Using Latent Semantic Indexing for Data Deduplication

Michael Spiz

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science

University of Washington

2006

Program Authorized to Offer Degree: Institute of Technology – Tacoma

University of Washington
Graduate School

This is to certify that I have examined this copy of a master's thesis by

Michael Spiz

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____
Isabelle Bichindaritz

_____
LouAnn Lyon-Banks

_____
Steve Hanks

Date: _____

Signature_____

Date_____

University of Washington

Abstract

Using Latent Semantic Indexing for Data Deduplication

Michael Spiz

Chair of the Supervisory Committee:
Professor Isabelle Bichindaritz
Computing and Software Systems

This paper presents a method for deduplicating records using latent semantic indexing (LSI). When merging two datasets from two different sources, there is often a problem with overlap between the records. Finding these duplicate records can be challenging since the format of the data is often different between databases. Existing methods for data deduplication focus primarily on using data cleaning and approximate string matching techniques. While these methods are effective for finding duplicates in records with few words, such as names and addresses, they do not work as well for records with more terms such as corporate names. The system described in this paper uses LSI techniques to discover duplicates in a dataset. This article shows the LSI deduplicator performs more accurately on test and real-world data than existing techniques, but at the expense of runtime and resource utilization.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express appreciation to all who helped with proofreading of the material presented here. A special thanks goes to professor Isabelle Bichindaritz, who in many evening meetings has helped assist me in submitting a paper for publication and then hone it into a thesis. Her participation was essential in helping me complete this part of my degree.

Chapter 1

INTRODUCTION

Many companies collect data from various independent sources, placing that information inside a data warehouse. When merging two datasets from two different sources, there is often a problem with overlap between the records. Finding these duplicate records can be challenging since the format of the data is often different between databases. Existing methods for data deduplication focus primarily on using data cleaning and approximate string matching techniques. While these methods are effective for finding duplicates in records with few words such as names and addresses, they do not work as well for records with more terms such as corporate names. The system described in this paper uses latent semantic indexing (LSI) techniques to discover duplicates in a dataset.

As an example, the company Diligenz Inc. acquires databases of corporate records from many states throughout the USA. Initially, these databases have different schemas, are in different formats, and adhere to different standards. All these databases are then carefully transformed into a standard schema to help ease the retrieval of data. However, searching for corporate entities in this data is still a challenging task. First, even though each corporate name has a unique identifier associated with it, there are no identifiers that associate a name with a corporate entity. This means that it is not possible to find all names belonging to a franchise operation, large institution, or any other entity with multiple names. Secondly, it is difficult to find distinct corporate entities that do business in multiple states because of the differences in the underlying schemas for the individual state databases. The desire to find a way to be able to autonomously generate associations between these names is the motivation for this paper.

Section 2.4 defines LSI and how it works. The problem statement is outlined in section 3. At this point, the LSI deduplication system is introduced in section 4. It describes both the

framework that was used and how the LSI indexer functions. The datasets used for testing are described in section 5. Using those datasets, section 6 shows an analysis of speed and accuracy of the indexer and compares it with two other methods. Similar existing projects are described in section 7. Possible future improvements are discussed in section 8. Lastly, the paper finishes with some conclusions from the project in section 10.

Chapter 2

BACKGROUND INFORMATION

2.1 Data Mining

Data mining is the process of extracting useful patterns and knowledge from large amounts of data usually stored in a data warehouse (a database designed to store gathered data). This is typically information that would otherwise not be easily extracted by just looking at the data. For example, a database with information about side-effects of medication could be mined to figure out what combinations of drugs cause adverse reactions in people. Another example could be for a database that has listings for upcoming airplane flights and their prices. One could use data mining to find which routes are the cheapest when flying to a given destination. Credit card companies use data mining to determine whether credit card purchases are likely to be fraudulent.

There are two main types of data mining: predictive and descriptive. [1] Predictive data mining is the process of predicting what events or behaviors are likely to occur based on existing data or information that has been gathered. A predictive data mining model is useful in figuring out consumer buying habits and tendencies or forecasting highway traffic levels given a location and a time.

Descriptive data mining is where one tries to analyze and organize a set of data in a way that allows one to find patterns or cluster it into distinctive groups. This type of data mining model is useful in finding traits in certain populations of people for marketing purposes or for finding inefficiencies in the productivity of a company.

The main focus for this project centers on devising a process that will aid in a certain form of descriptive data mining. This process is called data cleaning.

2.2 Data Cleaning

When collecting data in a data warehouse, the information can come from multiple independent sources. An example would be a warehouse collecting names and address of people for a mass marketing campaign. Names and addresses could be drawn from hundreds of different locations. Since there is no way to guarantee consistency between these data sources, discrepancies may arise between what would otherwise be identical sources. Discrepancies could include the use of different naming conventions; data entry mistakes due to typing errors; extra or omitted data; the use of abbreviations versus expanded representations; and corrupted data. People who are faced with this problem are interested in ways to detect and correct these erroneous bits of information in the data they collect. The process of correcting this data is called data cleaning.

The implementation of data cleaning algorithms can take several forms. One of these involves having a reference set of data, such as a dictionary. A spell-check program is a commonplace example of a data cleaning algorithm. It uses a dictionary of known words as a reference and verifies that each word in a document matches up with an entry in the dictionary. If a word is not found, it tries to find the closest match and suggests it to the user. In most cases, the suggested word will end up being the intended word in the document. The process of spell checking cleans up the document into an acceptable form.

Another form of data cleaning involves rearranging data into a standard format. The best example for this would again be cleaning a list of addresses. There are many different ways to write out an address, however, in order to make the information easy to process by a computer, the data needs to be in a specific order. A sufficiently complex data cleaning algorithm can rearrange and categorize the independent components of an address. Once this process is completed, the data can easily be imported into a database ready to be utilized for statistical analysis, searches, or anything else.

2.3 Min-edit Distance

When comparing two similar strings, one needs to have a quantitative way to measure their degree of similarity. There needs to be a way to say that the strings “appropriate meaning”

```
1  for i ← 0 to m do  M[i, 0] ← i;
2  for i ← 0 to n do  M[0, i] ← i;
3  for i ← 1 to m do
4      for j ← 1 to n do
6          M[i, j] ← min(
           M[i − 1, j] + 1, M[i, j − 1] + 1, M[i − 1, j − 1] + delta(x[i], y[j]));
7      end
8  end
9  return M[m, n];
10 delta(i, j) = if i = j then return 0 else return 1;
```

Figure 2.1: Min Edit distance algorithm.

and "approximate matching" are more similar to each other than "appropriate meaning" and "application programming." One way to accomplish this is through a min-edit distance algorithm. Min-edit distance is the minimum number of single character changes necessary in order to transform one string into another. A change would consist of changing a letter, adding a letter, or removing a letter. [22]

The simplest way of calculating the min edit distance for two strings is to construct a matrix of integers M of size $m * n$ where m is the length of the first string x and n is the length of the second string y. The algorithm to populate the matrix is shown in algorithm 2.1. To put it in English, the first row and column of the matrix is filled with consecutive integers starting at 0. Then, the matrix is filled out using the technique in line 1. This will cause the diagonal of the matrix to end up with the edit distance as the matrix fills out. By the end, the lower right-hand corner of the matrix will have the edit distance. This technique is used as a foundation to many data cleaning and data mining algorithms.

2.4 Latent Semantic Indexing

Latent Semantic Indexing (LSI) is a method for uncovering the semantics in a body of text in such a way that it is easier to process and search by a computer. It is also called Latent

Semantic Analysis (LSA). In bodies of text, it is common that different words are used to describe the same concept. Likewise, a single word can also have multiple meanings. This makes performing a search for a document by searching for instances of a word an inefficient way of finding all the documents related to that word [19, 21].

This is where LSI comes in. LSI creates a matrix whose columns are comprised of words and rows comprised of documents. The values of the matrix are weights that are proportional to the number of times those words appear in the document. The weights are structured so that words that are rarer have a greater weight. This is because words that are used commonly throughout a document such as “the”, “and”, or “this”, do not carry as much relevance in the document as other words do. Afterwards, a singular value decomposition (SVD) is applied to the matrix. This turns each document into a multi-dimensional vector that represents the content inside the document. Documents with similar terms and content will end up having vectors which are closer to each other than unrelated documents. Performing queries for words are more successful after performing LSI on the set of data because the search results will also return documents that have words that are synonymous with the queried word.

Chapter 3

PROBLEM STATEMENT

Consider an example where one has a data warehouse that contains millions of filled-out forms. These forms contain filings for bank loans made out by various corporations. In order to gather meaningful statistical data on the data filled out in these forms, one needs to make sure every form is entered in a consistent and uniform manner. This, unfortunately, is never the case. Figure 3.1 shows an example listing of what possible names for the University of Washington could look like. The names can be manually clustered into two, possibly three different groups: two branches of the University of Washington and Washington University. A human can quickly discern the differences between the variations and abbreviations and properly cluster the list. Having a computer perform this kind of operation accurately and autonomously is a much more difficult task.

Figure 3.2 shows an actual sampling of names from a database of corporate information in California. Again, just by glancing over the list, one can make educated guesses as to which names belong to the same corporate entity. For instance, "BANK ADMINISTRATION INSTITUTE" and all of its chapters should be grouped together. In contrast, all the "Bank of..." rows should be grouped separately except for "BANK OF CANTON OF CALIFORNIA" since each (minus the exception) describes a separate location for a bank

University of Washington
University of Washington, Tacoma
Univ of WA
Washington University
UW
UWT

Figure 3.1: Sample list of university names.

BANK ADMINISTRATION INSTITUTE
BANK ADMINISTRATION INSTITUTE-DESERT-SEA CHAPTER
BANK ADMINISTRATION INSTITUTE-GOLDEN GATE CHAPTER
BANK AND GOLDBERG PRODUCTIONS, INC.
BANK AUDI (U.S.A.)
BANK AUSTRIA CREDITANSTALT AMERICAN CORPORATION
BANK COMPLIANCE ASSOCIATES, INC.
BANK OF BERKELEY
BANK OF BURLINGAME
BANK OF CANTON OF CALIFORNIA
BANK OF CANTON OF CALIFORNIA LEASING CORPORATION
BEVERLY MANOR INC. OF BURBANK
BEVERLY MANOR INC. OF BURBANK SOUTH
BURBANK ENTERTAINMENT VILLAGE, L.L.C.
BURBANK ENTERTAINMENT VILLAGE ASSOCIATES, L.L.C.
TERRA BURBANK PARTNERS ONE, LLC
TERRA BURBANK PARTNERS TWO, LLC
RIVERBANK DENVER, INC.
RIVERBANK PARAGON, INC.

Figure 3.2: Sample of corporate names from the California corporate database.

and likely are separate corporations.

Current implementations of data deduplication systems focus mostly on the analysis of substrings and their similarity to one another [5, 8, 11]. Many use methods that involve more complexity, such as using stemming algorithms, soundex and metaphone, or Hidden Markov Models in order to clean and normalize the data as much as possible and to decrease word complexity. However, after performing all their preprocessing operations, these systems ultimately end up using some form of substring comparison in order to make the final determination of whether two records are duplicates.

This method for finding duplicates works very well with person names, addresses and other data that involves single words or simple structures. For longer, more complex strings, such as company names, the standard methods become less effective. The reason is that company name variations involve the addition, subtraction, and transposition of entire words.

Existing deduplication methods focus more on the character and word level variations and not phrase-level changes. These also do not detect synonymous word replacements such as "company" and "corporation". Replacements of that sort occur often in corporate name databases.

Having a system that can deal with such information would be a valuable asset for finding distinct corporate entities within a raw dataset of business names. Using such information one can compile these associations and either sell this to customers or perform analysis on them.

Chapter 4

LSI DEDUPLICATION SYSTEM

4.1 Febrl

Febrl (Freely Extensible Biomedical Record Linkage) [13] is an open-source Python project designed to perform data cleaning, and deduplication of database information. Febrl was specifically designed by the biomedical community to help sort through patient name and address datasets. These datasets often contain typos and duplicate information. It is very tedious to sift through the data by hand, so this project was created to automate the process.

Febrl is divided into several modular components. The first component is the standardizer, which is responsible for cleaning and normalizing the incoming dataset. For example, there are several ways to write out the word "avenue" such as "ave." or "av." The standardizer uses lookup tables to convert all these word variations into a single word. Febrl comes with standardization lookup tables for names, titles, and addresses, however other industry-specific lookup tables can be added to this without much problem.

Often in a database, names and addresses are stored in only two columns. For more effective deduplication, the program needs to separate these into their individual components (first name, middle initial, last name, house number, street, etc.). The problem with parsing out these columns is that certain components of the name and address may be missing or transposed depending on how the data was originally entered. To handle this, Febrl uses Hidden Markov Models (HMMs) to parse the names and addresses. The names and addresses are parsed out into tokens. Each token is assigned a tag based on information in lookup tables (known names, cities, postal codes, etc.). Once each token is tagged, they are passed into a trained HMM to determine what token most likely corresponds with what output field.

At this point, all the records are cleaned and partitioned into separate fields. The last step is to link similar records together. Comparing every record to every other record is quite

computationally expensive especially when there are a large number of records to compare. To help this situation, each record should only be compared to a subset (or block) of the entire record set. There are several blocking algorithms in Febrl that can be used to perform this task. The trick to these is to avoid having a matching record outside of the comparison subset.

Finally, a comparison algorithm is used in order to get a measure of similarity between two records. Febrl implements several approaches, one of which, the Naïve Bayes classifier, adds up weights for each of the attributes and uses that as a similarity measure.

As is apparent, there are many pieces to making Febrl work. The design of the system is still very developer-oriented. There is no user interface, and the configuration of a project is done using a controller class. While this gives the program a high learning curve, it also makes it modular and capable of accepting enhancements easily. Febrl makes for an excellent platform for researching algorithms and techniques for data cleaning and record linkage.

4.2 LSI Indexer

Instead of using a traditional string similarity based approach to deduplicate records, the system presented here uses LSI to index and then deduplicate the test records. The reasoning for using such an indexer is that datasets containing strings of multiple terms, such as company names, may contain enough semantic meaning in them to allow one to cluster possible duplicate records together.

The LSI indexer is built on top of the Febrl framework [13]. Since Febrl is written in Python, the indexer is also written in the same language. Figure 4.1 shows an overview of how the data is processed within the system.

The basic functionality of the LSI indexer is as follows. The program is controlled through a Python script. It first defines a schema for the data to be imported and then uses Febrl to import a comma separated data file. The fields that are to be indexed are defined in the script. As each record from the dataset is read, the parser combines the marked fields into a single large string labeled with an identifier (ID). This tuple is then passed to the indexer for processing. The indexer then tokenizes the string into words, stems each word

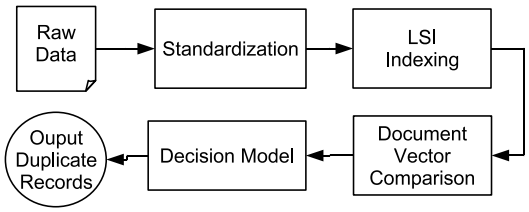

Figure 4.1: Overall data flow for the LSI indexer.

using the Porter stemming algorithm [24], and then adds each unique stemmed word into a hash table.

Once all the records are read in, a matrix is constructed where each column represents a stemmed word in the hash table, and each row is a normalized vector of the word frequencies. A Singular Value Decomposition (SVD) transformation is then performed on the matrix. The resulting vectors in the matrix, at this point, represent the semantic position within the dataset. The cosine distance is then measured between each vector. If the distance is within a specified tolerance, those two vectors are marked as duplicates. The result is a collection of sets of vectors that are potential duplicates.

Chapter 5

DATASETS

The evaluation was performed on two very different datasets in order to compare the LSI indexer against the other indexers. The first dataset is a sample listing of corporation names and addresses in the state of California. This database holds a collection of all registered businesses in the state of California. It holds approximately 30 million records, taking up a total of 18GB of space. There is, however, much less than this amount of businesses in California. So why are there so many records? Many businesses have franchises, branches, or separate divisions. These are required by the state to have separate names. These names, however, are not linked in any way to the parent company, so it is difficult to resolve these to a list of distinct corporate entities. This problem becomes multiplied when merging this type of corporate data from different states into one database. For this reason, this data is a good candidate for deduplication. By having a list of distinct corporate entities, one can gain a better picture of who is in control of what businesses. Table 5.1 shows a sample of the dataset. For the entire dataset, see appendix A.1.

The second dataset that was chosen comes from a dataset generator that is included with Febrl. It is designed to create real-world name and address datasets that contain duplicate information [10]. The generator allows one to control the number of duplicates in the dataset and the number and type of errors that are introduced into the data. A major advantage for having such a dataset is that one is guaranteed to know which records are duplicates. This makes benchmarking different algorithms much simpler. Table 5.2 shows a sample of the dataset. For the entire dataset, see appendix A.2.

In order to be able to get a good gauge for the number of duplicates that exist in the corporate dataset, the sample was limited to 150 records. This way, it was feasible to manually look through the records and find the actual number of duplicates.

Table 5.1: Raw data from the California corporate database.

EntityID	EntityName	MailAddress2	MailCity	MailState
630	(THE) UNIVERSITY HEIGHTS IMPROVEMENT ASSOCIATION INC.			
6141	222 UNIVERSITY AVENUE - NORTH	P O BOX 4456	BURLINGAME	CA
6142	222 UNIVERSITY AVENUE CORPORATION	P O BOX 4456	BURLINGAME	CA
14061	810 UNIVERSITY AVENUE INC.	810 UNIVERSITY AVE	BERKELEY	CA
14858	921 UNIVERSITY INC.	921 UNIVERSITY AVE	BERKELEY	CA
35927	ABLE UNIVERSITY PRESS INC.	4084 N BURGE RD	STOCKTON	CA
38101	ACADEMIC CREDIT UNIVERSITY	5181 OVERLAND AVE	CULVER CITY	CA
38153	ACADEMIC RESEARCH UNIVERSITY	4860 LONG BEACH BLVD	LONG BEACH	CA
38445	ACADEMY OF INTERNATIONAL SOCIETY OF PEOPLE UNIVERSITY	2315 CYPRESS CIRCLE DRIVE	LOMITA	CA
42920	ACHIEVEMENT UNIVERSITY	AOKI 1102-1-8-10 HIGASHIGOTANDA SHINAGAWAKU	TOKYO	JAPAN
44588	ACP UNIVERSITY CORPORATION	30129 VIA RIVERA	RANCHO PALOS VERDES	CA
48350	ADAM SMITH UNIVERSITY	3463 STATE STREET SUITE 363	SANTA BARBARA	CA

Table 5.2: Raw data for generated address list (columns truncated)

rec_id	given_name	surname	street_number	address_1	...
rec-359-dup-0	joel	baynes	16	beasley street	...
rec-74-org	alaiyah	leslie	33	becker place	...
rec-305-dup-0	finn	maspwn	556	bisdee street	...
rec-195-org	flynn	lock	41	hallett place	...
rec-232-org		white	237	westgarth street	...
rec-316-org	april	grubb	38	chuculba crescent	...
rec-90-org	ellie	carich	31	wynn street	...
rec-368-org	hannah	george	50	bural court	...
rec-97-org	courtney	highet	46	sugarloaf circuit	...
rec-393-org	thomas	penno	7	arabana street	...
rec-172-dup-0	chloe	vreugdemburg	35	chaton place	...
rec-25-org	blade	coleman	144	britten-jones drive	...

Chapter 6

ANALYSIS

The following describes the procedures used for discovering duplicate records using the different methods. Both the LSI indexer and the other indexers use the same code for reading the data. Since the LSI indexer only focuses on term frequencies, all the data for each row is concatenated and analyzed as a single phrase during the reading process.

The LSI indexer groups semantically similar rows together based on the cosine distance between the reference row and its potential duplicates. The higher the cosine distance, the more similar the two rows are. For the analysis of the California corporate data, a cutoff cosine distance of 0.55 was chosen.

The indexers built into Febrl require more information in the setup process. Each column of data can be processed and indexed differently in order to maximize the usefulness of the indexes that are generated. The first "bigram" indexer represents Febrl's default settings. For the address list, names were encoded for comparison using metaphone encoding [6], addresses were compared using the Jaro-Winkler approximate string matching algorithm [6], and cities were compared using keying error distance [6]. The second "edit distance" indexer changes the blocking method to sorting, and changes the comparators to use edit distance.

Since the California data sample is from real-world data, the actual number of known duplicate records is unknown. Thus, the only way to measure the effectiveness of the algorithms is to compare the results against a human. After manually analyzing the sample dataset, 22 duplicate records were found.

The results in table 6.1 show the percentage of duplicate records found compared to a manual human-based search of the data. Appendix B.1 shows the full list of human-marked duplicates. The algorithms that most closely match what a human would pick receives the highest percentage. The false-positive (FP) rate shows what percentage of marked duplicate

Table 6.1: Benchmark results from the LSI deduplicator.

	Percentage of duplicates detected	FP Rate	Run time (seconds)
Human (manual) comparison	100%	0%	900
LSI with address dataset	96%	0%	130
LSI with corporate dataset	36%	50%	190
Bigram with address dataset	54%	0%	3
Bigram with corporate dataset	9%	50%	3
Edit distance with address dataset	76%	0%	3
Edit distance with corporate dataset	13%	40%	13

records were not human-labeled as being a duplicate. For a full listing of the outputs from each algorithm, see appendix B.

The main drawbacks of the current implementation of the indexer are its execution time and memory consumption. All of the algorithms perform the same number of disk accesses. The LSI indexer needs to store and process its information in matrices, so it scales in $O(n^2)$ time. The memory consumption has the same problem. A matrix needs to be built for the index that is $m * n$ in size where m is the number of unique terms in the dataset and n is the number of items in the dataset. Both of these factors make scaling the size of the dataset difficult. In contrast the other indexers only need to look at subsets of the datasets. This makes them scale in $O(n \cdot log(n))$ time.

Regardless, the performance of the LSI algorithm was lower than expected. The source of the problem was traced to the dot multiplication functions in Python. In order to get a more accurate gauge of the true performance of the algorithm, a version of the system was designed that piped all the matrix and vector operations through Mathematica 5.1. As is apparent in Table 6.2, the LSI indexer performance is faster than the Febrl indexers for the size of the dataset being used, but slows down quickly due to the complexity of the algorithm.

The LSI algorithm achieved the highest accuracy of the algorithms tested even though it needed the least amount of data cleaning and pre-processing. This means that there is still more potential to get greater accuracy by performing more complex data cleaning up

Table 6.2: Performance of LSI Indexer through Mathematica

Dataset Size	Run Time (seconds)
150	1.2
250	5.2
500	32.6

front. Greater accuracy can also be achieved by normalizing the data using domain-specific transformations such as street names, surnames, or dates.

Chapter 7

DISCUSSION

7.1 ALIAS

Sunita Sarawagi created a system called ALIAS which provides a framework for interactive deduplication [27]. The goal of the system is to utilize an active learning approach to label duplicate records in a database. The system starts with a small initial training set. Duplicate training records are classified with a "1" and non-duplicate records with a "0". Each of these pairs are also run through a set of various similarity functions (edit-distance, soundex, etc.) that may also include domain-specific functions to enhance accuracy. The system then relies on the values generated from these various functions to determine the best way to find duplicate records within the dataset. The program finds record pairs which have the greatest level of uncertainty as to their similarity and presents those to the user for classification. Doing this iteratively, the program learns how to distinguish between similar and dissimilar records in the database. This system provides a novel way for finding duplicates in a dataset using training data provided by a person. The system, however, only uses syntactic measures for similarity. In certain situations, it would likely benefit from semantic similarity measures that LSI can provide.

7.2 Iterative Record Linkage

Indrajit Bhattacharya and Lise Getoor worked on a method for deduplication that tries to find linked records by examining them in the context of the other records in the database [3]. To create these links, the authors use an iterative approach: they make several passes over all the records in the database, creating new associations between records each time. Again, the approach used is syntactic in nature. As described in section 3, there are certain types of data that do not lend themselves well to being analyzed in such a manner. The LSI indexer attempts to provide a semantic approach to finding duplicates.

Chapter 8

FUTURE WORK

The programming done for this project was meant as a proof-of-concept. Therefore, not much emphasis was placed on performance and scalability. A possible first step to improving the current program would be to incorporate a distributed LSI algorithm. This would help in dividing the work of creating the index among multiple computers. Work has already been done in this area [28, 21]. Some of these also use sparse matrices for the document-term matrix in order to reduce memory usage. There also exist much faster implementations of algorithms that approximate the SVD algorithm [12].

An additional solution to the scaling issue would be to use a two-stage approach to the indexing by using LSI as a blocking algorithm. In the first stage, entries would be indexed and clustered by their syntactical similarity, just like Febrl currently does. In the second stage, LSI would be performed only on subsets, or blocks of syntactically similar data. This would reduce the complexity of the system to $O(n \cdot log(n))$, but may have negative effects on accuracy.

A problem with the current implementation is that it is sensitive to character transpositions, misspellings, and alternate word forms. Since the base unit for LSI is words, it treats each spelling variation as a separate word with a separate semantic meaning. It would benefit the program to create a type of stemming algorithm that would stem words and also account for character transpositions and misspellings.

The datasets that this type of program analyzes contain a very diverse set of terms due to the large number of names that are present in addresses and businesses. The diversity of terms increases the size of the document-term matrix, which in turn slows down indexing. One way to improve the performance of the program would be to implement a pass prior to indexing that would remove all terms that only occur once in the document space. These terms do not provide the LSI algorithm any semantic information.

It would be interesting to see this algorithm included into a hybrid deduplication system such as ALIAS [27]. Having a system that can take both string similarity measures and semantic similarity into account at the same time would likely allow for finding results with greater accuracy.

Chapter 9

EDUCATIONAL STATEMENT

9.1 Graduate Work Contribution

The process of writing my thesis involved the study and implementation of different types of text processing algorithms. I was able to use both work and personal experience to help guide me toward finding the correct combination of algorithms necessary to accomplish the goals of this project. Research and paper-writing techniques were applied from the CSS 598 Master's seminar course. The knowledge gained from the CSS 540 Formal Models in Computer Science course helped me get more of a mathematical understanding of the algorithms I was using. Finally, the CSS 555 Data Mining course directly applied to the subject matter of my paper. The mining techniques taught there helped hone my understanding of the concepts used in this paper.

9.2 New Learning

This project has expanded my knowledge of the field of data mining and data cleaning. Through the research I have performed, I learned about the large quantity and variety of techniques and algorithms having to do with analyzing tuples of data for similarity. This information has given me a thorough understanding of the of the rich field of data cleaning and processing. I have never developed an intelligent data processing system in my programming career. This type of programming is an order of magnitude more complex than just writing a simple parser or writing in business logic to some program. The project also exposed me to both Ruby and Python programming languages. I have acquired a much better understanding of the research paper writing process. Prior to this project, I was unaware of the amount of work involved in just writing a proposal and research paper. In addition, I also discovered what is needed to modify and submit a paper to a conference for publication. Now that I know what to expect, I feel more informed about what a doctorate

degree entails.

Chapter 10

CONCLUSION

This paper has shown that LSI provides an effective means for deduplicating records on both generated and real-world data. Both Febrl's existing algorithms and the LSI algorithm that were tested were not completely accurate, especially with the corporate data. However, the LSI deduplicator found a higher percentage of duplicates in both datasets than Febrl's built-in algorithms. The main drawback of using LSI is its performance and scalability. Future work will involve looking into ways to improve the scalability and to combine different deduplication approaches into a smarter system.

BIBLIOGRAPHY

[1] Data mining glossary, 2004.

[2] Eugene Agichtein and Venkatesh Ganti. Mining reference tables for automatic text segmentation. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 20–29, Seattle, WA, USA, 2004. ACM Press.

[3] Indrajit Bhattacharya and Lise Getoor. Iterative record linkage for cleaning and integration. In *Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 11–18, Paris, France, 2004. ACM Press.

[4] Christian Bhm, Bernhard Braunmller, Markus Breunig, and Hans-Peter Kriegel. High performance clustering based on the similarity join. In *Proceedings of the ninth international conference on Information knowledge management*, pages 298–305, McLean, VA USA, 2000.

[5] Mikhail Bilenko and Raymond J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 39–48, Washington, D.C., 2003. ACM Press.

[6] Paul E. Black. Dictionary of algorithms and data structures, 2006.

[7] Elaine Y. Chan, Wai Ki Ching, Michael K. Ng, and Joshua Z. Huang. An optimization algorithm for clustering using weighted dissimilarity measures. *Pattern Recognition*, 37(5):943–952, 2003.

[8] Surajit Chaudhuri, Kris Ganjam, Venkatesh Ganti, and Rajeev Motwani. Robust and efficient fuzzy match for online data cleaning. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 313–324, San Diego, California, 2003. ACM Press.

[9] Jianzhong Chen, Mary Shapcott, Sally McClean, and Kenny Adamson. Hierarchical model-based clustering of relational data with aggregates. In *Proceedings of the 2004 ACM symposium on Applied computing*, pages 620–621, Nicosia, Cyprus, 2004. ACM Press.

[10] Peter Christen. Probabilistic data generation for deduplication and data linkage. In *Sixth International Conference on Intelligent Data Engineering and Automated Learning*, Brisbane, 2005.

[11] Peter Christen and Tim Churches. Febrl - freely extensible biomedical record linkage, 2005.

[12] Genevieve Gorrell and Brandyn Webb. Generalized hebbian algorithm for latent semantic analysis. In *Proc. Interspeech*, 2005.

[13] ANU Data Mining Group. Febrl 0.3, 2005.

[14] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. Rock: A robust clustering algorithm for categorical attributes. *Information Systems*, 25(5):345–366, 2000.

[15] David Heckerman. Bayesian networks for data mining. 1(1):79–119, 1997.

[16] Mauricio A. Hernandez and Salvatore J. Stolfo. The merge/purge problem for large databases. In *Proceedings of the 1995 ACM SIGMOD international conference on Management of data*, pages 127–138, San Jose, California, United States, 1995. ACM Press.

[17] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, 1999.

[18] Vipin Kumar and Mohammed Zaki. High performance data mining (tutorial pm-3). In *Tutorial notes of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 309–425, Boston, Massachusetts, United States, 2000. ACM Press.

[19] Thomas K Landauer, Peter W. Foltz, and Darrell Laham. Introduction to latent semantic analysis. *Discourse Processes*, (25):259–284, 1998.

[20] Mong Li Lee, Tok Wang Ling, and Wai Lup Low. Intelliclean: a knowledge-based intelligent data cleaner. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 290–294, Boston, Massachusetts, United States, 2000. ACM Press.

[21] Todd A. Letsche and Michael W. Berry. Large-scale information retrieval with latent semantic indexing, 1996.

[22] Gonzalo Navarro. A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1):31–88, 2001.

[23] Witold Pedrycz, Vincenzo Loiac, and Sabrina Senatorec. P-fcm: a proximity–based fuzzy clustering. *Fuzzy Sets and Systems*, 148(1):21–41, 2004.

[24] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–7, 1980.

[25] Yu Qian and Kang Zhang. A customizable hybrid approach to data clustering. In *Proceedings of the 2003 ACM symposium on Applied computing*, pages 485–489, Melbourne, Florida, 2003. ACM Press.

[26] Tae-Wan Ryu and Christoph F. Eick. A database clustering methodology and tool. *Information Sciences*, In Press, Corrected Proof, 2004.

[27] Sunita Sarawagi and Anuradha Bhamidipaty. Interactive deduplication using active learning. In *Proc. of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Canada, 2002.

[28] Chunqiang Tang, Sandhya Dwarkadas, and Zhichen Xu. On scaling latent semantic indexing for large peer-to-peer systems. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 112–121, New York, NY, USA, 2004. ACM Press.

[29] Cheng-Fa Tsai, Chun-Wei Tsai, Han-Chang Wu, and Tzer Yang. Acodf: a novel data clustering approach for data mining in large databases. *Journal of Systems and Software*, 73(1):133, 2004.

[30] Haixun Wang, Wei Wang, Jiong Yang, and Philip S. Yu. Clustering by pattern similarity in large data sets. In *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, pages 394–405, Madison, Wisconsin, 2002.

[31] M. Zait and H. Messatfa. A comparative study of clustering methods. *Future Generation Computer Systems*, 13(2-3):149–159, 1997.

Appendix A

RAW DATASETS

A.1 California Corporate Raw Data

The following is the raw data from the California corporate records database.

```
EntityID,EntityName,MailAddress2,MailCity,MailState
630,"(THE) UNIVERSITY HEIGHTS IMPROVEMENT, ASSOCIATION, INC.",,,
6141,222 UNIVERSITY AVENUE - NORTH,P O BOX 4456,BURLINGAME,CA
6142,222 UNIVERSITY AVENUE CORPORATION,P O BOX 4456,BURLINGAME,CA
14061,"810 UNIVERSITY AVENUE, INC.",810 UNIVERSITY AVE,BERKELEY,CA
14858,"921 UNIVERSITY, INC.",921 UNIVERSITY AVE,BERKELEY,CA
35927,"ABLE UNIVERSITY PRESS, INC.",4084 N BURGE RD,STOCKTON,CA
38101,ACADEMIC CREDIT UNIVERSITY,5181 OVERLAND AVE,CULVER CITY,CA
38153,ACADEMIC RESEARCH UNIVERSITY,4860 LONG BEACH BLVD,LONG BEACH,CA
38445,ACADEMY OF INTERNATIONAL SOCIETY OF PEOPLE UNIVERSITY,2315 CYPRESS CIRCLE
DRIVE,LOMITA,CA
42920,ACHIEVEMENT UNIVERSITY,AOKI 1102-1-8-10 HIGASHI-GOTANDA
SHINAGAWA-KU,TOKYO,JAPAN
44588,ACP UNIVERSITY CORPORATION,30129 VIA RIVERA,RANCHO PALOS VERDES,CA
48350,ADAM SMITH UNIVERSITY,3463 STATE STREET SUITE 363,SANTA BARBARA,CA
48861,ADAMS UNIVERSITY,2900 BRISTOL STREET #207,COSTA MESA,CA
48978,ADAMSON UNIVERSITY EDUCATIONAL FOUNDATION USA,51 E COLORADO BLVD,PASADENA,CA
62524,"AFFILIATES OF THE UNIVERSITY OF CALIFORNIA, SANTA BARBARA",UNIVERSITY OF
CALIFORNIA 6550 HOLLISTER AVENUE,SANTA BARBARA,CA
64072,"AFTER SCHOOL UNIVERSITY, INC.",757 VIA JOSEFA,CORONA,CA
64698,AGAPE CHRISTIAN UNIVERSITY,933 S LAKE STREET,LOS ANGELES,CA
72419,AIWA INTERNATIONAL SOFTCHIROPRACTIC UNIVERSITY,800 W 1ST ST 906,LOS
ANGELES,CA
75042,"ALABAMA STATE UNIVERSITY ALUMNI ASSOCIATION, LOS ANGELES CHAPTER",P O BOX
59289,LOS ANGELES,CA
82165,ALEXENDRIA UNIVERSITY,1500 W 6TH ST,CORONA,CA
85400,ALL AMERICAN UNIVERSITY EDUCATIONAL INVESTMENT FOUNDATION,4534 BANDERA RM
D,MONTCLAIR,CA
85401,"ALL AMERICAN UNIVERSITY, INCORPORATED",354 W 6TH ST,SAN BERNARDINO,CA
90857,"ALLERGY MEDICAL CLINIC-UNIVERSITY ASSOCIATES, INC.",11645 WILSHIRE BLVD
#600,LOS ANGELES,CA
95372,ALMAR PROFESSIONAL MUSICALITY SCIENCE UNIVERSITY PERUSAL HALLS,262 W
WILLOW,POMONA,CA
97649,ALPHA RHO CHI FRATERNITY - ANDRONICUS CHAPTER AT THE UNIVERSITY OF SOUTHERN
CALIFORNIA,715 WEST 28TH ST,LOS ANGELES,CA
102030,ALUMNAE OF THE JAPANSES WOMEN STUDENT CLUB OF THE UNIVERSITY OF
CALIFORNIA,,,
```

```
102032,ALUMNI & FRIENDS OF XAVIER UNIVERSITY - ATENEO,P O BOX 7151,GLENDALE,CA
102034,ALUMNI AND FRIENDS OF FISK UNIVERSITY,300 YAMPA CIR,SACRAMENTO,CA
102037,ALUMNI ASSOCIATION INCORPORATED OF WINDSOR UNIVERSITY,4625 CRENSHAW
BLVD,LOS ANGELES,CA
102038,ALUMNI ASSOCIATION OF CALIFORNIA COAST UNIVERSITY,700 NORTH MAIN,SANTA
ANA,CA
102039,"ALUMNI ASSOCIATION OF CALIFORNIA STATE UNIVERSITY, NORTHRIDGE",18111
NORDHOFF ST RM #507,NORTHRIDGE,CA
102046,"ALUMNI ASSOCIATION OF THE SCHOOL OF MEDICINE OF LOMA LINDA UNIVERSITY,
INC.",11245 ANDERSON ST STE 200,LOMA LINDA,CA
102050,"ALUMNI ASSOCIATION, COLLEGE OF ARTS AND SCIENCES, LOMA LINDA UNIVERSITY,
INC.",4700 PIERCE STREET,RIVERSIDE,CA
102053,"ALUMNI ASSOCIATION, SCHOOL OF SOCIAL WELFARE, UNIVERSITY OF CALIFORNIA
BERKELEY",UNIVERSITY OF CALIFORNIA,BERKELEY,CA
102059,ALUMNI COUNCIL OF THE NEWMAN CLUB OF THE UNIVERSITY OF CALIFORNIA,,,
102074,ALUMNI RESEARCH FOUNDATION OF LOMA LINDA UNIVERSITY,TAX OFFICER,LOMA
LINDA,CA
107423,"AMERICA NATIONAL UNIVERSITY, INC.",3930 MAXSON RD #F,EL MONTE,CA
108652,"AMERICAN AND MEXICAN UNIVERSITY STUDENT CLUB, INCORPORATED",PO BX 2326,SAN
DIEGO,CA
108746,AMERICAN ARAB UNIVERSITY,2000 MISSION AVE,CARMICHAEL,CA
108779,"AMERICAN ARMSTRONG UNIVERSITY, INC.",1641 W MAIN STREET #222,ALHAMBRA,CA
109040,AMERICAN ASSOCIATION OF UNIVERSITY WOMEN LAGUNA BEACH FOUNDATION,915 HYDE
CT,COSTA MESA,CA
109043,AMERICAN ASSOCIATION OF UNIVERSITY WOMEN PASADENA BRANCH,160 NORTH OAKLAND
AVE,PASADENA,CA
109044,"AMERICAN ASSOCIATION OF UNIVERSITY WOMEN, CAMARILLO, CALIFORNIA BRANCH,
INC.",P O BOX 862,CAMARILLO,CA
109051,"AMERICAN ASSOCIATION OF UNIVERSITY WOMEN, SAN DIEGO BRANCH",P O BOX
241,LAFAYETTE,CA
109052,"AMERICAN ASSOCIATION OF UNIVERSITY WOMEN, SAN JOSE BRANCH",1165 MINNESOTA
AVE,SAN JOSE,CA
109053,"AMERICAN ASSOCIATION OF UNIVERSITY WOMEN, VENTURA COUNTY BRANCH",PO BOX
7507,VENTURA,CA
109054,"AMERICAN ASSOCIATION OF UNIVERSITY WOMEN-SACRAMENTO BRANCH, INC.",4300
WINDING WOODS WAY,FAIR OAKS,CA
109113,AMERICAN AUSTIN UNIVERSITY,911 WILSHIRE BLVD,LOS ANGELES,CA
109277,AMERICAN B.E.S.T UNIVERSITY,17485 DORIC ST,GRANADA HILLS,CA
110210,"AMERICAN BUSINESS UNIVERSITY, INC.",555 WEST 8TH AVE STE 107,"VANCOUVER,
BC",CANADA
110434,AMERICAN CAPITAL UNIVERSITY AND COLLEGE,8760 JARWOOD RD,BALTIMORE,MD
111324,AMERICAN COMMONWEALTH UNIVERSITY,2801 CAMINO DEL RIO SOUTH STE 201,SAN
DIEGO,CA
112040,"AMERICAN CULTURE UNIVERSITY, INC.",110 E 9TH ST #A-1075,LOS ANGELES,CA
114160,AMERICAN FLOATING UNIVERSITY,,,
114430,AMERICAN FRIENDS OF HAIFA UNIVERSITY,41 EAST 42ND STREET SUITE 828,NEW
YORK,NY
114461,AMERICAN FRIENDS OF THE JEWISH UNIVERSITY OF MOSCOW,CALIFORNIA PLAZA 29TH
FLOOR 300 SOUTH GRAND AVE,LOS ANGELES,CA
114841,"AMERICAN GLOBAL UNIVERSITY, INC.",1603 CAPITOL AVE #207,CHEYENNE,WY
```

116105,AMERICAN INDEPENDENT UNIVERSITY,300 S MENTOR #4,PASADENA,CA
116194,"AMERICAN INDIAN NATIONAL COLLEGE FUND, AND AMERICAN INDIAN UNIVERSITY CONSORTIUM, INC.",2617 EAST 14TH ST,OAKLAND,CA
116437,"AMERICAN INSTITUTE OF CHEMICAL ENGINEERS, UNIVERSITY OF CALIFORNIA SANTA BARBARA CHAPTER",AICHE STUDENT CHAPTER CHEM ENGR DEPT UC SANTA BARBARA ENGINEERING II BLDG,SANTA BARBARA,CA
116876,AMERICAN INTERNATIONAL UNIVERSITY,150 S LOS ROBLES AVE RM 680,PASADENA,CA
116877,AMERICAN INTERNATIONAL UNIVERSITY,RONALD BUMANN 2400 W VALLEY PKWY SP 99,ESCONDIDO,CA
116878,AMERICAN INTERNATIONAL UNIVERSITY,1516 SOUTH WESTERN AVENUE,LOS ANGELES,CA
118027,"AMERICAN MANAGEMENT UNIVERSITY, INC.",536 PLYMOUTH ST,SAN FRANCISCO,CA
119301,AMERICAN NATIONAL UNIVERSITY,P O BOX 951,SAN BERNARDINO,CA
119302,AMERICAN NATIONAL UNIVERSITY,10221 RIVERSIDE DR #201,N HOLLYWOOD,CA
119372,AMERICAN NETWORK UNIVERSITY,201 W LA HABRA BLVD,LA HABRA,CA
119837,AMERICAN PACIFIC INTERNATIONAL UNIVERSITY,PO BOX 5187,TORRANCE,CA
119919,AMERICAN PACIFIC UNIVERSITY,12782 WINDWARD CT,VICTORVILLE,CA
119920,AMERICAN PACIFIC UNIVERSITY,17526 VON KARMAN,IRVINE,CA
120886,AMERICAN PURLINTON UNIVERSITY,3179 W TEMPLE AVE,POMONA,CA
120945,AMERICAN R & D UNIVERSITY,17485 DORIC ST,GRANADA HILLS,CA
121908,AMERICAN SAN DIGO UNIVERSITY,5777 W CENTURY BLVD,LOS ANGELES,CA
121909,AMERICAN SAN LUIS UNIVERSITY,8635 W 3RD ST,LOS ANGELES,CA
122902,AMERICAN STATES UNIVERSITY CORPORATION,PO BX 5580,BUENA PARK,CA
124227,AMERICAN UNIVERSITY ACADEMIC ALLIANCE,12608 WEBSTER ST,OAKLAND,CA
124228,AMERICAN UNIVERSITY ALLIANCE,608 S HILL ST,LOS ANGELES,CA
124229,AMERICAN UNIVERSITY CHURCHES OF THE UNITED STATES OF AMERICA,,,
124230,AMERICAN UNIVERSITY COLLEGES ABROAD,530 BROADWAY STE 1146,SAN DIEGO,CA
124231,AMERICAN UNIVERSITY FOUNDATION IN ASIA,300 MONTGOMERY STREET #535,SAN FRANCISCO,CA
124232,AMERICAN UNIVERSITY OF ARMENIA CORPORATION,300 LAKESIDE DR 4TH FLOOR,OAKLAND,CA
124233,AMERICAN UNIVERSITY OF CALIFORNIA,3921 WILSHIRE BLVD SUITE 200,LOS ANGELES,CA
124234,AMERICAN UNIVERSITY OF LOS ANGELES,9914 WYSTONE AVE,NORTHRIDGE,CA
124236,AMERICAN UNIVERSITY OF ORIENTAL STUDIES,309 S SALTAIR AVE,LOS ANGELES,CA
124237,AMERICAN UNIVERSITY OF THE SCIENCE AND PHILOSOPHY OF LIFE,,,
124238,"AMERICAN UNIVERSITY VILLAGE, INC.",1608 WEBSTER ST,OAKLAND,CA
124828,AMERICAN WESTVALE UNIVERSITY,1072 S DE ANZA BLVD A108,SAN JOSE,CA
124998,AMERICAN WORLD WIDE UNIVERSITY,691 SOUTH IROLO STREET SUITE 504,LOS ANGELES,CA
125247,"AMERICAN-HUNGARIAN FOUNDATION FOR THE PAZMANY PETER CATHOLIC UNIVERSITY, BUDAPEST",302 PORTOLA RD,PORTOLA VALLEY,CA
125621,AMERICANTOWN UNIVERSITY,110 CAMBRIDGE ST,LONG BEACH,CA
129011,"AMPAC INTERNATIONAL UNIVERSITY, INC.",9245 E LAS TUNAS DRIVE,TEMPLE CITY,CA
134731,ANDHRA RESEARCH UNIVERSITY,,,
143189,"ANTIOCH-PITTSBURG, CALIFORNIA BRANCH OF THE AMERICAN ASSOCIATION OF UNIVERSITY WOMEN",274 PEBBLE BEACH LOOP,PITTSBURG,CA
152258,ARABIC AND ISLAMIC UNIVERSITY RESEARCH CENTER,4283 FOUNTAIN AVE 2ND FLOOR,LOS ANGELES,CA
154142,ARCATA UNIVERSITY LIONS CLUB,3447 H ST,EUREKA,CA

156937,"ARGENTINE-AMERICAN UNIVERSITY OF HEALTH SCIENCES, INC.",P O BOX 7000-R,REDONDO BEACH,CA
159466,"ARLINGTON-UNIVERSITY MEDICAL GROUP, INC.",6391 MAGNOLIA AVE,RIVERSIDE,CA
160870,ARMSTRONG UNIVERSITY ALUMNI ASSOCIATION,1608 WEBSTER ST,OAKLAND,CA
160871,ARMSTRONG UNIVERSITY DEVELOPMENT FOUNDATION,2222 HAROLD WAY,BERKELEY,CA
163863,"ART ALLIANCE, CALIFORNIA STATE UNIVERSITY, FULLERTON",PO BX 6850,FULLERTON,CA
163865,ART ALUMNI ASSOCIATION OF SAN JOSE STATE UNIVERSITY,1 WASHINGTON SQUARE ART 116,SAN JOSE,CA
169981,"ASHER STUDENT FOUNDATION - UNIVERSITY OF CALIFORNIA, LOS ANGELES",950 HILGARD AVE,LOS ANGELES,CA
170893,ASIA INTERNATIONAL UNIVERSITY PRESS,1746 FIRST ST,MANHATTAN BEACH,CA
171372,"ASIAN AMERICAN UNIVERSITY, INC.",2043 EL CAJON BLVD,SAN DIEGO,CA
171692,"ASIAN PACIFIC AMERICAN STAFF AND FACULTY ASSOCIATION, CALIFORNIA STATEUNIVERSITY, SACRAMENTO",6000 J STREET ED 210,SACRAMENTO,CA
171860,ASIAN UNIVERSITY CORPORATION,18 BLISS RIVER CT,SACRAMENTO,CA
174253,"ASSOCIATED ALUMNI, UNIVERSITY OF CALIFORNIA, IRVINE, CALIFORNIA COLLEGE OF MEDICINE",401 BERKELEY AVE STE 4650P,IRVINE,CA
176058,"ASSOCIATED STUDENTS CALIFORNIA STATE UNIVERSITY, FULLERTON, INC.",800 N STATE COLLEGE BLVD TSU ROOM 218,FULLERTON,CA
176059,"ASSOCIATED STUDENTS CALIFORNIA STATE UNIVERSITY, SAN BERNARDINO",5500 UNIVERSITY PKWY,SAN BERNARDINO,CA
176061,"ASSOCIATED STUDENTS INCORPORATED OF CALIFORNIA STATE UNIVERSITY, STANISLAUS",801 W MONTE VISTA AVE,TURLOCK,CA
176063,"ASSOCIATED STUDENTS OF CALIFORNIA STATE UNIVERSITY, CHANNEL ISLANDS, INC.",ONE UNIVERSITY DR,CAMARILLO,CA
176064,"ASSOCIATED STUDENTS OF CALIFORNIA STATE UNIVERSITY, CHICO",BELL MEMORIAL UNION ROOM 219,CHICO,CA
176065,"ASSOCIATED STUDENTS OF CALIFORNIA STATE UNIVERSITY, LOS ANGELES, INC.",5154 STATE UNIVERSITY DR,LOS ANGELES,CA
176074,ASSOCIATED STUDENTS OF SAN FRANCISCO STATE UNIVERSITY,1650 HOLLOWAY AVE M106,SAN FRANCISCO,CA
176081,"ASSOCIATED STUDENTS, CALIFORNIA STATE UNIVERSITY, BAKERSFIELD, INC.",9001 STOCKDALE HWY,BAKERSFIELD,CA
176553,ASSOCIATION FOR ARMENIAN STUDIES IN THE UNIVERSITY OF CALIFORNIA AT BERKELY,,,
176741,"ASSOCIATION FOR THE DEVELOPMENT OF THE CATHOLIC UNIVERSITY OF PORTUGAL, INC.",2121 COMMONWEALTH AVENUE,BRIGHTON,MA
176772,ASSOCIATION FOR UNIVERSITY AND COLLEGE COUNSELING CENTER DIRECTORS,UNIVERSITY COUNSELING CENTER COLORADO STATE UNIVERSITY,FORT COLLINS,CO
177305,ASSOCIATION OF PARENTS OF UNIVERSITY STUDENTS,763 SANTA ROSITA,SOLANO BEACH,CA
185944,AUGUST VOLLMER UNIVERSITY,217 NO MAIN ST STE LL7,SANTA ANA,CA
186128,AUM UNIVERSITY,,,
193676,AVENUE OF THE STARS UNIVERSITY CORPORATION,11999 SAN VICENTE BL #200,LOS ANGELES,CA
198318,AZUSA PACIFIC UNIVERSITY,PO BOX 7000,AZUSA,CA
205507,B.H. HEBREW UNIVERSITY OF JUDAIC STUDIES,1800 S ROBERTSON #238,LOS ANGELES,CA
210987,"BAKER UNIVERSITY CENTER FOR PROFESSIONAL STUDIES, INC.",P O BOX

```
25037,SHAWNEE MISSION,KS
237738,BEHR UNIVERSITY,15720 VENTURA BLVD #415,ENCINO,CA
237837,BEI JING UNIVERSITY OF TRADITIONAL CHINESE MEDICINE SOUTH CALIFORNIA
ACADEMIC ACUPUNCTURE REHABILITATION RESEARCH INSTITUTE CORPORATION,9000 SUNSET
BLVD STE 707,LOS ANGELES,CA
247363,BERKELEY INTERNATIONAL UNIVERSITY,10687 SNATA MONICA BLVD,LOS ANGELES,CA
247364,"BERKELEY INTERNATIONAL UNIVERSITY, CORPORATION",235 MONTGOMERY STREET,SAN
FRANCISCO,CA
248647,BERNARD UNIVERSITY SCHOOL OF LAW,110 NORTH THIRD STREET,SAN JOSE,CA
252418,"BETA THETA PI HALL ASSOCIATION OF STANFORD UNIVERSITY, CALIFORNIA",557
LASUEN STREET,STANFORD,CA
252419,BETA THETA PI HOUSE CORPORATION OF THE UNIVERSITY OF CALIFORNIA AT SANTA
BARBARA,416 S HOWARD,VENTURA,CA
255261,BEVERLY HILLS CONSERVATORY OF MUSIC & ARTS OF LOS ANGELES UNIVERSITY,P O
BOX 626,LONG BEACH,CA
255827,"BEVERLY HILLS UNIVERSITY APPAREL MARKETING, INC.",11500 OLYMPIC BLVD
#340,LOS ANGELES,CA
255828,"BEVERLY HILLS UNIVERSITY, INC.",3320 S BROADWAY,LOS ANGELES,CA
257332,BHAGWAT UNIVERSITY,1191 A ST,HAYWARD,CA
265546,"BIOLA UNIVERSITY, INC.",13800 BIOLA AVE,LA MIRADA,CA
269085,"BLACK ALUMNI ASSOCIATION OF THE CALIFORNIA STATE UNIVERSITY,
NORTHRIDGE",11233 BORDEN AVE,PACOIMA,CA
277069,"BM UNIVERSITY CITY, INC.",9320 FUERTE DR #105,LA MESA,CA
278075,BOARD OF COUNCILLORS OF LOMA LINDA UNIVERSITY,,,
283508,BOND UNIVERSITY USA,235 MONTGOMERY ST SUITE 1613,SAN FRANCISCO,CA
287798,"BOUCHER K-9 UNIVERSITY, INC.",11601 WILSHIRE BLVD STE 2490,LOS ANGELES,CA
303565,BROOKFIELD UNIVERSITY COMMONS INC.,12865 POINTE DEL MAR WAY STE 200,DEL
MAR,CA
315861,BURLINGAME UNIVERSITY CLUB,,,
322110,BYZANTIUM UNIVERSITY,16000 VILLA YORBA RM 123,HUNTINGTON BEACH,CA
333550,CA HUDSON UNIVERSITY,640 BERCUT DRIVE SUITE A,SACRAMENTO,CA
334729,CABRILHO CULTURAL CENTER---SAN JOSE STATE UNIVERSITY (CENTRO CULTURAL
CABRILHO---SAN JOSE STATE UNIVERSITY),1 WASHINGTON SQ,SAN JOSE,CA
334899,CABRILLO PACIFIC UNIVERSITY,5000 BIRCH STREET STE 6200,NEWPORT BEACH,CA
338380,CAL POLY UNIVERSITY CLUB,CALIFORNIA POLYTECHNIC STATE UNIVERSITY CLUB #107
PAYROLL SERVICES,SAN LUIS OBISPO,CA
```

A.2 Generated Address List Raw Data

The following is the generated address list used for the benchmark.

```
rec_id, given_name, surname, street_number, address_1, address_2, suburb, postcode,
state, date_of_birth, age, phone_number, soc_sec_id, blocking_number
rec-38-org, taylah, tiller, 22, namatjira drive, , laverton, 3976, nsw, 19900501,
32, 07 49698530, 6065899, 7
rec-97-dup-1, chloe, rizzo, 38, carmichael street, , norwod, 4069, , 19411231, ,
02 87935901, 8239567, 3
rec-64-org, lillian, warneke, 4, arabana street, nuffield village, caulfield east,
```

```
4740, nsw, 19970628, 30, 07 58724793, 4208139, 2
rec-67-org, lachlan, tolfts, 16, norman street, , kyabram, 5161, sa, 19940807, 22,
07 55159826, 5589328, 7
rec-90-org, isaiah, george, 104, woodfull loop, , andergrove, 6210, vic, 19020204,
, 04 17319806, 8199609, 2
rec-56-org, talia, mckane, 51, morgan crescent, , mount ommaney, 2171, wa,
19191220, 31, 08 05106631, 3254151, 7
rec-57-dup-2, matthew, lewan, 1, officer crescent, torbanlea, clifton gardens,
3081, nsw, 19520910, 03, , 2038840, 9
rec-57-dup-0, mattheow, lewan, 1, officer crescent, , clifton gardens, 3081, nsw,
19520910, 30, 08 22846946, 2038840, 9
rec-49-org, madeline, clarke, 53, , , paralowie, 4165, , 19730816, , 08 72943942,
1696780, 5
rec-75-org, sarsha, gao, 56, bimbiang crescent, barwite, wanneroo, 2333, qld,
19260817, 31, 07 24316937, 3094404, 6
rec-33-org, jessica, doody, 8, marcus clarke street, park hall village, white
hills, 3012, vic, 19011011, 25, 02 02784154, 3016126, 5
rec-14-dup-0, jenna, miels, 12, rischbieth crescent, , emerald, 4560, vic,
19710730, 34, 04 80481905, 4730620, 5
rec-59-org, troy, godding, 20, mainwaring rich circuit, grasmere park, flemington,
3116, , 19251006, 25, 07 38730952, 6108471, 9
rec-31-dup-0, mcfadden, joel, 3, ella close, , frankston, 2428, nsw, 19080217, ,
04 04876382, 9442009, 7
rec-16-org, victoria, crouch, 28, jerrabomberra avenue, , vaucluse, 2170, ,
19550922, 30, 04 76942418, 2211840, 6
rec-89-org, jayden, roebuck, 39, follett street, the family practice at k-mart
plaza, frankston, 2551, qld, 19270506, 32, 07 71019324, 2834727, 5
rec-99-org, bailee, nicolle, 11, edmunds place, , kellyville, 7007, vic, 19830322,
32, 04 17961741, 6365150, 7
rec-47-org, caleb, daish, 104, allchin circuit, cosy corner, geelong south, 7315,
vic, 19041018, , 03 07443742, 2330667, 1
rec-22-org, jett, rundle, 301, rosella street, , rostrevor, 6014, vic, 19650201, ,
, 9718474, 9
rec-15-org, dylan, lomman, 36, burkitt street, , lismore, 3085, , 19310204, , 07
23834008, 2432243, 0
rec-68-org, timothy, nguyen, 138, hayball place, , taree, 2483, vic, , 30, 08
71554478, 5066307, 4
rec-96-org, simone, dinh, 10, , sec 1, westmeadows, 3500, nsw, 19150210, , 07
84060229, 9225671, 3
rec-86-dup-1, isabhella, dixon, 73, balfour crescent, chest base hospital,
armidale, 2471, nsw, 19690305, , 07 53144526, 7688359, 0
rec-48-org, jessica, berry, 55, buckley circuit, , peterhead, 3143, sa, 19960501,
40, 04 71402314, 8603978, 5
rec-63-org, alexandra, guarino, 861, molloy crescent, ascotvale, ruby, 3350, qld,
19690811, 26, 07 83957444, 8045577, 3
rec-65-org, tenille, campbell, 98, wedgewood close, gillin park, morwell, 2203,
vic, 19280421, , 02 89129004, 7146372, 8
rec-76-dup-1, thomas, tuit, 5, jenyns place, horse-shoe bend, corwoa, 2440, vic,
19481031, 32, 02 76390775, 8002692, 1
rec-97-dup-3, chloe, rizoz, 38, carmichael street, warra creek, norwood, 4069, ,
```

19411231, , 02 87935901, 8239657, 3
rec-41-org, jack, sherriff, 128, david street, wymea park, glendale, 4165, vic, , 30, 02 36216989, 3466538, 0
rec-24-dup-2, jameds, crouch, 3, jessop place, , laverton, 2131, ss, 19890329, 34, 03 64490829, 4809064, 5
rec-38-dup-0, taylah, , 22, oxley street, , laverton, 3976, nsw, 19900501, 32, 07 49698530, 6065899, 7
rec-25-org, esther, , 9, cavill close, , harbord, 4680, sa, 19830303, 35, 03 90379367, 6494469, 7
rec-35-org, jacqueline, agius, 80, colebatch place, , sunnybank, 3168, , 19800901, 30, 07 49053174, 9988364, 5
rec-14-org, jenna, miels, 11, rischbieth crescent, , emerald, 4560, vic, 19710730, 34, 04 80451905, 4730620, 5
rec-35-dup-1, , agius, 80, colebatch place, , sunnybank, 3168, , 19800901, 30, 07 49053174, 9988364, 5
rec-60-org, ethan, brazzalotto, 521, monson place, , caulfield north, 4160, wa, 19651230, 26, 04 60605535, 3519484, 1
rec-81-dup-0, james, morrison, 9, albermarle place, gnorang, hectorville, 4210, sa, 19260120, , 02 52333351, 4615134, 6
rec-10-org, , green, 9, dumaresq street, , melton south, 7310, qld, , , 03 62148572, 4648140, 6
rec-80-dup-3, wilon, secomb, 38, higinbothamstreet, , richlands, 4178, wa, 19060309, 22, 04 06611182, 3083896, 8
rec-97-dup-0, chlo, rizzo, 38, , , norwood, 4069, , 19411231, , 02 87935901, 8239657, 3
rec-78-org, jordan, burford, 109, wellington street, malsah, mount annan, 4069, wa, 19440119, 34, 03 00454535, 1385262, 5
rec-7-org, isabella, woffenden, 47, osborne place, , conder, 4216, vic, 19251003, 29, 04 34762621, 3529657, 3
rec-57-dup-1, matthew, lewan, 1, officervcrescent, , clifton gardens, 3081, nsw, 19520910, 30, 08 22846946, 2038840, 9
rec-63-dup-3, alexandra, guarino, 861, molloy crescent, ascotvale, ruby, 3359, qld, 19690911, 26, 07 83957444, 8045577, 3
rec-66-dup-0, april, duvnjak, 728, atherton street, , little mountain, 2224, qld, 19003126, 21, 03 46596852, 7693518, 4
rec-88-org, ethan, mildren, 9, santry place, , kimba, 3194, vic, 19681011, 36, 07 24537502, 3804267, 9
rec-84-org, hayley, colquhoun, 7, aronson crescent, , clifton springs, 5353, nsw, 19830124, , 03 71630294, 5133046, 1
rec-46-dup-0, rhiannon, green, 122, callaghan street, mount marion, beecwrot, 3380, qld, 19750104, , , 1032798, 3
rec-18-org, cooper, matthews, 2, howse street, , carine, 2112, nsw, 19371107, , 02 10636359, 5215441, 7
rec-80-dup-2, wilson, secomb, 38, higinbotham street, , richlsdns, 4178, wa, 19060309, 22, 04 06611182, 3083896, 8
rec-31-org, joel, mcfadden, 3, ella close, , frankston, 2428, nsw, 19080217, , 04 04876382, 9442009, 7
rec-80-org, wilson, secomb, 38, higinbotham street, , richlands, 4178, wa, 19060309, 22, 04 06611182, 3083896, 8
rec-57-org, matthew, lewan, 1, officer crescent, torbanlea, clifton gardens, 3081,

```
nsw, 19520910, 30, 08 22846946, 2038840, 9
rec-5-org, caitlin, byers, 25, captain cook crescent, auroch lodge, arana hills,
3122, nsw, 19510817, 22, 03 55892476, 7243125, 5
rec-1-dup-1, sarah, gib, 31, mullan street, glenairne, newstead, 2263, nsw,
19751220, , 04 94924282, 1275236, 6
rec-63-dup-2, alexandra, guarino, 816, molloy crescent, ascotvale, ruby, 3350,
qld, , 26, 07 83957444, 8045577, 3
rec-6-dup-2, jayden, pascale, 28, goyder street, , wetherill park, 4051, nswb,
19280102, , 07 59020279, 9078297, 5
rec-4-org, jye, rapson, 1, jackie howe crescent, , kialla, 4054, vic, 19870502,
27, 07 76008800, 1870598, 5
rec-11-org, james, berry, 22, mattingley court, mt butler street, kilsyth, 3418,
nsw, 19891025, 36, 08 20918314, 7188813, 3
rec-32-dup-0, brookwlyn, van limbeek, 10, hair place, , west lempsey, 4744, vic,
19890127, 24, 08 61616652, 8687619, 0
rec-34-org, lucas, white, 22, boniwell street, whylackie, beauty point, 4570, nsw,
19170506, 35, 02 84793129, 3361126, 3
rec-63-dup-0, alexandra, guarino, 861, molloy crescent, , ruby, 3350, qld,
19690811, 26, 07 83955444, 8045577, 3
rec-93-org, jade, demarco, 45, corinna street, , berwick, 2474, qld, 19890218, 23,
03 42040582, 7438582, 9
rec-24-dup-0, james, crouch, 3, jessop place, , laverton, 2131, sa, 19985329, 34,
03 64490829, 4809064, 5
rec-29-org, chelsea, riding, 33, debenham street, the gums, acacia ridge, 4051,
nsw, 19770519, 35, , 7224750, 8
rec-1-dup-2, sarah, gibb, 31, mullangstreet, glenairne, newstead, 2263, nsw,
19751220, , , 1275236, 6
rec-85-org, james, petersen, 53, , antelle park, port macquarie, 2477, vic,
19880308, 26, 04 32506694, 2499660, 6
rec-24-org, james, crouch, 3, jessop place, , laverton, 2131, sa, 19890329, 34, 03
64490829, 4809064, 5
rec-27-org, juliana, harrington, 235, knoke avenue, , plumpton, 2477, tas,
19620726, 25, 04 19906412, 4079012, 7
rec-86-dup-2, dixon, isabella, 73, balfour crescent, chest base hospital,
armidale, 2471, nsw, 19690305, , 07 53144526, 7688359, 0
rec-80-dup-0, wilson, secomb, 38, higinbotham street, , richlands, 4167, wa,
19060309, 22, 04 06611182, 3083896, 8
rec-0-org, wilson, , 4, holroyd street, earlville shoppingtown, enoggera, 2447,
qld, 19541014, 33, 04 87178608, 3762030, 9
rec-63-dup-1, alexandra, guarkno, 861, molloy crescent, ascotvale, ruby, 3350,
qld, 19690811, 26, 07 83597444, 8045577, 3
rec-36-org, natalee, hage, 55, ashby circuit, callomonda, scotts creek, 4152, ,
19110223, 20, 02 89103888, 9922270, 7
rec-37-org, dylan, gioni, 4, musgrave street, , beenleigh, 2074, sa, 19230903, 29,
08 61931021, 9829697, 9
rec-45-org, jacqueline, white, 52, william wilkins crescent, , angaston, 5097,
vic, 19921215, 29, 03 70431787, 8372098, 8
rec-51-org, eliza, laundy, 28, moroney street, oxonia, beachmere, 3073, wa, , 28,
04 19035995, 5988086, 2
rec-97-dup-2, chlod, , 38, carmichael street, , norwood, 4069, , 19411231, , 02
```

```
87935901, 8239657, 3
rec-60-dup-0, ethan, brazzaotto, 521, chappell street, , caulfield north, 4160,
wa, 19651230, 26, 04 60605535, 3519484, 1
rec-19-org, alexandra, dudley, 9, heysen street, , broken hill, 2007, nsw,
19880125, 32, , 2064207, 2
rec-86-dup-3, isaylla, dixon, 73, balfour crescent, chest base hospital, armidale,
2471, nsw, 19690305, , 07 53144526, 7688359, 0
rec-91-org, james, oddy, 12, northcote crescent, , mount stuart, 2713, nsw,
19381209, , 02 71414811, 9952307, 5
rec-55-org, annabel, reid, 10, buntine crescent, blck 1089, coonabarabran, 2486,
qld, 19570203, 31, 07 08593763, 5646951, 2
rec-98-org, sarah, madigan, 27, catchpole street, , camden, 6066, sa, , 36, 08
69494566, 3107421, 4
rec-69-org, liam, gomulka, 14, kalgoorlie crescent, shiluvane, mount evelyn, 2571,
nsw, 19180619, 35, , 7737548, 1
rec-31-dup-2, vendula, mcfadden, 3, ella close, , frankston, 2482, nsw, 19080217,
, 04 04876382, 9442009, 7
rec-82-org, jessica, weller, 53, tenison-woods circuit, , murray bridge, 6020,
nsw, 19100522, , , 4371299, 5
rec-54-org, lynae, stuber, 37, beardsmore place, furlough house, vermont south,
2620, , 19700603, 21, 03 22285744, 3804790, 0
rec-17-org, abbey, coleman, 41, kelleway avenue, , east melbourne, 5097, vic,
19500326, 37, 03 63436504, 2240946, 7
rec-61-org, tony, , 12, perrin circuit, , oak park, 2444, wa, 19850918, , 04
99785568, 2694241, 7
rec-32-org, brooklyn, van limbeek, 10, hair place, , west kempsey, 4744, vic,
19890127, 24, 08 61616652, 8687619, 0
rec-60-dup-2, ethan, brazzalotto, 521, , , caulfield north, 4160, wa, 19651230, ,
04 60605535, 3519484, 1
rec-1-dup-0, sarah, gibb, 19, mullan street, glenairne, newstead, 2263, nsw,
19751220, , 04 94924282, 1275236, 6
rec-8-org, victoria, coleman, 60, mckenzie street, skilbister, hurstbridge, 3740,
nsw, , 22, , 5942454, 6
rec-20-org, nicholas, safai, 15, jennings street, , kempsey, 2409, wa, 19800903,
43, 03 19872208, 7411795, 1
rec-52-org, jessica, szklarz, 8, , , macgregor, 2010, nsw, 19450606, 33, 04
07039780, 8534361, 0
rec-86-dup-0, isabella, dixon, 7, balfour crescent, chest base hospital, armidale,
2471, nsw, 19690305, , 07 53144526, 7687359, 0
rec-30-org, isaiah, dixon, 12, freney place, , malvern east, 2222, nsw, 19500321,
27, 08 08432583, 4960551, 9
rec-86-org, isabella, dixon, 73, balfour crescent, chest base hospital, armidale,
2471, nsw, 19690305, , 07 53144526, 7688359, 0
rec-35-dup-0, jacqudline, agius, 80, colebatch place, , sunnybank, 3168, ,
19800901, 30, 07 46053174, 9988364, 5
rec-1-org, sarah, gibb, 31, mullan street, glenairne, newstead, 2263, nsw,
19751220, , 04 94924282, 1275236, 6
rec-3-org, hayley, pekarsky, 57, kitchener street, newland park, waterloo, 4350,
vic, 19580311, , 03 37326554, 6319579, 9
rec-76-dup-2, thomqd, tuit, 4, jenyns place, horse-shoe bend, corowa, 2440, vic,
```

19481031, 32, 02 76390775, 8002692, 1
rec-76-org, thomas, tuit, 4, jenyns place, horse-shoe bend, corowa, 2440, vic,
19481031, 32, 02 76390775, 8002692, 1
rec-21-org, toby, , 91, taronga place, business centre, gawler east, 3130, nsw,
19930503, , 04 72568836, 6126748, 5
rec-28-org, paige, vasilakis, 20, hamilton row, , berkeley vale, 5271, nsw,
19900413, 25, 03 95824112, 6560374, 9
rec-71-org, caleb, hawes, 68, , , casula, 2330, qld, 19481102, 27, 08 19515809,
3403981, 7
rec-20-dup-0, nicholas, safi, 15, jenningsmstreet, , kempsey, 2409, wa, 19800903,
43, 03 19872208, 7411795, 1
rec-81-org, james, morrison, 9, albermarle place, gnorang, hectorville, 4210, sa,
19260120, , 02 52333351, 4615524, 6
rec-44-org, lucy, szpakowska, 8, port arthur street, , frenchs forest, 2777, vic,
19290326, 33, 03 59709336, 6810608, 5
rec-94-org, lily, whillas, 24, alexandria street, , bethania, 2164, qld, 19641128,
24, 03 63453642, 5275322, 7
rec-42-org, rosie, , 1, henry street, , o'sullivan beach, 5043, nsw, 19980511, 30,
07 57747492, 1965093, 8
rec-87-org, jessica, margrie, 75, lomandra street, lashbrooke, talwood, 4575, qld,
19521110, 32, , 6236664, 5
rec-6-dup-0, jayden, pascale, 27, goyder rtreet, , wetherill park, 4051, nsw,
19280102, , 07 59020279, 9078296, 5
rec-27-dup-0, juliana, hand, 235, knoke acenue, , plumpton, 2477, tas, 19620726,
25, 04 19906412, 4079012, 7
rec-66-org, april, duvnjak, 728, atherton street, , little mountain, 2224, qld,
19001226, 21, 03 46596852, 7693518, 4
rec-46-org, rhiannon, green, 122, callaghan street, mount marion, beecroft, 3380,
qld, 19750104, , , 1032798, 3
rec-24-dup-1, james, crouch, 2, jessop place, , laverton, 2131, sa, 19890329, 34,
03 64490829, 4809064, 5
rec-77-org, daniel, , 2, yanda street, highland grove, oyster bay, 5280, sa,
19240221, 27, 02 17217060, 3593418, 7
rec-70-org, anthony, nguyen, 10, , fairlane estate, salter point, 2770, qld,
19900306, , 08 69013070, 3196548, 9
rec-97-org, chloe, rizzo, 38, carmichael street, , norwood, 4069, , 19411231, , 02
87935901, 8239657, 3
rec-39-org, julia, atsalas, 1, chalker circuit, , corrimal, 2222, tas, 19680612,
25, 07 12039247, 2043181, 4
rec-23-org, meggie, neville, 14, lind close, riverwalk one, toowoomba, 3148, nsw,
19390307, 26, 04 53557976, 4728580, 0
rec-31-dup-1, joel, mcfadden, 33, ella close, , frankston, 2428, nsw, 19080217, ,
04 04876382, 9442009, 7
rec-58-org, ruby, hoffman, 6, smith street, , alice springs, 2787, nsw, 19420217,
, 08 13952068, 8666812, 7
rec-73-org, katelyn, chandler, 28, gelane street, , ovingham, 4030, vic, 19070424,
23, 08 88509000, 9982188, 7
rec-12-org, clain, stapenell, 102, handasyde street, brileen, normanhurst, 4740,
nsw, , 22, , 7737465, 1
rec-6-org, jayden, pascale, 27, goyder street, , wetherill park, 4051, nsw,

19280102, , 07 59020279, 9078297, 5
rec-9-org, richard, copp, 16, conner close, northern tablelands tennis academy, farrer, 4820, , 19091005, , , 6947840, 7
rec-50-org, callum, pascale, 3, rolfe place, , landsdale, 6062, qld, 19380126, , 02 88562943, 4656566, 5
rec-43-org, peta, matthews, 23, harrison street, , revesby, 2150, nsw, , 35, 03 05609998, 7056403, 3
rec-26-org, zac, cochrane, 12, gruner street, parraweena, tooborac, 2756, wa, 19230626, 27, 08 40263171, 2366360, 3
rec-76-dup-0, thomas, tuit, 3, jenyns place, horse-shoe bend, corowa, 2440, vic, 19481331, 32, 02 76390775, 8002692, 1
rec-83-org, daniel, bitmead, 7, glasgow street, , toowong, 3644, vic, 19791013, 28, 03 03235752, 7468486, 0
rec-92-org, emma, stubbs, 16, maxworthy street, agarabi, samsonvale, 2250, nsw, 19661214, 29, 04 93333156, 2471400, 0
rec-72-org, nathan, dunstone, 94, crawford crescent, , carlingford, 5074, nsw, 19140417, 23, , 1608449, 5
rec-60-dup-1, ethan, brazzalotto, 521, monson place, , caulfield north, 4610, wa, 19651230, 26, 04 60605535, 3519684, 1
rec-13-org, jack, stephenson, 9, livingston avenue, , waratah west, 6058, tas, 19040108, 24, 02 73709713, 8681720, 7
rec-53-org, sophie, stancombe, 6, glenelg street, , yagoona, 2131, qld, 19870123, 28, 03 48904216, 1409948, 9
rec-95-org, chelsea, wooley, 7, owen crescent, , hackham west, 2099, vic, 19390420, 26, 02 78688185, 6698563, 0
rec-79-org, stirling, farnham, 36, archibald street, , bossley park, 3016, nsw, , 18, 08 84428029, 3685137, 8
rec-80-dup-1, wilson, secomb, 17, higinbotham street, , richlands, 4187, wa, 19060309, 22, 04 06611182, 3083896, 8
rec-6-dup-1, jayden, pascale, 27, goyderstreet, , wetheriull park, 4051, nsw, 19280102, , 07 59020279, 9078297, 5
rec-2-org, alisa, kelley, 32, duterrau crescent, berkeley village, dandenong north, 2144, nsw, 19960517, 29, 03 66396974, 9268197, 9
rec-35-dup-2, jacquelioe, agius, 80, colebatch place, , sunnybannk, 3168, , 19800901, 30, 07 49053174, 9988364, 5
rec-46-dup-1, rhiannon, green, 121, callaghan street, , beecroft, 3380, qld, 19750104, , , 1032798, 3
rec-62-org, samuel, manthorpe, 9, mcculloch street, albion grove, springwood, 5280, nsw, , 38, 04 06908287, 6340210, 3
rec-27-dup-1, juliana, harrington, 235, knoke avenue, , plumpton, 2477, tam, 19620726, 25, 04 19906412, 4079012, 7
rec-74-org, nicholas, matthews, 33, forsyth place, cleveland, albury, 4555, act, , 27, 08 58513493, 4764723, 4
rec-40-org, jenna, britten, 2, schlich street, ararat retirement village, dalby, 4670, sa, 19570104, 27, 04 98513995, 5502747, 5

Appendix B

DEDUPLICATOR OUTPUTS

B.1 Human-labeled Matching Records for California Data

The following are the EntityIDs that correspond to each other. This is the baseline benchmark for the computer-based algorithms.

```
6141 6142
14061 14858
109040 109043 109044 109051 109052 109053 109054
109277 120945
116876 116877 116878
119301 119302
119919 119920
109277 120945
176058 176059 176061 176063 176064 176065 176081
247363 247364
269085 102039

Total 22 duplicate records
```

B.2 Bigram Indexer Matching Pairs for California Data

```
Resulting record pairs:
-----------------------
  Output threshold:  0.000000
  Data set A:        CaliDataTemp
  Data set B:        CaliDataTemp

--------------------------------------------------------------------------------
Weight: 0.000000 [assigned]
Fields       | [RecID A: 68/CaliDataTemp]    | [RecID B: 69/CaliDataTemp]
EntityID     | 119919                        | 119920
EntityName   | american pacific university   | american pacific university
```

```
MailAddress2 | 12782 windward ct             | 17526 von karman
MailCity     | victorville                   | irvine
MailState    | ca                            | ca

-------------------------------------------------------------------------------
Weight: 0.000000 [assigned]
Fields       | [RecID A: 64/CaliDataTemp]    | [RecID B: 65/CaliDataTemp]
EntityID     | 119301                        | 119302
EntityName   | american national university  | american national university
MailAddress2 | p o box 951                   | 10221 riverside dr #201
MailCity     | san bernardino                | n hollywood
MailState    | ca                            | ca

-------------------------------------------------------------------------------
Weight: 0.000000
Fields       | [RecID A: 61/CaliDataTemp]    | [RecID B: 62/CaliDataTemp]
EntityID     | 116877                        | 116878
EntityName   | american international univer | american international univer
MailAddress2 | ronald bumann 2400 w valley p | 1516 south western avenue
MailCity     | escondido                     | los angeles
MailState    | ca                            | ca

-------------------------------------------------------------------------------
Weight: 0.000000 [assigned]
Fields       | [RecID A: 60/CaliDataTemp]    | [RecID B: 62/CaliDataTemp]
EntityID     | 116876                        | 116878
EntityName   | american international univer | american international univer
MailAddress2 | 150 s los robles ave rm 680   | 1516 south western avenue
MailCity     | pasadena                      | los angeles
MailState    | ca                            | ca

-------------------------------------------------------------------------------
Weight: 0.000000
Fields       | [RecID A: 60/CaliDataTemp]    | [RecID B: 61/CaliDataTemp]
EntityID     | 116876                        | 116877
EntityName   | american international univer | american international univer
MailAddress2 | 150 s los robles ave rm 680   | ronald bumann 2400 w valley p
MailCity     | pasadena                      | escondido
MailState    | ca                            | ca
```

B.3 Bigram Indexer Matching Pairs for Generated Data

```
Resulting record pairs:
```

```
-----------------------
  Output threshold:  10.000000
  Data set A:        example2tmp
  Data set B:        example2tmp

--------------------------------------------------------------------------------
Weight: 11.569103 [assigned]
Fields       | [RecID A: 99/example2tmp]     | [RecID B: 144/example2tmp]
address_hmm_ | 0.000120259775997             | 0.000120259775997
dob_day      | 01                            | 01
dob_month    | 09                            | 09
dob_year     | 1980                          | 1980
given_name   | jacqudline                    | jacquelioe
locality_nam | sunnybank                     | sunnybannk
postcode     | 3168                          | 3168
rec_id       | rec-35-dup-0                  | rec-35-dup-2
soc_sec_id   | 30                            | 30
surname      | agius                         | agius
wayfare_name | colebatch                     | colebatch
wayfare_numb | 80                            | 80
wayfare_type | place                         | place

--------------------------------------------------------------------------------
Weight: 11.569103 [assigned]
Fields       | [RecID A: 70/example2tmp]     | [RecID B: 141/example2tmp]
address_hmm_ | 0.00439198188176              | 0.00439198188176
dob_day      | 09                            | 09
dob_month    | 03                            | 03
dob_year     | 1906                          | 1906
gender_guess | male                          | male
given_name   | wilson                        | wilson
locality_nam | richlands                     | richlands
postcode     | 4167                          | 4187
rec_id       | rec-80-dup-0                  | rec-80-dup-1
soc_sec_id   | 22                            | 22
surname      | secomb                        | secomb
territory    | western_australia             | western_australia
wayfare_name | higinbotham                   | higinbotham
wayfare_numb | 38                            | 17
wayfare_type | street                        | street

--------------------------------------------------------------------------------
Weight: 11.569103
Fields       | [RecID A: 51/example2tmp]     | [RecID B: 141/example2tmp]
```

```
address_hmm_ | 0.00439198188176              | 0.00439198188176
dob_day      | 09                           | 09
dob_month    | 03                           | 03
dob_year     | 1906                         | 1906
gender_guess | male                         | male
given_name   | wilson                       | wilson
locality_nam | richlands                    | richlands
postcode     | 4178                         | 4187
rec_id       | rec-80-org                   | rec-80-dup-1
soc_sec_id   | 22                           | 22
surname      | secomb                       | secomb
territory    | western_australia            | western_australia
wayfare_name | higinbotham                  | higinbotham
wayfare_numb | 38                           | 17
wayfare_type | street                       | street

--------------------------------------------------------------------------------
Weight: 11.569103
Fields       | [RecID A: 51/example2tmp]     | [RecID B: 70/example2tmp]
address_hmm_ | 0.00439198188176              | 0.00439198188176
dob_day      | 09                           | 09
dob_month    | 03                           | 03
dob_year     | 1906                         | 1906
gender_guess | male                         | male
given_name   | wilson                       | wilson
locality_nam | richlands                    | richlands
postcode     | 4178                         | 4167
rec_id       | rec-80-org                   | rec-80-dup-0
soc_sec_id   | 22                           | 22
surname      | secomb                       | secomb
territory    | western_australia            | western_australia
wayfare_name | higinbotham                  | higinbotham
wayfare_numb | 38                           | 38
wayfare_type | street                       | street

--------------------------------------------------------------------------------
Weight: 11.569103
Fields       | [RecID A: 49/example2tmp]     | [RecID B: 141/example2tmp]
address_hmm_ | 0.00011953581888              | 0.00439198188176
dob_day      | 09                           | 09
dob_month    | 03                           | 03
dob_year     | 1906                         | 1906
gender_guess | male                         | male
given_name   | wilson                       | wilson
```

```
locality_nam | richlsdns                    | richlands
postcode     | 4178                        | 4187
rec_id       | rec-80-dup-2                | rec-80-dup-1
soc_sec_id   | 22                          | 22
surname      | secomb                      | secomb
territory    | western_australia           | western_australia
wayfare_name | higinbotham                 | higinbotham
wayfare_numb | 38                          | 17
wayfare_type | street                      | street

--------------------------------------------------------------------------------
Weight: 11.569103
Fields       | [RecID A: 49/example2tmp]     | [RecID B: 70/example2tmp]
address_hmm_ | 0.00011953581888              | 0.00439198188176
dob_day      | 09                          | 09
dob_month    | 03                          | 03
dob_year     | 1906                        | 1906
gender_guess | male                        | male
given_name   | wilson                      | wilson
locality_nam | richlsdns                   | richlands
postcode     | 4178                        | 4167
rec_id       | rec-80-dup-2                | rec-80-dup-0
soc_sec_id   | 22                          | 22
surname      | secomb                      | secomb
territory    | western_australia           | western_australia
wayfare_name | higinbotham                 | higinbotham
wayfare_numb | 38                          | 38
wayfare_type | street                      | street

--------------------------------------------------------------------------------
Weight: 11.569103 [assigned]
Fields       | [RecID A: 49/example2tmp]     | [RecID B: 51/example2tmp]
address_hmm_ | 0.00011953581888              | 0.00439198188176
dob_day      | 09                          | 09
dob_month    | 03                          | 03
dob_year     | 1906                        | 1906
gender_guess | male                        | male
given_name   | wilson                      | wilson
locality_nam | richlsdns                   | richlands
postcode     | 4178                        | 4178
rec_id       | rec-80-dup-2                | rec-80-org
soc_sec_id   | 22                          | 22
surname      | secomb                      | secomb
territory    | western_australia           | western_australia
```

```
wayfare_name | higinbotham                    | higinbotham
wayfare_numb | 38                             | 38
wayfare_type | street                         | street

-------------------------------------------------------------------------------
Weight: 11.569103
Fields       | [RecID A: 32/example2tmp]     | [RecID B: 144/example2tmp]
address_hmm_ | 0.000120259775997             | 0.000120259775997
dob_day      | 01                            | 01
dob_month    | 09                            | 09
dob_year     | 1980                          | 1980
gender_guess | female                        |
given_name   | jacqueline                    | jacquelioe
locality_nam | sunnybank                     | sunnybannk
postcode     | 3168                          | 3168
rec_id       | rec-35-org                    | rec-35-dup-2
soc_sec_id   | 30                            | 30
surname      | agius                         | agius
wayfare_name | colebatch                     | colebatch
wayfare_numb | 80                            | 80
wayfare_type | place                         | place

-------------------------------------------------------------------------------
Weight: 11.569103
Fields       | [RecID A: 32/example2tmp]     | [RecID B: 99/example2tmp]
address_hmm_ | 0.000120259775997             | 0.000120259775997
dob_day      | 01                            | 01
dob_month    | 09                            | 09
dob_year     | 1980                          | 1980
gender_guess | female                        |
given_name   | jacqueline                    | jacqudline
locality_nam | sunnybank                     | sunnybank
postcode     | 3168                          | 3168
rec_id       | rec-35-org                    | rec-35-dup-0
soc_sec_id   | 30                            | 30
surname      | agius                         | agius
wayfare_name | colebatch                     | colebatch
wayfare_numb | 80                            | 80
wayfare_type | place                         | place

-------------------------------------------------------------------------------
Weight: 11.569103 [assigned]
Fields       | [RecID A: 24/example2tmp]     | [RecID B: 61/example2tmp]
address_hmm_ | 2.06441163958e-007            | 0.00011953581888
```

```
dob_day      | 11                          | 11
dob_month    | 08                          | 08
dob_year     | 1969                        | 1969
gender_guess | female                      | female
given_name   | alexandra                   | alexandra
locality_nam | ascotvale ruby              | ruby
postcode     | 3350                        | 3350
rec_id       | rec-63-org                  | rec-63-dup-0
soc_sec_id   | 26                          | 26
surname      | guarino                     | guarino
territory    | queensland                  | queensland
wayfare_name | molloy                      | molloy
wayfare_numb | 861                         | 861
wayfare_type | crescent                    | crescent

--------------------------------------------------------------------------------
Weight: 11.569103 [assigned]
Fields       | [RecID A: 22/example2tmp]    | [RecID B: 98/example2tmp]
address_hmm_ | 3.89400108333e-006           | 3.89400108333e-006
dob_day      | 05                          | 05
dob_month    | 03                          | 03
dob_year     | 1969                        | 1969
gender_guess |                             | female
given_name   | isabhella                   | isabel
institution_ | chest base                  | chest base
institution_ | hospital                    | hospital
locality_nam | armidale                    | armidale
postcode     | 2471                        | 2471
rec_id       | rec-86-dup-1                | rec-86-org
surname      | dickson                     | dickson
territory    | new_south_wales             | new_south_wales
wayfare_name | balfour                     | balfour
wayfare_numb | 73                          | 73
wayfare_type | crescent                    | crescent

--------------------------------------------------------------------------------
Weight: 11.569103
Fields       | [RecID A: 7/example2tmp]     | [RecID B: 52/example2tmp]
address_hmm_ | 0.00439198188176            | 6.42884838319e-007
dob_day      | 10                          | 10
dob_month    | 09                          | 09
dob_year     | 1952                        | 1952
gender_guess |                             | male
given_name   | mattheow                    | matthew
```

```
locality_nam | clifton_gardens              | torbanlea clifton_gardens
postcode     | 3081                        | 3081
rec_id       | rec-57-dup-0                | rec-57-org
soc_sec_id   | 30                          | 30
surname      | lewan                       | lewan
territory    | new_south_wales             | new_south_wales
wayfare_name | officer                     | officer
wayfare_numb | 1                           | 1
wayfare_type | crescent                    | crescent

------------------------------------------------------------------------------
Weight: 11.569103 [assigned]
Fields       | [RecID A: 7/example2tmp]     | [RecID B: 42/example2tmp]
address_hmm_ | 0.00439198188176             | 6.73473358668e-005
dob_day      | 10                          | 10
dob_month    | 09                          | 09
dob_year     | 1952                        | 1952
gender_guess |                             | male
given_name   | mattheow                    | matthew
locality_nam | clifton_gardens             | clifton_gardens
postcode     | 3081                        | 3081
rec_id       | rec-57-dup-0                | rec-57-dup-1
soc_sec_id   | 30                          | 30
surname      | lewan                       | lewan
territory    | new_south_wales             | new_south_wales
wayfare_name | officer                     | officervcrescent
wayfare_numb | 1                           | 1
wayfare_type | crescent                    |

------------------------------------------------------------------------------
Weight: 11.569103 [assigned]
Fields       | [RecID A: 1/example2tmp]     | [RecID B: 120/example2tmp]
address_hmm_ | 0.000120259775997            | 0.00441858149492
dob_day      | 31                          | 31
dob_month    | 12                          | 12
dob_year     | 1941                        | 1941
gender_guess | female                      | female
given_name   | chloe                       | chloe
locality_nam | norwod                      | norwood
postcode     | 4069                        | 4069
rec_id       | rec-97-dup-1                | rec-97-org
surname      | rizzo                       | rizzo
wayfare_name | carmichael                  | carmichael
wayfare_numb | 38                          | 38
```

```
wayfare_type | street                         | street

--------------------------------------------------------------------------------
Weight: 11.567734 [assigned]
Fields       | [RecID A: 94/example2tmp]     | [RecID B: 107/example2tmp]
address_hmm_ | 0.197852210408                | 0.00303389668379
dob_day      | 03                            | 03
dob_month    | 09                            | 09
dob_year     | 1980                          | 1980
gender_guess | male                          | male
given_name   | nicholas                      | nicholas
locality_nam | kempsey                       | kempsey
postcode     | 2409                          | 2409
rec_id       | rec-20-org                    | rec-20-dup-0
soc_sec_id   | 43                            | 43
surname      | safai                         | safi
territory    | western_australia             | western_australia
wayfare_name | jennings                      | jenningsmstreet
wayfare_numb | 15                            | 15
wayfare_type | street                        |

--------------------------------------------------------------------------------
Weight: 11.567734 [assigned]
Fields       | [RecID A: 59/example2tmp]     | [RecID B: 90/example2tmp]
address_hmm_ | 4.58073698972e-006            | 0.000106716314383
dob_day      | 27                            | 27
dob_month    | 01                            | 01
dob_year     | 1989                          | 1989
gender_guess |                               | female
given_name   | brookwlyn van                 | brooklyn van
locality_nam | lempsey                       | kempsey
locality_qua | west                          | west
postcode     | 4744                          | 4744
rec_id       | rec-32-dup-0                  | rec-32-org
soc_sec_id   | 24                            | 24
surname      | limbeek                       | limbeek
territory    | victoria                      | victoria
wayfare_name | hair                          | hair
wayfare_numb | 10                            | 10
wayfare_type | place                         | place

--------------------------------------------------------------------------------
Weight: 11.567734 [assigned]
Fields       | [RecID A: 54/example2tmp]     | [RecID B: 100/example2tmp]
```

```
address_hmm_ | 2.89609997772e-005           | 2.89609997772e-005
dob_day      | 20                          | 20
dob_month    | 12                          | 12
dob_year     | 1975                        | 1975
gender_guess | female                      | female
given_name   | sarah                       | sarah
locality_nam | glenairne newstead          | glenairne newstead
postcode     | 2263                        | 2263
rec_id       | rec-1-dup-1                 | rec-1-org
surname      | gib                         | gibb
territory    | new_south_wales             | new_south_wales
wayfare_name | mullan                      | mullan
wayfare_numb | 31                          | 31
wayfare_type | street                      | street

--------------------------------------------------------------------------------
Weight: 11.567734 [assigned]
Fields       | [RecID A: 35/example2tmp]    | [RecID B: 136/example2tmp]
address_hmm_ | 8.21495319393e-007           | 8.21495319393e-007
dob_day      | 30                          | 30
dob_month    | 12                          | 12
dob_year     | 1965                        | 1965
gender_guess | male                        | male
given_name   | ethan                       | ethan
locality_nam | caulfield                   | caulfield
locality_qua | north                       | north
postcode     | 4160                        | 4610
rec_id       | rec-60-org                  | rec-60-dup-1
soc_sec_id   | 26                          | 26
surname      | brazzalotto                 | brazzalotto
territory    | western_australia           | western_australia
wayfare_name | monson                      | monson
wayfare_numb | 521                         | 521
wayfare_type | place                       | place

--------------------------------------------------------------------------------
Weight: 11.426492 [assigned]
Fields       | [RecID A: 116/example2tmp]   | [RecID B: 145/example2tmp]
address_hmm_ | 2.74468561915e-008           | 0.00134789165914
dob_day      | 04                          | 04
dob_month    | 01                          | 01
dob_year     | 1975                        | 1975
gender_guess | female                      | female
given_name   | rhiannon                    | rhiannon
```

```
locality_nam | marion beecroft               | beecroft
locality_qua | mount                        |
postcode     | 3380                         | 3380
rec_id       | rec-46-org                   | rec-46-dup-1
surname      | green                        | green
territory    | queensland                   | queensland
wayfare_name | callaghan                    | callaghan
wayfare_numb | 122                          | 121
wayfare_type | street                       | street

--------------------------------------------------------------------------------
Weight: 11.426492
Fields       | [RecID A: 47/example2tmp]     | [RecID B: 145/example2tmp]
address_hmm_ | 8.81364841948e-009            | 0.00134789165914
dob_day      | 04                           | 04
dob_month    | 01                           | 01
dob_year     | 1975                         | 1975
gender_guess | female                       | female
given_name   | rhiannon                     | rhiannon
locality_nam | marion beecwrot               | beecroft
locality_qua | mount                        |
postcode     | 3380                         | 3380
rec_id       | rec-46-dup-0                 | rec-46-dup-1
surname      | green                        | green
territory    | queensland                   | queensland
wayfare_name | callaghan                    | callaghan
wayfare_numb | 122                          | 121
wayfare_type | street                       | street

--------------------------------------------------------------------------------
Weight: 11.426492
Fields       | [RecID A: 47/example2tmp]     | [RecID B: 116/example2tmp]
address_hmm_ | 8.81364841948e-009            | 2.74468561915e-008
dob_day      | 04                           | 04
dob_month    | 01                           | 01
dob_year     | 1975                         | 1975
gender_guess | female                       | female
given_name   | rhiannon                     | rhiannon
locality_nam | marion beecwrot               | marion beecroft
locality_qua | mount                        | mount
postcode     | 3380                         | 3380
rec_id       | rec-46-dup-0                 | rec-46-org
surname      | green                        | green
territory    | queensland                   | queensland
```

```
wayfare_name | callaghan                    | callaghan
wayfare_numb | 122                          | 122
wayfare_type | street                       | street

--------------------------------------------------------------------------------
Weight: 11.372753
Fields       | [RecID A: 43/example2tmp]    | [RecID B: 61/example2tmp]
address_hmm_ | 2.06441163958e-007           | 0.00011953581888
dob_day      | 11                           | 11
dob_month    | 09                           | 08
dob_year     | 1969                         | 1969
gender_guess | female                       | female
given_name   | alexandra                    | alexandra
locality_nam | ascotvale ruby               | ruby
postcode     | 3359                         | 3350
rec_id       | rec-63-dup-3                 | rec-63-dup-0
soc_sec_id   | 26                           | 26
surname      | guarino                      | guarino
territory    | queensland                   | queensland
wayfare_name | molloy                       | molloy
wayfare_numb | 861                          | 861
wayfare_type | crescent                     | crescent

--------------------------------------------------------------------------------
Weight: 11.372753
Fields       | [RecID A: 24/example2tmp]    | [RecID B: 43/example2tmp]
address_hmm_ | 2.06441163958e-007           | 2.06441163958e-007
dob_day      | 11                           | 11
dob_month    | 08                           | 09
dob_year     | 1969                         | 1969
gender_guess | female                       | female
given_name   | alexandra                    | alexandra
locality_nam | ascotvale ruby               | ascotvale ruby
postcode     | 3350                         | 3359
rec_id       | rec-63-org                   | rec-63-dup-3
soc_sec_id   | 26                           | 26
surname      | guarino                      | guarino
territory    | queensland                   | queensland
wayfare_name | molloy                       | molloy
wayfare_numb | 861                          | 861
wayfare_type | crescent                     | crescent
```

B.4 LSI Matching Records for Generated Data

The following shows the output for the LSI algorithm and the matching sets it created. The columns in the groups are separated by semicolons. The first column is the row number, the second is the data it used for comparison, and the third is the cosine distance between the item and the first item in the cluster.

```
0 ; taylah tiller 22 namatjira drive laverton 3976 nsw
30 ; taylah 22 oxley street laverton 3976 nsw ; 0.617213399848

1 ; chloe rizzo 38 carmichael street norwod 4069
120 ; chloe rizzo 38 carmichael street norwood 4069 ; 0.833333333333
27 ; chloe rizoz 38 carmichael street warra creek norwood 4069 ;
0.57735026919

52 ; matthew lewan 1 officer crescent torbanlea clifton gardens 3081 nsw
7 ; mattheow lewan 1 officer crescent clifton gardens 3081 nsw ;
0.824957911384
42 ; matthew lewan 1 officervcrescent clifton gardens 3081 nsw ;
0.755928946018

33 ; jenna miels 11 rischbieth crescent emerald 4560 vic
11 ; jenna miels 12 rischbieth crescent emerald 4560 vic ; 1.0

50 ; joel mcfadden 3 ella close frankston 2428 nsw
13 ; mcfadden joel 3 ella close frankston 2428 nsw ; 1.0
123 ; joel mcfadden 33 ella close frankston 2428 nsw ; 1.0
85 ; vendula mcfadden 3 ella close frankston 2482 nsw ; 0.714285714286

22 ; isabhella dixon 73 balfour crescent chest base hospital armidale 2471
nsw
96 ; isabella dixon 7 balfour crescent chest base hospital armidale 2471
nsw ; 0.9
69 ; dixon isabella 73 balfour crescent chest base hospital armidale 2471
nsw ; 0.9
98 ; isabella dixon 73 balfour crescent chest base hospital armidale 2471
nsw ; 0.9
80 ; isaylla dixon 73 balfour crescent chest base hospital armidale 2471
nsw ; 0.9

24 ; alexandra guarino 861 molloy crescent ascotvale ruby 3350 qld
61 ; alexandra guarino 861 molloy crescent ruby 3350 qld ; 0.942809041582
55 ; alexandra guarino 816 molloy crescent ascotvale ruby 3350 qld ;
```

```
0.888888888889
43 ; alexandra guarino 861 molloy crescent ascotvale ruby 3359 qld ;
0.888888888889
72 ; alexandra guarkno 861 molloy crescent ascotvale ruby 3350 qld ;
0.888888888889

26 ; thomas tuit 5 jenyns place horse-shoe bend corwoa 2440 vic
132 ; thomas tuit 3 jenyns place horse-shoe bend corowa 2440 vic ;
0.888888888889
103 ; thomas tuit 4 jenyns place horse-shoe bend corowa 2440 vic ;
0.888888888889
102 ; thomqd tuit 4 jenyns place horse-shoe bend corowa 2440 vic ;
0.777777777778

67 ; james crouch 3 jessop place laverton 2131 sa
29 ; jameds crouch 3 jessop place laverton 2131 ss ; 1.0
117 ; james crouch 2 jessop place laverton 2131 sa ; 1.0

32 ; jacqueline agius 80 colebatch place sunnybank 3168
34 ; agius 80 colebatch place sunnybank 3168 ; 0.912870929175
99 ; jacqudline agius 80 colebatch place sunnybank 3168 ; 0.833333333333
144 ; jacquelioe agius 80 colebatch place sunnybannk 3168 ; 0.666666666667

35 ; ethan brazzalotto 521 monson place caulfield north 4160 wa
136 ; ethan brazzalotto 521 monson place caulfield north 4610 wa ; 0.875
91 ; ethan brazzalotto 521 caulfield north 4160 wa ; 0.866025403784
78 ; ethan brazzaotto 521 chappell street caulfield north 4160 wa ; 0.625

39 ; chlo rizzo 38 norwood 4069
120 ; chloe rizzo 38 carmichael street norwood 4069 ; 0.612372435696

47 ; rhiannon green 122 callaghan street mount marion beecwrot 3380 qld
116 ; rhiannon green 122 callaghan street mount marion beecroft 3380 qld ;
0.9
145 ; rhiannon green 121 callaghan street beecroft 3380 qld ; 0.67082039325

49 ; wilson secomb 38 higinbotham street richlsdns 4178 wa
51 ; wilson secomb 38 higinbotham street richlands 4178 wa ; 0.833333333333
141 ; wilson secomb 17 higinbotham street richlands 4187 wa ;
0.666666666667
70 ; wilson secomb 38 higinbotham street richlands 4167 wa ; 0.666666666667

54 ; sarah gib 31 mullan street glenairne newstead 2263 nsw
92 ; sarah gibb 19 mullan street glenairne newstead 2263 nsw ; 0.875
```

```
100 ; sarah gibb 31 mullan street glenairne newstead 2263 nsw ; 0.875
65 ; sarah gibb 31 mullangstreet glenairne newstead 2263 nsw ;
0.668153104781

56 ; jayden pascale 28 goyder street wetherill park 4051 nswb
127 ; jayden pascale 27 goyder street wetherill park 4051 nsw ; 0.875
113 ; jayden pascale 27 goyder rtreet wetherill park 4051 nsw ; 0.75

59 ; brookwlyn van limbeek 10 hair place west lempsey 4744 vic
90 ; brooklyn van limbeek 10 hair place west kempsey 4744 vic ;
0.777777777778

67 ; james crouch 3 jessop place laverton 2131 sa
29 ; jameds crouch 3 jessop place laverton 2131 ss ; 1.0
117 ; james crouch 2 jessop place laverton 2131 sa ; 1.0

68 ; juliana harrington 235 knoke avenue plumpton 2477 tas
147 ; juliana harrington 235 knoke avenue plumpton 2477 tam ; 0.875
114 ; juliana hand 235 knoke acenue plumpton 2477 tas ; 0.75

77 ; chlod 38 carmichael street norwood 4069
120 ; chloe rizzo 38 carmichael street norwood 4069 ; 0.73029674334
27 ; chloe rizoz 38 carmichael street warra creek norwood 4069 ;
0.632455532034

142 ; jayden pascale 27 goyderstreet wetheriull park 4051 nsw
113 ; jayden pascale 27 goyder rtreet wetherill park 4051 nsw ;
0.668153104781
127 ; jayden pascale 27 goyder street wetherill park 4051 nsw ;
0.668153104781
```

B.5 LSI Matching Records for California Data

The following shows the output for the LSI algorithm and the matching sets it created. The columns in the groups are separated by semicolons. The first column is the row number, the second is the data it used for comparison, and the third is the cosine distance between the item and the first item in the cluster.

```
1 ; 222 university avenue - north p o box 4456 burlingame ca
2 ; 222 university avenue corporation p o box 4456 burlingame ca ;
0.857142857143
```

```
3 ; 810 university avenue, inc.  810 university ave berkeley ca
4 ; 921 university, inc.  921 university ave berkeley ca ; 0.647756404767

18 ; alabama state university alumni association, los angeles chapter p
o box 59289 los angeles ca
112 ; associated students of california state university, los angeles, inc.
5154 state university dr los angeles ca ; 0.656734351438

21 ; all american university, incorporated 354 w 6th st san bernardino
ca
64 ; american national university p o box 951 san bernardino ca ;
0.571428571429

30 ; alumni association of california state university, northridge 18111
nordhoff st rm #507 northridge ca
137 ; black alumni association of the california state university,
northridge 11233 borden ave pacoima ca ; 0.61236744599

33 ; alumni association, school of social welfare, university of california
berkeley university of california berkeley ca
106 ; associated alumni, university of california, irvine, california
college of medicine 401 berkeley ave ste 4650p irvine ca ; 0.559751415835

35 ; alumni research foundation of loma linda university tax officer loma
linda ca
139 ; board of councillors of loma linda university ; 0.561654779077

42 ; american association of university women, camarillo, california
branch, inc.  p o box 862 camarillo ca
43 ; american association of university women, san diego branch p o box 241
lafayette ca ; 0.559208620996

44 ; american association of university women, san jose branch 1165
minnesota ave san jose ca
43 ; american association of university women, san diego branch p o box 241
lafayette ca ; 0.577002338724

45 ; american association of university women, ventura county branch po
box 7507 ventura ca
43 ; american association of university women, san diego branch p o box 241
lafayette ca ; 0.585202781897

47 ; american austin university 911 wilshire blvd los angeles ca
81 ; american university of california 3921 wilshire blvd suite 200 los
```

angeles ca ; 0.67082039325
72 ; american san digo university 5777 w century blvd los angeles ca ;
0.589255650989
141 ; boucher k-9 university, inc. 11601 wilshire blvd ste 2490 los
angeles ca ; 0.559016994375

48 ; american b.e.s.t university 17485 doric st granada hills ca
71 ; american r & d university 17485 doric st granada hills ca ;
0.925820099773

73 ; american san luis university 8635 w 3rd st los angeles ca
72 ; american san digo university 5777 w century blvd los angeles ca ;
0.589255650989

75 ; american university academic alliance 12608 webster st oakland ca
85 ; american university village, inc. 1608 webster st oakland ca ;
0.571428571429

82 ; american university of los angeles 9914 wystone ave northridge ca
83 ; american university of oriental studies 309 s saltair ave los angeles
ca ; 0.589255650989

97 ; armstrong university alumni association 1608 webster st oakland ca
85 ; american university village, inc. 1608 webster st oakland ca ;
0.571428571429

100 ; art alumni association of san jose state university 1 washington
square art 116 san jose ca
146 ; cabrilho cultural center---san jose state university (centro cultural
cabrilho---san jose state university) 1 washington sq san jose ca ;
0.558988247056

101 ; asher student foundation - university of california, los angeles
950 hilgard ave los angeles ca
112 ; associated students of california state university, los angeles, inc.
5154 state university dr los angeles ca ; 0.614950662799

110 ; associated students of california state university, channel islands,
inc. one university dr camarillo ca
112 ; associated students of california state university, los angeles, inc.
5154 state university dr los angeles ca ; 0.615183829039
114 ; associated students, california state university, bakersfield, inc.
9001 stockdale hwy bakersfield ca ; 0.572003283075

```
128 ; berkeley international university, corporation 235 montgomery street
san francisco ca
140 ; bond university usa 235 montgomery st suite 1613 san francisco ca ;
0.555555555556

133 ; beverly hills university apparel marketing, inc.  11500 olympic blvd
#340 los angeles ca
134 ; beverly hills university, inc.  3320 s broadway los angeles ca ;
0.612372435696
```

B.6 Sorting Indexer Matching Pairs for California Data

```
Resulting record pairs:
-----------------------
  Output threshold:  6.000000
  Data set A:        CaliDataTemp
  Data set B:        CaliDataTemp

------------------------------------------------------------------------------
Weight: 22.784554 [assigned]
Fields       | [RecID A: 1/CaliDataTemp]     | [RecID B: 2/CaliDataTemp]
EntityID     | 6141                          | 6142
EntityName   | 222 university avenue - north | 222 university avenue corpora
MailAddress2 | p o box 4456                  | p o box 4456
MailCity     | burlingame                    | burlingame
MailState    | ca                            | ca

------------------------------------------------------------------------------
Weight: 22.306027 [assigned]
Fields       | [RecID A: 48/CaliDataTemp]    | [RecID B: 71/CaliDataTemp]
EntityID     | 109277                        | 120945
EntityName   | american b.e.s.t university   | american r & d university
MailAddress2 | 17485 doric st                | 17485 doric st
MailCity     | granada hills                 | granada hills
MailState    | ca                            | ca

------------------------------------------------------------------------------
Weight: 19.022152 [assigned]
Fields       | [RecID A: 75/CaliDataTemp]    | [RecID B: 85/CaliDataTemp]
EntityID     | 124227                        | 124238
EntityName   | american university academic  | american university village,
MailAddress2 | 12608 webster st              | 1608 webster st
```

```
MailCity     | oakland                        | oakland
MailState    | ca                             | ca

--------------------------------------------------------------------------------
Weight: 16.515843 [assigned]
Fields       | [RecID A: 3/CaliDataTemp]      | [RecID B: 4/CaliDataTemp]
EntityID     | 14061                          | 14858
EntityName   | 810 university avenue, inc.    | 921 university, inc.
MailAddress2 | 810 university ave             | 921 university ave
MailCity     | berkeley                       | berkeley
MailState    | ca                             | ca

--------------------------------------------------------------------------------
Weight: 11.982026 [assigned]
Fields       | [RecID A: 72/CaliDataTemp]     | [RecID B: 73/CaliDataTemp]
EntityID     | 121908                         | 121909
EntityName   | american san digo university   | american san luis university
MailAddress2 | 5777 w century blvd            | 8635 w 3rd st
MailCity     | los angeles                    | los angeles
MailState    | ca                             | ca

--------------------------------------------------------------------------------
Weight: 9.690593
Fields       | [RecID A: 47/CaliDataTemp]     | [RecID B: 81/CaliDataTemp]
EntityID     | 109113                         | 124233
EntityName   | american austin university     | american university of califo
MailAddress2 | 911 wilshire blvd              | 3921 wilshire blvd suite 200
MailCity     | los angeles                    | los angeles
MailState    | ca                             | ca

--------------------------------------------------------------------------------
Weight: 9.372111
Fields       | [RecID A: 47/CaliDataTemp]     | [RecID B: 73/CaliDataTemp]
EntityID     | 109113                         | 121909
EntityName   | american austin university     | american san luis university
MailAddress2 | 911 wilshire blvd              | 8635 w 3rd st
MailCity     | los angeles                    | los angeles
MailState    | ca                             | ca

--------------------------------------------------------------------------------
Weight: 9.372111
Fields       | [RecID A: 47/CaliDataTemp]     | [RecID B: 72/CaliDataTemp]
EntityID     | 109113                         | 121908
EntityName   | american austin university     | american san digo university
```

```
MailAddress2 | 911 wilshire blvd             | 5777 w century blvd
MailCity     | los angeles                   | los angeles
MailState    | ca                            | ca

--------------------------------------------------------------------------------
Weight: 8.702418 [assigned]
Fields       | [RecID A: 81/CaliDataTemp]    | [RecID B: 83/CaliDataTemp]
EntityID     | 124233                        | 124236
EntityName   | american university of califo | american university of orient
MailAddress2 | 3921 wilshire blvd suite 200  | 309 s saltair ave
MailCity     | los angeles                   | los angeles
MailState    | ca                            | ca

--------------------------------------------------------------------------------
Weight: 7.883206
Fields       | [RecID A: 72/CaliDataTemp]    | [RecID B: 87/CaliDataTemp]
EntityID     | 121908                        | 124998
EntityName   | american san digo university  | american world wide universit
MailAddress2 | 5777 w century blvd           | 691 south irolo street suite
MailCity     | los angeles                   | los angeles
MailState    | ca                            | ca

--------------------------------------------------------------------------------
Weight: 7.754122
Fields       | [RecID A: 76/CaliDataTemp]    | [RecID B: 81/CaliDataTemp]
EntityID     | 124228                        | 124233
EntityName   | american university alliance  | american university of califo
MailAddress2 | 608 s hill st                 | 3921 wilshire blvd suite 200
MailCity     | los angeles                   | los angeles
MailState    | ca                            | ca

--------------------------------------------------------------------------------
Weight: 7.710936 [assigned]
Fields       | [RecID A: 57/CaliDataTemp]    | [RecID B: 60/CaliDataTemp]
EntityID     | 116105                        | 116876
EntityName   | american independent universi | american international univer
MailAddress2 | 300 s mentor #4               | 150 s los robles ave rm 680
MailCity     | pasadena                      | pasadena
MailState    | ca                            | ca

--------------------------------------------------------------------------------
Weight: 6.935722
Fields       | [RecID A: 73/CaliDataTemp]    | [RecID B: 87/CaliDataTemp]
EntityID     | 121909                        | 124998
```

```
EntityName   | american san luis university  | american world wide universit
MailAddress2 | 8635 w 3rd st                 | 691 south irolo street suite
MailCity     | los angeles                   | los angeles
MailState    | ca                            | ca

--------------------------------------------------------------------------------
Weight: 6.935722
Fields       | [RecID A: 47/CaliDataTemp]    | [RecID B: 87/CaliDataTemp]
EntityID     | 109113                        | 124998
EntityName   | american austin university    | american world wide universit
MailAddress2 | 911 wilshire blvd             | 691 south irolo street suite
MailCity     | los angeles                   | los angeles
MailState    | ca                            | ca

--------------------------------------------------------------------------------
Weight: 6.849587
Fields       | [RecID A: 47/CaliDataTemp]    | [RecID B: 62/CaliDataTemp]
EntityID     | 109113                        | 116878
EntityName   | american austin university    | american international univer
MailAddress2 | 911 wilshire blvd             | 1516 south western avenue
MailCity     | los angeles                   | los angeles
MailState    | ca                            | ca

--------------------------------------------------------------------------------
Weight: 6.514151 [assigned]
Fields       | [RecID A: 62/CaliDataTemp]    | [RecID B: 87/CaliDataTemp]
EntityID     | 116878                        | 124998
EntityName   | american international univer | american world wide universit
MailAddress2 | 1516 south western avenue     | 691 south irolo street suite
MailCity     | los angeles                   | los angeles
MailState    | ca                            | ca

--------------------------------------------------------------------------------
Weight: 6.164491
Fields       | [RecID A: 76/CaliDataTemp]    | [RecID B: 83/CaliDataTemp]
EntityID     | 124228                        | 124236
EntityName   | american university alliance  | american university of orient
MailAddress2 | 608 s hill st                 | 309 s saltair ave
MailCity     | los angeles                   | los angeles
MailState    | ca                            | ca
```

B.7 Sorting Indexer Matching Pairs for Generated Data

```
Resulting record pairs:
-----------------------
  Output threshold:  10.000000
  Data set A:        CaliDataTemp
  Data set B:        CaliDataTemp

--------------------------------------------------------------------------------
Weight: 25.956545 [assigned]
Fields       | [RecID A: 1/CaliDataTemp]     | [RecID B: 2/CaliDataTemp]
EntityID     | 6141                          | 6142
EntityName   | 222 university avenue - north | 222 university avenue corpora
MailAddress2 | p o box 4456                  | p o box 4456
MailCity     | burlingame                    | burlingame
MailState    | ca                            | ca

--------------------------------------------------------------------------------
Weight: 24.119768 [assigned]
Fields       | [RecID A: 75/CaliDataTemp]    | [RecID B: 85/CaliDataTemp]
EntityID     | 124227                        | 124238
EntityName   | american university academic  | american university village,
MailAddress2 | 12608 webster st              | 1608 webster st
MailCity     | oakland                       | oakland
MailState    | ca                            | ca

--------------------------------------------------------------------------------
Weight: 22.955229 [assigned]
Fields       | [RecID A: 48/CaliDataTemp]    | [RecID B: 71/CaliDataTemp]
EntityID     | 109277                        | 120945
EntityName   | american b.e.s.t university   | american r & d university
MailAddress2 | 17485 doric st                | 17485 doric st
MailCity     | granada hills                 | granada hills
MailState    | ca                            | ca

--------------------------------------------------------------------------------
Weight: 17.568604 [assigned]
Fields       | [RecID A: 3/CaliDataTemp]     | [RecID B: 4/CaliDataTemp]
EntityID     | 14061                         | 14858
EntityName   | 810 university avenue, inc.   | 921 university, inc.
MailAddress2 | 810 university ave            | 921 university ave
MailCity     | berkeley                      | berkeley
MailState    | ca                            | ca
```

```
--------------------------------------------------------------------------------
Weight: 15.336607
Fields       | [RecID A: 47/CaliDataTemp]    | [RecID B: 81/CaliDataTemp]
EntityID     | 109113                        | 124233
EntityName   | american austin university    | american university of califo
MailAddress2 | 911 wilshire blvd             | 3921 wilshire blvd suite 200
MailCity     | los angeles                   | los angeles
MailState    | ca                            | ca

--------------------------------------------------------------------------------
Weight: 12.378031 [assigned]
Fields       | [RecID A: 47/CaliDataTemp]    | [RecID B: 73/CaliDataTemp]
EntityID     | 109113                        | 121909
EntityName   | american austin university    | american san luis university
MailAddress2 | 911 wilshire blvd             | 8635 w 3rd st
MailCity     | los angeles                   | los angeles
MailState    | ca                            | ca

--------------------------------------------------------------------------------
Weight: 11.865449
Fields       | [RecID A: 76/CaliDataTemp]    | [RecID B: 81/CaliDataTemp]
EntityID     | 124228                        | 124233
EntityName   | american university alliance  | american university of califo
MailAddress2 | 608 s hill st                 | 3921 wilshire blvd suite 200
MailCity     | los angeles                   | los angeles
MailState    | ca                            | ca

--------------------------------------------------------------------------------
Weight: 11.727286
Fields       | [RecID A: 72/CaliDataTemp]    | [RecID B: 73/CaliDataTemp]
EntityID     | 121908                        | 121909
EntityName   | american san digo university  | american san luis university
MailAddress2 | 5777 w century blvd           | 8635 w 3rd st
MailCity     | los angeles                   | los angeles
MailState    | ca                            | ca

--------------------------------------------------------------------------------
Weight: 11.625660 [assigned]
Fields       | [RecID A: 81/CaliDataTemp]    | [RecID B: 83/CaliDataTemp]
EntityID     | 124233                        | 124236
EntityName   | american university of califo | american university of orient
MailAddress2 | 3921 wilshire blvd suite 200  | 309 s saltair ave
MailCity     | los angeles                   | los angeles
```

```
MailState    | ca                              | ca

--------------------------------------------------------------------------------
Weight: 11.278359 [assigned]
Fields       | [RecID A: 133/CaliDataTemp]     | [RecID B: 134/CaliDataTemp]
EntityID     | 255827                          | 255828
EntityName   | beverly hills university appa   | beverly hills university, inc
MailAddress2 | 11500 olympic blvd #340         | 3320 s broadway
MailCity     | los angeles                     | los angeles
MailState    | ca                              | ca

--------------------------------------------------------------------------------
Weight: 10.814595
Fields       | [RecID A: 47/CaliDataTemp]      | [RecID B: 72/CaliDataTemp]
EntityID     | 109113                          | 121908
EntityName   | american austin university      | american san digo university
MailAddress2 | 911 wilshire blvd               | 5777 w century blvd
MailCity     | los angeles                     | los angeles
MailState    | ca                              | ca

--------------------------------------------------------------------------------
Weight: 10.240638 [assigned]
Fields       | [RecID A: 72/CaliDataTemp]      | [RecID B: 87/CaliDataTemp]
EntityID     | 121908                          | 124998
EntityName   | american san digo university    | american world wide universit
MailAddress2 | 5777 w century blvd             | 691 south irolo street suite
MailCity     | los angeles                     | los angeles
MailState    | ca                              | ca

--------------------------------------------------------------------------------
Weight: 10.225572 [assigned]
Fields       | [RecID A: 62/CaliDataTemp]      | [RecID B: 76/CaliDataTemp]
EntityID     | 116878                          | 124228
EntityName   | american international univer   | american university alliance
MailAddress2 | 1516 south western avenue       | 608 s hill st
MailCity     | los angeles                     | los angeles
MailState    | ca                              | ca

--------------------------------------------------------------------------------
Weight: 10.099210
Fields       | [RecID A: 76/CaliDataTemp]      | [RecID B: 83/CaliDataTemp]
EntityID     | 124228                          | 124236
EntityName   | american university alliance    | american university of orient
MailAddress2 | 608 s hill st                   | 309 s saltair ave
```

```
MailCity     | los angeles                   | los angeles
MailState    | ca                            | ca

--------------------------------------------------------------------------------
Weight: 10.054012
Fields       | [RecID A: 62/CaliDataTemp]    | [RecID B: 81/CaliDataTemp]
EntityID     | 116878                        | 124233
EntityName   | american international univer | american university of califo
MailAddress2 | 1516 south western avenue     | 3921 wilshire blvd suite 200
MailCity     | los angeles                   | los angeles
MailState    | ca                            | ca

--------------------------------------------------------------------------------
Weight: 10.000221
Fields       | [RecID A: 47/CaliDataTemp]    | [RecID B: 76/CaliDataTemp]
EntityID     | 109113                        | 124228
EntityName   | american austin university    | american university alliance
MailAddress2 | 911 wilshire blvd             | 608 s hill st
MailCity     | los angeles                   | los angeles
MailState    | ca                            | ca
```

Appendix C

SOURCE CODE FOR THE LSI DEDUPLICATOR

This chapter shows the source code that is related to the LSI deduplicator. This does not include code from Febrl or any of the libraries used.

C.1 LSI.py

```
import wordHash
from wordList import *
from contentNode import *
from scipy import *
from array import *
import time
#import Numeric
#from Matrix import *


# This class implements a Latent Semantic Indexer, which can search,
#classify and cluster data based on underlying semantic relations. For more
#information on the algorithms used, please consult Wikipedia
#[http://en.wikipedia.org/wiki/Latent_Semantic_Indexing].

class LSI:

    # Create a fresh index.
    # If you want to call #build_index manually, use
    #      Classifier::LSI.new :auto_rebuild => false
    #
    def __init__(self,options = {}):
        self.auto_rebuild = True
        self.word_list, self.items = WordList(), {}
        self.version, self.built_at_version = 0, -1


    # Returns true if the index needs to be rebuilt.  The index needs
    # to be built after all informaton is added, but before you start
    # using it for search, classification and cluster detection.
    def needs_rebuild(self):
        return len(self.items) > 1 and self.version != self.built_at_version
```

```
    # Adds an item to the index. item is assumed to be a string, but
    # any item may be indexed so long as it responds to #to_s or if
    # you provide an optional block explaining how the indexer can
    # fetch fresh string data. This optional block is passed the item,
    # so the item may only be a reference to a URL or file name.
    #
    # For example:
    #   lsi = Classifier::LSI.new
    #   lsi.add_item "This is just plain text"
    #   lsi.add_item "/home/me/filename.txt" { |x| File.read x }
    #   ar = ActiveRecordObject.find( :all )
    #   lsi.add_item ar, *ar.categories { |x| ar.content }
    #
    def add_item( self, item, block, lsiIndex=-1, *categories ):
        if block:
            clean_word_hash = block.call(item).clean_word_hash
        else:
            clean_word_hash = wordHash.clean_word_hash(item)

        self.items[item] = ContentNode(clean_word_hash, categories, lsiIndex)
        self.version += 1

        if self.auto_rebuild:
            self.build_index()

    # A less flexible shorthand for add_item that assumes
    # you are passing in a string with no categorries. item
    # will be duck typed via to_s .
    #
    def add_quick( self, item, lsiIndex=-1 ):
        self.add_item(item, None, lsiIndex)

    # Returns the categories for a given indexed items. You are free to
    #add and remove items from this as you see fit. It does not invalide
    # an index to change its categories.
    def categories_for(self, item):
        if item in self.items:
            return self.items[item].categories
        else:
             return []

    # Removes an item from the database, if it is indexed.
    #
    def remove_item( item ):
        if item in self.items:
            self.items.remove(item)
            self.version += 1


    # Returns an array of items that are indexed.
```

```
def items(self):
  return self.items.allkeys()

# This function rebuilds the index if needs_rebuild? returns true.
# For very large document spaces, this indexing operation may take some
# time to complete, so it may be wise to place the operation in another
# thread.
#
# As a rule, indexing will be fairly swift on modern machines until
# you have well over 500 documents indexed, or have an incredibly diverse
# vocabulary for your documents.
#
# The optional parameter "cutoff" is a tuning parameter. When the index is
# built, a certain number of s-values are discarded from the system. The
# cutoff parameter tells the indexer how many of these values to keep.
# A value of 1 for cutoff means that no semantic analysis will take place,
# turning the LSI class into a simple vector search engine.
def build_index( self, cutoff=0.75 ):
    if not self.needs_rebuild():
        return

    print "Building index.."
    self.make_word_list()

    doc_list = self.items.values()
    tda = [ node.raw_vector_with(self.word_list) for node in doc_list ]

    tdm = mat(tda)
    tdm = transpose(tdm)
    ntdm = self.build_reduced_matrix(tdm, cutoff)

    print
    file = open('testmat.txt','w')
    file.write(str(ntdm))
    file.close()

    numrows,numcols = ntdm.shape

    for col in range(numrows):
        if col in doc_list:
            doc_list[col].lsi_vector = ntdm.column(col)
            doc_list[col].lsi_norm = ntdm.column(col).normalize

    self.built_at_version = self.version

# This method returns max_chunks entries, ordered by their average semantic
# rating. Essentially, the average distance of each entry from all other
# entries is calculated, the highest are returned.
#
# This can be used to build a summary service, or to provide more
```

```
    # information about your dataset's general content. For example, if you
    # were to use categorize on the results of this data, you could gather
    # information on what your dataset is generally about.
    def highest_relative_content( self, max_chunks=10 ):
        if self.needs_rebuild():
            return []

        avg_density = Dictonary()

        for x in self.items.keys():
            avg_density[x] = proximity_array_for_content(x).inject(0.0)

        return avg_density.sort().reverse

    # This function is the primitive that find_related and classify
    # build upon. It returns an array of 2-element arrays. The first element
    # of this array is a document, and the second is its "score", defining
    # how "close" it is to other indexed items.
    #
    # These values are somewhat arbitrary, h aving to do with the vector space
    # created by your content, so the magnitude is interpretable but not always
    # meaningful between indexes.
    #
    # The parameter doc is the content to compare. If that content is not
    # indexed, you can pass an optional block to define how to create the
    # text data. See add_item for examples of how this works.
    def proximity_array_for_content( self, doc, block ):
        if self.needs_rebuild():
            return []

        content_node = self.node_for_content( doc, block )

        keys = self.items.keys()
        result = []

        for item in keys:
            tmp1 = content_node.search_vector()
            tmp2 = self.items[item].search_vector()
            val = dot(tmp1, tmp2)
            #val = (content_node.search_vector() *
                   self.items[item].search_vector())[0]
            result.append([item, val])

        result.sort(self.result_compare)
        result.reverse()
        return result

    # Similar to proximity_array_for_content, this function takes similar
    # arguments and returns a similar array. However, it uses the normalized
    # calculated vectors instead of their full versions. This is useful when
```

```
# you're trying to perform operations on content that is much smaller than
# the text you're working with. search uses this primitive.
def proximity_norms_for_content( self, doc, *block ):
    if self.needs_rebuild():
        return []

    content_node = self.node_for_content( doc, block )
    result = self.items.keys()
    retval = []

    for item in result:
        start = time.time()
        val = (mat(content_node.search_norm()) *
              self.items[item].search_norm())[0]
        element = [item, val]
        retval.append(element)
        stop = time.time()
        print stop - start

    retval.sort(self.result_compare)
    retval.reverse()
    return retval

def result_compare(self,x,y):
    if float(x[1]) > float(y[1]):
        return 1
    elif float(x[1]) < float(y[1]):
        return -1
    else:
        return 0


# This function allows for text-based search of your index. Unlike other
# functions like find_related and classify, search only takes short
# strings. It will also ignore factors like repeated words. It is best for
# short, google-like search terms.
# A search will first priortize lexical relationships, then semantic ones.
#
# While this may seem backwards compared to the other functions that LSI
# supports, it is actually the same algorithm, just applied on a smaller
# document.
def search( self, string, max_nearest=3 ):
    if self.needs_rebuild():
        return []

    carry = self.proximity_norms_for_content( string )
    #print carry
    result = [x[0] for x in carry]

    return result[0:max_nearest-1]
```

```
# This function takes content and finds other documents
# that are semantically "close", returning an array of documents sorted
# from most to least relavant.
# max_nearest specifies the number of documents to return. A value of
# 0 means that it returns all the indexed documents, sorted by relavence.
#
# This is particularly useful for identifing clusters in your
# document space. For example you may want to identify several "What's
# Related" items for weblog articles, or find paragraphs that relate to
# each other in an essay.
def find_related( self, doc, max_nearest=3, *block ):
    carry = self.proximity_array_for_content( doc, block )
    carry = [ pair for pair in carry if pair[0] == doc ]
    result = [x[0] for x in carry]

    return result[0:max_nearest-1]

# This function uses a voting system to categorize documents, based on
# the categories of other documents. It uses the same logic as the
# find_related function to find related documents, then returns the
# most obvious category from this list.
#
# cutoff signifies the number of documents to consider when clasifying
# text. A cutoff of 1 means that every document in the index votes on
# what category the document is in. This may not always make sense.
#
def classify( self, doc, cutoff=0.30, *block ):
    icutoff = round(self.items.size * cutoff)
    carry = self.proximity_array_for_content( doc, block )
    carry = carry[0..icutoff-1]
    votes = {}

    for pair in carry:
        categories = self.items[pair[0]].categories
        for category in categories:
            #votes[category] ||= 0.0
            votes[category] += pair[1]

    ranking = votes.keys.sort()
    return ranking[-1]

# Prototype, only works on indexed documents.
# I have no clue if this is going to work, but in theory
# it's supposed to.
def highest_ranked_stems( self, doc, count=3 ):
    if not self.items[doc]:
        raise "Requested stem ranking on non-indexed content!"

    arr = node_for_content(doc).lsi_vector.to_a
```

```
    top_n = arr.sort().reverse()[0..count-1]

    top_n = top_n.copy()
    for x in top_n:
        self.word_list.word_for_index(arr.index(x))

    return top_n


def build_reduced_matrix( self, matrix, cutoff=0.75 ):
  # TODO: Check that M>=N on these dimensions! Transpose helps assure this
  u, s, v =  linalg.svd(matrix, 1)
  numrows, numcols = matrix.shape
  u = u.take(range(numcols), 1)
  # TODO: Better than 75% term, please. :\
  s.sort()
  sTemp = s

  #reverse the array
  i, j = 0, len(sTemp) - 1
  while i <= j:
      temp = sTemp[i]
      sTemp[i] = sTemp[j]
      sTemp[j] = temp
      i += 1
      j -= 1

  s_cutoff = sTemp[int(round(len(sTemp) * cutoff)) - 1]
  for ord in range(len(s)):
      if s[ord] < s_cutoff:
          s[ord] = 0.0
  # Reconstruct the term document matrix, only with reduced rank
  sIdent = []

  for i in range(len(s)):
    sIdent.append([])
    for j in range(len(s)):
        sIdent[i].append(0.0)

  for i in range(len(s)):
    sIdent[i][i] = s[i]

  #s = diagonal(s)
  v = transpose(v, None)

  return u * mat(sIdent) * v
  #return u *  matrix_base.diag(s)  * matrix_base.transpose(v, None)

def node_for_content( self, item, block):
    if item in self.items:
```

```
            return self.items[item]
        else:
            if block:
                clean_word_hash = block.call(item).clean_word_hash
            else:
                clean_word_hash = wordHash.clean_word_hash(item)

        # make the node and extract the data
        cn = ContentNode(clean_word_hash, block)

        if not self.needs_rebuild():
            # make the lsi raw and norm vectors
            cn.raw_vector_with( self.word_list )

        return cn

    def make_word_list(self):
      self.word_list = WordList()

      for node in self.items.itervalues():
          for key in node.word_hash.keys():
              self.word_list.add_word(key)


    # Similar to proximity_array_for_content, this function takes similar
    # arguments and returns a similar array. However, it uses the normalized
    # calculated vectors instead of their full versions. This is useful when
    # you're trying to perform operations on content that is much smaller than
    # the text you're working with. search uses this primitive.
    def proximity_norms_for_ref( self, doc, *block ):
        if self.needs_rebuild():
            return []

        content_node = self.node_for_content( doc, block )

        result = self.items.keys()
        retval = []

        for item in result:
            val = (mat(content_node.search_norm()) *
                  self.items[item].search_norm())[0]
            element = [item, self.items[item].lsi_index, val]
            retval.append(element)

        retval.sort(self.result_compareref)
        retval.reverse()
        return retval


    def result_compareref(self,x,y):
```

```
    xVal = float(x[2])
    yVal = float(y[2])
    if xVal > yVal:
        return 1
    elif xVal < yVal:
        return -1
    else:
        return 0


def searchrefs( self, string, max_nearest=3 ):
    if self.needs_rebuild():
        return []

    carry = self.proximity_norms_for_ref( string )
    print carry[0:5]
    result = [[x[0],[x[1]]] for x in carry]
    return result[0:max_nearest-1]


def ProximityNormsScore( self, doc, minScore, *block ):
    if self.needs_rebuild():
        return []

    content_node = self.node_for_content( doc, block )

    result = self.items.keys()
    retval = []

    docSearchNorm = content_node.search_norm()
    for item in result:
        itemSearchNorm = self.items[item].search_norm()
        val = vdot(docSearchNorm, itemSearchNorm)

        if val >= minScore:
            element = [item, self.items[item].lsi_index, val]
            retval.append(element)

    retval.sort(self.result_compareref)
    retval.reverse()
    return retval

def SearchRefsScore( self, string, minScore=.5 ):
    if self.needs_rebuild():
        return []

    carry = self.ProximityNormsScore( string, minScore )
    return carry
```

C.2 ContentNode.py

```
import array
import vector
import math

class ContentNode:
    def __init__(self, word_hash, categories, lsiIndex=-1 ):
        self.categories = categories
        self.word_hash = word_hash
        self.lsi_index = lsiIndex

    def search_vector(self):
        return self.raw_vector

    def search_norm(self):
        return self.raw_norm

    def raw_vector_with(self, word_list ):
        vec = [0] * word_list.size()

        for word in self.word_hash.iterkeys():
            if not word_list[word] == None:
                vec[word_list[word]] = self.word_hash[word]

        total_words = 0.0
        for item in vec:
            total_words += item

        #Perform first-order association transform if this vector has more
        #than one word in it.

        if total_words > 1.0:
            weighted_total = 0.0
            for term in vec:
                if ( term > 0 ):
                    weighted_total += (( term / total_words ) *
                     math.log( term / total_words ))


        vec = [math.log(val + 1.0) / -weighted_total for val in vec]

        self.raw_norm = vector.normalize(vec)
        self.raw_vector = vec
        return vec
```

C.3 Vector.py

```
import math
```

```
def magnitude(vec):
    sumsqs = 0.0
    for i in range(len(vec)):
        sumsqs += vec[i] ** 2.0

    return math.sqrt(sumsqs)

def normalize(vec):
    nv = []
    mag = magnitude(vec)
    for i in range(len(vec)):
        nv.append(vec[i] / mag)

    return nv
```

C.4 WordHash.py

```
import re
from stemmer import *

# These are extensions to the String class to provide convenience
# methods for the Classifier package.

  # Removes common punctuation symbols, returning a new string.
  # E.g.,
  #   "Hello (greeting's), with {braces} < >...?".without_punctuation
  #   => "Hello  greetings   with  braces         "
def without_punctuation(st):
    badpunc = ',?.!;:"@#$%^&*()_=+[]{}\|<>/`~'
    table = string.maketrans(badpunc, ' ' * len(badpunc))
    return st.translate(badpunc, "'\-")

  # Return a Hash of strings => ints. Each word in the string is stemmed,
  # interned, and indexes to its frequency in the document.
def word_hash(st):
    r1 = re.compile(r"[^\w\s]")
    r2 = re.compile(r"[\w]")

    return word_hash_for_words(r1.sub("",st).split() +
           r2.sub(" ", st).split())


# Return a word hash without extra punctuation or short symbols,
# just stemmed words
def clean_word_hash(st):
    r1 = re.compile(r"[^\w\s]")
    return word_hash_for_words(r1.sub("",st).split())


def word_hash_for_words(words):
```

```
    d = {}
    r1 = re.compile(r"[\w]+")
    r2 = re.compile(r"[^\w]")

    stemmer = PorterStemmer()

    for word in words:
        if r1.match(word):
            word = word.lower()

        key = stemmer.stem(word, 0, len(word)-1)
        if r2.match(word) or word not in CORPUS_SKIP_WORDS and
          len(word) > 2:
            if not d.has_key(key):
                d[key] = 0
            d[key] += 1

    return d

CORPUS_SKIP_WORDS = [
  "a",
  "again",
  "all",
  "along",
  "are",
  "also",
  "an",
  "and",
  "as",
  "at",
  "but",
  "by",
  "came",
  "can",
  "cant",
  "couldnt",
  "did",
  "didn",
  "didnt",
  "do",
  "doesnt",
  "dont",
  "ever",
  "first",
  "from",
  "have",
  "her",
  "here",
  "him",
  "how",
```

```
"i",
"if",
"in",
"into",
"is",
"isnt",
"it",
"itll",
"just",
"last",
"least",
"like",
"most",
"my",
"new",
"no",
"not",
"now",
"of",
"on",
"or",
"should",
"sinc",
"so",
"some",
"th",
"than",
"this",
"that",
"the",
"their",
"then",
"those",
"to",
"told",
"too",
"true",
"try",
"until",
"url",
"us",
"were",
"when",
"whether",
"while",
"with",
"within",
"yes",
"you",
"youll",
```

```
 ]
```

C.5 WordList.py

```
 # This class keeps a word => index mapping. It is used to map stemmed words
 # to dimensions of a vector.

class WordList:
    def __init__(self):
        self.location_table = {} #hash table

    # Adds a word (if it is new) and assigns it a unique dimension.
    def add_word(self, word):
        if not self.location_table.has_key(word):
            self.location_table[word] = len(self.location_table)

    # Returns the dimension of the word or nil if the word is not in
 the space.
    def __getitem__(self, lookup):
      return self.location_table[lookup]

    def word_for_index(self, ind):
        for [key,val] in self.location_table:
            if val == ind:
                return key

    # Returns the number of words mapped.
    def size(self):
      return len(self.location_table)
```

C.6 LatentProj.py

```
import time
from dataset import *
from lsi import *

startTime = time.time()
lsiIndexer = LSI()
lsiIndexer.auto_rebuild = False

calidata = DataSetCSV(name = 'datasetTiny',
             description = 'Addresses',
             access_mode = 'read',
            header_lines = 1,
               file_name = 'dsgen\datasetTiny.csv',
                  fields = {'rec_id':0,
                            'given_name':1,
                            'surname':2,
                            'street_number':3,
                            'address_1':4,
```

```
                            'address_2':5,
                            'suburb':6,
                            'postcode':7,
                            'state':8,
                            'date_of_birth':9,
                            'age':10,
                            'phone_number':11,
                            'soc_sec_id':12,
                            'blocking_number':13},
          fields_default = '',
            strip_fields = True,
          missing_values = ['','missing'])

parseCols = ['given_name','surname','street_number','address_1',
              'address_2', 'suburb', 'postcode', 'state']

match_cutoff = .55

print "Reading records..."
records = calidata.read_records(0, 149)

print "adding items to index..."
for record in records:
    combined = ""
    for col in parseCols:
        if col in record:
            combined += " " + record[col]
    print combined
    lsiIndexer.add_quick(combined, record["_rec_num_"])
    #store the combined in the dataset
    record["combined"] = combined

print "building index..."
lsiIndexer.build_index()

print 'finding sematic duplicates...'

duplicateSets = []
markedMatches = []

for record in records:
    if record["_rec_num_"] not in markedMatches:
        similarItems = lsiIndexer.SearchRefsScore(record["combined"],
 match_cutoff)
        if len(similarItems) > 1:
            duplicateSets.append(similarItems)
            for item in similarItems:
                markedMatches.append(item[1])

#print "found " & len(duplicateSets) & " sets"
```

```
for set in duplicateSets:
    print str(set[0][1]) + ' ~ ' + set[0][0]
    for i in range(1,len(set)):
        print str(set[i][1]) + ' ~ ' + set[i][0] + ' ~ ' + str(set[i][2])
    print ''


stopTime = time.time()
print stopTime - startTime
```