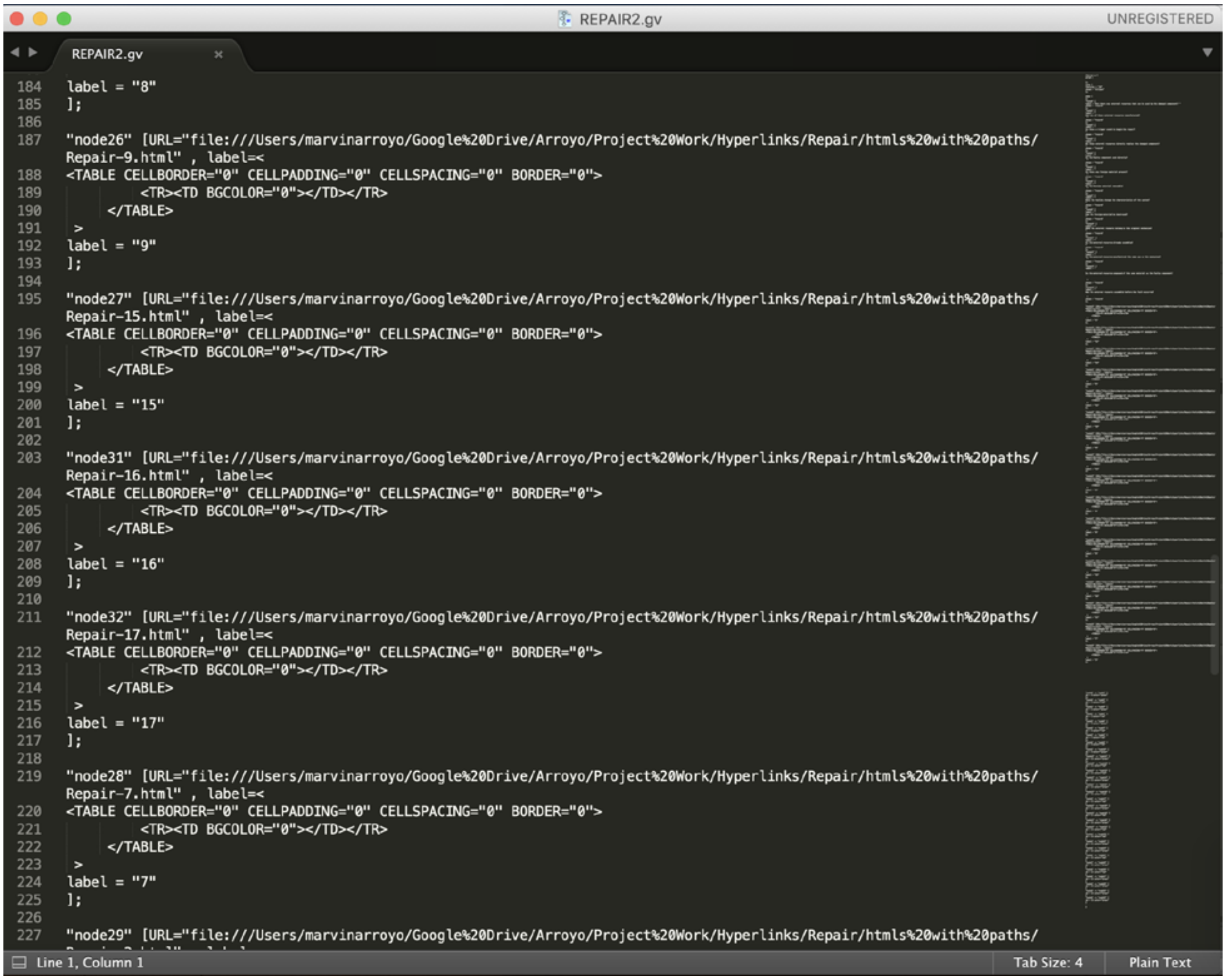
```

"node15" [URL="file:///Users/marvinarroyo/Google%20Drive/Arroyo/Project%20Work/Hyperlinks/Repair/htmls%20with%20paths/
Repair-1.html" , label=<
<TABLE CELLBORDER="0" CELLPADDING="0" CELLSPACING="0" BORDER="0">
          <TR><TD BGCOLOR="0"></TD></TR>
     </TABLE>
 >
label = "1"
];

"node16" [URL="file:///Users/marvinarroyo/Google%20Drive/Arroyo/Project%20Work/Hyperlinks/Repair/htmls%20with%20paths/
Repair-12.html" , label=<
<TABLE CELLBORDER="0" CELLPADDING="0" CELLSPACING="0" BORDER="0">
          <TR><TD BGCOLOR="0"></TD></TR>
     </TABLE>
 >
label = "12"
];

"node17" [URL="file:///Users/marvinarroyo/Google%20Drive/Arroyo/Project%20Work/Hyperlinks/Repair/htmls%20with%20paths/
Repair-14.html" , label=<
<TABLE CELLBORDER="0" CELLPADDING="0" CELLSPACING="0" BORDER="0">
          <TR><TD BGCOLOR="0"></TD></TR>
     </TABLE>
 >
label = "14"
];

"node18" [URL="file:///Users/marvinarroyo/Google%20Drive/Arroyo/Project%20Work/Hyperlinks/Repair/htmls%20with%20paths/
Repair-5.html" , label=<
<TABLE CELLBORDER="0" CELLPADDING="0" CELLSPACING="0" BORDER="0">
          <TR><TD BGCOLOR="0"></TD></TR>
     </TABLE>
 >
label = "5"
];

"node19" [URL="file:///Users/marvinarroyo/Google%20Drive/Arroyo/Project%20Work/Hyperlinks/Repair/htmls%20with%20paths/
Repair-11.html" , label=<
<TABLE CELLBORDER="0" CELLPADDING="0" CELLSPACING="0" BORDER="0">
          <TR><TD BGCOLOR="0"></TD></TR>
     </TABLE>
 >
label = "11"
];

"node20" [URL="file:///Users/marvinarroyo/Google%20Drive/Arroyo/Project%20Work/Hyperlinks/Repair/htmls%20with%20paths/
Repair-10.html" , label=<
<TABLE CELLBORDER="0" CELLPADDING="0" CELLSPACING="0" BORDER="0">
          <TR><TD BGCOLOR="0"></TD></TR>
```

Figure 8 (Cont.) Graphviz code to generate the Repair graph.



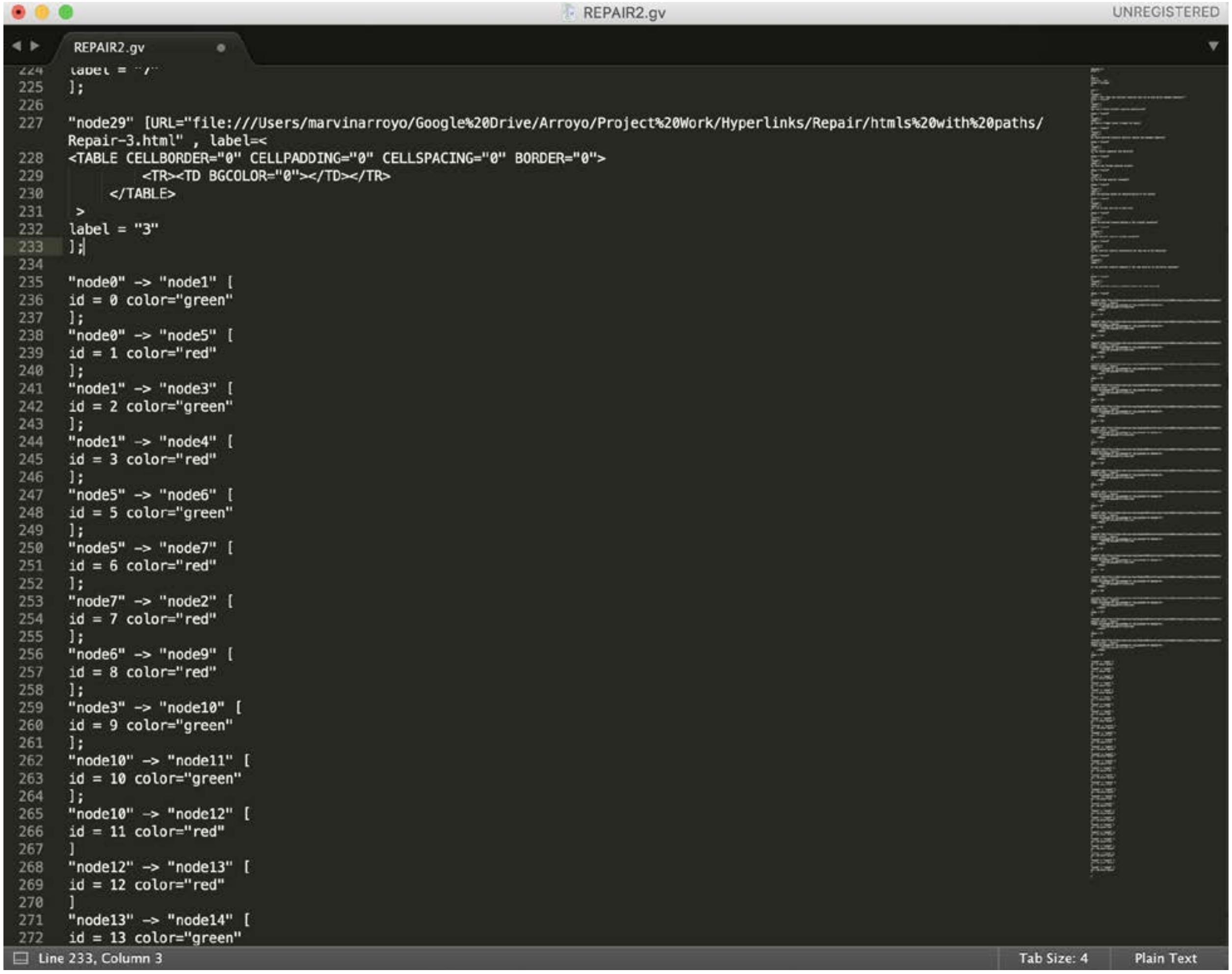
```
          <TR><TD BGCOLOR="0"></TD></TR>
     </TABLE>
 >
label = "10"
];

"node21" [URL="file:///Users/marvinarroyo/Google%20Drive/Arroyo/Project%20Work/Hyperlinks/Repair/htmls%20with%20paths/
Repair-4.html" , label=<
<TABLE CELLBORDER="0" CELLPADDING="0" CELLSPACING="0" BORDER="0">
          <TR><TD BGCOLOR="0"></TD></TR>
     </TABLE>
 >
label = "4"
];

"node22" [URL="file:///Users/marvinarroyo/Google%20Drive/Arroyo/Project%20Work/Hyperlinks/Repair/htmls%20with%20paths/
Repair-13.html" , label=<
<TABLE CELLBORDER="0" CELLPADDING="0" CELLSPACING="0" BORDER="0">
          <TR><TD BGCOLOR="0"></TD></TR>
     </TABLE>
 >
label = "13"
];

"node23" [URL="file:///Users/marvinarroyo/Google%20Drive/Arroyo/Project%20Work/Hyperlinks/Repair/htmls%20with%20paths/
Repair-2.html" , label=<
<TABLE CELLBORDER="0" CELLPADDING="0" CELLSPACING="0" BORDER="0">
          <TR><TD BGCOLOR="0"></TD></TR>
     </TABLE>
 >
label = "2"
];

"node24" [URL="file:///Users/marvinarroyo/Google%20Drive/Arroyo/Project%20Work/Hyperlinks/Repair/htmls%20with%20paths/
Repair-6.html" , label=<
<TABLE CELLBORDER="0" CELLPADDING="0" CELLSPACING="0" BORDER="0">
          <TR><TD BGCOLOR="0"></TD></TR>
     </TABLE>
 >
label = "6"
];

"node25" [URL="file:///Users/marvinarroyo/Google%20Drive/Arroyo/Project%20Work/Hyperlinks/Repair/htmls%20with%20paths/
Repair-8.html" , label=<
<TABLE CELLBORDER="0" CELLPADDING="0" CELLSPACING="0" BORDER="0">
          <TR><TD BGCOLOR="0"></TD></TR>
     </TABLE>
 >
label = "8"
```

Figure 8 (Cont.) Graphviz code to generate the Repair graph.



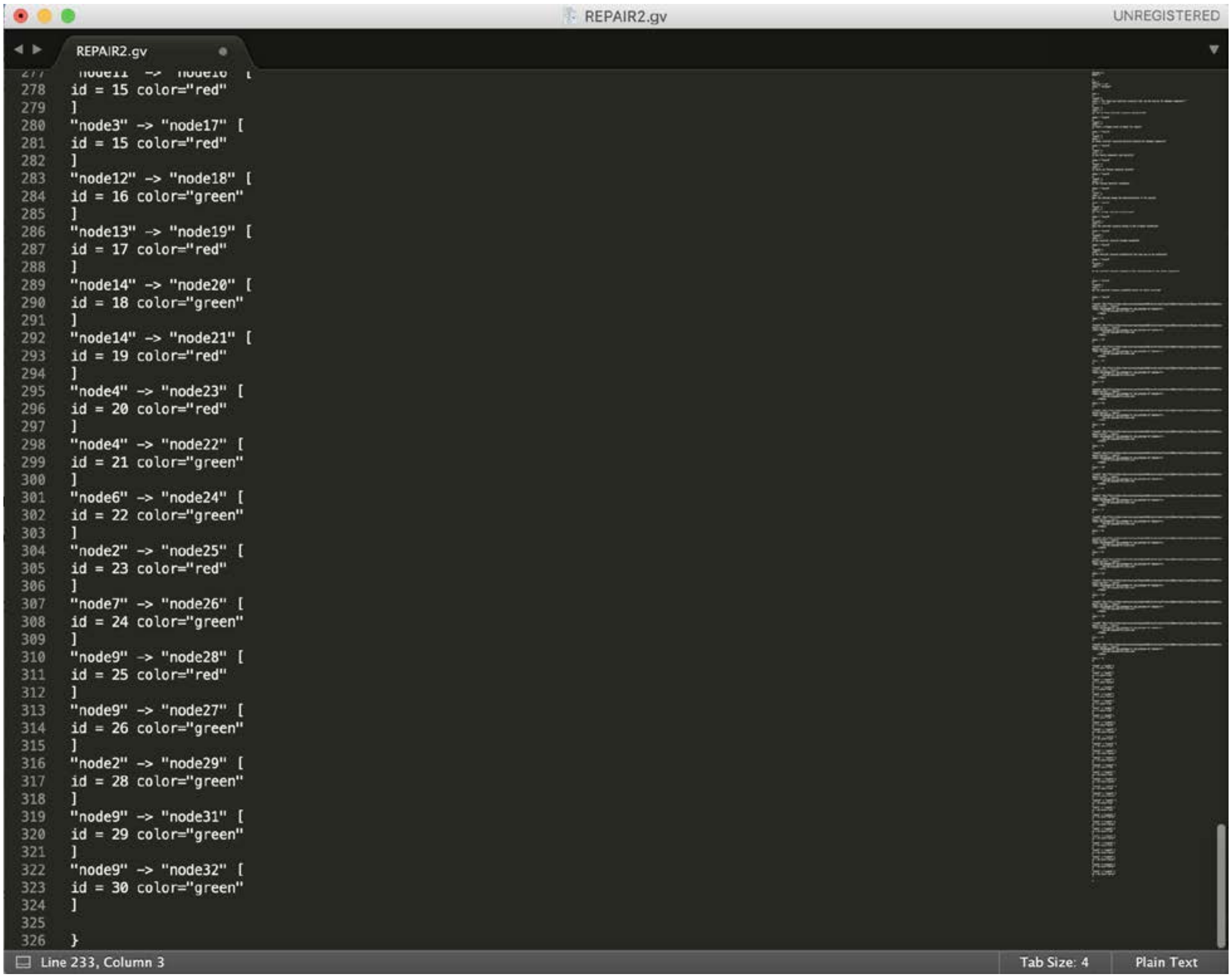
```
label = "8"
];

"node26" [URL="file:///Users/marvinarroyo/Google%20Drive/Arroyo/Project%20Work/Hyperlinks/Repair/htmls%20with%20paths/
Repair-9.html" , label=<
<TABLE CELLBORDER="0" CELLPADDING="0" CELLSPACING="0" BORDER="0">
          <TR><TD BGCOLOR="0"></TD></TR>
     </TABLE>
 >
label = "9"
];

"node27" [URL="file:///Users/marvinarroyo/Google%20Drive/Arroyo/Project%20Work/Hyperlinks/Repair/htmls%20with%20paths/
Repair-15.html" , label=<
<TABLE CELLBORDER="0" CELLPADDING="0" CELLSPACING="0" BORDER="0">
          <TR><TD BGCOLOR="0"></TD></TR>
     </TABLE>
 >
label = "15"
];

"node31" [URL="file:///Users/marvinarroyo/Google%20Drive/Arroyo/Project%20Work/Hyperlinks/Repair/htmls%20with%20paths/
Repair-16.html" , label=<
<TABLE CELLBORDER="0" CELLPADDING="0" CELLSPACING="0" BORDER="0">
          <TR><TD BGCOLOR="0"></TD></TR>
     </TABLE>
 >
label = "16"
];

"node32" [URL="file:///Users/marvinarroyo/Google%20Drive/Arroyo/Project%20Work/Hyperlinks/Repair/htmls%20with%20paths/
Repair-17.html" , label=<
<TABLE CELLBORDER="0" CELLPADDING="0" CELLSPACING="0" BORDER="0">
          <TR><TD BGCOLOR="0"></TD></TR>
     </TABLE>
 >
label = "17"
];

"node28" [URL="file:///Users/marvinarroyo/Google%20Drive/Arroyo/Project%20Work/Hyperlinks/Repair/htmls%20with%20paths/
Repair-7.html" , label=<
<TABLE CELLBORDER="0" CELLPADDING="0" CELLSPACING="0" BORDER="0">
          <TR><TD BGCOLOR="0"></TD></TR>
     </TABLE>
 >
label = "7"
];

"node29" [URL="file:///Users/marvinarroyo/Google%20Drive/Arroyo/Project%20Work/Hyperlinks/Repair/htmls%20with%20paths/
```

Figure 8 (Cont.) Graphviz code to generate the Repair graph.



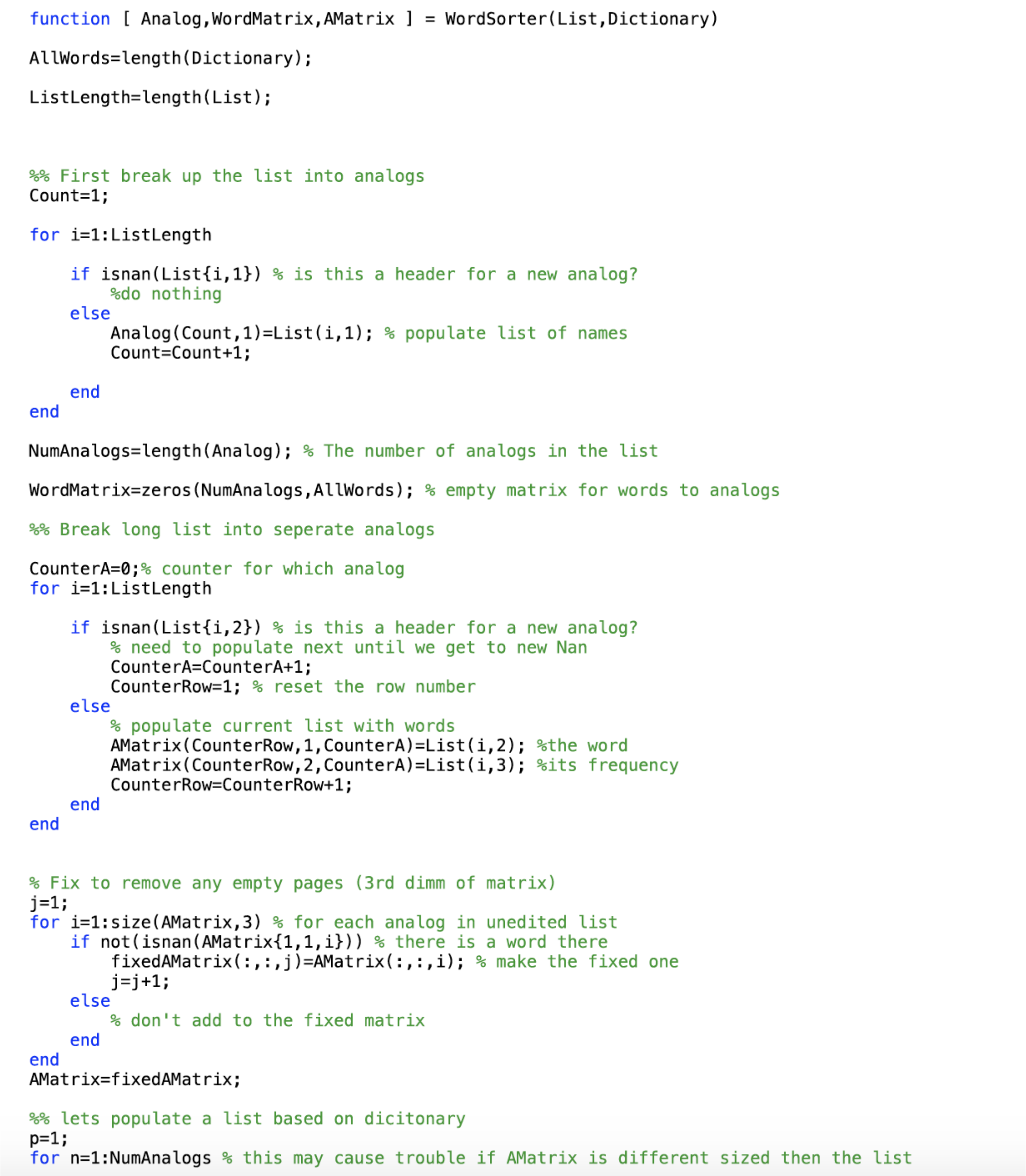
```
label = "7"
];

"node29" [URL="file:///Users/marvinarroyo/Google%20Drive/Arroyo/Project%20Work/Hyperlinks/Repair/htmls%20with%20paths/
Repair-3.html" , label=<
<TABLE CELLBORDER="0" CELLPADDING="0" CELLSPACING="0" BORDER="0">
        <TR><TD BGCOLOR="0"></TD></TR>
    </TABLE>
 >
label = "3"
];

"node0" -> "node1" [
id = 0 color="green"
];
"node0" -> "node5" [
id = 1 color="red"
];
"node1" -> "node3" [
id = 2 color="green"
];
"node1" -> "node4" [
id = 3 color="red"
];
"node5" -> "node6" [
id = 5 color="green"
];
"node5" -> "node7" [
id = 6 color="red"
];
"node7" -> "node2" [
id = 7 color="red"
];
"node6" -> "node9" [
id = 8 color="red"
];
"node3" -> "node10" [
id = 9 color="green"
];
"node10" -> "node11" [
id = 10 color="green"
];
"node10" -> "node12" [
id = 11 color="red"
]
"node12" -> "node13" [
id = 12 color="red"
]
"node13" -> "node14" [
id = 13 color="green"
```

Figure 8 (Cont.) Graphviz code to generate the Repair graph.



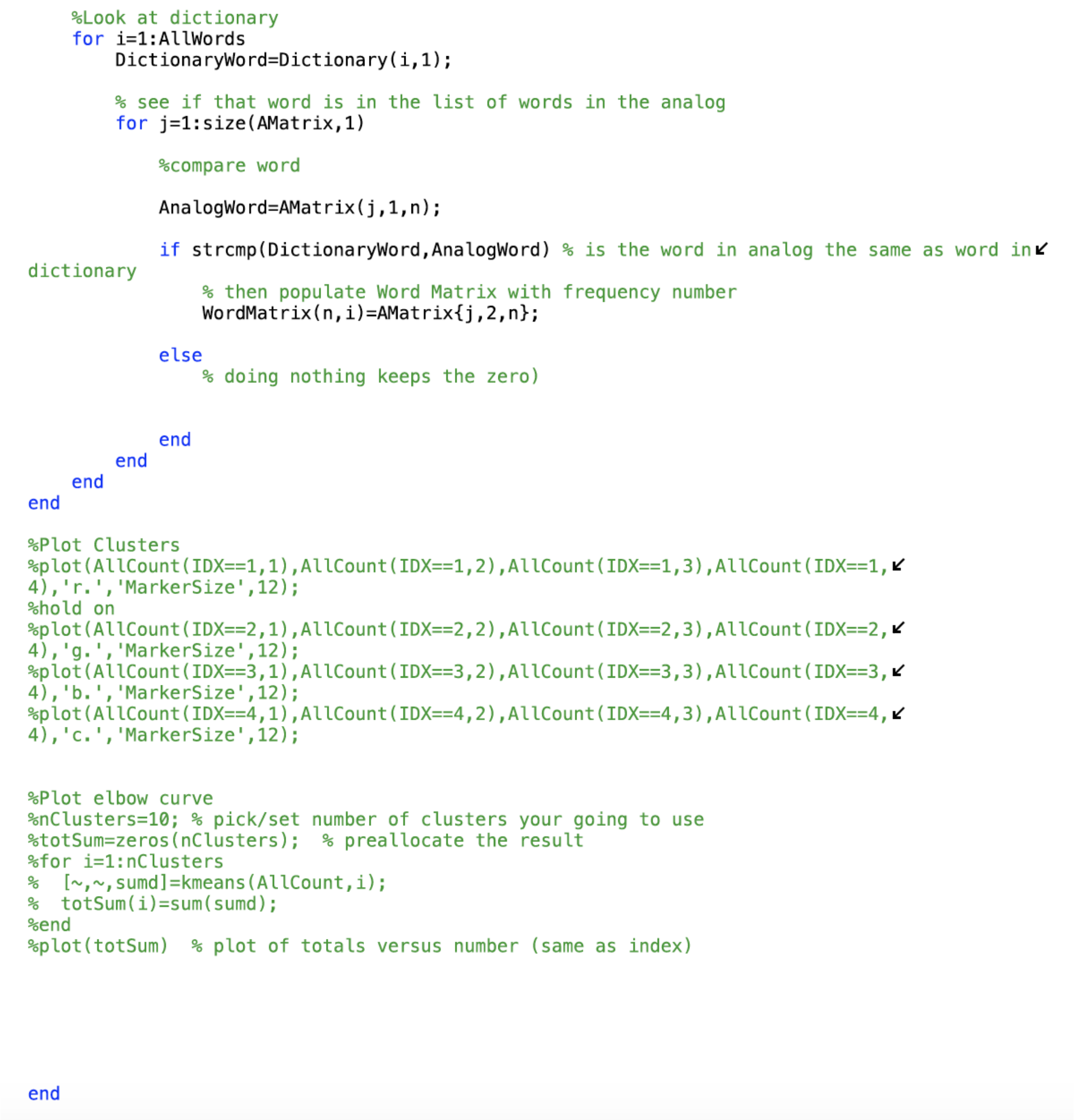
```
"node11" -> "node16" [
id = 15 color="red"
]
"node3" -> "node17" [
id = 15 color="red"
]
"node12" -> "node18" [
id = 16 color="green"
]
"node13" -> "node19" [
id = 17 color="red"
]
"node14" -> "node20" [
id = 18 color="green"
]
"node14" -> "node21" [
id = 19 color="red"
]
"node4" -> "node23" [
id = 20 color="red"
]
"node4" -> "node22" [
id = 21 color="green"
]
"node6" -> "node24" [
id = 22 color="green"
]
"node2" -> "node25" [
id = 23 color="red"
]
"node7" -> "node26" [
id = 24 color="green"
]
"node9" -> "node28" [
id = 25 color="red"
]
"node9" -> "node27" [
id = 26 color="green"
]
"node2" -> "node29" [
id = 28 color="green"
]
"node9" -> "node31" [
id = 29 color="green"
]
"node9" -> "node32" [
id = 30 color="green"
]

}
```

Figure 8 (Cont.) Graphviz code to generate the Repair graph.



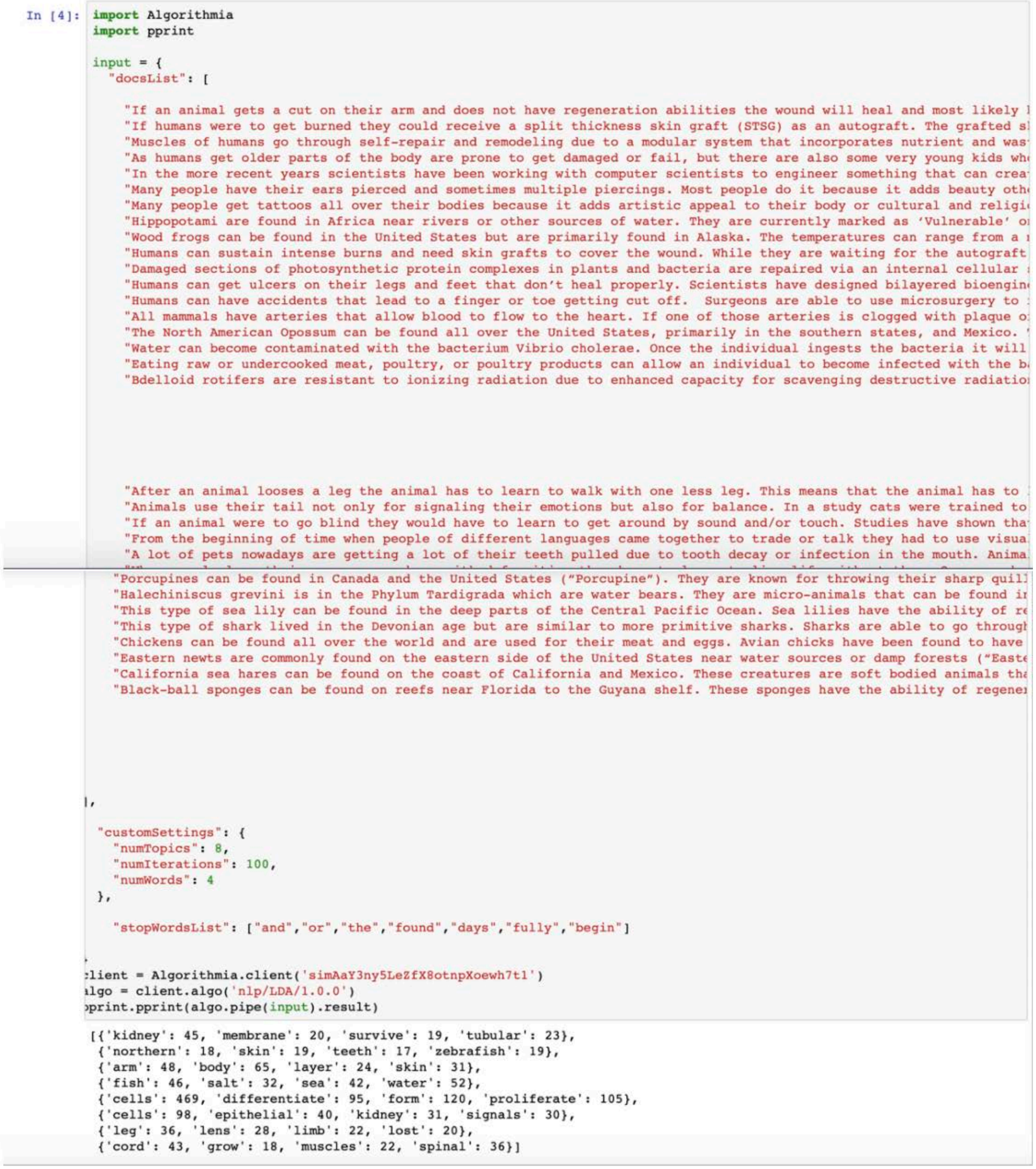
```
function [ Analog,WordMatrix,AMatrix ] = WordSorter(List,Dictionary)

AllWords=length(Dictionary);

ListLength=length(List);


%% First break up the list into analogs
Count=1;

for i=1:ListLength

    if isnan(List{i,1}) % is this a header for a new analog?
        %do nothing
    else
        Analog(Count,1)=List(i,1); % populate list of names
        Count=Count+1;

    end
end

NumAnalogs=length(Analog); % The number of analogs in the list

WordMatrix=zeros(NumAnalogs,AllWords); % empty matrix for words to analogs

%% Break long list into seperate analogs

CounterA=0;% counter for which analog
for i=1:ListLength

    if isnan(List{i,2}) % is this a header for a new analog?
        % need to populate next until we get to new Nan
        CounterA=CounterA+1;
        CounterRow=1; % reset the row number
    else
        % populate current list with words
        AMatrix(CounterRow,1,CounterA)=List(i,2); %the word
        AMatrix(CounterRow,2,CounterA)=List(i,3); %its frequency
        CounterRow=CounterRow+1;
    end
end


% Fix to remove any empty pages (3rd dimm of matrix)
j=1;
for i=1:size(AMatrix,3) % for each analog in unedited list
    if not(isnan(AMatrix{1,1,i})) % there is a word there
        fixedAMatrix(:,:,j)=AMatrix(:,:,i); % make the fixed one
        j=j+1;
    else
        % don't add to the fixed matrix
    end
end
AMatrix=fixedAMatrix;

%% lets populate a list based on dicitonary
p=1;
for n=1:NumAnalogs % this may cause trouble if AMatrix is different sized then the list
```

Figure 9. MATLAB Code for the k-means algorithm.



```
    %Look at dictionary
    for i=1:AllWords
        DictionaryWord=Dictionary(i,1);

        % see if that word is in the list of words in the analog
        for j=1:size(AMatrix,1)

            %compare word

            AnalogWord=AMatrix(j,1,n);

            if strcmp(DictionaryWord,AnalogWord) % is the word in analog the same as word in↙
dictionary
                % then populate Word Matrix with frequency number
                WordMatrix(n,i)=AMatrix{j,2,n};

            else
                % doing nothing keeps the zero)


            end
        end
    end
end

%Plot Clusters
%plot(AllCount(IDX==1,1),AllCount(IDX==1,2),AllCount(IDX==1,3),AllCount(IDX==1,↙
4),'r.','MarkerSize',12);
%hold on
%plot(AllCount(IDX==2,1),AllCount(IDX==2,2),AllCount(IDX==2,3),AllCount(IDX==2,↙
4),'g.','MarkerSize',12);
%plot(AllCount(IDX==3,1),AllCount(IDX==3,2),AllCount(IDX==3,3),AllCount(IDX==3,↙
4),'b.','MarkerSize',12);
%plot(AllCount(IDX==4,1),AllCount(IDX==4,2),AllCount(IDX==4,3),AllCount(IDX==4,↙
4),'c.','MarkerSize',12);


%Plot elbow curve
%nClusters=10; % pick/set number of clusters your going to use
%totSum=zeros(nClusters);  % preallocate the result
%for i=1:nClusters
%   [~,~,sumd]=kmeans(AllCount,i);
%   totSum(i)=sum(sumd);
%end
%plot(totSum)  % plot of totals versus number (same as index)




end
```

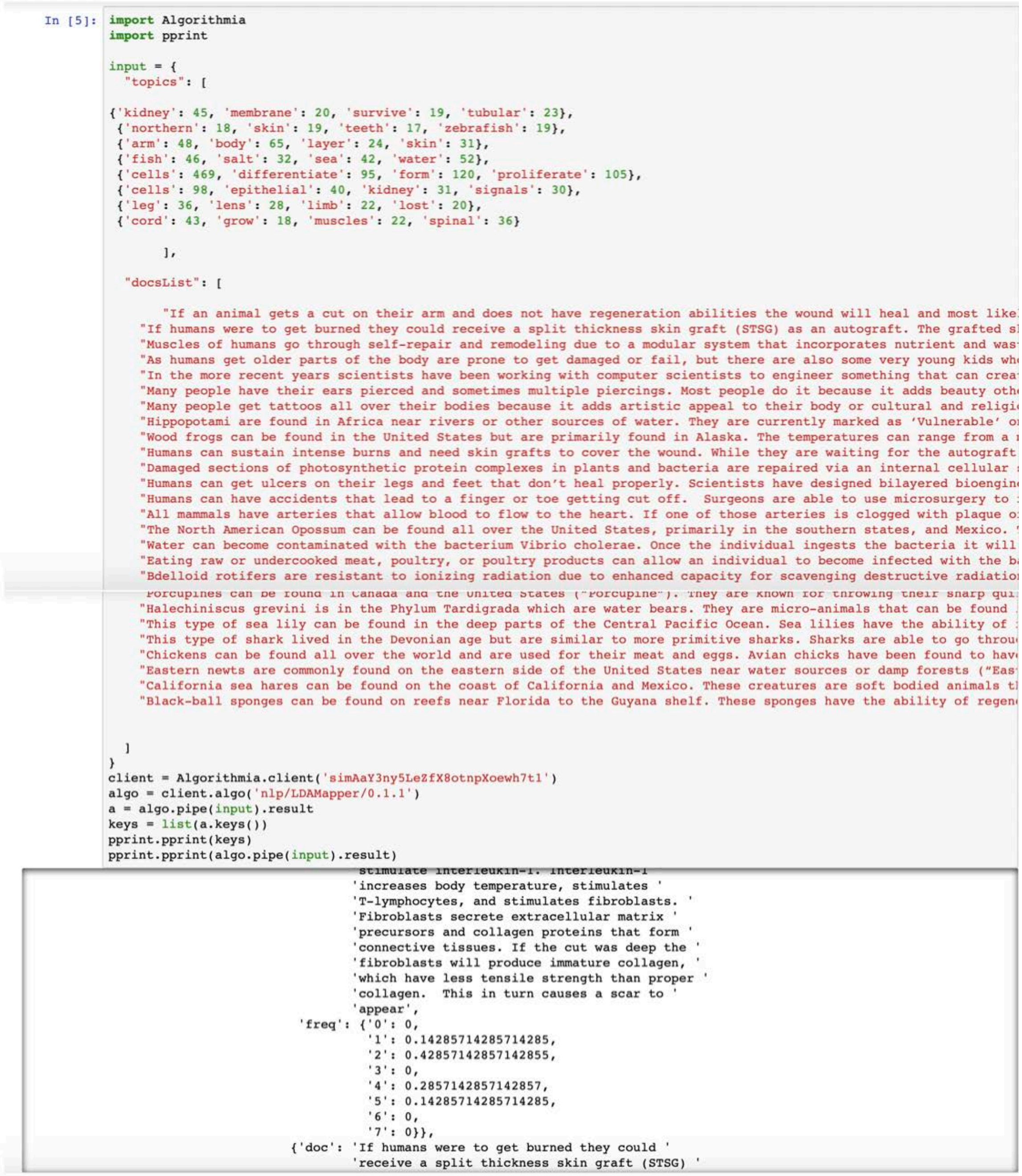
Figure 9 (Cont.) MATLAB Code for the k-means algorithm.



```
In [4]: import Algorithmia
        import pprint

        input = {
          "docsList": [

            "If an animal gets a cut on their arm and does not have regeneration abilities the wound will heal and most likely l
            "If humans were to get burned they could receive a split thickness skin graft (STSG) as an autograft. The grafted sl
            "Muscles of humans go through self-repair and remodeling due to a modular system that incorporates nutrient and was
            "As humans get older parts of the body are prone to get damaged or fail, but there are also some very young kids wh
            "In the more recent years scientists have been working with computer scientists to engineer something that can crea
            "Many people have their ears pierced and sometimes multiple piercings. Most people do it because it adds beauty oth
            "Many people get tattoos all over their bodies because it adds artistic appeal to their body or cultural and religi
            "Hippopotami are found in Africa near rivers or other sources of water. They are currently marked as 'Vulnerable' o
            "Wood frogs can be found in the United States but are primarily found in Alaska. The temperatures can range from a
            "Humans can sustain intense burns and need skin grafts to cover the wound. While they are waiting for the autograft
            "Damaged sections of photosynthetic protein complexes in plants and bacteria are repaired via an internal cellular
            "Humans can get ulcers on their legs and feet that don't heal properly. Scientists have designed bilayered bioengin
            "Humans can have accidents that lead to a finger or toe getting cut off.  Surgeons are able to use microsurgery to
            "All mammals have arteries that allow blood to flow to the heart. If one of those arteries is clogged with plaque o
            "The North American Opossum can be found all over the United States, primarily in the southern states, and Mexico.
            "Water can become contaminated with the bacterium Vibrio cholerae. Once the individual ingests the bacteria it will
            "Eating raw or undercooked meat, poultry, or poultry products can allow an individual to become infected with the b
            "Bdelloid rotifers are resistant to ionizing radiation due to enhanced capacity for scavenging destructive radiatio




            "After an animal looses a leg the animal has to learn to walk with one less leg. This means that the animal has to
            "Animals use their tail not only for signaling their emotions but also for balance. In a study cats were trained to
            "If an animal were to go blind they would have to learn to get around by sound and/or touch. Studies have shown tha
            "From the beginning of time when people of different languages came together to trade or talk they had to use visua
            "A lot of pets nowadays are getting a lot of their teeth pulled due to tooth decay or infection in the mouth. Anima
          "Porcupines can be found in Canada and the United States ("Porcupine"). They are known for throwing their sharp quill
          "Halechiniscus grevini is in the Phylum Tardigrada which are water bears. They are micro-animals that can be found in
          "This type of sea lily can be found in the deep parts of the Central Pacific Ocean. Sea lilies have the ability of re
          "This type of shark lived in the Devonian age but are similar to more primitive sharks. Sharks are able to go through
          "Chickens can be found all over the world and are used for their meat and eggs. Avian chicks have been found to have
          "Eastern newts are commonly found on the eastern side of the United States near water sources or damp forests ("Easte
          "California sea hares can be found on the coast of California and Mexico. These creatures are soft bodied animals tha
          "Black-ball sponges can be found on reefs near Florida to the Guyana shelf. These sponges have the ability of regener




        ],

          "customSettings": {
            "numTopics": 8,
            "numIterations": 100,
            "numWords": 4
          },

            "stopWordsList": ["and","or","the","found","days","fully","begin"]

        }
        client = Algorithmia.client('simAaY3ny5LeZfX8otnpXoewh7t1')
        algo = client.algo('nlp/LDA/1.0.0')
        pprint.pprint(algo.pipe(input).result)
```

```
[{'kidney': 45, 'membrane': 20, 'survive': 19, 'tubular': 23},
 {'northern': 18, 'skin': 19, 'teeth': 17, 'zebrafish': 19},
 {'arm': 48, 'body': 65, 'layer': 24, 'skin': 31},
 {'fish': 46, 'salt': 32, 'sea': 42, 'water': 52},
 {'cells': 469, 'differentiate': 95, 'form': 120, 'proliferate': 105},
 {'cells': 98, 'epithelial': 40, 'kidney': 31, 'signals': 30},
 {'leg': 36, 'lens': 28, 'limb': 22, 'lost': 20},
 {'cord': 43, 'grow': 18, 'muscles': 22, 'spinal': 36}]
```

Figure 10. Python Code for LDA model.



```
In [5]: import Algorithmia
        import pprint

        input = {
          "topics": [

        {'kidney': 45, 'membrane': 20, 'survive': 19, 'tubular': 23},
         {'northern': 18, 'skin': 19, 'teeth': 17, 'zebrafish': 19},
         {'arm': 48, 'body': 65, 'layer': 24, 'skin': 31},
         {'fish': 46, 'salt': 32, 'sea': 42, 'water': 52},
         {'cells': 469, 'differentiate': 95, 'form': 120, 'proliferate': 105},
         {'cells': 98, 'epithelial': 40, 'kidney': 31, 'signals': 30},
         {'leg': 36, 'lens': 28, 'limb': 22, 'lost': 20},
         {'cord': 43, 'grow': 18, 'muscles': 22, 'spinal': 36}

               ],

          "docsList": [

               "If an animal gets a cut on their arm and does not have regeneration abilities the wound will heal and most like
            "If humans were to get burned they could receive a split thickness skin graft (STSG) as an autograft. The grafted s
            "Muscles of humans go through self-repair and remodeling due to a modular system that incorporates nutrient and was
            "As humans get older parts of the body are prone to get damaged or fail, but there are also some very young kids wh
            "In the more recent years scientists have been working with computer scientists to engineer something that can crea
            "Many people have their ears pierced and sometimes multiple piercings. Most people do it because it adds beauty oth
            "Many people get tattoos all over their bodies because it adds artistic appeal to their body or cultural and religi
            "Hippopotami are found in Africa near rivers or other sources of water. They are currently marked as ‘Vulnerable’ o
            "Wood frogs can be found in the United States but are primarily found in Alaska. The temperatures can range from a
            "Humans can sustain intense burns and need skin grafts to cover the wound. While they are waiting for the autograft
            "Damaged sections of photosynthetic protein complexes in plants and bacteria are repaired via an internal cellular
            "Humans can get ulcers on their legs and feet that don’t heal properly. Scientists have designed bilayered bioengin
            "Humans can have accidents that lead to a finger or toe getting cut off.  Surgeons are able to use microsurgery to
            "All mammals have arteries that allow blood to flow to the heart. If one of those arteries is clogged with plaque o
            "The North American Opossum can be found all over the United States, primarily in the southern states, and Mexico.
            "Water can become contaminated with the bacterium Vibrio cholerae. Once the individual ingests the bacteria it will
            "Eating raw or undercooked meat, poultry, or poultry products can allow an individual to become infected with the b
            "Bdelloid rotifers are resistant to ionizing radiation due to enhanced capacity for scavenging destructive radiatio
             Porcupines can be found in Canada and the United States ("Porcupine"). They are known for throwing their sharp qui
            "Halechiniscus grevini is in the Phylum Tardigrada which are water bears. They are micro-animals that can be found
            "This type of sea lily can be found in the deep parts of the Central Pacific Ocean. Sea lilies have the ability of
            "This type of shark lived in the Devonian age but are similar to more primitive sharks. Sharks are able to go throu
            "Chickens can be found all over the world and are used for their meat and eggs. Avian chicks have been found to hav
            "Eastern newts are commonly found on the eastern side of the United States near water sources or damp forests (“Eas
            "California sea hares can be found on the coast of California and Mexico. These creatures are soft bodied animals t
            "Black-ball sponges can be found on reefs near Florida to the Guyana shelf. These sponges have the ability of regen


          ]
        }
        client = Algorithmia.client('simAaY3ny5LeZfX8otnpXoewh7t1')
        algo = client.algo('nlp/LDAMapper/0.1.1')
        a = algo.pipe(input).result
        keys = list(a.keys())
        pprint.pprint(keys)
        pprint.pprint(algo.pipe(input).result)
```

```
                                           'stimulate interleukin-1. Interleukin-1 '
                                           'increases body temperature, stimulates '
                                           'T-lymphocytes, and stimulates fibroblasts. '
                                           'Fibroblasts secrete extracellular matrix '
                                           'precursors and collagen proteins that form '
                                           'connective tissues. If the cut was deep the '
                                           'fibroblasts will produce immature collagen, '
                                           'which have less tensile strength than proper '
                                           'collagen.  This in turn causes a scar to '
                                           'appear',
                                    'freq': {'0': 0,
                                             '1': 0.14285714285714285,
                                             '2': 0.42857142857142855,
                                             '3': 0,
                                             '4': 0.2857142857142857,
                                             '5': 0.14285714285714285,
                                             '6': 0,
                                             '7': 0}},
                                   {'doc': 'If humans were to get burned they could '
                                           'receive a split thickness skin graft (STSG) '
```

Figure 10 (Cont.) Python Code for LDA model.

APPENDIX B

Table 9. Combination 1 containing 1 topic and 1 word.

Topics	Words
1	Cells

Table 10. Combination 2 containing 2 topics and 2 words.

Topics	Words	
1	Cells	Proliferate
2	Body	Water

Table 11. Combination 3 containing 2 topics and 5 words.

Topics	Words				
1	Cells	Differentiate	Form	Proliferate	Wound
2	Body	Fish	Lens	Survive	Water

Table 12. Combination 4 containing 2 topics and 10 words.

Topics	Words				
1	Body	Fish	Food	Growth	Layer
	Ocean	Skin	Structure	Survive	Water
2	Ability	Blastema	Cells	Differentiate	Epithelial
	Form	Proliferate	Proliferating	Regenerating	Wound

Table 13. Combination 6 containing 4 topics and 6 words.

Topics	Words					
1	Cells	Damaged	Kidney	Proliferate	Regenerate	Stop
2	Body	Fish	Ocean	Salt	Sea	Water
3	Animal	Change	Fish	Male	Skin	Survive
4	Cells	Differentiate	Form	Proliferate	Tissue	Wound

Table 14. Combination 7 containing 4 topics and 8 words.

Topics	Words			
1	Body	Environment	Make	Salt
	Skin	Survive	Temperature	Water
2	Body	Change	Female	Fish
	Liver	Male	Protein	South
3	Cells	Damaged	Epithelial	Form
	Kidney	Proliferate	Tail	Wound
4	Blastema	Blood	Cells	Differentiate
	Grow	Proliferate	Tail	Wound

Table 15. Combination 8 containing 4 topics and 10 words.

Topics	Words				
1	Ability	Change	Female	Fish	Food
	Layer	Lens	Male	Matrix	Sea
2	Ability	Blastema	Cells	Differentiate	Form
	Grow	Proliferate	Regenerating	Tail	Wound
3	Animal	Body	Environment	Fish	Ocean
	Salt	Skin	Survive	Water	Waters
4	Cell	Cells	Damaged	Form	Kidney
	Migrate	Proliferate	Regeneration	Site	Size

Table 16. Combination 9 containing 5 topics and 4 words.

Topics	Words			
1	Body	Fish	Salt	Water
2	Climate	Lens	South	Temperature
3	Cells	Differentiate	Form	Proliferate
4	Arm	Blood	Sea	Skin
5	Cells	Epithelial	Kidney	Migrate

Table 17. Combination 10 containing 5 topics and 6 words.

Topics	Words					
1	Cells	Create	Epithelial	Kidney	Signals	Skin
2	Cells	Differentiate	Form	Proliferate	Regeneration	Wound
3	Body	Head	Internal	Lens	Nest	Sea
4	Back	Bone	Legs	Liver	Lost	Tooth
5	Environment	Fish	Leg	Salt	Survive	Water

Table 18. Combination 11 containing 6 topics and 4 words.

Topics	Words			
1	Change	Fish	Male	Survive
2	Arm	Leg	Limb	Size
3	Body	Kidney	Salt	Water
4	Damaged	Humans	Regeneration	Zebrafish
5	Cells	Kidney	Migrate	Proliferate
6	Cells	Differentiate	Form	Proliferate

Table 19. Combination 12 containing 6 topics and 6 words.

Topics	Words					
1	Change	Fish	Lens	Male	Salt	Water
2	Ability	Body	Cells	Proliferate	Regenerating	Regeneration
3	Blastema	Cells	Differentiate	Form	Tail	Wound
4	Animals	Body	Plant	Protein	System	Tooth
5	Cells	Layer	Layers	Skin	Stage	Zebrafish
6	Cells	Kidney	Leg	Proliferate	Size	Stop

Table 20. Combination 13 containing 6 topics and 8 words.

Topics	Words			
1	Animal	Blood	Body	Cord
	Muscles	Skin	Spinal	Tissue
2	Ants	Create	Hair	Layer
	Layers	Nest	Shell	Snails
3	Cells	Form	Kidney	Mesenchymal
	Migrate	Start	Tail	Wound
4	Ability	Cells	Damaged	Differentiate
	Epithelial	Proliferate	Proliferating	Regenerating
5	Body	Change	Environment	Fish
	Male	Salt	Survive	Water
6	Complex	Liver	Regeneration	Sea
	System	Teeth	Tissue	Tooth

Table 21. Combination 15 containing 8 topics and 8 words.

Topics	Words			
1	Animal	Change	Climate	Environment
	Genes	Northern	South	Zebrafish
2	Cells	Epithelial	Epithelium	Form
	Kidney	Phase	Signals	Tooth
3	Claw	Exoskeleton	Leg	Legs
	Lens	Limb	Nest	Proteins
4	Body	Grow	Muscle	Muscles
	Sea	Skin	Tissue	Tissues
5	Blastema	Cells	Damaged	Form
	Proliferate	Proliferating	Regenerate	Regenerating
6	Ability	Animals	Food	Internal
	Produce	Sea	Sharks	Water
7	Bacteria	Eating	Fish	Place
	Protein	Quill	Small	Toxin
8	Blood	Cells	Cord	Differentiate
	Layer	Proliferate	Tail	Wound

Table 22. Probabilities of documents belonging to each topic of combination 2.

Probabilities	
Topic 1	Topic 2
0.5	0.5
0	1
0.4	0.6
0	1
0	1
1	0
0	1
1	0
0.75	0.25
0.8571	0.1429
1	0
0.8333	0.1667
1	0
0.5455	0.4545
0.7857	0.2143
1	0
0.8889	0.1111
0.5556	0.4444
1	0
1	0
1	0
1	0
0.8333	0.1667
0.2	0.8
0.8571	0.1429
1	0
0.4444	0.5556

Table 23. Probabilities of documents belonging to each topic of combination 3.

Probabilities	
Topic 1	Topic 2
1	0
0.875	0.125
0.8	0.2
0.5	0.5
0	1
0	1
0	1
0	1
0	1
0	1
0.1538	0.8462
0.7333	0.2667
1	0
1	0
0.5	0.5
1	0
0.5	0.5
1	0
0.6667	0.3333
1	0
1	0
1	0
0.7778	0.2222
1	0
0.8421	0.1579
0.9333	0.0667
1	0
0.9091	0.0909
1	0
0.9474	0.0526
0.9474	0.0526
1	0

Table 24. Probabilities of documents belonging to each topic of combination 4.

Probabilities	
Topic 1	Topic 2
0.5	0.5
0.6	0.4
0.3333	0.6667
0.5714	0.4286
0.9231	0.0769
1	0
1	0
0.8333	0.1667
1	0
0.875	0.125
0.15	0.85
0.4	0.6
0.2857	0.7143
0.5	0.5
0.2	0.8
0.1765	0.8235
0.2105	0.7895
0	1
0.3478	0.6522
0	1
0.0909	0.9091
0.1714	0.8286
0	1
0.2308	0.7692
0.1081	0.8919
0	1
0.2273	0.7727
0	1
0.1515	0.8485
0.2857	0.7143
0.069	0.931

Table 25. Probabilities of documents belonging to each topic of combination 5.

Probabilities			
Topic 1	Topic 2	Topic 3	Topic 4
0	0.3333	0.3333	0.3333
0	0.5	0.5	0
0	0.6667	0.1667	0.1667
0.25	0.5	0.25	0
0	0	0.3333	0.6667
0	0	0.4	0.6
0	0.5	0	0.5
0	0	0.8333	0.1667
0	0	0.5	0.5
0	0.1667	0.8333	0
0.3333	0.4444	0.2222	0
0.04	0.92	0	0.04
0.0909	0.9091	0	0
0.3333	0.3333	0.3333	0
0.4375	0.5	0.0625	0
0.2353	0.6471	0.0588	0.0588
0.3529	0.6471	0	0
0	0.6296	0.1111	0.2593
0	0.83333	0.1667	0
0	0.52	0.24	0.24
0.1071	0.5357	0.0357	0.3214
0	0.7727	0.0909	0.1364
0.0667	0.5333	0	0.4
0.0667	0.9333	0	0
0	0.9167	0.0833	0
0	1	0	0
0	1	0	0
0	1	0	0
0	0.8333	0.1667	0
0	1	0	0
0.125	0.75	0.125	0

Table 26. Probabilities of documents belonging to each topic of combination 6.

Probabilities			
Topic 1	Topic 2	Topic 3	Topic 4
0	0	0.3333	0.6667
0.3889	0.1667	0.0556	0.3889
0.4	0.0667	0	0.5333
0.3636	0	0.0909	0.5455
0	0.4	0.4	0.2
0	0	1	0
0	0	1	0
0	0	1	0
0.0769	0.4615	0.3077	0.1538
0	0.3333	0.6667	0
0	1	0	0
0.3809	0.0476	0	0.5714
0.25	0.0357	0	0.7143
0.1429	0.2857	0	0.5714
0.3333	0.2222	0	0.4444
0.2609	0.2174	0.0435	0.4783
0.3077	0.3461	0.1538	0.1923
0.3333	0.2222	0.2222	0.2222
0.3658	0	0	0.6341
0.1786	0.0714	0.25	0.5
0.2	0.2667	0	0.5333
0.3548	0.1613	0.129	0.3548
0.3888	0.1111	0	0.5
0.5	0	0	0.5
0.36	0.04	0.16	0.44
0.35	0.25	0	0.4
0.4167	0.1667	0	0.4167
0.3636	0	0	0.6364
0.5	0	0.125	0.375
0.4375	0.0625	0.0625	0.4375
0.1739	0.3043	0.2609	0.2609

Table 27. Probabilities of documents belonging to each topic of combination 7.

Probabilities			
Topic 1	Topic 2	Topic 3	Topic 4
0.0625	0.0625	0.4375	0.4375
0.0588	0	0.4118	0.5294
0.4167	0.0833	0.3333	0.1667
0.4	0	0	0.6
1	0	0	0
0.25	0.25	0.5	0
0.0667	0	0.7333	0.2
0.75	0.1429	0.1071	0
0.5	0.5	0	0
0.6667	0	0	0.3333
0	0.6667	0.3333	0
0.0645	0.2903	0.2258	0.4193
0.0741	0.0741	0.3333	0.5185
0	0.0833	0.3333	0.5833
0.05	0.05	0.55	0.35
0	0.5454	0.1818	0.2727
0.1667	0.1333	0.2667	0.4333
0.0714	0.0476	0.3095	0.5714
0	0	0.4737	0.5263
0.0303	0.0303	0.5454	0.3939
0.1143	0.1714	0.3143	0.4
0	0.0435	0.5652	0.3913
0	0	0.6667	0.3333
0.1111	0	0.3889	0.5
0	0.8571	0.1429	0
0	0	0.44	0.56
0.05	0.05	0.45	0.45
0.4545	0	0.4545	0.0909
0.0455	0.0455	0.5	0.4091
0.0909	0	0.4545	0.4545
0.2083	0.25	0.3333	0.2083
0.069	0.1034	0.4483	0.3793

Table 28. Probabilities of documents belonging to each topic of combination 8.

Probabilities			
Topic 1	Topic 2	Topic 3	Topic 4
0.0526	0.2632	0.1579	0.5263
0	0.32	0.12	0.56
0.3333	0.6667	0	0
0	0.32	0.2	0.48
0.125	0.0313	0.6875	0.1563
0.0938	0	0.75	0.1563
0.3333	0	0.6667	0
0.5	0	0	0.5
0.1667	0.1667	0.6667	0
0.1667	0.1667	0.5	0.1667
0.3	0.05	0.65	0
0.0357	0.5	0	0.4643
0	0.375	0	0.625
0.1136	0.2955	0.0909	0.5
0.0606	0.4545	0.0152	0.4697
0.0588	0.4706	0.0294	0.4412
0.1351	0.2432	0.1351	0.4865
0.0541	0.3243	0.027	0.5946
0.0638	0.4468	0.0426	0.4468
0.0217	0.3696	0	0.6087
0.0545	0.4545	0.0545	0.4364
0.0286	0.3429	0.0857	0.5429
0.1053	0.4035	0.1228	0.3684
0.0204	0.5714	0.0204	0.3878
0.1	0.4	0.08	0.4
0.0976	0.3902	0.1219	0.3902
0.0263	0.421	0	0.5526
0.1389	0.4167	0.0556	0.3889
0.0811	0.4865	0	0.4324
0.0909	0.2727	0	0.6364
0.027	0.4054	0	0.5676
0	0.3809	0.0476	0.5714

Table 29. Probabilities of documents belonging to each topic of combination 9.

Probabilities				
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
0.1111	0	0.2222	0.5556	0.1111
0	0	0.4211	0.2632	0.3158
0.2143	0	0.4286	0	0.3571
0.1667	0	0	0.8333	0
0.6667	0	0	0.3333	0
0	0	0.125	0.875	0
0.0417	0	0.625	0	0.3333
0.3333	0	0.5	0	0.1667
0	0	0.6667	0	0.3333
0.25	0	0.45	0	0.3
0.4286	0	0.2143	0.2143	0.1429
0.25	0	0.25	0.25	0.25
0	0	0	1	0
0	0	0.6757	0.027	0.2973
0.0417	0	0.375	0.4167	0.1667
0.1	0	0.4	0.4	0.1
0.1923	0	0.4231	0	0.3846
0.0588	0	0.5294	0	0.4118
0	0	0.3636	0.4545	0.1818
0	0	0.5	0.2273	0.2727
0.1905	0	0.3809	0.2857	0.1429
0	0	0.5	0.2	0.3
0	0.3	0.4667	0	0.2333
0	0	0.4286	0.1429	0.4286
0.0769	0	0.5385	0	0.3846
0.375	0	0.375	0	0.25
0	0	0.4444	0.1111	0.4444

Table 30. Probabilities of documents belonging to each topic of combination 10.

Probabilities				
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
0.2727	0.3636	0	0.2727	0.0909
0.3333	0.3333	0.3333	0	0
0	0.5	0.5	0	0
0.1667	0.3333	0.3333	0	0.1667
0	0	0	0	1
0	0	0	0.5	0.5
0	0.5	0.25	0.25	0
0	0	1	0	0
0	0	0.5	0	0.5
0	0	0.1111	0	0.8889
0.3889	0.2778	0.3333	0	0
0.3043	0.4783	0.0435	0.0435	0.1304
0.3333	0.6667	0	0	0
0.3333	0.6667	0	0	0
0.4	0.3	0.1	0	0.2
0.4286	0.5	0.0714	0	0
0.4091	0.3182	0.0455	0	0.2273
0.2857	0.5714	0.1429	0	0
0.1333	0.5333	0.0667	0.0667	0.2
0.2333	0.6	0.1667	0	0
0.3667	0.4333	0	0.2	0
0.6667	0.3333	0	0	0
0.3214	0.5	0	0.0357	0.1429
0.5714	0.4286	0	0	0
0.4865	0.4324	0	0	0.0811
0.2857	0.6667	0.0476	0	0
0.4	0.6	0	0	0
0.303	0.6364	0	0	0.0606
0.4762	0.4762	0	0.0476	0
0.2593	0.6667	0	0	0.0741
0.2593	0.6667	0	0	0.0741

Table 31. Probabilities of documents belonging to each topic of combination 11.

Probabilities					
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
0.2	0.1	0.3	0	0.2	0.2
0	0.1667	0	0.6667	0	0.1667
0	0.0714	0.0714	0	0.3571	0.5
1	0	0	0	0	0
0.5	0.25	0	0	0	0.25
0	0.875	0	0.125	0	0
0	0.5714	0.2857	0.1428	0	0
0	0.75	0	0.125	0	0.125
0	0.75	0.25	0	0	0
0.3529	0.0588	0	0	0.2353	0.3529
0	0	0	0	0.5	0.5
0.1154	0.0769	0.0769	0.0385	0.2692	0.4231
0.1	0	0.05	0.05	0.3	0.5
0	0	0	0	0.375	0.625
0.0556	0	0.0556	0	0.2778	0.6111
0	0.25	0	0.25	0.25	0.25
0	0.0435	0.2174	0.0435	0.2609	0.4348
0	0.0645	0.0968	0.0323	0.3548	0.4516
0	0	0	0.0588	0.4118	0.5294
0	0	0.0357	0.0714	0.3571	0.5357
0	0	0.1818	0	0.2273	0.5909
0	0	0	0.0526	0.3684	0.5789
0	0	0.2424	0.0606	0.4545	0.2424
0	0	0	0.1538	0.3077	0.5385
0	0.6667	0	0	0	0.3333
0	0.2222	0	0	0.2963	0.4815
0	0.24	0.04	0.16	0.24	0.32
0	0	0.5	0	0.125	0.375
0	0	0.0526	0.1053	0.3158	0.5263
0.1111	0	0	0	0.3333	0.5556

Table 32. Probabilities of documents belonging to each topic of combination 12.

Probabilities					
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
0	0.25	0.15	0.2	0.15	0.25
0.1111	0.1111	0.2222	0.4444	0.1111	0
0	0.2308	0.3846	0.1538	0.1538	0.0769
0	0.2963	0.1481	0	0.2593	0.2963
0.5	0	0	0.5	0	0
0.4	0.2	0	0.2	0.2	0
0.6667	0	0	0.1667	0.1667	0
1	0	0	0	0	0
0	0.2222	0.1111	0	0.5556	0.1111
0.3333	0.2778	0	0.2778	0.0556	0.0556
0.1667	0.3333	0	0.1667	0.1667	0.1667
0	0.2414	0.2759	0	0.2069	0.2759
0.1081	0.2162	0.2432	0	0.1351	0.2973
0	0.1774	0.4355	0.0161	0.2097	0.1613
0	0.2059	0.3235	0	0.1765	0.2941
0.1786	0.25	0.1786	0	0.1071	0.2857
0.0278	0.2222	0.2222	0	0.1944	0.3333
0	0.2826	0.3478	0.0217	0.1522	0.1957
0	0.2619	0.2619	0	0.1429	0.3333
0.0345	0.2931	0.2931	0.0345	0.1379	0.2069
0	0.2791	0.2326	0.0465	0.186	0.2558
0.0952	0.2222	0.2381	0.0159	0.1746	0.254
0	0.2407	0.4074	0.0185	0.1296	0.2037
0	0.2745	0.2941	0.0588	0.1569	0.2157
0	0.275	0.175	0	0.325	0.225
0	0.2609	0.2826	0	0.1957	0.2609
0.0256	0.2564	0.2308	0.0256	0.2564	0.2051
0	0.3	0.275	0	0.175	0.25
0	0.2353	0.1765	0.2353	0.1765	0.1765
0	0.2821	0.2564	0.0513	0.1538	0.2564
0.0455	0.2273	0.2727	0.2273	0.0455	0.1818
0	0.3913	0.2609	0.087	0.087	0.1739

Table 33. Probabilities of documents belonging to each topic of combination 13.

Probabilities					
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
0.0952	0.0952	0.3809	0.2857	0.0476	0.0952
0.2105	0.1579	0.3158	0.3158	0	0
0.4615	0.0769	0.1538	0.1538	0.0769	0.0769
0.5	0	0	0.1667	0.1667	0.1667
0	0	0	0	1	0
0.2	0	0.3	0	0.4	0.1
0.3	0	0.1	0	0.5	0.1
0.25	0	0	0	0.75	0
0.625	0	0	0.125	0.25	0
0.2	0.0667	0.1333	0.0667	0.3333	0.2
0.1	0	0.6	0	0.3	0
0.16	0.08	0.16	0.28	0.04	0.28
0.129	0.1613	0.2581	0.2903	0.0323	0.129
0.1026	0.0256	0.359	0.359	0.0256	0.1282
0.1	0.1	0.3	0.2	0.2	0.1
0.4	0	0.1	0.3	0	0.2
0.1875	0	0.3125	0.3125	0.1563	0.0313
0.1111	0	0.1111	0.2222	0.2222	0.3333
0.3636	0	0.0909	0.3636	0.1818	0
0	0.6	0	0.4	0	0
0.0455	0.0455	0.3636	0.5455	0	0
0.3056	0.0556	0.3611	0.1389	0.0278	0.1111
0	0	0.375	0.3125	0.125	0.1875
0.0313	0	0.3438	0.4375	0.1563	0.0313
0.05	0	0.4	0.45	0.05	0.05
0.1111	0	0.4444	0.3889	0	0.0556
0.16	0	0.32	0.36	0.12	0.04
0.3103	0.0345	0.2414	0.2069	0.1379	0.069
0	0.44	0.2	0.28	0	0.08
0	0.04	0.36	0.56	0	0.04
0.1429	0	0.4286	0.4286	0	0
0.0769	0.0385	0.2692	0.3846	0.0385	0.1923

Table 34. Probabilities of documents belonging to each topic of combination 15.

Probabilities							
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
0	0.1176	0	0.4706	0.1176	0.1176	0.0588	0.1176
0	0	0	0.8	0	0	0.1	0.1
0	0.1053	0	0	0.2632	0.1053	0.5263	0
0	0.125	0.05	0.25	0.175	0.05	0.075	0.275
0.2273	0.1364	0	0.2273	0	0.4091	0	0
0.1429	0.0952	0	0.0952	0	0.3809	0.2857	0
0.1667	0.1667	0	0	0	0.3889	0.2778	0
0.2273	0.1818	0	0.0909	0	0.4091	0.0909	0
0.4667	0	0	0.0667	0	0.1333	0.3333	0
0.1	0.1	0	0.05	0.1	0.05	0.5	0.1
0.2667	0	0	0.1333	0	0.3333	0.2667	0
0	0.2439	0	0.0488	0.3171	0	0	0.3902
0.0417	0.1667	0	0.25	0.375	0	0	0.1667
0	0.1176	0	0.0588	0.2353	0.1765	0.0588	0.3529
0.0313	0.125	0	0.1875	0.3438	0.125	0	0.1875
0	0.1389	0	0.2222	0.1944	0.1667	0	0.2778
0	0.25	0	0.125	0.4063	0.0938	0	0.125
0.0645	0.129	0	0.1613	0.3226	0.0645	0.0645	0.1935
0	0.1639	0	0.1311	0.3443	0.0984	0.0164	0.2459
0	0.2632	0	0	0.3158	0.0526	0.0526	0.3158
0	0.4091	0	0.0909	0.1818	0.0455	0.0909	0.1818
0	0.2708	0.0208	0	0.4167	0.0625	0.0208	0.2083
0	0.2857	0	0	0.2857	0	0	0.4286
0.0185	0.3333	0	0	0.3148	0.0185	0.0556	0.2593
0	0.2195	0.0976	0.0732	0.3171	0.0244	0	0.2683
0.0435	0.1739	0	0	0.4783	0.0435	0	0.2609
0	0.2537	0	0.0298	0.3433	0.0448	0	0.3284
0	0.3235	0	0	0.3529	0.0588	0	0.2647
0.0172	0.1897	0	0.1207	0.3103	0.0517	0	0.3103
0.0172	0.1897	0	0.1207	0.3103	0.0517	0	0.3103
0.0208	0.375	0	0	0.3125	0	0	0.2917
0	0.3571	0	0.0238	0.3095	0.0238	0	0.2857

Table 35. Code number for each analogue.

Number	Analogue name
1	General wound healing
2	Skin Grafts, same individual
3	Muscles (Humans)
4	Transplants, in vitro
5	Transplants, 3D printing
6	Ear piercing
7	Tattoos
8	Common Hippopotamus
9	Wood frog
10	Skin graft, allograft
11	Photosynthetic systems
12	Bioengineered skin grafts
13	Sewing finger or toe back on
14	Stents
15	North American Opossum
16	Vibrio cholerae
17	Salmonella
18	Scavenging reactive species. Predatory Mites
19	Losing a limb
20	Losing a tail
21	Vision loss
22	Hearing loss
23	Losing teeth
24	Losing an arm
25	Having one kidney
26	Bull Sharks
27	Pink Salmon
28	American Shad
29	Red drum
30	Atlantic striped bass
31	Rainbow smelt
32	Qhino checkerspot
33	Alpine chipmunk
34	European wasp spider
35	Nine-banded armadillo
36	Red-billed gull
37	Wild salamanders in North America
38	Tawny owls
39	Colonies of grass-eating ants
40	Spider Monkey
41	Prosthetics

Table 35 (Cont.) Code number for each analogue.

Number	Analogue name
42	Wheelchairs
43	Robotics
44	Self-driving wheelchair
45	Jellyfish
46	Blue streak cleaner wrasse
47	Cuttlefish
48	Clown fish
49	Parrotfish,
50	Hawkfish
51	Scarlet macaw
52	Spanish shawl
53	Puffer fish
54	Asian tiger keelback
55	Great barracuda
56	Mexican Axolotl
57	Snapping Shrimp or Pistol Shrimp
58	Sea Cucumbers
59	Planarian
60	Zebrafish
61	Stick insects
62	Colonies of Temnothorax ants
63	American Cockroach
64	Hairy Frog
65	Brittle Stars
66	Brown ghost knifefish
67	Japanese Spiky Sea Cucumber
68	Common Octopus
69	Deer
70	Colonial Sea Squirt
71	Painted turtle
72	Italian crested newt
73	Human
74	Human
75	Immortal jellyfish
76	Snail
77	Zebrafish
78	Catfish
79	American alligator
80	Cichlid
81	Rat
82	Brown Hydra

Table 35 (Cont.) Code number for each analogue.

Number	Analogue name
83	Tropical clawed frog
84	Mammal
85	North American Opossum
86	Sea cucumbers
87	Chinese Mystery Snail
88	Humans
89	Percival’s Spiny mouse
90	Purple sea urchin
91	Noble Feather Star
92	Antarctic brittle star
93	Fatty membranes of cells
94	Eastern (Red-spotted) newt
95	Hermann’s tortoise
96	Human
97	Zebrafish
98	Pacific blood star
99	Underground storage swelling enables regeneration
100	Freshwater pearl mussel
101	Rabbits
102	Killer whale
103	Long-spined sea urchin
104	Zebrafish
105	Japanese sea lily
106	Human
107	Cattle
108	Senegal (gray) bichir
109	Hair regrowth
110	European lesser spotted dogfish
111	Rat
112	Chicken
113	American bullfrog
114	House mouse
115	Mexican Axolotl
116	Mexican Axolotl
117	Little skate
118	Rabbit
119	Zebrafish
120	Zebrafish
121	Atlantic sand fiddler
122	Zebrafish

Table 35 (Cont.) Code number for each analogue.

Number	Analogue name
123	Atlantic tomcod
124	Japanese Spiky Sea Cucumber
125	Whitespotted Bamboo Shark
126	Goldfish
127	Green Anole
128	Rat
129	Laver spire snail
130	African clawed frog
131	Deepwater Spiny Dogfish
132	Common leopard gecko
133	Eastern (red-spotted) newt
134	Japanese rice fish
135	Megarian Banded Centipede
136	American five-lined skink
137	Chicken
138	Common fruit fly
139	Eastern (red-spotted) newt
140	Chicken
141	Oyster toadfish
142	House mouse
143	Blackback land crab
144	Catfish
145	Rainbow trout
146	Brown garden snail
147	Nile tilapia
148	Pharaoh cuttlefish
149	Snail fur
150	Rusty crayfish
151	Eastern glass lizard
152	Star ascidian
153	Himalayan Newt
154	Porcupine
155	Halechiniscus grevini
156	Vityazicrinus petrachenkoi
157	Shark
158	Chicken
159	Eastern (Red-spotted) newt
160	California sea hare
161	Black-ball sponge

Table 36. Final classification for analogues organized by probability (topics 5 through 8).

Probabilities							
Number	Topic 5	Number	Topic 6	Number	Topic 7	Number	Topic 8
125	0.71	107	0.56	125	0.71	107	0.56
159	0.71	6	0.5	159	0.71	6	0.5
133	0.62	145	0.49	133	0.62	145	0.49
142	0.61	114	0.47	142	0.61	114	0.47
112	0.6	123	0.47	112	0.6	123	0.47
78	0.59	147	0.46	78	0.59	147	0.46
106	0.59	84	0.44	106	0.59	84	0.44
129	0.57	88	0.44	129	0.57	88	0.44
135	0.57	140	0.44	135	0.57	140	0.44
84	0.56	149	0.44	84	0.56	149	0.44
140	0.56	141	0.42	140	0.56	141	0.42
87	0.55	144	0.42	87	0.55	144	0.42
74	0.54	106	0.41	74	0.54	106	0.41
154	0.54	5	0.4	154	0.54	5	0.4
113	0.53	87	0.4	113	0.53	87	0.4
114	0.53	112	0.4	114	0.53	112	0.4
128	0.53	128	0.4	128	0.53	128	0.4
152	0.53	25	0.39	152	0.53	25	0.39
146	0.52	126	0.39	146	0.52	126	0.39
156	0.52	74	0.38	156	0.52	74	0.38
17	0.5	117	0.38	17	0.5	117	0.38
89	0.5	134	0.38	89	0.5	134	0.38
98	0.5	17	0.37	98	0.5	17	0.37
103	0.5	133	0.37	103	0.5	133	0.37
108	0.5	135	0.37	108	0.5	135	0.37
111	0.5	66	0.36	111	0.5	66	0.36
118	0.5	110	0.36	118	0.5	110	0.36
139	0.5	118	0.36	139	0.5	118	0.36
105	0.49	124	0.36	105	0.49	124	0.36
132	0.49	154	0.35	132	0.49	154	0.35
136	0.49	81	0.33	136	0.49	81	0.33
148	0.49	113	0.33	148	0.49	113	0.33
63	0.48	131	0.33	77	0.04	104	0.12
69	0.48	138	0.33	108	0.04	71	0.1
5	0.47	67	0.32	149	0.04	136	0.1

Table 36 (Cont.) Final classification for analogues organized by probability (topics 5 through 8).

Probabilities							
Number	Topic 5	Number	Topic 6	Number	Topic 7	Number	Topic 8
67	0.47	96	0.32	156	0.04	56	0.09
70	0.47	4	0.31	12	0.03	59	0.09
101	0.47	108	0.31	151	0.03	65	0.09
138	0.47	130	0.31	1	0	52	0.08
151	0.47	73	0.3	3	0	63	0.08
160	0.47	132	0.3	4	0	68	0.08
155	0.46	14	0.29	5	0	102	0.08
56	0.45	82	0.29	6	0	111	0.08
72	0.45	83	0.29	7	0	150	0.08
91	0.45	120	0.29	8	0	155	0.08
158	0.45	125	0.29	9	0	156	0.08
66	0.44	142	0.29	10	0	5	0.07
77	0.44	146	0.29	11	0	91	0.07
80	0.44	91	0.28	14	0	132	0.07
97	0.44	97	0.28	15	0	143	0.07
115	0.44	122	0.28	16	0	153	0.07
116	0.44	151	0.28	17	0	161	0.07
124	0.44	156	0.28	18	0	4	0.06
137	0.44	2	0.27	21	0	70	0.06
82	0.42	59	0.27	23	0	78	0.06
86	0.42	76	0.27	25	0	123	0.06
99	0.42	98	0.27	26	0	126	0.06
59	0.41	158	0.27	27	0	137	0.06
83	0.41	161	0.27	29	0	141	0.06
104	0.41	69	0.26	30	0	144	0.06
127	0.41	85	0.26	31	0	145	0.06
149	0.4	150	0.26	32	0	67	0.05
161	0.4	90	0.25	34	0	119	0.05
85	0.39	99	0.25	35	0	120	0.05
92	0.38	103	0.25	36	0	146	0.05
117	0.38	152	0.25	37	0	147	0.05
120	0.38	155	0.25	38	0	25	0.04
4	0.37	56	0.24	39	0	69	0.04
130	0.37	80	0.24	44	0	77	0.04
2	0.36	92	0.24	45	0	80	0.04

Table 36 (Cont.) Final classification for analogues organized by probability (topics 5 through 8).

Probabilities							
Number	Topic 5	Number	Topic 6	Number	Topic 7	Number	Topic 8
65	0.36	63	0.23	46	0	82	0.04
110	0.36	72	0.23	47	0	97	0.04
143	0.36	129	0.23	48	0	118	0.04
119	0.35	148	0.23	49	0	64	0.03
6	0.33	109	0.22	50	0	105	0.03
15	0.33	136	0.22	51	0	113	0.03
16	0.33	137	0.22	52	0	134	0.03
18	0.33	75	0.21	53	0	148	0.03
68	0.33	89	0.21	54	0	152	0.03
79	0.33	111	0.21	55	0	1	0
107	0.33	77	0.2	58	0	2	0
109	0.33	101	0.2	59	0	6	0
131	0.33	119	0.2	60	0	7	0
75	0.32	139	0.2	62	0	9	0
94	0.32	95	0.19	64	0	10	0
95	0.31	105	0.19	65	0	11	0
122	0.31	157	0.19	66	0	12	0
73	0.3	9	0.18	69	0	13	0
147	0.3	70	0.18	70	0	14	0
153	0.3	78	0.18	71	0	15	0
1	0.29	100	0.18	73	0	16	0
14	0.29	127	0.18	74	0	17	0
64	0.29	16	0.17	75	0	18	0
76	0.27	68	0.17	76	0	21	0
100	0.27	115	0.17	78	0	23	0
71	0.26	116	0.17	80	0	24	0
134	0.26	30	0.16	82	0	26	0
150	0.26	71	0.16	83	0	27	0
58	0.25	86	0.16	84	0	28	0
88	0.25	94	0.16	85	0	29	0
90	0.25	55	0.15	86	0	30	0
126	0.24	153	0.15	88	0	31	0
57	0.23	1	0.14	89	0	32	0
96	0.23	11	0.14	90	0	34	0
123	0.23	102	0.14	91	0	35	0

Table 36 (Cont.) Final classification for analogues organized by probability (topics 5 through 8).

Probabilities							
Number	Topic 5	Number	Topic 6	Number	Topic 7	Number	Topic 8
145	0.23	143	0.14	93	0	36	0
24	0.22	159	0.14	95	0	37	0
141	0.22	47	0.13	96	0	38	0
12	0.2	64	0.13	97	0	40	0
13	0.2	79	0.13	98	0	41	0
93	0.2	29	0.12	99	0	43	0
157	0.19	58	0.12	100	0	44	0
9	0.18	104	0.12	101	0	45	0
39	0.17	160	0.12	102	0	46	0
81	0.17	12	0.11	103	0	47	0
102	0.17	31	0.11	104	0	48	0
144	0.16	13	0.1	105	0	49	0
7	0.15	60	0.1	106	0	50	0
60	0.15	26	0.09	107	0	51	0
11	0.14	65	0.09	109	0	53	0
47	0.13	27	0.08	110	0	62	0
121	0.11	28	0.08	112	0	66	0
25	0.09	52	0.08	113	0	72	0
42	0.08	57	0.08	114	0	74	0
52	0.08	61	0.08	115	0	76	0
55	0.08	93	0.07	116	0	84	0
61	0.08	121	0.04	117	0	87	0
3	0	3	0	118	0	88	0
8	0	7	0	119	0	89	0
10	0	8	0	120	0	90	0
19	0	10	0	122	0	92	0
20	0	15	0	123	0	93	0
21	0	18	0	124	0	94	0
23	0	19	0	125	0	96	0
26	0	20	0	126	0	98	0
27	0	21	0	127	0	100	0
28	0	23	0	128	0	106	0
29	0	24	0	129	0	107	0
30	0	32	0	131	0	108	0
31	0	34	0	133	0	109	0

Table 36 (Cont.) Final classification for analogues organized by probability (topics 5 through 8).

Probabilities							
Number	Topic 5	Number	Topic 6	Number	Topic 7	Number	Topic 8
145	0.23	143	0.14	93	0	36	0
24	0.22	159	0.14	95	0	37	0
141	0.22	47	0.13	96	0	38	0
12	0.2	64	0.13	97	0	40	0
13	0.2	79	0.13	98	0	41	0
93	0.2	29	0.12	99	0	43	0
157	0.19	58	0.12	100	0	44	0
9	0.18	104	0.12	101	0	45	0
39	0.17	160	0.12	102	0	46	0
81	0.17	12	0.11	103	0	47	0
102	0.17	31	0.11	104	0	48	0
144	0.16	13	0.1	105	0	49	0
7	0.15	60	0.1	106	0	50	0
60	0.15	26	0.09	107	0	51	0
11	0.14	65	0.09	109	0	53	0
47	0.13	27	0.08	110	0	62	0
121	0.11	28	0.08	112	0	66	0
25	0.09	52	0.08	113	0	72	0
42	0.08	57	0.08	114	0	74	0
52	0.08	61	0.08	115	0	76	0
55	0.08	93	0.07	116	0	84	0
61	0.08	121	0.04	117	0	87	0
3	0	3	0	118	0	88	0
8	0	7	0	119	0	89	0
10	0	8	0	120	0	90	0
19	0	10	0	122	0	92	0
20	0	15	0	123	0	93	0
21	0	18	0	124	0	94	0
23	0	19	0	125	0	96	0
26	0	20	0	126	0	98	0
27	0	21	0	127	0	100	0
28	0	23	0	128	0	106	0
29	0	24	0	129	0	107	0
30	0	32	0	131	0	108	0
31	0	34	0	133	0	109	0

Table 36 (Cont.) Final classification for analogues organized by probability (topics 5 through 8).

Probabilities							
Number	Topic 5	Number	Topic 6	Number	Topic 7	Number	Topic 8
32	0	35	0	134	0	110	0
34	0	36	0	135	0	112	0
35	0	37	0	136	0	114	0
36	0	38	0	139	0	117	0
37	0	39	0	140	0	124	0
38	0	40	0	141	0	125	0
40	0	41	0	142	0	128	0
41	0	42	0	144	0	129	0
43	0	43	0	145	0	130	0
44	0	44	0	146	0	131	0
45	0	45	0	147	0	133	0
46	0	46	0	148	0	135	0
48	0	48	0	152	0	138	0
49	0	49	0	153	0	140	0
50	0	50	0	155	0	142	0
51	0	51	0	158	0	149	0
53	0	53	0	159	0	154	0
54	0	54	0	160	0	157	0
62	0	62	0	161	0	159	0
22		22		22		22	
33		33		33		33	

Table 37. Data classification by analogue.

Probabilities								
Number	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
1	0	0.14	0.43	0	0.29	0.14	0	0
2	0	0.14	0.14	0	0.36	0.27	0.09	0
3	0	0	0	0	0	0	0	1
4	0.06	0	0.19	0	0.37	0.31	0	0.06
5	0	0	0.07	0	0.47	0.4	0	0.07
6	0	0.08	0.08	0	0.33	0.5	0	0
7	0	0.31	0.54	0	0.15	0	0	0
8	0	0.17	0.33	0.33	0	0	0	0.17
9	0.27	0	0.27	0.09	0.18	0.18	0	0
10	0	0.5	0.5	0	0	0	0	0
11	0.71	0	0	0	0.14	0.14	0	0
12	0	0.26	0.4	0	0.2	0.11	0.03	0
13	0.1	0.2	0.3	0	0.2	0.1	0.1	0
14	0	0	0.43	0	0.29	0.29	0	0
15	0	0	0.67	0	0.33	0	0	0
16	0.17	0	0.33	0	0.33	0.17	0	0
17	0	0	0.12	0	0.5	0.37	0	0
18	0.5	0	0	0.17	0.33	0	0	0
19	0	0	0	0	0	0	0.67	0.33
20	0	0	0	0	0	0	0.25	0.75
21	1	0	0	0	0	0	0	0
22								
23	0	0.67	0	0.33	0	0	0	0
24	0	0	0.67	0	0.22	0	0.11	0
25	0.39	0	0	0.09	0.09	0.39	0	0.04
26	0.09	0	0.14	0.68	0	0.09	0	0
27	0.08	0.08	0.2	0.56	0	0.08	0	0
28	0.08	0	0.08	0.72	0	0.08	0.04	0
29	0.21	0	0	0.67	0	0.12	0	0
30	0.16	0	0.12	0.56	0	0.16	0	0
31	0.18	0	0.07	0.64	0	0.11	0	0
32	1	0	0	0	0	0	0	0
33								
34	1	0	0	0	0	0	0	0
35	0.11	0.22	0.56	0.11	0	0	0	0

Table 37 (Cont.) Data classification by analogue.

Probabilities								
Number	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
36	0.17	0.17	0.5	0.17	0	0	0	0
37	0.14	0.14	0.43	0.29	0	0	0	0
38	1	0	0	0	0	0	0	0
39	0	0	0.33	0.33	0.17	0	0	0.17
40	0	0	0.86	0	0	0	0.14	0
41	0	0	0.43	0	0	0	0.57	0
42	0	0	0.23	0.31	0.08	0	0.23	0.15
43	0	0	0.31	0	0	0	0.69	0
44	0	0	0	1	0	0	0	0
45	0	0	0	1	0	0	0	0
46	0.14	0	0	0.86	0	0	0	0
47	0	0	0.07	0.67	0.13	0.13	0	0
48	0	0	0.17	0.83	0	0	0	0
49	0	0	0	1	0	0	0	0
50	0.14	0	0.14	0.71	0	0	0	0
51	1	0	0	0	0	0	0	0
52	0.08	0.31	0.31	0.08	0.08	0.08	0	0.08
53	0.08	0.08	0.43	0.43	0	0	0	0
54	0	0.17	0.33	0.17	0	0	0	0.33
55	0.15	0	0	0.46	0.08	0.15	0	0.15
56	0	0	0.03	0.03	0.45	0.24	0.15	0.09
57	0	0	0	0.15	0.23	0.08	0.08	0.46
58	0	0	0	0	0.25	0.12	0	0.62
59	0	0	0.18	0.04	0.41	0.27	0	0.09
60	0	0.15	0.05	0.4	0.15	0.1	0	0.15
61	0	0.08	0.16	0	0.08	0.08	0.4	0.2
62	0	0	1	0	0	0	0	0
63	0	0	0	0	0.48	0.23	0.21	0.08
64	0	0.23	0.26	0.06	0.29	0.13	0	0.03
65	0	0	0.27	0.18	0.36	0.09	0	0.09
66	0	0	0.04	0.16	0.44	0.36	0	0
67	0	0	0.05	0	0.47	0.32	0.1	0.05
68	0	0	0.37	0	0.33	0.17	0.04	0.08
69	0	0.09	0.09	0.04	0.48	0.26	0	0.04
70	0	0	0.23	0.06	0.47	0.18	0	0.06

Table 37 (Cont.) Data classification by analogue.

Probabilities								
Number	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
71	0	0	0.37	0.1	0.26	0.16	0	0.1
72	0.03	0	0	0	0.45	0.23	0.29	0
73	0	0.1	0.1	0	0.3	0.3	0	0.2
74	0	0	0.08	0	0.54	0.38	0	0
75	0	0	0	0.32	0.32	0.21	0	0.16
76	0	0.13	0.33	0	0.27	0.27	0	0
77	0	0.08	0.04	0.16	0.44	0.2	0.04	0.04
78	0	0	0	0.18	0.59	0.18	0	0.06
79	0	0.2	0	0	0.33	0.13	0.13	0.2
80	0	0.12	0.08	0.08	0.44	0.24	0	0.04
81	0	0	0	0	0.17	0.33	0.17	0.33
82	0	0	0.21	0.04	0.42	0.29	0	0.04
83	0	0	0.06	0.03	0.41	0.29	0	0.21
84	0	0	0	0	0.56	0.44	0	0
85	0	0	0.03	0	0.39	0.26	0	0.32
86	0	0	0.13	0.1	0.42	0.16	0	0.19
87	0	0	0	0	0.55	0.4	0.05	0
88	0.28	0	0	0.03	0.25	0.44	0	0
89	0	0.14	0.14	0	0.5	0.21	0	0
90	0.25	0	0	0.25	0.25	0.25	0	0
91	0	0	0.21	0	0.45	0.28	0	0.07
92	0	0	0.29	0.05	0.38	0.24	0.05	0
93	0.2	0	0	0.53	0.2	0.07	0	0
94	0	0	0.06	0.03	0.32	0.16	0.42	0
95	0.06	0	0.18	0.12	0.31	0.19	0	0.12
96	0.18	0	0.23	0.04	0.23	0.32	0	0
97	0	0.08	0.04	0.12	0.44	0.28	0	0.04
98	0.03	0	0.2	0	0.5	0.27	0	0
99	0	0	0	0	0.42	0.25	0	0.33
100	0	0	0.36	0.18	0.27	0.18	0	0
101	0	0.07	0.07	0.07	0.47	0.2	0	0.13
102	0	0.11	0.36	0.14	0.17	0.14	0	0.08
103	0	0	0	0.12	0.5	0.25	0	0.12
104	0	0.12	0.06	0.18	0.41	0.12	0	0.12
105	0	0	0.16	0.13	0.49	0.19	0	0.03

Table 37 (Cont.) Data classification by analogue.

Probabilities								
Number	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
106	0	0	0	0	0.59	0.41	0	0
107	0	0	0.11	0	0.33	0.56	0	0
108	0.04	0	0	0.11	0.5	0.31	0.04	0
109	0	0	0.44	0	0.33	0.22	0	0
110	0.22	0	0	0.07	0.36	0.36	0	0
111	0	0	0.04	0	0.5	0.21	0.17	0.08
112	0	0	0	0	0.6	0.4	0	0
113	0	0	0.03	0.07	0.53	0.33	0	0.03
114	0	0	0	0	0.53	0.47	0	0
115	0	0	0	0.03	0.44	0.17	0	0.36
116	0	0	0	0.03	0.44	0.17	0	0.36
117	0.24	0	0	0	0.38	0.38	0	0
118	0	0	0.09	0	0.5	0.36	0	0.04
119	0	0.1	0.05	0.25	0.35	0.2	0	0.05
120	0	0.09	0.05	0.14	0.38	0.29	0	0.05
121	0.04	0	0	0.11	0.11	0.04	0.46	0.25
122	0	0.05	0.03	0.08	0.31	0.28	0	0.26
123	0.23	0	0	0	0.23	0.47	0	0.06
124	0.04	0	0.08	0.08	0.44	0.36	0	0
125	0	0	0	0	0.71	0.29	0	0
126	0.15	0	0	0.15	0.24	0.39	0	0.06
127	0	0	0	0	0.41	0.18	0	0.41
128	0	0	0.07	0	0.53	0.4	0	0
129	0	0.06	0.11	0.03	0.57	0.23	0	0
130	0	0	0.06	0	0.37	0.31	0.25	0
131	0.21	0	0	0.12	0.33	0.33	0	0
132	0	0.02	0.07	0	0.49	0.3	0.05	0.07
133	0	0	0	0	0.62	0.37	0	0
134	0.29	0	0	0.03	0.26	0.38	0	0.03
135	0	0	0.03	0.03	0.57	0.37	0	0
136	0	0.1	0.1	0	0.49	0.22	0	0.1
137	0	0	0.22	0	0.44	0.22	0.06	0.06
138	0	0.03	0.08	0	0.47	0.33	0.08	0
139	0	0	0	0	0.5	0.2	0	0.3
140	0	0	0	0	0.56	0.44	0	0

Table 37 (Cont.) Data classification by analogue.

Probabilities								
Number	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
141	0.19	0	0	0.11	0.22	0.42	0	0.06
142	0	0.03	0.08	0	0.61	0.29	0	0
143	0.04	0	0.04	0	0.36	0.14	0.36	0.07
144	0.19	0	0	0.16	0.16	0.42	0	0.06
145	0.2	0	0	0.03	0.23	0.49	0	0.06
146	0.05	0	0.09	0	0.52	0.29	0	0.05
147	0.19	0	0	0	0.3	0.46	0	0.05
148	0	0	0.2	0.06	0.49	0.23	0	0.03
149	0.04	0	0.08	0	0.4	0.44	0.04	0
150	0	0.02	0.12	0.12	0.26	0.26	0.12	0.08
151	0	0	0.03	0	0.47	0.28	0.03	0.19
152	0	0	0.06	0.12	0.53	0.25	0	0.03
153	0	0.18	0.3	0	0.3	0.15	0	0.07
154	0	0	0	0	0.54	0.35	0.11	0
155	0	0	0.17	0.04	0.46	0.25	0	0.08
156	0	0	0	0.08	0.52	0.28	0.04	0.08
157	0	0.5	0	0.06	0.19	0.19	0.06	0
158	0	0	0	0	0.45	0.27	0	0.27
159	0	0	0	0.14	0.71	0.14	0	0
160	0	0	0	0.12	0.47	0.12	0	0.29
161	0	0	0.27	0	0.4	0.27	0	0.07
141	0.19	0	0	0.11	0.22	0.42	0	0.06
142	0	0.03	0.08	0	0.61	0.29	0	0
143	0.04	0	0.04	0	0.36	0.14	0.36	0.07
144	0.19	0	0	0.16	0.16	0.42	0	0.06
145	0.2	0	0	0.03	0.23	0.49	0	0.06
146	0.05	0	0.09	0	0.52	0.29	0	0.05
147	0.19	0	0	0	0.3	0.46	0	0.05
148	0	0	0.2	0.06	0.49	0.23	0	0.03
149	0.04	0	0.08	0	0.4	0.44	0.04	0
150	0	0.02	0.12	0.12	0.26	0.26	0.12	0.08
151	0	0	0.03	0	0.47	0.28	0.03	0.19
152	0	0	0.06	0.12	0.53	0.25	0	0.03
153	0	0.18	0.3	0	0.3	0.15	0	0.07
154	0	0	0	0	0.54	0.35	0.11	0

Table 37 (Cont.) Data classification by analogue.

Probabilities								
Number	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
155	0	0	0.17	0.04	0.46	0.25	0	0.08
156	0	0	0	0.08	0.52	0.28	0.04	0.08
157	0	0.5	0	0.06	0.19	0.19	0.06	0
158	0	0	0	0	0.45	0.27	0	0.27
159	0	0	0	0.14	0.71	0.14	0	0
160	0	0	0	0.12	0.47	0.12	0	0.29
161	0	0	0.27	0	0.4	0.27	0	0.07
155	0	0	0.17	0.04	0.46	0.25	0	0.08
156	0	0	0	0.08	0.52	0.28	0.04	0.08
157	0	0.5	0	0.06	0.19	0.19	0.06	0
158	0	0	0	0	0.45	0.27	0	0.27
159	0	0	0	0.14	0.71	0.14	0	0
160	0	0	0	0.12	0.47	0.12	0	0.29
161	0	0	0.27	0	0.4	0.27	0	0.07