INFORMATION THEORETIC CLUSTERING OF ASTROBIOLOGY DOCUMENTS

A THESIS SUBMITTED TO THE GRADUATE DIVISION OF THE
UNIVERSITY OF HAWAIʻI AT MĀNOA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER SCIENCE

AUGUST 2012

By
Lisa J. Miller

Thesis Committee:

Susanne Still, Chairperson
Kim Binsted
Rich Gazan
Nancy Reed

Acknowledgments

I would like to acknowledge the considerable assistance and guidance of Susanne Still, my collaborator, principal advisor, and committee chair for this project. From helping me comprehend the information theory required to undertake such an endeavor, to providing source code for the clustering implementation, this project would not have been possible without her support. I am particularly grateful for her contribution of the concept and algorithm development for the greedy word ranking algorithm introduced here.

I thank Rich Gazan for the opportunity to work on the AIRFrame project and for his advice and assistance in social informatics, the analysis of interdisciplinarity, and astrobiology in general.

Kim Binsted, Steve Freeland, Karen Meech, and the rest of the University of Hawai‘i - NASA Astrobiology Institute Team gave me the opportunity to present parts of this work at the Computational Astrobiology Summer School, the astrobiology seminar, and the UH-NAI annual retreat. These experiences resulted in greatly appreciated, useful feedback. Nancy Reed started me out on this path through graduate school and has given me much valuable advice and support in the past and as a committee member on this project.

I would also like thank my fellow AIRFrame team-member, Mike Gowanlock, the students in the Spring 2012 Machine Learning graduate class, and Rob Shaw for their suggestions and encouragement.

I am very grateful for my family, my parents, George and Rebecca Miller, and my daughter, Veronica Gibson. Without their support and encouragement I would not have completed this project.

This work was funded by the National Aeronautics and Space Administration through the NASA Astrobiology Institute, University of Hawai‘i under Cooperative Agreement No. NNA08DA77A issued through the Office of Space Science.

Abstract

Astrobiology is a new and highly interdisciplinary field encompassing research from a diversity of disciplines including astrophysics, biology, chemistry, and geology. The AIRFrame project has been funded by NASA as part of an attempt to connect astrobiology research and researchers across disciplinary and institutional boundaries. One of the major tasks in building the AIRFrame system is to identify the major topics of research in astrobiology across disciplines and how existing work fits into these topics. While there are two astrobiology-specific scholarly journals, most researchers choose to publish in journals of their own discipline. In this work, an unsupervised learning method was applied to a corpus of astrobiology-related journal articles originating from a variety of disciplines with a goal of discovering common themes and topics.

Unsupervised learning, or clustering, discovers groupings within a dataset without the aid of labeled data samples. The Information Bottleneck method [43] was employed for this project because it has been shown to be one of the most accurate and robust methods of clustering unlabeled multi-dimensional data such as text [40, 4]. Within this same framework, it also is possible to determine the maximum number of meaningful clusters that can be resolved in a finite dataset [41]. This work was the first application of this method to document clustering. Additionally, a new related algorithm was developed for preprocessing data. This new method was evaluated on its ability to indicate which words from the document data are best for use in clustering.

These methods were combined to produce a dataset grouped by common topics present in 479 abstracts and full-text articles listed as publications in the NASA Astrobiology Institute 2009 Annual Report. The resulting clusters revealed several themes present in the data and groups of documents that are strongly connected on many levels through different numbers of clusters.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Astrobiology is a new and highly interdisciplinary field encompassing research from a diversity of disciplines, including astrophysics, biology, chemistry, and geology. While there are two astrobiology-specific scholarly journals, most researchers choose to publish in journals of their own discipline. It is a concern that progress in astrobiological research is being hindered or that research is being duplicated because of a lack of integration of resources between involved disciplines. The AIRFrame project has been funded by NASA as part of an attempt to connect astrobiology research and researchers across disciplinary and institutional borders [1].

In this work, an unsupervised learning method was applied to a corpus of astrobiology-related journal articles originating from a variety of disciplines. The resulting document clusters will be used to assist the AIRFrame team in evaluating the topics and connections present in current astrobiology research and will allow the building of a labeled dataset for further analysis on additional documents.

Unsupervised learning, or clustering, discovers groupings within a dataset without the aid of any labeled samples from the data. Four tasks commonly make up the clustering process:

1. Feature selection, or preprocessing, so the data is in a usable form for the clustering algorithm.
2. Determining the appropriate number of clusters to make.
3. The clustering itself.
4. Evaluation of the discovered clusters.

Cluster evaluation and selecting the number of clusters are often combined: Evaluation is done on a range of numbers of clusters and the amount of clusters that provide the best representation of the data for project purposes are retained.

1.1 Feature Selection

In classification tasks such as clustering, the items to be grouped are often represented as *pattern* or *feature vectors* [12, 10]. With document data, each document is commonly represented as a single vector whose dimensions or *features* are normalized word counts (all indices sum to one). A feature vector must contain an index for each dimension in the entire *feature space* (every word used to cluster the entire dataset) whether or not it appears in that particular document. The similarities or differences between feature vectors are what determine cluster membership. An example of this representation appears below in Table 1.1.

Table 1.1: Feature Vector Example Using Normalized Word Counts

Two very short documents:

1	The quick brown fox jumped over the lazy dog. The fat cat was lazy.
2	The lazy cat was lazier than the lazy dog.

Feature vector representations of normalized word counts:

	the	quick	brown	fox	jumped	over	lazy	dog	fat	cat	was	lazier	than
1	.21	.07	.07	.07	.07	.07	.14	.07	.07	.07	.07	0	0
2	.22	0	0	0	0	0	.22	.11	0	.11	.11	.11	.11

With actual documents, feature vectors can become extremely long, often with more than 10,000 word-features. Removing or combining words to reduce dimensionality of the feature space while maintaining the most "important" elements is known as *feature selection* or *feature extraction*. Important elements are those features which define subsets in the data relevant to the intended goal of a particular clustering project [12, 10, 39]. Feature selection is nearly always employed in document clustering to make computation tractable. In this project, we introduce and evaluate a new feature selection algorithm.

Feature selection was extremely important here because all available full-texts of journal articles in the astrobiology corpus were used. It is more common to use only the article abstracts, but full-text includes terms rarely included in an abstract but that are often shared in longer discussions in related work, such as: common names and acronyms; equipment and measurements used; and references cited [25]. For this project, it was desirable to take advantage of the additional opportunities the full-text might provide for discovering commonality among articles from diverse disciplines. Unfortunately, the use of full-text enormously increases the number of different words

present in the data and the dimensionality of feature vectors. Therefore, the project required a feature selection technique able to accurately identify the words carrying the most information about the contents of the documents so that only those words were retained in the feature vectors.

One focus of this project was to analyze the effects of two commonly used feature selection techniques on clustering performance:

1. *Term frequency * inverse document frequency* (TF*IDF) ranking removes words that occur in a large percentage of the documents while retaining those that appear very commonly in a few documents [17].

2. Removal of words providing little information about the documents as a whole.

Removal of words which provide little information about the documents is a feature selection method especially common in information theory-based clustering [38, 39, 9]. This process is covered in detail in Sec. 2.5.

We propose that this method can be improved upon by taking a well known bias in the calculation of mutual information into account, similar to [41, 44, 28, 27, 14, 23, 31, 7]. A new algorithm was developed that corrects for this bias, to better identify words which do not contribute information to the document data. This new method was evaluated in its ability to enhance clustering performance by reducing retention of uninformative words and preventing the elimination of words most important for clustering on both synthetic and document datasets.

1.2 Clustering of Documents

Unsupervised learning is a harder task than supervised learning where data is organized into classes learned from a labeled set of training data [10]. This project required the use of an unsupervised method because there was no existing labeled dataset for astrobiology documents. Employing domain experts to hand-label documents was not a feasible option as there is a lack of willing, qualified participants whose knowledge encompasses the entire domain of astrobiology.

The Information Bottleneck (IB) method [43] was used for the clustering task. Despite being more than ten years old, it remains one of the most accurate and robust methods for clustering unlabeled multi-dimensional data such as text [40, 4]. For analysis purposes, the performance of our implementation of the IB method was compared with previous work on accepted benchmark datasets using a commonly accepted evaluation method. The benchmark datasets are labeled with

class labels. Clustering algorithms do not have access to these labels but comparison of clustering outcomes with the labeling gives a way to evaluate results.

Results were compared with a well-known supervised classification algorithm trained on the benchmark dataset labels. Supervised learning is generally assumed to be more accurate because the classes are learned from a labeled training set. These comparisons not only demonstrated how our implementation compared with previous work, they also indicated the quality of the clustering results on the astrobiology data by showing that the clusters discovered by this method resemble classes identified by human experts.

1.3 Number of Clusters

One of the major hurdles in clustering unlabeled data is to determine the number of clusters to make. This is often considered a separate task and not always integrated into a clustering algorithm. There have been numerous approaches developed to determine the number of clusters in a dataset, many are heuristic and may not be related to the clustering method employed [10, 12, 6, 24].

It has been shown that it is possible to determine, as a function of dataset size, the maximum number of meaningful clusters that can be resolved in a finite dataset within the same framework of the Information Bottleneck clustering method [41]. This method, detailed in Sec. 2.4, looks at the growth of the bias-corrected mutual information between words and clusters as more clusters are added. If adding a cluster fails to increase this value, the maximum number of clusters resolvable has been reached.

This project is the first use of this method for document clustering.

1.4 Cluster Evaluation

Synthetic data and labeled benchmark datasets allow standard measures of classification precision to be used to evaluate the performance of clustering processes. Additionally, on controlled synthetic data, the method described above for determining the number of resolvable clusters was shown in this project to indicate clustering accuracy.

The astrobiology data clusters were evaluated for content and quality both by manual investigation of the documents and words present in the clusters, and by examination of graphical representations of the cluster memberships.

1.5 Project Goals

- Produce a dataset for the astrobiology field grouped by topics present in the data.
  - The dataset and groupings will be a valuable tool for further work in discovering relationships to astrobiology in documents originating from multiple disciplines.
- Evaluation of the performance of information theoretic methods in document clustering.
  - A new method for ranking words which determines the words which provide the most information in a dataset was developed, implemented, and tested.
  - An existing information theoretic approach to determining the number of clusters resolvable in a dataset was implemented and evaluated. This was the first use of this method in a document clustering task.

Chapter 2

Previous Work

2.1 Clustering of Documents

There are a wide array of different types of clustering methods based on diverse principles [18, 30, 12, 10]. Despite an enormous body of work produced since the 1955 publication of the k-means algorithm, a 2010 overview of the unsupervised learning field states, "there is no best clustering algorithm"[15]. In 2003, Kleinberg's *Impossibility Theorem* showed that there is no clustering function that can satisfy three desirable properties: scale invariance, richness (all partitions of the data are achievable), and consistency [18]. Clustering of data remains a hard problem where algorithm design is difficult and the problem is poorly posed. Issues such as selecting appropriate criteria for clustering, determining the number of clusters to use, and stability to initial conditions are all difficult if not impossible to address with many methods.

The Information Bottleneck [43] (detailed in Sec. 2.3) yields optimal quality clusters while eliminating the need to make assumptions about the data prior to clustering. There is no need to measure of difference (or distance) between data points [42]. One only needs to define the part of the data relevant to the desired clustering outcome [43]. Combined with Deterministic Annealing [32] and the method to determine the number of clusters resolvable in data [41] described in Sec. 2.4, this information theoretic approach addresses nearly all of the issues arising with unsupervised learning, giving a rigorous and consistent theoretical treatment to the problem and a robust implementation [41, 42, 40, 4].

For this project, the Information Bottleneck was employed particularly because the intent was to discover groupings of related research within a body of multi-disciplinary literature without making any ad-hoc judgments about the data.

2.2 Information Theory

The field of information theory has its roots in signal processing. It was founded in 1948 by Claude Shannon with his landmark paper, *A Mathematical Theory of Communication,* in which he developed a quantitative model for communication over noisy channels and a mathematical definition of information. Shannon's information measure is separated from any type of semantic meaning as this was "irrelevant to the engineering problem" [35] he was trying to solve.

Using the notation from [8], let X be a discrete random variable with alphabet $\mathcal{X}$ and the probability distribution $p(x) = Pr(X = x), x \in \mathcal{X}$. The information contained in X is the *entropy* of X, or the uncertainty one has about the value X takes measured in bits:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{2.2.1}$$

Let Y be a second similar discrete random variable and $H(X|Y)$ be the conditional entropy of X given Y, or the uncertainty about X when Y is known:

$$H(X|Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x|y) \tag{2.2.2}$$

The mutual information between X and Y then is defined as the reduction in uncertainty about X when Y is known:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)\, p(y)} \end{aligned} \tag{2.2.3}$$

Rate Distortion Theory

Let D be the expected distortion:

$$D = \sum_{(x,y)} p(x)\, p(y|x)\, d(x, y) = \langle d(x, y) \rangle_{p(x,y)} \tag{2.2.4}$$

where $d(x, y)$, the distortion measure, is an externally defined cost of representing the symbol x by the symbol y. Let R be the transmission rate: the rate that X can be compressed to (in bits per data sample transmitted) and successfully reconstructed if one knows Y at distortion, D. Shannon's Rate Distortion Theory:

$$R(D) = \min_{p(y|x):\langle d(x,y) \rangle \leq D} I(X;Y) \tag{2.2.5}$$

uses mutual information to give theoretical limits to R.

The rate distortion function can be computed as a constrained variational problem by introducing a Lagrange multiplier, β on the expected distortion:

$$\min_{p(y|x)} \left(I\left(X;Y\right) + \beta \left\langle d\left(x,y\right)\right\rangle_{p(x,y)} \right) \tag{2.2.6}$$

With the formal solution:

$$p(y|x) = \frac{p\left(y\right)}{Z\left(x,\beta\right)} \exp\left(-\beta d\left(x,y\right)\right) \tag{2.2.7}$$

where $Z\left(x,\beta\right)$ is the normalization or partition function:

$$Z\left(x,\beta\right) = \sum_{y\in\mathcal{Y}} p\left(y\right) \exp\left(-\beta d\left(x,y\right)\right). \tag{2.2.8}$$

2.2.1 Finite Sampling Bias

The calculation of mutual information from finite data suffers from a much-studied systematic error known as the *finite sampling bias* which causes the quantity of information to be inflated due to statistical errors in the data [41, 44, 28, 27, 14, 23, 31, 7]. This bias arises from the fact that information is defined in terms of the *true* underlying continuous probability distributions. With real-world data, the true values of probabilities such as $p\left(x,y\right)$, $p\left(x\right)$, and $p\left(y\right)$ are generally not known. Empirical estimates must be calculated from the data, such as:

$$\hat{p}\left(x,y\right) = \frac{N_{(x,y)}}{N} \tag{2.2.9}$$

where $N_{(x,y)}$ is the frequency of x and y occurring together and N is the total number of *events*. Use of these estimates in place of true distributions results in a maximum likelihood estimate of the mutual information [23].

Taylor expansion of maximum likelihood estimates of entropies shows that the maximum likelihood estimate of the mutual information overestimates the true value, this is the finite sampling bias. A well-known approximation of the bias is the Miller-Madow approximation [23]:

$$\frac{\left(N_X - 1\right)\left(N_Y - 1\right)}{2\log(2)N} \tag{2.2.10}$$

where N_X is the number of unique x's and N_Y is the number of unique y's actually seen in the data. Treves and Panzeri gave a much more detailed treatment to the calculation of the bias term using the replica trick arriving at the same equation as a suitable approximation [44].

This calculation of the bias becomes inaccurate for data that is *undersampled* (where the size of the data N is small compared to the actual size of $\mathcal{X}$ and $\mathcal{Y}$) because the frequency of

occurrence for many possible events will be zero or very close zero. In his 1955 paper, Miller gives a heuristic limit that N must be greater than $5N_X N_Y$ or the correction becomes inaccurate [23]. Carleton stated in 1969 that the number of observations of each event with probability greater than zero must be much greater than one for this approximation to work [7]. Treves and Panzeri suggested in 1995 that it will only be "well behaved" when the approximation, Eq.(2.2.10), is much smaller than one [44].

Unfortunately many, if not most, real datasets fall into the regime where this simple correction may not be accurate. Much work has been done to develop a bias correction that is applicable in more situations starting as far back as 1955 and continuing to the present day, but this method remains the most usable, [31, 7, 28, 27, 14].

2.2.2 Notation

In order to make the application to the document clustering problem more clear, some changes will now be made to the variable notation:

- Let C be a discrete random variable representing clusters of documents with alphabet $\mathcal{C}$ and the probability distribution $p(c) = Pr(C = c), c \in \mathcal{C}$.
- Let D be a discrete random variable representing documents with alphabet $\mathcal{D}$ and the probability distribution $p(d) = Pr(D = d), d \in \mathcal{D}$.
- Let W be a discrete random variable representing words with alphabet $\mathcal{W}$ and the probability distribution $p(w) = Pr(W = w), w \in \mathcal{W}$
- Let N_C be the number of clusters.
- Let N_D be the number of documents.
- Let N_W be the number of unique words in the documents.
- Let N be the total data size, the total number of word occurrences in all documents.

2.3 Information Bottleneck Method

The Information Bottleneck method, [43], based on Shannon's Rate Distortion Theory, replaces the need for an externally defined distortion measure with another mutual information term. One defines a new variable, the *relevant* data or the part of the data it is desirable to retain

information about. Instead of minimizing distortion, the Information Bottleneck maximizes the information kept about the relevant variable, while minimizing the overall mutual information.

In the document clustering implementation for this project, the relevant variable is W, the words contained in the documents. The mutual information retained about the documents, D, in the clusters, C, is minimized, while the mutual information about the words in the clusters is maximally retained. This is lossy compression of documents into clusters in such a way that the information kept about the words in those documents is maximized [41]:

$$\min_{p(c|d)} \left(I[C;D] - \beta I[C;W] \right) \tag{2.3.1}$$

β controls a trade-off between information preservation and data compression.

Like the Rate Distortion Theory, there is an exact formal solution to the optimization problem, Eq. (2.3.1) [43], with optimal assignment of documents to clusters being:

$$p(c|d) = \frac{p\left(c\right)}{Z\left(d,\beta\right)} \exp\left(-\beta D_{KL}[p\left(w|d\right)||p\left(w|c\right)]\right) \tag{2.3.2}$$

where $Z\left(d,\beta\right)$ is the normalization or partition function:

$$Z\left(d,\beta\right) = \sum_{c\in\mathcal{C}} p\left(c\right) \exp\left(-\beta D_{KL}[p\left(w|d\right)||p\left(w|c\right)]\right) \tag{2.3.3}$$

and

$$D_{KL}[p\left(w|d\right)||p\left(w|c\right)] = \sum_{w\in\mathcal{W}} p\left(w|d\right) \log \frac{p\left(w|d\right)}{p\left(w|c\right)} \tag{2.3.4}$$

is the relative entropy, or Kullback-Leibler divergence [19].

Many different variations have been made on the Information Bottleneck (IB) method, including Agglomerative IB [37], Sequential IB [36], Geometric Clustering [42], Multivariate IB [11], Double-Clustering [38], Co-clustering [9], and DataLoom [4].

2.3.1 Information Bottleneck with Annealing Algorithm

The algorithm used in this project to implement the Information Bottleneck method, Algorithm 1, is a version of the Blahut-Arimoto algorithm [5], an alternating iterative process. It utilizes the Deterministic Annealing (DA) approach developed to improve robustness to initial conditions and avoid local minima [32]. When using DA, an analogy is often made to free energy calculations in statistical mechanics and terminology from that domain is commonly used.

Execution begins with the Lagrange multiplier, β, set to a very small number, or the "temperature" set very high. (β though not a physical temperature, is often called an *inverse temperature*

Algorithm 1 Information Bottleneck with Deterministic Annealing

$\{\beta = \frac{1}{\text{“temperature”}}$ (starts very small)$\}$
while solution not deterministic and $\beta < \beta_{MAX}$ **do**
  while $D_{KL}[P_{new}(C)\,||P_{prev}(C)] > \theta$ **do**
    for all $c \in \mathcal{C}, d \in \mathcal{D}$ **do**
      $p(c|d) = \dfrac{p(c) * e^{-\beta D_{KL}[p(w|d)\|p(w|c)]}}{Z(d,\beta)}$
    end for
    for all $c \in \mathcal{C}, d \in \mathcal{D}$ **do**
      $p(d|c) = \dfrac{p(c|d) * p(d)}{\sum_d p(c|d) * p(d)}$
    end for
    for all $c \in \mathcal{C}, w \in \mathcal{W}$ **do**
      $p(w|c) = \sum_d p(w|d) * p(d|c)$
    end for
    Calculate $P_{new}(C)$
  end while
  increase β by α
  Check if solution deterministic
end while
Return $P(C|D)$ and $P(W|C)$

due to the similarity of its position in Eq. 2.3.2 to that of the temperature in a Boltzmann distribution.) In this regime, the level of randomness is high in a physical system and correspondingly the assignments of documents to clusters, the distribution $P(C|D)$, is quite uniform.

The first step of the algorithm keeps the *cluster centers*, $P(W|C)$, fixed and calculates the cluster assignments of the documents, $P(C|D)$ using the relative entropy between the probabilities of words in documents $P(W|D)$ and $P(W|C)$. This is associated with calculating the free energy in a physical system. (*Cluster centers* refers to the fact that $P(W|C)$ is a summary or average of all documents in each cluster.)

The second step fixes the just calculated cluster assignments and minimizes Eq.(2.3.1) for the current value of β by calculating new cluster centers, $P(W|C)$. This step is analogous to minimizing the free energy so that it better characterizes the thermodynamic equilibrium of a physical system at that temperature [32].

The two steps alternate until the relative entropy between cluster probabilities from the previous iteration, $P_{prev}(C)$, and the current iteration, $P_{new}(C)$, falls below a specified threshold, θ. This can be viewed as a physical system coming to equilibrium for that temperature. β is then increased by the annealing rate, α (temperature is lowered). Iterations begin again, forcing slightly less random assignments of the documents to clusters, causing a little more compression to occur and a little less information to be retained. This is similar to cooling down a physical system, thereby forcing it into a less random configuration. [32].

Execution stops when a *deterministic* solution is reached or β reaches a set limit. A deterministic or "hard clustering" solution is where each document is assigned to a particular cluster with probability one. The limit for β, β_{max} is a large value, corresponding to a temperature close to zero. By this point, any meaningful change has ceased to occur in $P(C|D)$, movement of documents between clusters has effectively stopped, and it is judged that a deterministic solution will not be attained.

2.4 Number of Clusters

Determining the number of clusters to make from unlabeled data is considered an unsolved problem with its own research area [10, 12, 6]. Many clustering algorithms assume the number of clusters is given and will then put the data into those clusters whether or not they are a good fit for the data's structure. Often the method used for determining the numbers of clusters is heuristic or unrelated to the clustering method itself [41, 12, 10, 6, 24].

A common practice in unsupervised learning is to run a clustering algorithm over a range of numbers of clusters then use the result with the best *goodness* or *stopping rule* score. Many different goodness measures or stopping rules for cluster membership have been developed. These include several methods of comparing the between- and within-cluster distances such as the *Gamma statistic* and the *Cubic Clustering Criterion*; the *Likelihood Ratio* criterion to compare the hypothesis of k clusters to $k-1$; the sum-of-squared-error within clusters, $\frac{Je(2)}{Je(1)}$; the *max-F* test (the ratio of the sum-of-squares between and within clusters); and average similarity/dissimilarity statistics (*U-statistics*) [10, 12, 6, 24].

It has been shown that the maximum number of meaningful clusters in a finite dataset can be determined using a correction similar to the Miller-Madow approximation [23] (see Sec. 2.2.1) within the Information Bottleneck (IB) framework [41]. This is particularly attractive because it requires few additional assumptions be made about the data and does not define a separate criterion for goodness. This method employs the same criterion the IB method seeks to maximize when compressing documents into clusters, the relevant information or the mutual information between clusters, C, and words, W:

$$I[C;W] = \sum_{c\in\mathcal{C},w\in\mathcal{W}} p(c)p(w|c)\log_2\left(\frac{p(w|c)}{p(w)}\right) \tag{2.4.1}$$

Calculated empirically from data, $I\left[C;W\right]$ always increases as more clusters are added as seen by the dashed red line in Fig. 2.1. However, as discussed in Sec. 2.2.1, the calculation of mutual information from finite data suffers from a finite sampling bias. A correction to the relevant information has been derived from the leading order term of a Taylor expansion of $I[C;W]$ [41]. This correction is subtracted from $I[C;W]$ resulting in $I_{corr}[C;W]$ as seen in Eq. (2.4.2), where N_W is the number of words (dimensions of the document feature vectors), N_C is the number of clusters, and N is the total data size (the total word counts for all documents)

$$I_{corr}[C;W] = I\left[C;W\right] - \frac{(N_W-1)N_C}{2\log\left(2\right)N}. \tag{2.4.2}$$

(In order to apply this method, the clustering solution must be deterministic or the meaning of N_C in Eq. 2.4.2 is unclear.)

$I_{corr}[C;W]$ has a maximum or at least a plateau, as seen by the solid blue line in Fig. 2.1. This maximum (or plateau) is the number of clusters at which the maximal meaningful structure of the data can be resolved. After this point, the uncorrected $I\left[C;W\right]$ will continue to increase when adding more clusters (red in the figure), however, this is not capturing any more meaningful information, it is clustering noise due to finite sampling effects or *overfitting* the data.

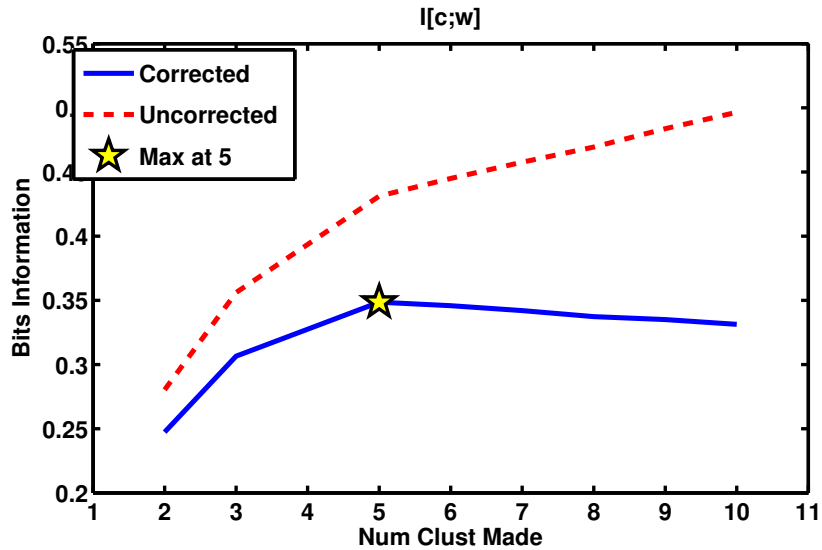


Figure 2.1: Finite Sampling Bias Correction on $I[C;W]$. $I[C;W]$, the mutual information between clusters and words is shown on synthetic data with and without the correction applied. The point at which the corrected term reaches its maximum is the maximal number of clusters resolvable in the data. In this case, five is indicated which is also the number of distributions this data was drawn from. (Note, after the correction is applied, the value is no longer a true information term.)

This method was applied to document data for the first time in this project. It will be referred to as the $I_{corr}[C;W]$ method. (Note that $I_{corr}[C;W]$ cannot be considered an information in the strictest sense.)

2.5 Feature Selection

As discussed in Sec.1.1, the data in document clustering tasks is very high-dimensional and there are usually many more unique words than documents in a corpus. Many words occur very rarely or in only one document, leading to long feature vectors mainly full of zeros. Conversely, some very common words appear in all documents, possibly providing no help in distinguishing between any of them. Feature selection methods to reduce dimensionality are employed widely in most text processing.

The SMART information retrieval project at Cornell University worked in part to evaluate and develop feature selection methods for more than 30 years [13]. The methods used in both the SMART research and information retrieval system for *automatic indexing* remain some of the most popular procedures [33]:

- *Stop word* removal removes extremely common function words such as “and”, “of”, “the”, and “but” from all feature vectors.

- *Stemming* or *suffix stripping* removes the endings of words, leaving only the *word stems* or *terms* in the data. This allows grouping of all forms of one word, such as “give”, “given”, and “giving”, into one vector index for the term, e.g. “giv”. One widely employed stemming method is the Porter stemming algorithm which does not require that a list of word stems be provided [29].

- TF*IDF or *term frequency * inverse document frequency* developed first in 1972, ranks words (or terms) such that those that occur very commonly in just a few documents rank highest, while any word appearing in all the documents ranks lowest [17].

$$TF * IDF\,(w) = p\,(w) * \log\left(\frac{N_D}{N_{d(w)}}\right) \tag{2.5.1}$$

  Where w is a word in the data, N_D is the total number of documents, and $N_{d(w)}$ is the number of documents in which w occurs.

A feature selection method particularly popular in Information Bottleneck-based clustering is the ranking of individual words, w, by contribution, $I\,[w]$, Eq. (2.5.2) to the mutual information between all words and documents, $I\,[W;D]$ [38, 39, 9]:

$$I[w] = \sum_{d\in\mathcal{D}} p(w|d)p(d)\log_2\left(\frac{p(w|d)}{p(w)}\right) \tag{2.5.2}$$

$$I[W;D] = \sum_{w\in\mathcal{W}} I[w]. \tag{2.5.3}$$

With this method, the number of words to keep is not indicated, a cutoff threshold must be decided upon. We developed the new method introduced in this work, Sec. 3.2.2, in part to address this issue.

This method will be referred to as the $I[w]$ method in later sections. (Note that this value is not a true mutual information term.)

2.6 Benchmark Document Datasets

The performance of our implementation of the Information Bottleneck was analyzed on two document classification benchmark datasets in order to examine its ability to reconstruct human-recognized labellings and to compare with previous work. A commonly used benchmark document

set containing article-length documents similar to the astrobiology data was searched for. Unfortunately, longer documents contain a greater variety of words increasing the dimensionality of feature vectors and exacerbating performance problems in classification algorithms. Researchers in this area usually do not choose to use that kind of data in published experiments. This results in benchmark datasets with much shorter documents than the astrobiology corpus and no labeled datasets that are both comparable and commonly cited.

The sets used here are considered the two standard document classification benchmarks [34]:

1. MOD-APTE split [2] of the Reuters 21578 document collection [21]. A set of 9133 newswire stories from 1987 hand labeled by Reuters and Carnegie Group personnel. Many of these documents are extremely short, less than a paragraph in length, and the topics are very interconnected all coming from economic subjects. Many documents have multiple labels or no label and there are several labels with very small membership. The singly-labeled documents from the 10 largest labels were used here. This gives 8008 documents, the vast majority of which are in only two labels, *earn*, "earnings" and, *acq*, "acquisitions"

2. 20 Newsgroups. A set of 18,828 posts to 20 online newsgroups collected in 1995 [20]. The 20 groups fall into 6 high-level categories: *computers, for sale, sports and automobile recreation, politics, religion,* and *science*. Different combinations of these categories are also common clustering targets. These documents are somewhat longer in general than the Reuters set, averaging about 260 words each. The set used in this project had all duplicate posts and headers removed except for "From" and "Subject." The forwarded text within posts was retained.

Chapter 3

Methodology and Experiments

The project experiments were divided into three components based on the different types of data used.

1. Synthetic data - testing the new feature selection algorithm.
   (a) Constructed synthetic data.
   (b) Feature selection performed using four methods: TF*IDF, $I[w]$, our new method, and random word lists.
   (c) Clustered that data using the Information Bottleneck method, Sec 2.3.
   (d) Compared the clustering outcomes by evaluating if the true number of clusters could be resolved using the $I_{corr}[C;W]$ method, Sec. 2.4.
2. Benchmark data - comparing with previous work.
   (a) Tested clustering implementation by comparing results with [36], using the same pre-processing as they did.
   (b) Examined how many clusters were resolvable using the $I_{corr}[C;W]$ method.
   (c) Applied the new feature selection algorithm.
3. Astrobiology data - building labeled dataset.
   (a) Performed feature selection using three different methods (TF*IDF, $I[w]$, and our new method) and compared resulting word lists.
   (b) Used the new feature selection method to determine how many words should be kept.

(c) Ran the clustering algorithm on a wide range of cluster numbers and examined how many clusters were resolvable using the $I_{corr}[C;W]$ method.

(d) Evaluated cluster memberships of selected results.

3.1 Datasets

3.1.1 Synthetic Data

Synthetic datasets were constructed in order to evaluate the performance of feature selection techniques using clearly understood, controlled data. This data was not intended to replicate document data but rather was inspired by [41] where synthetic data was used to evaluate the $I_{corr}[C;W]$ algorithm. The sizes of the datasets were kept much smaller than document sets in order to run many experiments quickly with a wide variety of situations.

Each dataset was made from several discrete Gaussian distributions. Each distribution represented a cluster made up of vectors representing documents. Each vector contained "words" represented by integers drawn from that distribution. Different data configurations were made, ranging in clustering difficulty by varying the length of the document vectors and γ, the distance between the distribution/cluster means. Six different realizations of each data configuration were made.

An example of one version of this synthetic data is shown in Fig. 3.1

3.1.2 Astrobiology Dataset

In order to gather a representative sampling of current research topics in astrobiology, a set of astrobiology-related documents were collected from the NASA Astrobiology Institute (NAI) 2009 Annual Report [26]. The NAI is the primary funding agency for major astrobiology grants. Currently there are 14 teams spread across 150 institutions with more than 700 involved researchers. The grants run for five years with each team required to submit a report on all funded projects annually. Each project report is labeled with the project's name, its primary researchers and the NAI Astrobiology Roadmap [22] Goals and Objectives (subgoals) judged relevant by the team. Additionally, all associated publications are listed.

According to research done for the Textpresso semantic search engine [25], abstracts of scientific journal articles are not enough to discover important connections among documents. Abstracts traditionally do not include important details that related research shares. By examining the

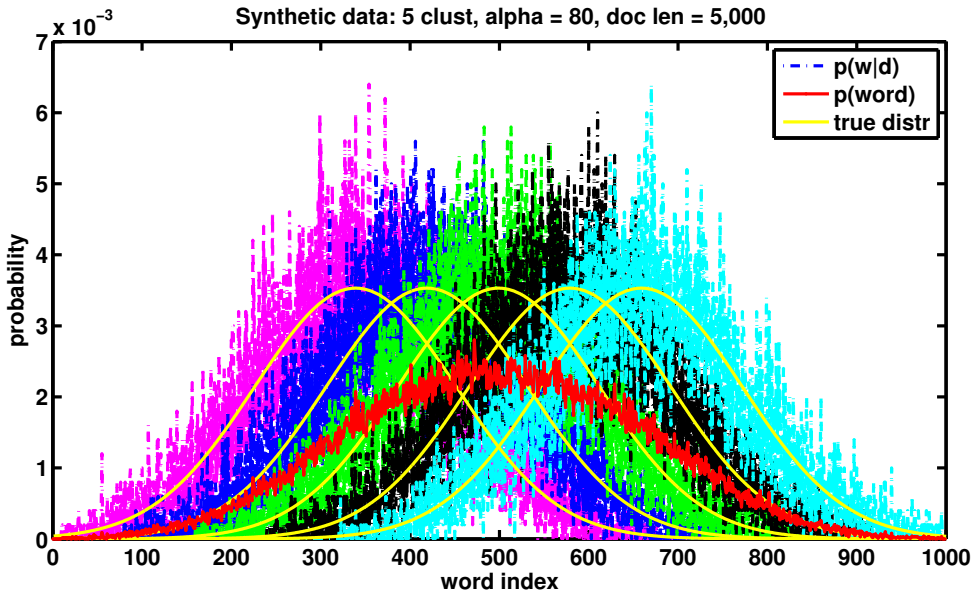


Figure 3.1: Five Cluster Synthetic Gaussian Data. Dashed lines are probability of word given document vectors, $p(w|d)$. Each color is a particular cluster/distribution. Each document is drawn from one of the true distributions shown in solid yellow. The solid red line is the probability of words over the entire dataset, $p(w)$.

full text of articles, the discovery of shared terms, datasets, research methods, and references can be enhanced. From the NAI Annual Report, 479 articles and reports were retrieved, most of these are full-text unless it was unavailable, then the abstract was used.

Ideally, the Goal and Objective labels from the project reports should have yielded the type of topic groupings required for the AIRFrame project, eliminating the need for clustering. However, this was not the case, because:

1. Labels are applied by the specific project participants, leading to labeling inconsistencies due to subjective and discipline-related judgments about scope and definition of the Goals and Objectives.

2. Publications are labeled with all Goals and Objectives associated with an entire project. Individual papers probably only fall under a few of those, but there is no simple way to judge this from the information available.

3. Some Roadmap labels are very general, apparently tempting researchers to label almost everything with them. Almost half of the retrieved documents are labeled with Objective 1.1,

Formation and evolution of habitable planets, and nearly 200 are labeled with 4.1, *Earth's early biosphere*.

4. The Goals and Objectives don't completely identify topics in astrobiology due to rapid changes in the field.

3.2 Clustering

3.2.1 Feature Selection

Four feature selection methods were compared including the new method introduced below. Those examined were:

- *Term frequency * inverse document frequency*, TF*IDF, Eq. (2.5.1) [17].
- Per-word contribution to the mutual information between words and documents, $I[w]$, Eq. (2.5.2) [36].
- Our new information theoretic word ranker.
- Words chosen randomly.

For all document data, the words were stemmed using the Porter Stemming Algorithm, [29] and a standard list of stop words were removed. For the TF*IDF method [17], the outcome was used to rank the words, it was not used to weight the words in the feature vectors as is the case in some work. The feature vectors used for clustering were always the normalized word counts calculated after stemming, stop word removal, and word list cutting.

3.2.2 Greedy Information Theoretic Word Ranker

Previous uses of word ranking by $I[w]$, Eq. (2.5.2), for feature selection do not appear to have taken the finite sampling bias (Sec. 2.2.1) into account when calculating the value. It has been shown that in order for this bias to be of negligible size, the the average document length must be larger than the number of words by a factor of 32 [27]. (If $\frac{N}{N_D} = \bar{W}_D$ is the average document length, then $\frac{\bar{W}_D}{N_W} > 32$.) The document datasets examined here are not at all close to that, the average length of documents is much less than the total number of words $\left(\frac{\bar{W}_D}{N_W} < 1\right)$. This indicates that the calculation of $I[w]$ could be affected by the bias, causing $I[W;D]$ to continue to grow as more and more words are kept even though some might actually be uninformative noise

and could be removed from the feature vectors (similar to the situation when adding more clusters to $I\left[C;W\right]$, as discussed in Sec. 2.4.)

We developed a new method for this project to rank words according to contribution to $I\left[W;D\right]$ that takes this bias into account (complete pseudo-code is given in Algorithm 2.) The new algorithm is based particularly on the method for determining the number of resolvable clusters also employed in the project, Eq. (2.4.2) [41] and on similar methods successfully applied in correcting neural response data from animal experiments [44, 28, 27]. A suboptimal greedy algorithm was necessary for the new method due to the combinatorial nature of the problem: choosing the combination of words which gives the most mutual information between words and documents as more words are added.

The algorithm greedily chooses an ordered list of words ranked by their contribution to the corrected mutual information between words and the documents:

$$I_{corr}[W;D] = I\left[W;D\right] - \frac{(N_D - 1) * N_W}{2\log\left(2\right)N} \tag{3.2.1}$$

Execution of the algorithm proceeds as follows:

1. All the words in the feature vectors start in an unordered list, $\mathbf{W_u}$, an empty ordered list is created for ranked words, $\mathbf{W_r}$.
2. The *uncorrected* per-word contribution to the mutual information between words and documents, $I[w]$ from Eq. (2.5.2), is calculated for each word and stored.
3. At each iteration, each as-yet unranked word is temporarily added to $\mathbf{W_r}$ one at a time.
4. Using the words in $\mathbf{W_r}$ and only the documents containing those words, $I_{corr}[W,D]$, Eq. (3.2.1) is calculated.
5. The unranked word giving the highest $I_{corr}[W,D]$ for the current iteration is chosen (greedily), inserted into the list of ranked words $\mathbf{W_r}$, and removed from $\mathbf{W_u}$. $I_{corr}[W,D]$ is recorded for this step.
6. Iterations continue until all words have been ranked.

Once the algorithm has finished executing, the cumulative $I_{corr}\left[W;D\right]$ is evaluated. All words that give a positive increase in this value are judged to be contributing information and kept for clustering. A plot showing $I_{corr}\left[W;D\right]$ versus the cumulative uncorrected $I\left[w\right]$ and the positioning of the words chosen can be seen in Fig. 3.2. Note in the figure, the cumulative uncorrected $I\left[w\right]$ continues to increase over all of the words. This value gives no indication of how many words to keep, leaving no choice but to pick an arbitrary number of words or percentage of information

Algorithm 2 Greedy Corrected $I[W;D]$

{Let $\mathbf{W_u}$ be a list of unranked words, all words to start}

{Let $\mathbf{W_r}$ be an ordered list of ranked words, empty to start.}

{Let N_{Wr} be the number of words in $\mathbf{W_r}$, 0 to start.}

{Let N be the total number of occurrences of words in $\mathbf{W_r}$, 0 to start.}

{Let $\mathbf{D_r}$ be a list of documents containing any words in $\mathbf{W_r}$, empty to start.}

{Let N_{Dr} be the number of documents in $\mathbf{D_r}$}

{Let N_w be number of occurrences of word w in the data.}

{Let N_{Dw} be number of documents word w occurs in, **not** already in D_r.}

{Let I_{sum} be a running sum of uncorrected per-word MI terms of words in $\mathbf{W_r}$, 0 to start.}

for all $w \in \mathbf{W_u}$ **do**

  {Calculate uncorrected per-word MI contribution}

  $I[w] = \sum_{d \in D} p(w|d)p(d)\log_2\left(\frac{p(w|d)}{p(w)}\right)$

end for

while $\mathbf{W_u}$ **not empty do**

  $I_{\text{max}} = -\infty$

  for all $w \in \mathbf{W_u}$ **do**

    Correction $= \frac{(N_{Dr}+N_{Dw}-1)*N_{Wr}}{2\log(2)(N_r+N_w)}$

    $I_{\text{temp}} = I_{\text{sum}} + I[w] -$ Correction

    if $I_{\text{temp}} > I_{\text{max}}$ **then**

      $I_{\text{max}} = I_{\text{temp}}$

      $w_{\text{max}} = w$

    end if

  end for

  add w_{max} to $\mathbf{W_r}$ and remove it from $\mathbf{W_u}$

  add $D_{w_{\text{max}}}$ to D_r

  add $N_{w_{\text{max}}}$ to N

  add $I[w_{\text{max}}]$ to I_{sum}

end while

return $\mathbf{W_r}$ and I_{sum}

as a cutoff point. The new method clearly indicates the maximum number of words that should be retained.

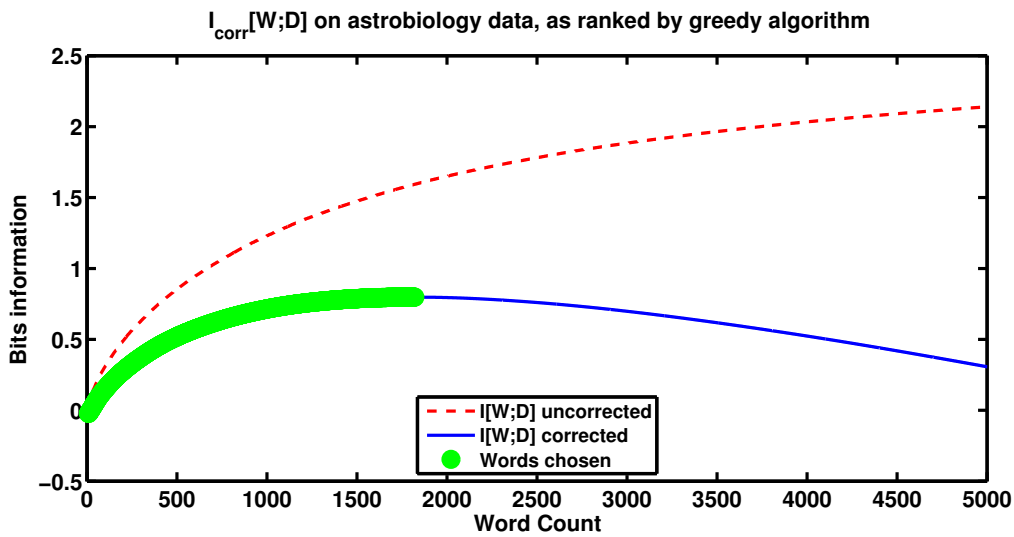


Figure 3.2: Corrected vs. Uncorrected $I[W;D]$ on Astrobiology Data. Words chosen due to their positive contribution to $I_{corr}[W;D]$ shown in green.

3.2.3 Clustering of Documents

As described in Secs. 1.2 and 2.3.1, the Information Bottleneck method [43] with Deterministic Annealing [32] was used for the clustering task. The implementation used takes in a matrix of document feature vectors containing normalized word counts, $\hat{p}(w|d)$. The initial cluster centers, $P_{init}(W|C)$, and initial document-cluster assignments, $P_{init}(C|D)$, are made by choosing a different random document to represent each initial cluster. Execution of the clustering algorithm proceeds as described in Sec. 2.3.1. When execution finishes, the cluster centers, $P(W|C)$, and the cluster assignments of the documents, $P(C|D)$, are recorded into files.

As seen in Sec.2.4, the clustering solution needs to be deterministic in order to apply the $I_{corr}[C;W]$ method for finding the maximum number of resolvable clusters. Because the document data used was not at all well separated, some solutions contained clusters sharing many words and some documents (especially short ones) were completely represented by more than one cluster. The value of β_{max} was set quite high to force a deterministic solution in more instances at the expense

of execution speed, however, a completely deterministic solution was impossible to reach in some cases.

3.2.4 Number of Clusters

The $I_{corr}[C;W]$ method (Sec. 2.4) was employed to determine the maximum number of resolvable clusters in the data by running the IB clustering algorithm over a range of numbers of clusters. In this process, $I_{corr}[C;W]$ is evaluated for each number of clusters in increasing order where the solution has deterministic cluster assignments. A point is looked for where the value of $I_{corr}[C;W]$ ceases to increase beyond a rounding error. The last number of clusters for which $I_{corr}[C;W]$ increases is the maximum clusters resolvable in the dataset given the size of the data.

The desired outcome for tests on synthetic data where the true distributions are known, is when the $I_{corr}[C;W]$ plot reaches its maximum at the actual number of distributions the data was drawn from. If this method indicates less than the true number of clusters, not enough data samples were used to resolve them all. This is not necessarily an unsatisfactory result, this outcome simply indicates that this is the most clusters resolvable using that number of data points.

An undesirable result would be if this method indicated more than the true number of clusters. In this case, something has happened to the distributions that has fractured them into clearly separate, multiple pieces such that artifact distributions have been introduced. Either the documents are too short to properly represent the distributions to begin with, or feature selection has removed critical elements from the word list.

3.2.5 Cluster Evaluation

Clustering accuracy on the labeled datasets was evaluated using the *micro-averaged precision*:

$$P(L,C) = \frac{\sum_{c \in C} \mu_l(c)}{\sum_{c \in C} \text{size}(c)} \tag{3.2.2}$$

where C is the clusters and $l \in L$ is a data label in the set of all labels. $\mu_l(c)$ is the maximum number of documents with the same label l in cluster c. Precision is the portion of data points with the majority label in a cluster. Micro-averaged precision is the sum of this over all clusters. This quantity is employed widely as a measure of classifier accuracy in the supervised learning field and when labeled datasets are used in clustering tests.

With the synthetic data, the $I_{corr}[C;W]$ method (Sec. 2.5) was used to evaluate the clustering outcomes. The goal of clustering this data was to evaluate the effects of word lists shortened

using different feature selection methods. Evaluation on both precision and the ability to resolve the actual number of clusters was important because perfect precision is still possible if clusters are split (e.g. precision is perfect if each cluster contains only one document.) With this type of data, and deterministic cluster assignments, if the actual number of distributions is indicated as the maximum number of clusters resolvable over several clustering runs, the accuracy of the clustering is also perfect.

For the unlabeled astrobiology data, the only real option was to evaluate the cluster quality by hand. The word distributions of the clusters (cluster centers), the documents assigned to those clusters, and the projects the documents originated from were all examined. The distributions of the originating NAI teams and NAI Goals in the clusters were also evaluated. Additionally, a graphical network evaluation tool was used to explore the continuity of sub-clusters of documents across several solutions with different numbers of clusters. This process is discussed further in the results in Sec. 4.3.3.

3.3 Experiments

3.3.1 Synthetic Data

The feasibility of clustering different synthetic datasets was evaluated first. The clustering algorithm was run with all words over a range of number of clusters on six realizations of each data configuration. $I_{corr}\left[C;W\right]$ method for determining the number of clusters resolvable was applied by examining the slope of the $I\left[C;W\right]$ curve as described above in Sec. 3.2.4. A plot of the outcomes from this step for five cluster data with 2,000 length documents can be seen in Fig. 3.3. If this procedure indicated less than the true number of distributions in the data for even one realization that dataset was judged to be too difficult to have satisfactory results with any feature selection so it was discarded.

To further explore the feasibility of clustering different data versions, word lists were created of progressively shorter lengths with words ranked by the uncorrected $I\left[w\right]$, Eq.(2.5.2). The clustering algorithm was then run with each word list. The purpose of this experiment was to discover datasets that were trivial for the IB method to cluster. The datasets where the $I_{corr}\left[C;W\right]$ method resolved the actual distributions with 10% of the words for all six data realizations were determined to be trivial and discarded.

The remaining datasets were used to explore the performance of the new greedy feature selection algorithm and the effects of the TF*IDF and $I[w]$ feature selection methods on the con-

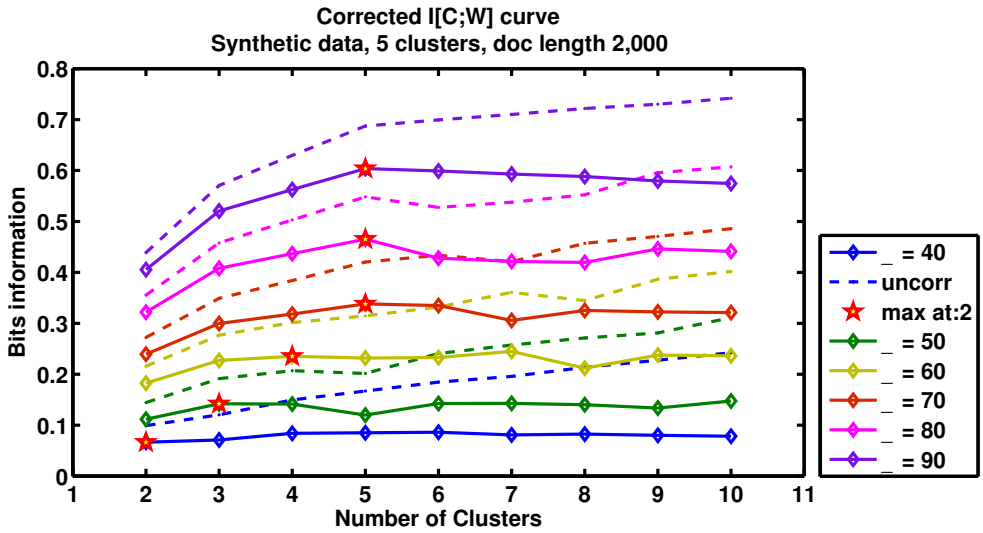


Figure 3.3: Corrected vs. Uncorrected $I[C;W]$ Curves on Synthetic Five Cluster Data. Each line color represents a different γ, the spacing between distribution means. Dashed lines are the uncorrected $I\,[C;W]$, solids are the corrected value used to determine the maximum resolvable clusters. Stars indicate the number of clusters that can be resolved. Note that as the spacing between the original distribution means decreases, the ability to resolve all of the distributions also decreases.

tinued ability of the clustering algorithm and the $I_{corr}\,[C;W]$ method to resolve the true number of distributions. Random word lists of various lengths were also created and the number of resolvable clusters resulting from those were evaluated.

3.3.2 Benchmark Datasets

In order to compare results with previous work, the two benchmark datasets introduced in Sec 2.6, Reuters and 20 Newsgroups, were preprocessed and feature selection done using the methods employed by [36]:

- Documents with less than 10 characters were removed.
- Documents with more than 1 label were removed.
- For the Reuters set, only documents with the 10 largest labels were kept.
- Stop words were removed.
- For 20 Newsgroups, any text with a symbol or number in it was removed.
- Words were stemmed using the Porter Stemming Algorithm [29].
- Words were ranked using $I\,[w]$, Eq.(2.5.2).

- The top 2000 words were kept for both datasets.

Six separate restarts of the clustering algorithm were made, with 10 clusters for the Reuters set and both 10 and 20 clusters for the 20 Newsgroups data ([36] used a combination of the 20 Newsgroups labels to evaluate it at 10 clusters.) Additionally, clustering runs were done with increasing numbers of clusters on both datasets to examine if the $I_{corr}[C;W]$ method would indicate the number of labels as the number of resolvable clusters or if the data were being overfit at that point (as discussed in Sec. 2.4).

More experiments on the benchmark data were intended, similar to those done the synthetic data. Comparison of the new greedy feature selection method's performance to the other methods and evaluation of the maximum number of resolvable clusters present were planned. Unfortunately, issues with both methods were encountered on these datasets and this part of the project was not possible. This is discussed in detail in Sec. 4.2.

3.3.3 Astrobiology Data

Preprocessing on the astrobiology dataset was nearly identical to that performed on the benchmark data. Stop words were removed and words stemmed, but words containing numerical digits were not removed due to the presence of many chemical formulations in the data. All three feature selection methods, TF*IDF, $I[w]$, and the new greedy ranking method were performed. Lists of 2,000 words were made with TF*IDF and $I[w]$ ranking. A second list was made with $I[w]$ with 3,431 words containing 75% of the total $I[W;D]$. The new feature selection method performed as expected on this data, indicating a list of 1,806 words to be retained. A plot of $I_{corr}[W;D]$ produced by the new method is shown in Fig. 3.2.

The IB clustering algorithm was run over a wide range of numbers of clusters to evaluate differences in cluster quality and to help determine how many clusters might be resolvable in the data.

Chapter 4

Results and Discussion

4.1 Synthetic Dataset Tests

By construction, the synthetic datasets were quite different from real text data. The experiments on this data were explorations of different feature selection methods and not necessarily indicative of their value when working with document data.

The $I_{corr}[C;W]$ method was used to analyze clustering results on the synthetic data, as discussed in Sec. 3.2.4. Table 4.1 summarizes the results of testing different feature selection methods. In the table, the first $I_{corr}[W;D]$ column lists the number of words the new greedy method indicated be kept. The second $I_{corr}[W;D]$ column lists the average number of clusters resolvable from the data using that word list. The next column lists the number of words required to successfully recover the actual number of clusters using the uncorrected $I[w]$ ranking. The last column lists the number of random words needed to recover the actual number of clusters.

These results clearly show that random word lists allowed the true number of clusters to be resolved with much fewer words than the information-theoretic methods in all cases. Because the synthetic data was constructed from overlapping Gaussian distributions, the data was easily clustered correctly if the word list included some words from from the tails of each distribution and some from near each mean. The random word lists contained such words, having been drawn randomly from across the entire range of allowed words.

The uncorrected $I[w]$ method and the new greedy $I_{corr}[W;D]$ method tended to pick words located near the center of the distributions, as seen in Figs. 4.2 and 4.3. Low count words that appeared in only one distribution were ranked last. When the word lists were shortened based on these rankings, these words were cut out causing the outermost distributions to appear to be part of their closest neighbor distribution to the clustering algorithm.

Table 4.1: Comparison of Feature Selection Methods on Synthetic Data. $\mathbf{I_{corr}}[\mathbf{W};\mathbf{D}]$ refers to the new greedy feature selection method, the *N Words* column is the number of words the method indicates should be kept, *Clust Made* is the average number of clusters that can be resolved using the word list the method creates. The $\mathbf{I}[\mathbf{w}]$ column lists the number of words the uncorrected $I[w]$ ranking needs for the true number of clusters to be resolvable. The **Random** column lists the number of randomly chosen words needed to resolve the true number clusters.

Data			$\mathbf{I_{corr}}[\mathbf{W};\mathbf{D}]$		$\mathbf{I}[\mathbf{w}]$	**Random**
N Clust	*Doc Len*	γ	*N Words*	*Clust Made*	*Words Needed*	*Words Needed*
5	2000	90	606	4.33	600	200
5	2000	80	569	4	700	200
5	5000	90	697	4.83	700	100
5	5000	80	693	4.67	700	100
5	5000	70	591	5	700	200
5	5000	60	766	4.5	800	300
5	10000	90	807	5	600	100

Additionally, the greedy configuration of the new algorithm caused it to be adversely affected by the structure of the data. When the documents were short it kept less words than when the documents were long. When the distributions were further apart (large γ) it kept more words than when they were close together. This is unsatisfactory because data with longer documents contains much less noise and more widely spaced distributions are more obvious. Less words, not more, should be needed to resolve clusters in data with either of these features.

The TF*IDF method is not included in the table due to issues employing it on the synthetic data. Because of the fraction inside the logarithm in the TF*IDF equation, (Eq. 2.5.1) if a word is in all, or nearly all documents, the TF*IDF value becomes very small or zero no matter if there is a large difference in the probability of the word for different documents. This resulted in scores of zero for more than half of the words in all of the synthetic datasets. Fig. 4.1 shows how words with non-zero TF*IDF rank were located on the tails of the overall $p(word)$ distribution. This rendered it impossible to correctly resolve distributions in the center that were overlapped by others on both sides. For the five cluster data in the figure, the cluster in the middle could not be resolved at all. None of the synthetic datasets we used were clustered correctly using TF*IDF for feature selection even using all words with non-zero scores.

One part of the process that performed very well with this data was the $I_{corr}[C;W]$ method to determine the maximum number of clusters resolvable. This method allowed evaluation of the performance of different feature selection methods and word lists in an intuitive way by evaluation of figures like Fig. 3.3.

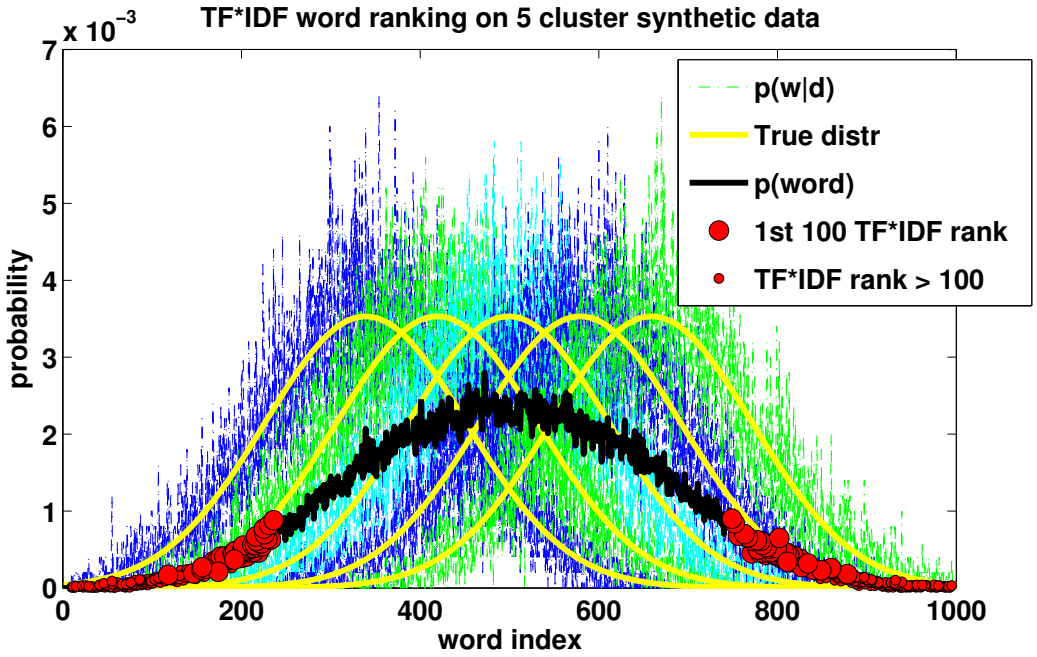


Figure 4.1: Words Chosen using TF*IDF Ranking on Synthetic Data. Five cluster dataset plotted on the data distributions. The first 100 words chosen are larger circles.

4.2 Benchmark Data Tests

The Reuters and 20 Newsgroups datasets were chosen for this project because of the availability of results from previous work. The IB clusterer was run on both sets using word lists and preprocessing common in IB-based clustering [36, 37, 9]. Results were evaluated using the micro-averaged precision, Eq. 3.2.2, on the 10 labels of the Reuters set, all labels of the 20 Newsgroups, and a combination of some of the 20 Newsgroups into 10 higher-level topics, these results are presented in Table 4.2.

The micro-averaged precision was used as a performance measure in [36] however, there the best result of 15 runs was reported. Here only six runs were completed for each test, with the average and best of six scores recorded. Therefore, scores were not expected to be as high as the published results. Additionally, the Reuters scores reported in [36] appear to include multi-labeled documents which gave a precision advantage over these results because only documents with single labels were used here.

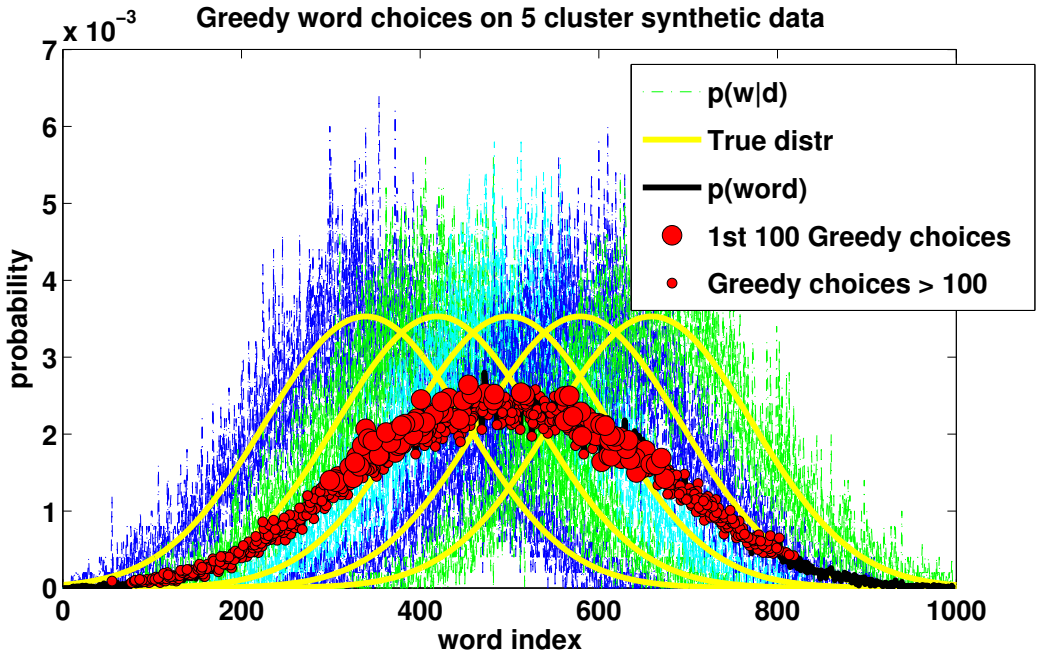


Figure 4.2: Words Chosen by the New Greedy Feature Selection Method on Synthetic Data. Five cluster dataset plotted on the data distributions. The first 100 words chosen are larger circles.

4.2.1 New Feature Selection Method on Benchmark Data

On large datasets with short documents like Reuters and 20 Newsgroups, the sheer number of documents overwhelmed the greedy feature selection algorithm in the early part of execution. Because of the greedy way the algorithm picks words to keep, a huge number of documents were added as N_D to the numerator of the correction term in Eq. 3.2.1 by the first, or first few words, while the total data size in the denominator, N, did not grow fast enough due to the low number of word instances per (short) document. This led to an undersampled regime as discussed in Sec. 2.2.1 where the bias correction equation does not work well. A very large correction term forced $I_{corr}[W;D]$ to have negative values from the very beginning. For Reuters, the set with the shortest documents, this method retained only a single word. For 20 Newsgroups with slightly longer documents, 49 words were kept.

While this behavior rendered the new algorithm unusable for feature selection on those datasets, it did provide an indication that there is a very large amount of noise due to undersampling in that data, so much so, that clustering with a high degree of accuracy may be impossible. This is further reflected in Fig 4.5c showing the word distributions of selected documents from the Reuters set. The documents contain so few words that each one is represented by just a tiny percentage of

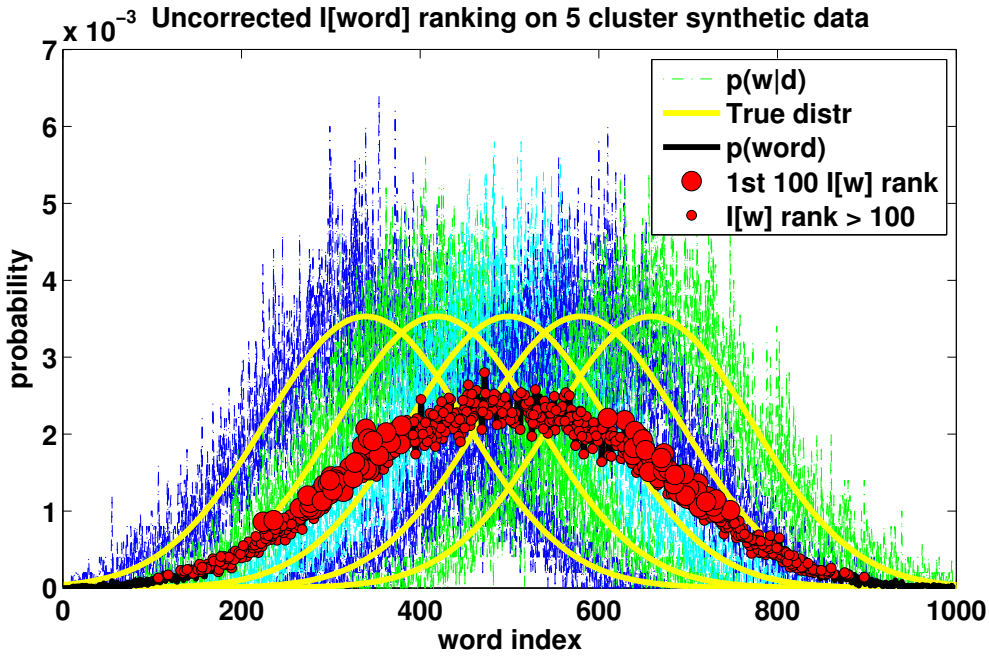


Figure 4.3: Words Chosen using $I[w]$ Ranking on Synthetic Data. Five cluster dataset plotted on the data distributions. The first 100 words chosen are larger circles

the total words in the data. Note that this is a log-log scale plot, words with zero probability are not shown.

4.2.2 Number of Clusters in Benchmark Data

The $I_{corr}[C;W]$ method indicated far more clusters resolvable in these large datasets than labels, and many more than would probably be a useful clustering outcome. Due to the size of the benchmark sets, only a few clustering runs were performed with more or less clusters than the number of labels in the data. In order for the calculation of the bias correction term used for this method to make sense, the clustering assignments must be deterministic. With both benchmark datasets completely deterministic assignments were not achieved due to the presence of many short documents, as discussed in Sec. 3.2.3. Only results that were nearly deterministic were examined where a very small percentage of the documents have multiple assignments. In all of these, the $I_{corr}[C;W]$ values continued to increase up to 30 clusters, the maximum attempted. This indicates that overfitting does not occur when clustering these sets at, or near the number of labels as is common practice.

Table 4.2: Micro-Averaged Precision Scores on Benchmark Data. Maximum possible value is 100. Top 2,000 words ranked by $I[w]$ used unless noted. Our results use the Information Bottleneck clustering method; six runs performed. The first column is the average score, and the second column, the best score. Results are compared to previously published scores: column 3 displays the best of 15 runs using the Sequential Information Bottleneck (sIB)[36]. Column 4: the parallel co-clustering algorithm, DataLoom [4] using all the words (no preprocessing). Column 5: supervised Naive Bayes classifier as reported in [36]. The value in parentheses in Naive Bayes/Reuters slot is the micro-averaged precision for Naive Bayes using 9,962 words as reported in a widely cited paper[16]. It appears that multiply-labeled documents in the Reuters set are used by [36] while we and [16] used only singly-labeled documents.

	Our results		Published clustering		Classifier
Dataset	**avg of 6**	**best of 6**	**sIB**	**DataLoom**	**Naive Bayes**
20 News, 10 clusters	69.8	74.5	79.5	N/A	80.8
20 News, 20 clusters	46.2	50.9	57.5	55.1	65.0
Reuters, 10 clusters	73.2	74.5	85.8	N/A	90.8 (72)

4.3 Astrobiology Document Clustering

Since this data was the primary focus of the overall project, much effort was spent to investigate the clustering outcomes, the behavior of the new greedy feature selection algorithm, and how many clusters could be resolved using the $I_{corr}[C;W]$ method.

4.3.1 Feature Selection on Astrobiology Data

Due to the much longer documents in this data compared to the benchmark sets, the new feature selection method was not overwhelmed and functioned as intended. This allowed several tests comparing different feature selection methods' effects on clustering outcomes to be conducted, similar to those performed on the synthetic data.

When comparing the contents of the word lists made from this data by TF*IDF, $I[w]$, and the new method, it was discovered that there were hardly any differences in the lists despite there being over 42,000 words to choose from. The words may have been chosen in a different order, but the contents of the resulting lists were nearly the same. The new greedy method indicated keeping 1,806 words. For that number of words, the $I[w]$ ranked list differed from our method's list by 2% and the TF*IDF ranked list differed by 6.7%. Fig. 4.4 shows a comparison of the words in these lists relative to their probability over the entire dataset.

Differences in Word Distributions

The reasons for why word rankings from the information theoretic methods and the TF*IDF method were so different on the synthetic data yet so similar with this data were inves-

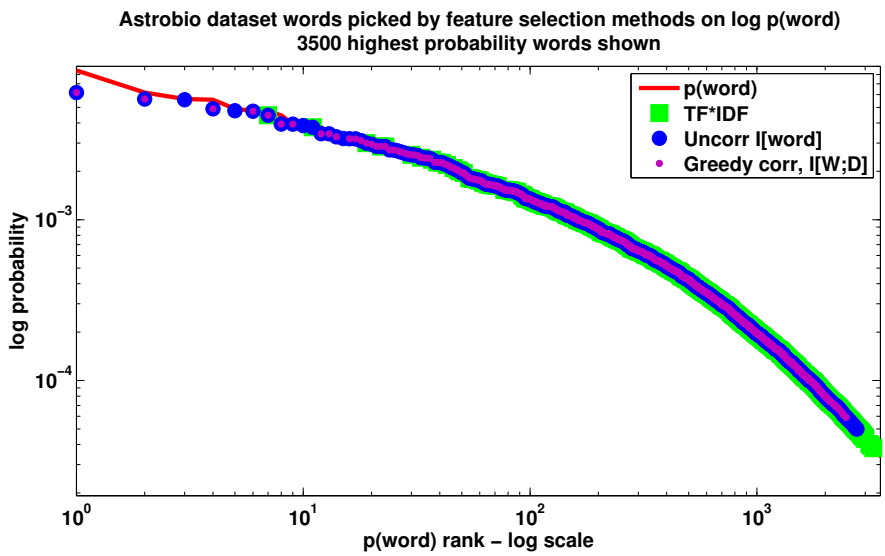


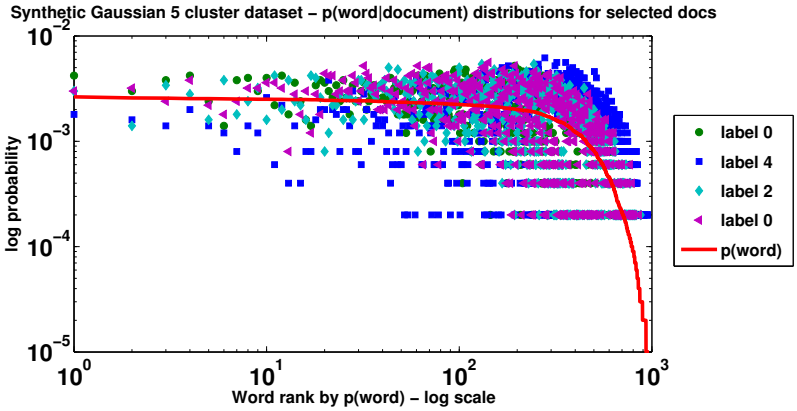
Figure 4.4: Comparison of Word Lists Produced by Different Feature Selection Methods. Astrobiology word lists plotted on log-log plot of $p\,(word)$ ranked from highest to lowest. TD*IDF and uncorrected $I\,[w]$ lists are 2,000 words long. The new greedy $I_{corr}\,[W;D]$ method chooses 1,806 words. (The log-log plot is used in order to show $p(word)$, which decreases exponentially, as a relatively straight line.)

tigated. Fig. 4.5 shows the overall $p\,(word)$ distributions and the word distributions for several individual documents,$p\,(w|d)$, for each dataset, ranked by $p\,(word)$.

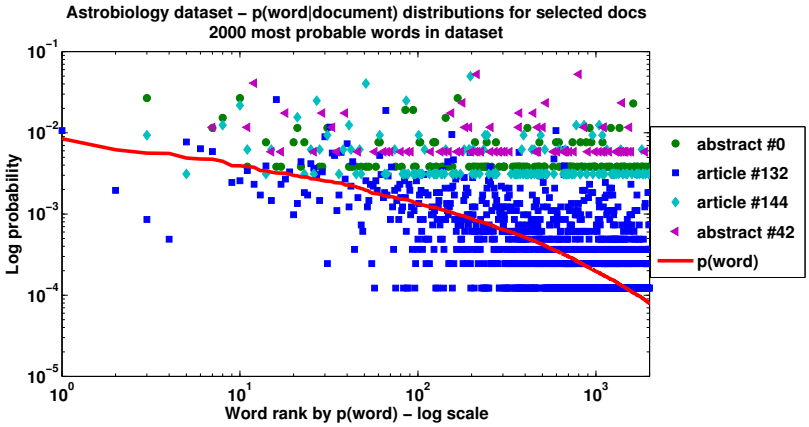
The plots show that the different dataset types, synthetic, the benchmarks, and the astrobiology set are quite different in how their words are distributed. Log-log scale plots were used in order to view $p\,(word)$ as a relatively straight line. Any difference in line slope in a log-log plot indicates an exponentially large difference in the data. For the document data sets, $p\,(word)$, when ranked from high to low drops exponentially while in the synthetic data it is much closer to a linear drop. (The steep drop in lowest end of $p\,(word)$ in the synthetic data indicates the ends of the tails of the outermost distributions.)

The exponential drop in $p\,(word)$ in the document data explains the similarities in the word lists from the different feature selection methods on the astrobiology data: only a tiny percentage of words occur with probability much above zero. The feature selection methods are all based on $p\,(word)$, therefore, they are all left to choose from the same small pool of words, where $p\,(word)$ is not tiny.

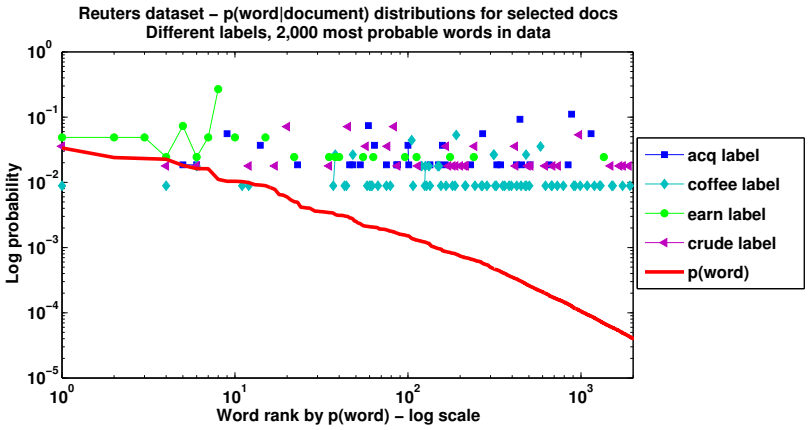
While the $p\,(word)$ distributions from the all of the text datasets have similar shape, the astrobiology dataset has a much richer structure in the word distributions in the documents, $p\,(w|d)$, than the benchmark data (as indicated by the relative sparseness of the Reuters data plot, Fig. 4.5c



(a) Synthetic data



(b) Astrobiology



(c) Reuters

Figure 4.5: Word Distributions of Selected Documents from Different Datasets. Log-log scale plots of $p\,(word)$ ranked from high to low and $p\,(word|doc)$. Zero probabilities are not plotted.

compared to that of the astrobiology data, Fig. 4.5b.) This difference is due to the longer length of the astrobiology documents. With more instances of each word per document, the astrobiology data is not so undersampled (the number of documents,N_D, is not nearly as large as the total data size,N). This allowed our new greedy feature selection method, Eq. 3.2.1, to successfully indicate informative words for the astrobiology data and not for the benchmark data.

4.3.2 Determining Number of Clusters

The maximum in the number of resolvable clusters in the astrobiology data using the $I_{corr}\,[C;W]$ method is greater than 200. Deterministic assignments were achieved on most runs. A plot of the $I_{corr}\,[C;W]$ curve, Fig. 4.6, indicates no discernible maximum or plateau, though the slope of the corrected curve does become noticeably shallower above approximately 60 clusters.

The intent of this project was to reveal connections between data in order to present simplified representations of relationships in interdisciplinary research. We need a small number of clusters that group a large number of documents for our work. The $I_{corr}[C;W]$ method definitely indicated that a number of clusters usable to summarize the data for our purposes will not be overfitting it in any way.

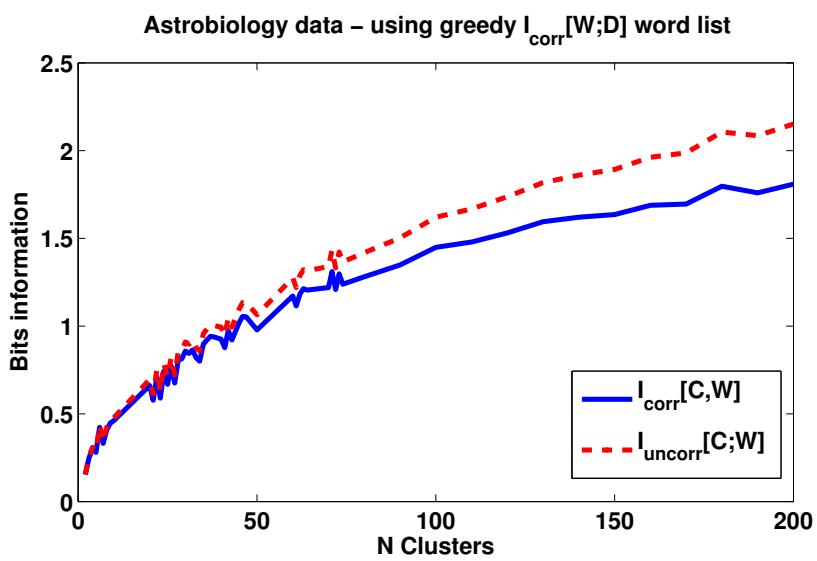


Figure 4.6: Corrected vs. Uncorrected $I[C;W]$ Curve for Astrobiology Data. 2 to 200 clusters using the IB clustering algorithm and the word list produced by the new greedy method. No clear maximum or plateau to indicate a maximum number of resolvable clusters is reached.

Jaggedness of $I[C;W]$ curve

The jaggedness of the $I[C;W]$ curve was investigated by running the clustering algorithm several times with 6, 7, and 8 clusters. Fig. 4.7 shows that low points occur at solutions where the clustering algorithm had difficulty reaching deterministic assignments. This is evidenced by the size of β when execution finished. Large β, or lower temperature, indicates more compression was required to squeeze the data into deterministic cluster assignments than solutions finishing with a small β. A solution with smaller β should therefore be considered a more optimal result. A slower annealing rate, α, could smooth out this curve by allowing the clustering algorithm to better avoid local minima, at the expense of dramatically slowing down execution.

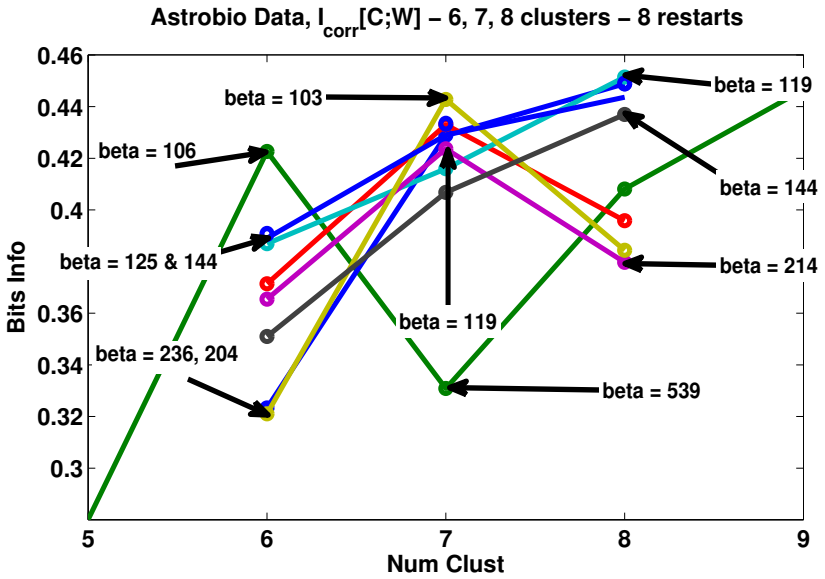


Figure 4.7: Comparison of $I_{corr}[C;W]$; 8 Clustering Runs on Astrobiology Data; 6, 7, and 8 Clusters. Reaching deterministic assignments while β is smaller gives a higher I[C;W] value, indicating a more optimal solution. Larger β indicates more compression was needed to achieve deterministic assignments, resulting in less information being retained.

4.3.3 Astrobiology Cluster Evaluation

Three different manageable quantities of clusters, two, six, and ten, were chosen for evaluation of the astrobiology data. Structures carried through all three sets of clusters were examined to see if there were strongly connected components within the data that maintained connections over random initializations at each level or if completely different structures were present with different divisions.

Fig. 4.8 is a force-directed graph created using the Gephi network visualization software [3]. This plot shows the strength of connections between one cluster and two, two clusters and six, and six clusters and ten. The nodes indicate the clusters, with the fractions on them being the cluster's index, or "name", over the number of clusters "family" it belongs to: $\left(\frac{\text{cluster index}}{\text{number of clusters}}\right)$. Node 1/1 represents all documents in a single cluster. The width of the edges connecting nodes represents the number of documents shared between the two.

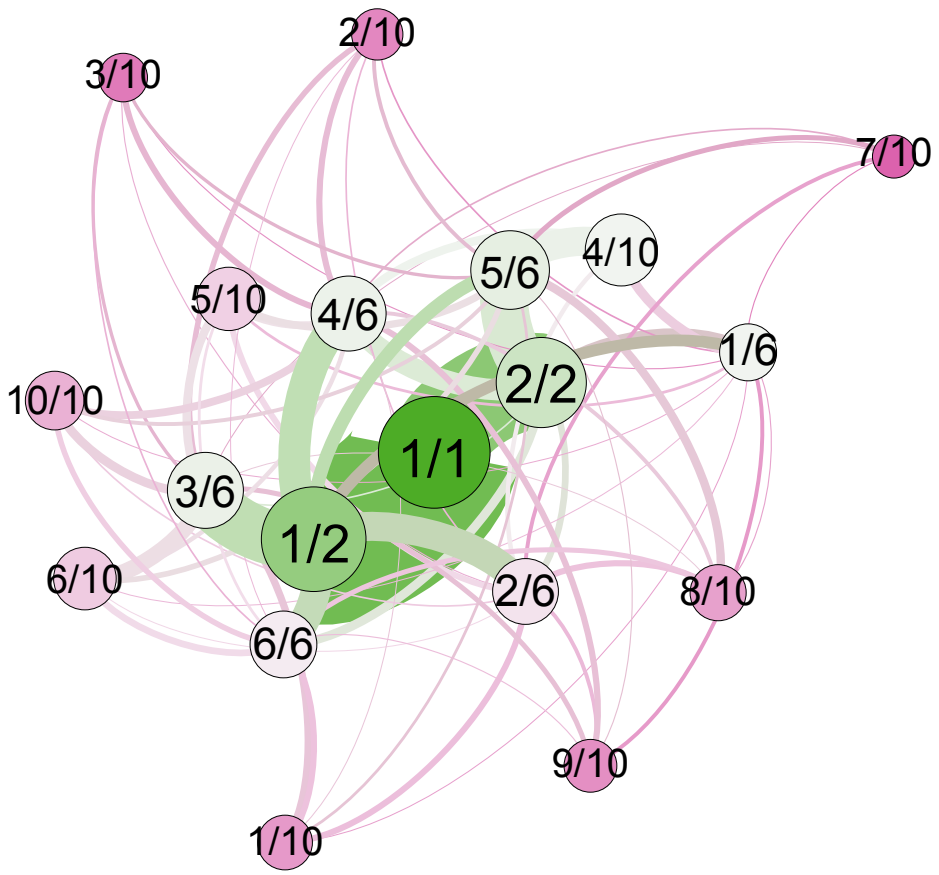


Figure 4.8: Connections Between 2, 6, and 10 Clusters in Astrobiology Data. Nodes represent clusters. Bottom number is number of clusters made, top is particular cluster's index. Size of a node reflects number of documents in cluster. Width of edges indicates the number of documents shared between clusters.

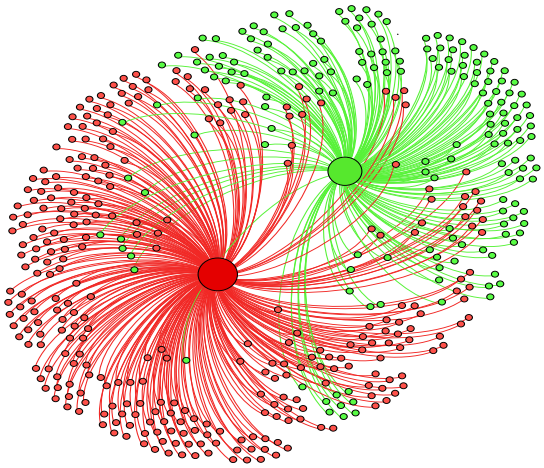
As can be seen, there are very strong connections between several clusters. For example, clusters 2/2, 5/6, and 4/10 are very closely related. 86 documents or almost half of cluster 2/2's 175 documents are in cluster 5/6, and a more than half of those, 47, continue on to cluster 4/10. Only 24 documents from cluster 1/2 are in cluster 5/6 and five continue to 4/10.

Analysis continued using the same force-directed graph with the document nodes shown and only clusters from the two, six or ten cluster runs displayed at each step. These can be seen in Fig. 4.9 with the document nodes colored to match their representative cluster. The force directing algorithm was run with all the cluster nodes present and all documents having an edge to their representative two, six, and ten cluster node. This resulted in the positioning of the document nodes so that they grouped by the connections that continue in all three numbers of clusters. Outlier documents that don't seem to easily group with others at any of these levels are isolated and are not pushed away in a single direction, leaving them closer to the cluster nodes. The 47 documents discussed above are the large group in the upper right of each plot.

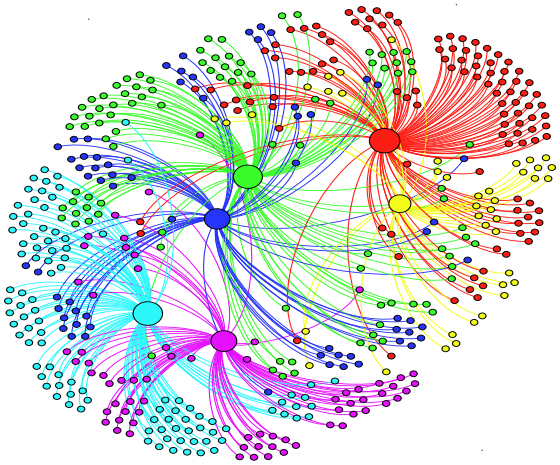
These results reveal that there are connections within the astrobiology documents at many levels. This is not a hierarchical structure, but a strongly interconnected network of relationships. While this increases the difficulty of evaluating the clustering outcome, it also provides opportunities for the AIRFrame development team to create much richer, multi-leveled experiences for AIRFrame users in the future.

Focusing on the data in six clusters, The NAI Team distribution over the clusters was investigated. Fig. 4.10 is a representation of the six cluster components with the document nodes colored to represent the originating team. The strongly connected component seen earlier is in the node/cluster at the top right. It is now shown to originate primarily from the University of Wisconsin and the NASA Ames teams. Exploring further by looking at the documents present in the cluster and their respective projects revealed that it is largely made up of articles related to microbes and minerals from the University of Wisconsin team's *Co-evolution of microbial metabolisms in the Neoarchean and Paleoproterozoic* and *Iron isotope biosignatures: Laboratory studies and modern environments* projects and the NASA Ames teams' project, *Mineralogical Traces of Early Habitable Environments*. Looking at the word distribution in that cluster showed that the most probable word is *Fe2* which is an iron isotope that plays a prominent role in many of the University of Wisconsin team's publications in these projects.

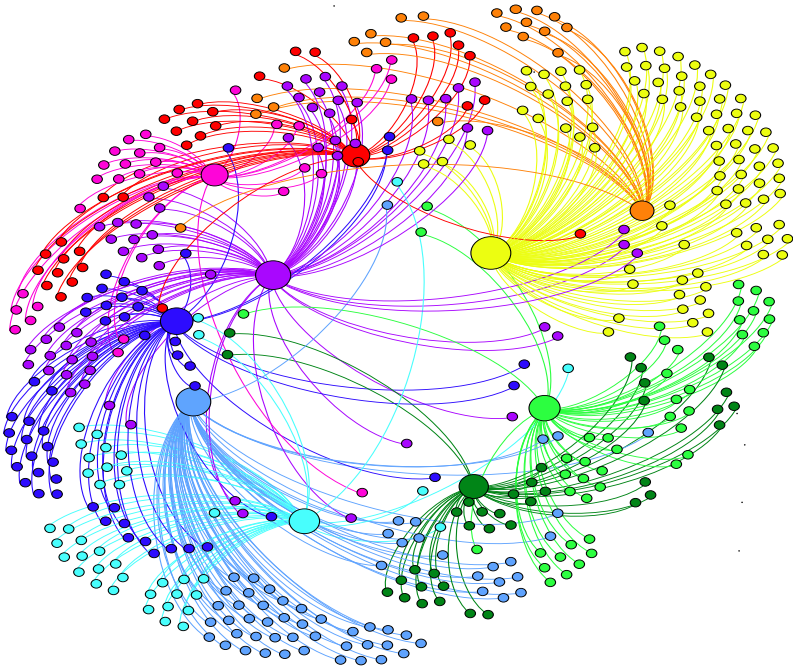
The cluster on the bottom of Fig. 4.10 is 2/6 in Fig. 4.8 and the magenta node in the six cluster plot from Fig. 4.9. This cluster also contains a strongly connected component, one relating to missions and observations in our solar system. Prominent are documents relating to observations done by the Messenger mission to Mercury with articles originating from the Carnegie Institute of Washington team's *Origin, Evolution, and Volatile Inventories of Terrestrial Planets* project and the *Astronomical Observations of Planetary Atmospheres and Exoplanets* project at



(a) Two clusters



(b) Six clusters



(c) Ten clusters

Figure 4.9: Comparison of Cluster-Document Relationships in Astrobiology Data, Two, Six and Ten Clusters. Small dots are documents, large dots are clusters with size indicating the number of member documents. Groups of document dots indicate components that remained together through all three clusterings. Single document dots close to clusters are outliers that don't consistently stay connected to other documents.

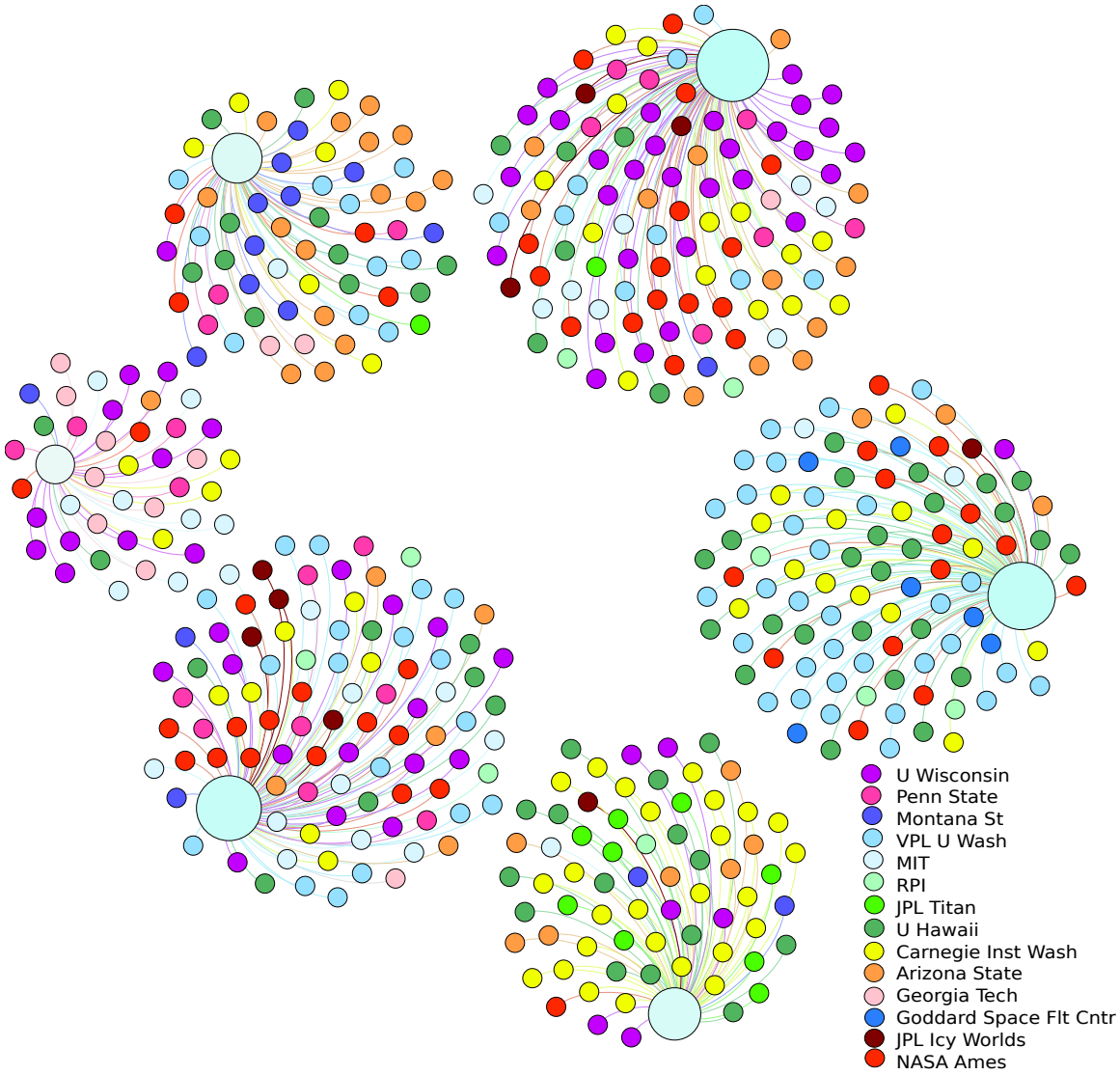


Figure 4.10: Astrobiology data in 6 clusters. Small circles are documents, large circles represent clusters with size indicating number of members. Colors on document dots indicate NAI team submitting the document.

NASA's Jet Propulsion Laboratory. High probability words and word stems in this cluster include *mantl, interstellar, hydrotherm, melt, anomoli, oxygen, europa, kuiper, cometari* and *graviti*

The same six cluster plot was attempted with the NAI Goals connected to the documents colorized. As was discussed in Sec. 3.1.2, most documents are labeled with multiple Goals and as anticipated, the resulting plot was not very satisfactory. It did reveal that almost no documents labeled with Goal 7: *Determine how to recognize signatures of life on other worlds and on early Earth*, are present in the 2/6 cluster discussed above nor in cluster 1/6 in Fig. 4.8 (the rightmost in Fig. 4.10) but it is unclear at this time if this reflects the subjects of the documents in the clusters or if it is just an artifact due to team labeling preferences.

Manual evaluation of the six cluster word distributions (the final cluster centers, $P(W|C)$, saved by the IB clusterer), lists of documents contained in the six clusters (from the final cluster assignments, $P(C|D)$), and projects represented by those documents suggested tentative topic labels for that set of clusters. Table 4.3 gives a sample of the most probable words in each cluster and initial topic labels.

Table 4.3: Topics and Highly Probable Words from Astrobiology Data in Six Clusters. Selected highly probable words/stems and initial topics derived from word distributions, documents contained in clusters, and projects represented.

Cluster 1	**Cluster 2**	**Cluster 3**
RNA/Bio	***Missions***	***Imaging Missions***
date	mantl	ioniz
molecul	interstellar	color
biochem	hydrotherm	sun
environment	melt	photometri
sea	anomoli	red
extrem	oxygen	spectroscopi
cycl	europa	co2
meteorit	kuiper	bright
sedimentari	cometari	window
pair	graviti	jet
Cluster 4	**Cluster 5**	**Cluster 6**
Habitability/Expeditions	***Microbes/Minerals***	***Astronomy***
cloud	fe2	residu
geochem	ecosystem	kinet
subseafloor	metal	bodi
prebiot	prokaryot	disk
reaction	bacteria	analog
biosynthesi	compound	star
hot	lake	ch4
amino	soil	infrar
thermal	detect	zone
natur	biofilm	crater

Chapter 5

Conclusions and Future Work

In this work progress was made in constructing a new labeled dataset for the astrobiology field. The data was clustered without any ad-hoc assumptions, using methods based solely on information theory – from feature selection to clustering to determining the maximum number of resolvable clusters.

A new feature selection method for determining informative words was developed during the project. This method was compared to two commonly used methods and its effectiveness on synthetic and benchmark datasets examined. It was then applied to the astrobiology data, where it provided theoretical guidance for which words, and how many of them to use for clustering.

The performance of the Information Bottleneck [43] clustering method implemented for this project was compared with previously published work using benchmark data. Additionally, an information theoretic method for determining the maximum number of clusters resolvable in finite sized data was used [41]. This work was the first application of that method to text.

Finally, the document clusters produced from the astrobiology data by this procedure were investigated for themes and connection structure.

5.1 Feature Selection

The new greedy feature selection algorithm developed in this work applies a correction to the per-word contribution to the mutual information between words and clusters to account for the finite sampling bias (Sec 3.2.1). Its performance was compared to two commonly used feature selection methods: the un-bias-corrected per-word contribution to the mutual information between words and documents, $I[w]$ (Sec 2.5) and *term frequency * inverse document frequency*, TF*IDF (Eq. 2.5.1), [17].

On the astrobiology dataset, all methods chose similar words, however, our new method is the only one that indicates the number of words to keep. The differences between word lists of the same length produced by the three methods were less than 7%, see Sec.4.3.1.

Synthetic data was constructed not intended to mimic word distributions found in documents but with highly overlapping distributions. The TF*IDF method eliminated words needed for successful cluster identification and none of the preprocessing methods tested clearly outperformed lists of randomly chosen words on this data.

5.2 Resolving the Number of Clusters

The $I_{corr}[C;W]$ method [41] (Sec. 2.4) for determining the maximum number of clusters resolvable in a dataset of finite size was applied on document data for the first time. With the astrobiology data, the value of $I_{corr}[C;W]$ increased up to 200 clusters, the maximum number attempted. This result allows us to be confident that any number of clusters that could feasibility be used to summarize the data into topics for the AIRFrame project will not be overfitting the data. On the benchmark document data, the value of $I_{corr}[C;W]$ continued to rise as the number of clusters was increased past the number of labels in the data. This shows that clustering these datasets with the number of labels as the number of clusters, as is standard practice, is also not overfitting the data.

Additionally, this method provided a fast, intuitive way to evaluate feature selection techniques and cluster quality (in place of micro-averaged precision or recall) on synthetic data.

5.3 Testing the Clustering Method

The Information Bottleneck with Deterministic Annealing clustering method (Algorithm 1, Sec. 2.3.1) was implemented for this project. Clustering results on two benchmark document datasets were compared with previously published results. Micro-averaged precision scores on the labeled data were quite similar to the published numbers despite not being able to completely reconstruct the previous experiments or number of tests. Results showed a $11 - 18\%$ loss of precision over a supervised method trained on the data labels and the best-of-six results differed by only five points of precision from the best-of-fifteen results reported for the *sequential Information Bottleneck* [36].

5.4 Astrobiology Clusters

The clusters found in the astrobiology data revealed a large amount of interconnected structure. Graphical evaluation of two, six, and ten cluster solutions showed a large number of connected components that carried through the different partitions. This indicates very strong connections between the members of the connected components on many levels. Graphical analysis of the originating NAI teams, examination of the word distributions in clusters, and analysis of the contents and project origins of the documents in the cluster results confirmed that this is the case and some themes within the clusters were immediately apparent.

5.5 Future Work

Future work will be in further analysis of the clustering results on the astrobiology data. The structure discovered in the the dataset is quite complex and interconnected offering opportunities to build the AIRFrame project into a system which allows viewing and exploration of astrobiological research in interesting and novel ways. It is planned to expand the investigation of the structure of the data into a 3-D environment in the near future.

Since this project began, the 2010 NAI Annual Report has been released. Adding this data will double the number of documents in the dataset. More documents should provide improved clustering results. Additionally, comparison of clustering results before and after adding a new year's worth of work will give the AIRFrame project a method of tracking the splitting and combining of topics in astrobiology research over time.

Bibliography

[1] AIRFrame. Astrobiology Integrative Research Framework Project. Website. http://airframe.ics.hawaii.edu/.

[2] C. Apté, F. Damerau, and S. M. Weiss. Automated learning of decision rules for text categorization. ACM Transactions on Information Systems (TOIS), 12(3):233–251, 1994.

[3] M. Bastian, S. Heymann, and M. Jacomy. Gephi: An open source software for exploring and manipulating networks. In Third International AAAI Conference on Weblogs and Social Media, March 2009.

[4] R. Bekkerman and M. Scholz. Data weaving: scaling up the state-of-the-art in data clustering. In Proceeding of the 17th ACM conference on Information and knowledge management, pages 1083–1092, Napa Valley, California, USA, 2008. ACM.

[5] R. Blahut. Computation of channel capacity and rate-distortion functions. Information Theory, IEEE Transactions on, 18(4):460473, 1972.

[6] H. H. Bock. Probabilistic models in cluster analysis. Computational Statistics & Data Analysis, 23(1):528, 1996.

[7] A. G. Carlton. On the bias of information estimates. Psychological Bulletin, 71(2):108–109, 1969.

[8] T. M. Cover and J. A. Thomas. Elements of information theory, volume 2. A. Wiley-Interscience, 2005.

[9] I. S. Dhillon and S. Mallela. Information-Theoretic Co-clustering. In KDD-2003: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24-27, 2003, Washington, DC, USA, page 89. Assn for Computing Machinery, 2003.

[10] R. O. Duda, P. E. Hart, D. G. Stork, et al. Pattern classification, volume 2. Wiley New York, 2001.

[11] N. Friedman, O. Mosenzon, N. Slonim, and N. Tishby. Multivariate information bottleneck. In Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, page 152–161, 2001.

[12] A. D. Gordon. Classification. CRC Press, June 1999.

[13] D. K. Harmon. Overview of the Third Text Retrieval Conference (Trec-3). DIANE Publishing, October 1995.

[14] R. A. A. Ince, R. Senatore, E. Arabzadeh, F. Montani, M. E. Diamond, and S. Panzeri. Information-theoretic methods for studying population codes. Neural Networks, 23(6):713–727, August 2010.

[15] A. K. Jain. Data clustering: 50 years beyond k-means. Pattern Recognition Letters, 31(8):651–666, June 2010.

[16] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In Machine Learning: ECML-98, pages 137–142. 1998.

[17] K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, 28(1):1121, 1972.

[18] J. Kleinberg. An impossibility theorem for clustering. In Advances in Neural Information Processing Systems 15: Proceedings of the 2002 Conference, page 463, 2003.

[19] S. Kullback and R. A. Leibler. On information and sufficiency. The Annals of Mathematical Statistics, 22(1):79–86, March 1951.

[20] K. Lang. Newsweeder: Learning to filter netnews. In Proceedings of the 12th International Machine Learning Conference (ML95), 1995.

[21] D. Lewis. Reuters-21578 text categorization test collection. Website. http://www.daviddlewis.com/resources/testcollections/reuters21578.

[22] D. J. Des Marais, J. A. Nuth III, L. J. Allamandola, A. P. Boss, J. D. Farmer, T. M. Hoehler, B. M. Jakosky, V. S. Meadows, A. Pohorille, B. Runnegar, et al. The NASA astrobiology roadmap. Astrobiology, 8(4):715–730, 2008.

[23] G. A. Miller. Note on the bias of information estimates. In Information Theory in Psychology: Problems and Methods, pages 95–100, Monticello, IL, 1955. The Free Press.

[24] G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50(2):159–179, 1985.

[25] H. M. Müller, E. E. Kenny, and P. W. Sternberg. Textpresso: An Ontology-Based information retrieval and extraction system for biological literature. PLoS Biol, 2(11):e309, 2004.

[26] NAI. Nasa astrobiology institute 2009 annual report. Website. http://astrobiology.nasa.gov/nai/library-of-resources/annual-reports/2009/.

[27] S. Panzeri, R. Senatore, M. A. Montemurro, and R. S. Petersen. Correcting for the sampling bias problem in spike train information measures. Journal of Neurophysiology, 98(3):1064 –1072, 2007.

[28] S. Panzeri and A. Treves. Analytical estimates of limited sampling biases in different information measures. Network-Computation in Neural Systems, 7(1):87–108, 1996.

[29] M. F. Porter. An algorithm for suffix stripping. Program: electronic library and information systems, 14(3):130–137, 1980.

[30] J. Li R. Datta, D. Joshi and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys (CSUR), 40:5:1–5:60, May 2008. ACM ID: 1348248.

[31] M. S. Rogers and B. F. Green. The moments of sample information when the alternatives are equally likely. In Information Theory in Psychology: Problems and Methods, pages 101–108, Monticello, IL, 1955. The Free Press.

[32] K. Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. Proceedings of the IEEE, 86(11):2210–2239, 1998.

[33] G. Salton. Automatic text processing: the transformation, analysis, and retrieval of information by computer. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[34] H. Schütze, C. D. Manning, and P. Raghavan. Introduction to Information Retrieval. Cambridge University Press, 2008.

[35] C. E. Shannon. A mathematical theory of communication. Bell System Technical Journal, 27:379–423 and 623–656, 1948.

[36] N. Slonim, N. Friedman, and N. Tishby. Unsupervised document classification using sequential information maximization. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pages 129–136, Tampere, Finland, 2002. ACM.

[37] N. Slonim and N. Tishby. Agglomerative information bottleneck. Advances in neural information processing systems, 12:617–23, 2000.

[38] N. Slonim and N. Tishby. Document clustering using word clusters via the information bottleneck method. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pages 208–215, 2000.

[39] N. Slonim and N. Tishby. The power of word clusters for text classification. In 23rd European Colloquium on Information Retrieval Research, volume 1, 2001.

[40] D. Srivastava and S. Venkatasubramanian. Information theory for data management. Proceedings of the 2010 international conference on Management of data, page 1255–1256, 2010. ACM ID: 1807337.

[41] S. Still and W. Bialek. How many clusters? an Information-Theoretic perspective. Neural Computation, 16(12):2483–2506, March 2004.

[42] S. Still, W. Bialek, and L. Bottou. Geometric clustering using the information bottleneck method. Advances in neural information processing systems, 16, 2004.

[43] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing, pages 368–377. University of Illinois, 1999.

[44] A. Treves and S. Panzeri. The upward bias in measures of information derived from limited data samples. Neural Computation, 7(2):399–407, March 1995.