12-2010

A Visual Approach to Automated Text Mining and Knowledge Discovery

Andrey A. Puretskiy
apuretsk@gmail.com

Recommended Citation

Puretskiy, Andrey A., "A Visual Approach to Automated Text Mining and Knowledge Discovery. " PhD diss., University of Tennessee, 2010.
https://trace.tennessee.edu/utk_graddiss/907

To the Graduate Council:

I am submitting herewith a dissertation written by Andrey A. Puretskiy entitled "A Visual Approach to Automated Text Mining and Knowledge Discovery." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Computer Science.

Michael Berry, Major Professor

We have read this dissertation and recommend its acceptance:

Jian Huang, Bradley Vander Zanden, Charles Collins

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a dissertation written by Andrey A. Puretskiy entitled “A Visual Approach to Automated Text Mining and Knowledge Discovery.” I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy with a major in Computer Science.

Michael W. Berry, Major Professor

We have read this dissertation
and recommend its acceptance:

Jian Huang

Brad Vander Zanden

Charles Collins

Accepted for the Council:

Carolyn R. Hodges
Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

A Visual Approach to Automated Text Mining and Knowledge Discovery

A Dissertation Presented for the
Doctor of Philosophy Degree
University of Tennessee, Knoxville

Andrey A. Puretskiy
December 2010

ACKNOWLEDGEMENTS

The author would like to thank his advisor, Dr. Michael W. Berry for his invaluable mentoring and advice throughout the author’s time in graduate school. Dr. Berry has been the model professor and advisor, and this dissertation would not have been possible without his expertise and support. The author thanks the committee members, Dr. Jian Huang, Dr. Bradley Vander Zanden, and Dr. Charles Collins, for providing valuable feedback that greatly improved the quality of this work. The author gratefully acknowledges the Electrical Engineering and Computer Science Department of the University of Tennessee, Knoxville, for providing the author with three years of Teaching Assistant funding. The author thanks his friend and occasional employer, Dr. Richard Sparks, whose advice and occasional consulting paychecks had greatly contributed to the author’s quality of life during graduate school. The author thanks his parents, Alexander and Svetlana Puretskiy, for supporting the author in every endeavor over the course of many years.

ABSTRACT

The focus of this dissertation has been on improving the non-negative tensor factorization technique of text mining. The improvements have been made in both pre-processing and post-processing stages, with the goal of making the non-negative tensor factorization algorithm accessible to the casual user. The improved implementation allows the user to construct and modify the contents of the tensor, experiment with relative term weights and trust measures, and experiment with the total number of algorithm output features. Non-negative tensor factorization output feature production is closely integrated with a visual post-processing tool, FutureLens, that allows the user to perform in depth analysis and has a great potential for discovery of interesting and novel patterns within a large collection of textual data. This dissertation necessitated a number of significant modifications and additions to FutureLens in order to facilitate its integration into the analysis environment.

Table of Contents

List of Figures

Chapter 1

Introduction to Visual Analytics and Nonnegative Tensor Factorization

One of the many direct consequences of the significant and ever-increasing information digitalization trend of recent decades has been a newfound ability to gather, organize, store, and analyze vast repositories of knowledge. As computing and digitalization increasingly permeate virtually every aspect of society, researchers and analysts in a wide variety of fields sometimes find themselves virtually overwhelmed with enormous quantities of information. The fields of data mining and visual analytics developed alongside the ever-increasing information stores in order to provide analytical knowledge discovery capabilities in wide-ranging fields such as biology, social science, law, and business.

1.1 Alternative Approaches to Visual Analytics

Visual analytics is a highly interdisciplinary field that is defined as the science of analytical reasoning supported by highly interactive visual interfaces. Visual analytics tools are designed to extract insight from large datasets. The goal of visual analytics tools is to facilitate extraction and verification of associations, interconnections, and relationships contained within the data. One significant challenge in this new and developing field has been the seamless integration of highly advanced mathematical text analysis techniques with visual software tools that would enable users to better understand and utilize the information extracted by these techniques [1]. The sections below describe some of the visual analytics methodologies alternative to the combined non-negative tensor factorization/FutureLens approach presented in this dissertation.

1.1.1 Visual Summarization with Tag Clouds

Tag clouds and various related concepts are an extremely popular and user-friendly way to summarize a large amount of textual data. One of the most important and advantageous properties of a tag cloud is its ability to quickly and easily generate a visual ranking of the relative significance of the terms within the summary. As shown in Figure 1, an additional advantage of the tag cloud technique is that the resulting product may serve as a useful visual element in applications such as web and graphic design. The fact that many tag cloud generators allow the user the ability to greatly customize the visual appearance of the tag cloud, altering elements such as font type, font color, and text orientation within the figure, greatly enhances the tag cloud's usefulness in design-oriented applications.

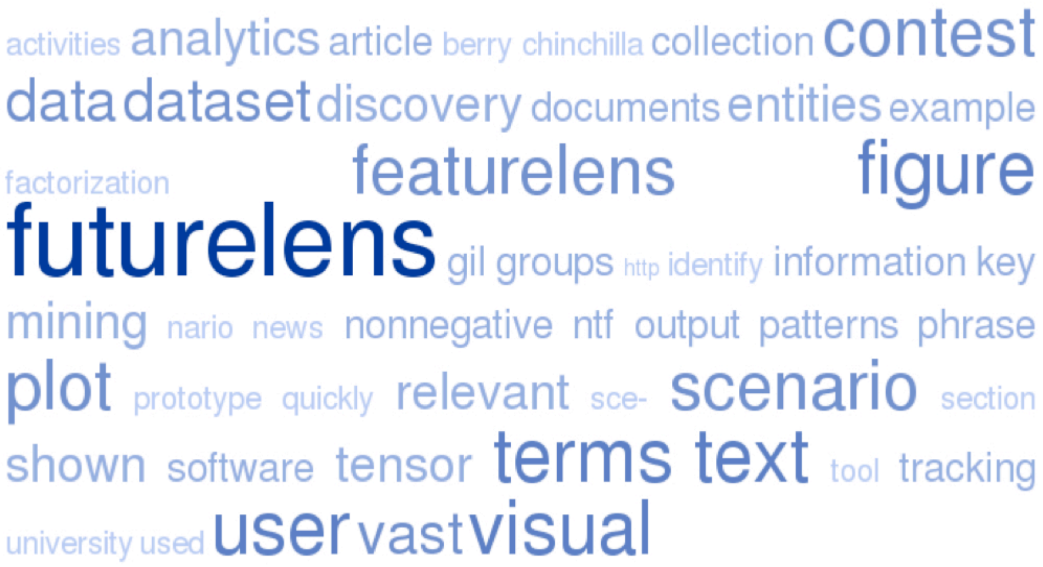

activities analytics article berry chinchilla collection contest
data dataset discovery documents entities example
factorization featurelens figure
futurelens gil groups http identify information key
mining nario news nonnegative ntf output patterns phrase
plot prototype quickly relevant sce- scenario section
shown software tensor terms text tool tracking
university used user vast visual

Figure 1. The tag cloud technique is capable of producing a quick summary that emphasizes the relative significance of the terms within it.

The most significant disadvantage and a major limitation of tag clouds as a visual analytics tool is their lack of depth. They do not provide the user with the capability to analyze the data summary any further. There is no inherent, built-in ability to establish connections within the dataset. For that purpose, other approaches are necessary [14,15].

1.1.2 Establishing Connections with TextArc

TextArc is a visual analysis environment that allows the user to trace the connections between terms, establishing a particular term's relevance to some part of the data space. This is accomplished via drawing an interconnected graph, as shown in Figure 2 (using Shakespeare's Hamlet). As demonstrated by this figure, this approach has an inherent scalability limitation that may be unconquerable without heavy modifications to the approach itself. Even a single, although vast, literary work contains an enormous number of different terms. Under the TextArc approach, all of these need to be displayed, and potentially visually connected to all other relevant terms. The user interacts with TextArc in real time, so the connections may need to be updated frequently. While the approach is highly functional for smaller input datasets, visual clutter and lack of clarity are potential drawbacks when analyzing larger datasets [14,16].

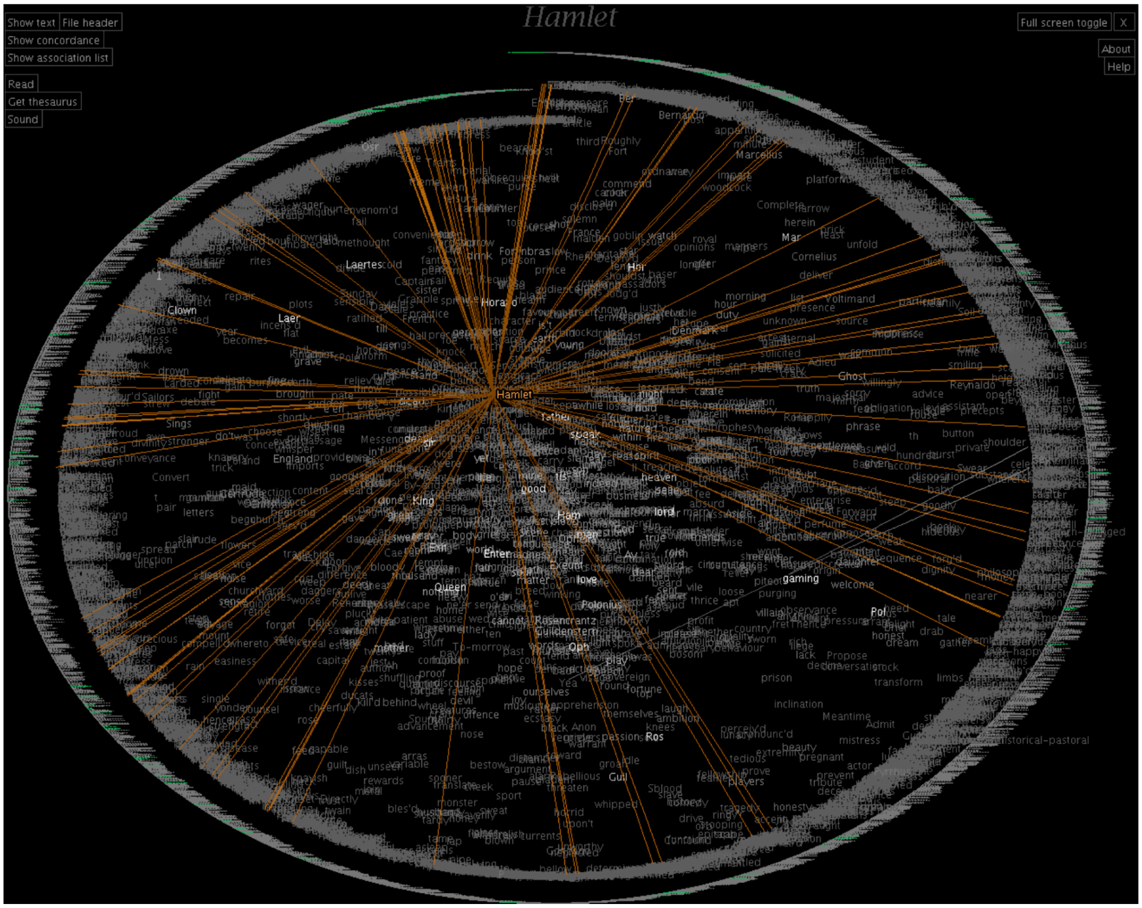


Figure 2. TextArc applied to Shakespeare's Hamlet.

1.1.3 Sentiment Tracking as Visual Analytics

Sentiment tracking is defined as the process of tracking changes in the author(s) attitude or mood through some particular written work. As with the UIMA-SEASR project shown in Figure 3, this may be accomplished by categorizing the terms within the text by parts of speech. The software can then utilize a thesaurus to connect the adjectives found in the text to basic core emotion adjectives. According to research psychological research such as that conducted by Parrot in 2001 [18], there are six core emotions: *surprise, anger, fear, sadness, joy, love*. The thesaurus technique simply needs to count the number of "steps" through the thesaurus between an adjective from the text and one of the core adjectives that is generally accepted as belonging to one of the above six categories. By monitoring the shortest path through the thesaurus, each text adjective may be labeled as belonging to one of these areas [14].

While this approach provides a powerful high-level summarization/labeling capability, it does not allow the user to analyze the details or specifics of the dataset. As demonstrated in this brief survey of alternative visual analytics techniques, a more helpful analytic methodology would provide the user with both high-level summarization/labeling tools, and the capability to analyze specific connections at a very detailed level. Latter chapters of this dissertation will demonstrate that the combined NTF/FutureLens approach (introduced in the next section) accomplishes just that.

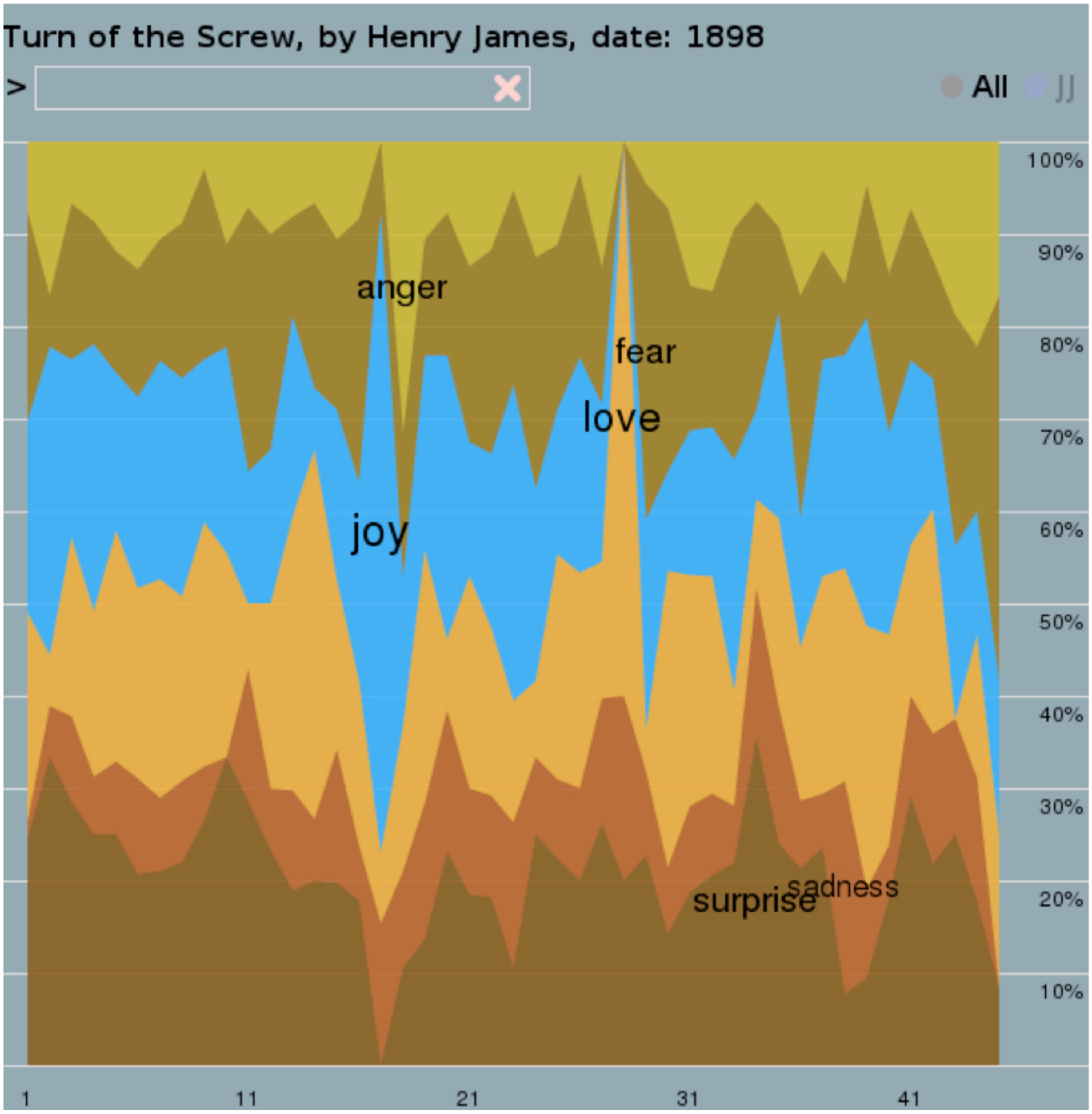


Figure 3. The UIMA-SEASR approach applied to Henry James' Turn of the Screw.

1.2 Scenario and Knowledge Discovery Motivations

Visual analytics is a broad field and analyst goals and motivations may vary greatly. This dissertation focuses specifically on analysis motivated by scenario and knowledge discovery. Scenario discovery focuses on answering the 5W's: *What, Where, When, Who, Why* about a particular topic of interest to the analyst. Knowledge discovery means the analyst is interested in learning new information about the subject of interest, rather than simply generating evidence for a pre-existing theory. For example, a research project might involve a dataset of news articles about South-East Asia in the 1970s. The goal of the research project might be learning more about how various nations of the region conducted diplomacy during this decade: who were the main actors, what was their focus, and so on.

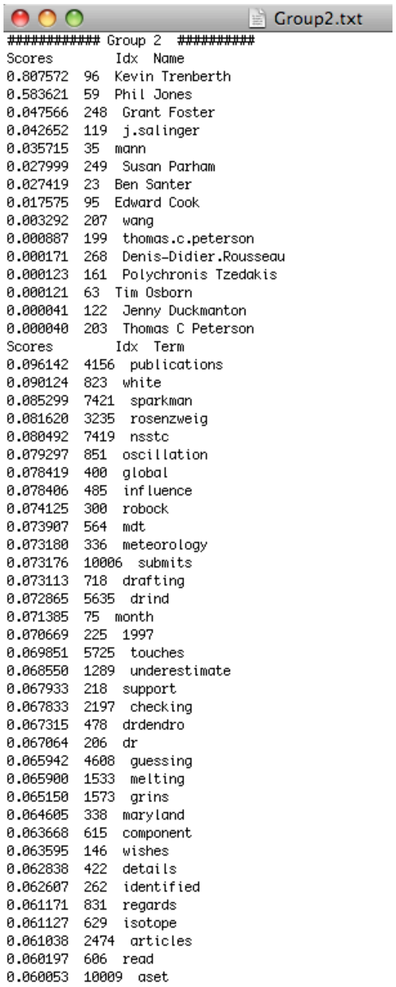
This type of analysis is, of course, possible without the use of the environment presented in this dissertation. It is also possible without the use of computers. The advantages provided by the analysis environment consist of speed and efficiency. Working without the aid of visual analytics, it may take a group of analyst several days to process a large collection consisting of hundreds, or even thousands of documents and extract scenarios from it. With the help of this analysis environment, just one user could perform analysis upon the same dataset within just a few hours.

1.3 The NTF/FutureLens Approach to Visual Analytics

Nonnegative tensor factorization (NTF) is an advanced mathematical technique that has been shown to be effective in analyzing large amounts of textual data in a number of studies. One of the major goals of this dissertation is to improve upon the standard NTF approach. This improvement is to be achieved in two different ways.

First, through incorporation of additional user input (feedback) into the factorization process, the user has been given an ability to adjust importance or “trustworthiness” of certain elements of the data. The analysis process may be greatly improved through such provision of greater control over the entire process to the user. The user has also been provided with an ability to create an additional dimension to the tensor. This is accomplished by allowing the user to create a list of special terms, or *entities*, that the integrated analysis environment then tags within the dataset. The tagged entities are then used to create an additional dimension within the tensor, providing for the potential establishment of connections that otherwise would not have been found with a lower-dimensional model.

Second, the nonnegative tensor factorization has been closely integrated with a visual post-processing tool (FutureLens). Without a post-processing step, the output of the NTF algorithm may be difficult to interpret or analyze in depths, since it consists of a simple list of terms. A sample NTF output feature descriptor file is shown in Figure 4 below.



```
############ Group 2  ##########
Scores        Idx  Name
0.807572  96  Kevin Trenberth
0.583621  59  Phil Jones
0.047566  248  Grant Foster
0.042652  119  j.salinger
0.035715  35  mann
0.027999  249  Susan Parham
0.027419  23  Ben Santer
0.017575  95  Edward Cook
0.003292  207  wang
0.000887  199  thomas.c.peterson
0.000171  268  Denis-Didier.Rousseau
0.000123  161  Polychronis Tzedakis
0.000121  63  Tim Osborn
0.000041  122  Jenny Duckmanton
0.000040  203  Thomas C Peterson
Scores        Idx  Term
0.096142  4156  publications
0.090124  823  white
0.085299  7421  sparkman
0.081620  3235  rosenzweig
0.080492  7419  nsstc
0.079297  851  oscillation
0.078419  400  global
0.078406  485  influence
0.074125  300  robock
0.073907  564  mdt
0.073180  336  meteorology
0.073176  10006  submits
0.073113  718  drafting
0.072865  5635  drind
0.071385  75  month
0.070669  225  1997
0.069851  5725  touches
0.068550  1289  underestimate
0.067933  218  support
0.067833  2197  checking
0.067315  478  drdendro
0.067064  206  dr
0.065942  4608  guessing
0.065900  1533  melting
0.065150  1573  grins
0.064605  338  maryland
0.063668  615  component
0.063595  146  wishes
0.062838  422  details
0.062607  262  identified
0.061171  831  regards
0.061127  629  isotope
0.061038  2474  articles
0.060197  606  read
0.060053  10009  aset
```

Figure 4: A sample NTF output feature descriptor (group) file. Such output files are difficult to analyze further without additional post-processing tools, such as FutureLens.

The ability to visualize the results of NTF and track their occurrence through the original dataset is crucial for an effective analysis process. The integrated environment provides the user with simple tools that can greatly facilitate the process of preparing a dataset and NTF output groups for analysis with FutureLens. The capabilities of the environment include the ability to add dates to the data files, insert tagged entities, and adjust term weights in accordance with some trust or interest model. These capabilities are described fully in Chapter 4.

Chapter 2

NTF-PARAFAC: Examples of Usage and Effectiveness

There exists a plethora of approaches to analyzing large amounts of textual information. The exact nature of the dataset and the goals of the analysis process influence which approach would be most effective in each individual case. For cases where the dataset contains tagged entities and a clearly defined time line, nonnegative tensor factorization (NTF) techniques have been shown to be highly effective. NTF allows the analyst to extract term-by-entity associations from the data. With the addition of a visual post-processing tool, such as Futurelens, it becomes possible to trace the progression of term-entity, term-term, and entity-entity relationships through the data space over time. One example of such a study involved scenario discovery using the fictional news article dataset from the IEEE VAST-2007 contest [7]. As shown by this example, NTF based on the well-known PARAFAC [6] model for multidimensional data can be highly effective in extracting important features from a large textual dataset. The example is described more fully in Section 2.2.

2.1 NTF-PARAFAC Results Visualization through FutureLens

As will be illustrated by the examples in the subsequent sections and chapters of this dissertation, a visual NTF results processing tool is integral to the analysis process. While it is possible to analyze NTF results without such a tool, the process would potentially be significantly slower and more prone to human error. FutureLens was created in 2008 as a proof of concept generic text visualization tool [9].

This dissertation includes significant modifications and improvements to the original version of FutureLens, as well as its integration into the overall analysis environment. Among the improvements to FutureLens are numerous bug fixes that enable the searching and color-coding capabilities to function as intended. This dissertation also added the capability to construct search phrases within the FutureLens GUI, and altered FutureLens' term collection constructing code to enable proper function. Functionally, a phrase differs from a collection of terms because order and term adjacency are important in a phrase, but not a term collection.

Perhaps most significantly, this dissertation adds an automated NTF results classification feature. The feature requires some initial user input, namely the creation of categories and category description files to be used as input for automated classification. However, once these have been created, they may be reused with very few, if any, modifications across a variety of research project types. The rest of the process is automated, meaning FutureLens takes the category descriptors as input and generates a color-coded labeling scheme for the loaded NTF output groups without any further user input. This feature is an important part of the dissertation and it is therefore discussed in much greater detail in Chapter 5.

2.2 NTF-PARAFAC and IEEE VAST-2007

The IEEE VAST 2007 Contest (http://www.cs.umd.edu/hcil/VASTcontest07/) was designed to promote the development of the field of visual analytics, focusing specifically on scenario extraction from textual data. The dataset associated with the contest consisted of approximately 1,400 fictitious news articles and blog posts. The topics of the articles and posts ranged widely, however, most had some direct or indirect connection to animals. There was no unifying theme beyond that—the topics were extremely diverse, including for example: pet adoptions from specific shelters, laws pertaining to treatment of horses in the United States, agricultural practices in Canada, salmon fishing statistics, arson investigations involving suspected eco-terrorists, environmental legislation in China, and many others.

Participants in the contest were given the task of identifying two significant law enforcement/counter-terrorism scenarios within the data. The "hidden" scenarios involved emergencies related to wildlife law enforcement, with endangered species issues and eco-terrorism playing an underlying role. The task could be decomposed into smaller goals, such as: (i) identifying entities of interest (e.g., persons, organizations, locations); (ii) depicting this information in a visual and interactive manner; (iii) answering specific questions pertaining to the scenarios [7]. The NTF-PARAFAC approach was shown to be effective in helping to achieve these goals [1].

The Parallel Factors (PARAFAC) model, also known as Canonical Decomposition and (more recently) Canonical Polyadic Decomposition (CP), was proposed by Harshman in 1970 [6,8,21]. Given a third-order tensor *X* of size $m \times n \times p$ and a desired approximation rank *r*, the PARAFAC model approximates *X* as a sum of *r* rank-1 tensors formed by the outer products of three vectors [1], as shown in Equation 1 below.

$$\chi \approx \sum_{i=1}^{r} a_i \circ b_i \circ c_i \qquad (1)$$

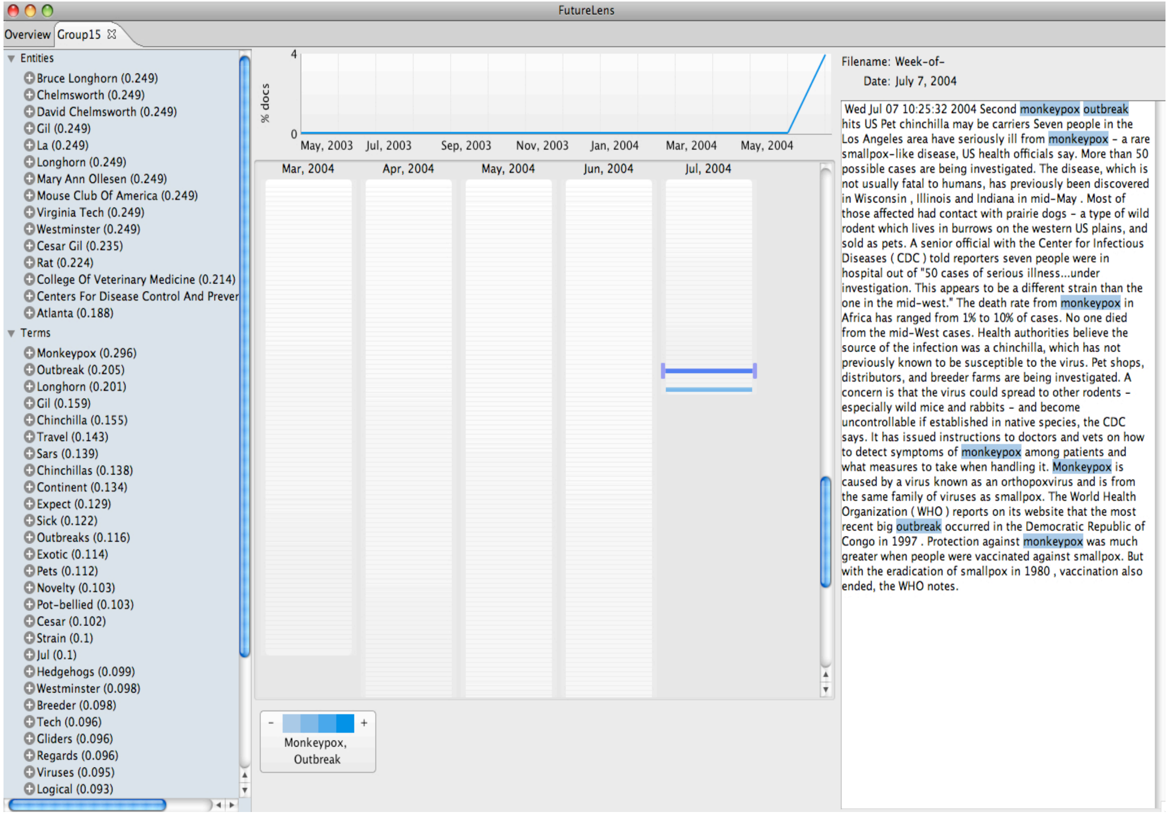
The goal of NTF is to find best fitting nonnegative matrices, *A*, *B*, and *C*, that fit the data in *X*. This is demonstrated by Equation 2, where the norm refers to the 2-norm [1]:

$$\min_{A,B,C} = \left\| \chi - \sum_{i=1}^{r} a_i \circ b_i \circ c_i \right\| \qquad (2)$$

In the study described in [1], NTF-PARAFAC was applied to a $12{,}121 \times 7{,}141 \times 15$ (term-by-entity-by-date) sparse tensor that contained 1,142,077 nonzeros. The dates were binned on a monthly basis. Applying the NTF algorithm resulted in twenty-five total output groups, each described by fifteen interrelated entities and thirty-five interrelated terms.

The groups corresponding to the two fictional “hidden” scenarios were correctly identified, although the identification process required a significant time commitment and several post-processing steps [1]. The study was subsequently replicated using a visual post-processing software tool, Futurelens, in order to improve the effectiveness and efficiency of processing NTF output group results [9]. The two figures below illustrate how FutureLens was used to identify and gather evidence for the IEEE VAST-2007 scenario involving a bioterrorism-induced monkeypox outbreak.

Figure 5 shows one of the NTF output files (“Group 15”) loaded into FutureLens. This group is described by a list of top 15 most relevant entities and 35 most relevant terms. In this figure, the user has selected two of the top terms (*monkeypox* and *outbreak*), and then combined them into a collection of terms (*monkeypox*, *outbreak*). In a collection of terms, term order and adjacency do not matter (unlike in a phrase). FutureLens located two articles containing the term collection (*monkeypox*, *outbreak*). The first article describes much of the bioterrorism scenario, however a few crucial details regarding the perpetrator are missing in this article.

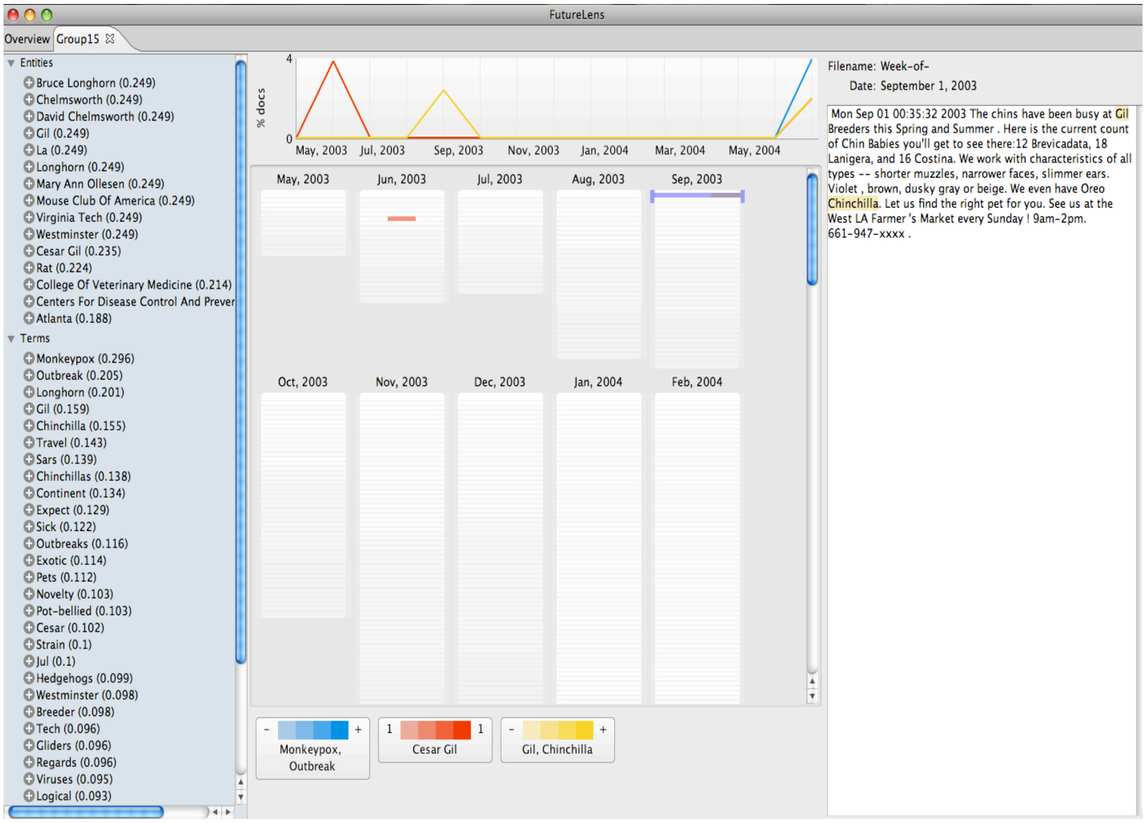


Filename: Week-of-
Date: July 7, 2004

Wed Jul 07 10:25:32 2004 Second monkeypox outbreak hits US Pet chinchilla may be carriers Seven people in the Los Angeles area have seriously ill from monkeypox - a rare smallpox-like disease, US health officials say. More than 50 possible cases are being investigated. The disease, which is not usually fatal to humans, has previously been discovered in Wisconsin , Illinois and Indiana in mid-May . Most of those affected had contact with prairie dogs - a type of wild rodent which lives in burrows on the western US plains, and sold as pets. A senior official with the Center for Infectious Diseases (CDC) told reporters seven people were in hospital out of "50 cases of serious illness...under investigation. This appears to be a different strain than the one in the mid-west." The death rate from monkeypox in Africa has ranged from 1% to 10% of cases. No one died from the mid-West cases. Health authorities believe the source of the infection was a chinchilla, which has not previously known to be susceptible to the virus. Pet shops, distributors, and breeder farms are being investigated. A concern is that the virus could spread to other rodents - especially wild mice and rabbits - and become uncontrollable if established in native species, the CDC says. It has issued instructions to doctors and vets on how to detect symptoms of monkeypox among patients and what measures to take when handling it. Monkeypox is caused by a virus known as an orthopoxvirus and is from the same family of viruses as smallpox. The World Health Organization (WHO) reports on its website that the most recent big outbreak occurred in the Democratic Republic of Congo in 1997 . Protection against monkeypox was much greater when people were vaccinated against smallpox. But with the eradication of smallpox in 1980 , vaccination also ended, the WHO notes.

*Figure 5: This figure demonstrates how FutureLens may be used to aid the interpretation of an NTF output file. Here, a collection including two top terms (*monkeypox *and* outbreak*) has been created by the user and relevant articles located within the data.*

In order to locate addition information pertaining to this scenario, the user adds the entity corresponding to the suspect's name (*Cesar Gil*), and a collection of terms (*chinchilla*, *Gil*), to the FutureLens display. Doing so allows the user to locate an article where Cesar Gil explains his philosophy regarding the trade in exotic animals (he states that breaking a few laws is an acceptable tactic in stopping such trade). In addition, as shown in Figure 6 below, the user is also able to locate an article corresponding to an advertisement of chinchillas for sale by a business called *Gil Breeders*. A complete storyline corresponding to this NTF output feature now emerges.



Entities
Bruce Longhorn (0.249)
Chelmsworth (0.249)
David Chelmsworth (0.249)
Gil (0.249)
La (0.249)
Longhorn (0.249)
Mary Ann Ollesen (0.249)
Mouse Club Of America (0.249)
Virginia Tech (0.249)
Westminster (0.249)
Cesar Gil (0.235)
Rat (0.224)
College Of Veterinary Medicine (0.214)
Centers For Disease Control And Prever
Atlanta (0.188)

Terms
Monkeypox (0.296)
Outbreak (0.205)
Longhorn (0.201)
Gil (0.159)
Chinchilla (0.155)
Travel (0.143)
Sars (0.139)
Chinchillas (0.138)
Continent (0.134)
Expect (0.129)
Sick (0.122)
Outbreaks (0.116)
Exotic (0.114)
Pets (0.112)
Novelty (0.103)
Pot-bellied (0.103)
Cesar (0.102)
Strain (0.1)
Jul (0.1)
Hedgehogs (0.099)
Westminster (0.098)
Breeder (0.098)
Tech (0.096)
Gliders (0.096)
Regards (0.096)
Viruses (0.095)
Logical (0.093)

Filename: Week-of-
Date: September 1, 2003

Mon Sep 01 00:35:32 2003 The chins have been busy at Gil Breeders this Spring and Summer . Here is the current count of Chin Babies you'll get to see there:12 Brevicadata, 18 Lanigera, and 16 Costina. We work with characteristics of all types -- shorter muzzles, narrower faces, slimmer ears. Violet , brown, dusky gray or beige. We even have Oreo Chinchilla. Let us find the right pet for you. See us at the West LA Farmer 's Market every Sunday ! 9am-2pm. 661-947-xxxx .

Figure 6: This figure shows how a more complete description of the scenario corresponding to this NTF output feature may be obtained using FutureLens. The article displayed here shows a company owned by Cesar Gil advertising chinchillas for sale – thus revealing Gil's method for distributing monkeypox-infected chinchillas.

The combined NTF and visual post processing approach thus proved to be effective in discovering the hidden plotlines of interest in the VAST-2007 contest dataset. The additional work presented in this dissertation focused on improving the approach by allowing user alteration of the tensor, integrating various pre-processing and post-processing steps into a single environment, and adding automated NTF output group classification capability. Subsequent chapters of this dissertation describe this work fully, as well as providing additional evidence for the potential effectiveness of this approach to text analysis. The following section provides some additional evidence of how NTF may be useful in discovering new and interesting information, in this case using a different input medium.

2.3 NTF-PARAFAC and the “Climategate” Emails

The so-called “climategate” event took place in late 2009 and began with the posting on the Internet of email correspondence between climate research scientists, many of them affiliated with the University of East Anglia’s Climatic Research Unit. It is difficult to ascertain how exactly the emails were originally obtained, or by whom. Several versions of events exist, but the media most commonly cites one involving a hacking of a UEA email server [10,25]. Regardless of how the emails were originally obtained, they have been posted or linked to by a large number of different websites, thus making the climategate email dataset part of public record by the time this study was undertaken.

The goal of the climategate email study was to demonstrate how nonnegative tensor factorization (NTF) techniques can extract term-by-author-by-time associations from the University of East Anglia climate research email dataset [11]. In particular, the goal was to demonstrate how NTF could potentially be used to automatically expose possibly unethical schemes or actions of specific individuals or groups. The term-by-author-by-time decomposition is illustrated in Figure 7 below [13].

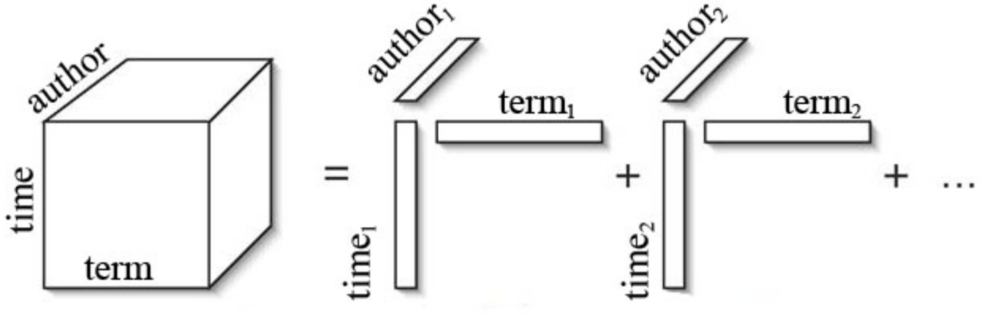
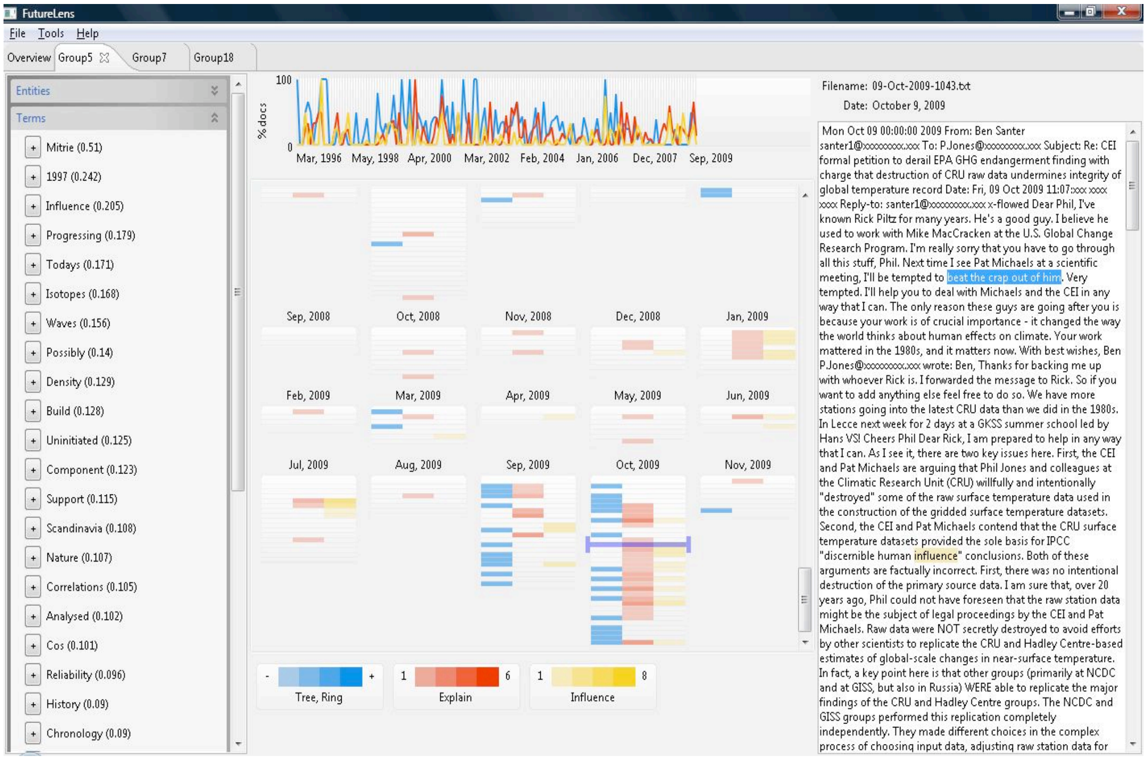


Figure 7: The 3-way NTF PARAFAC decomposition model produces a number of data features, each one corresponding to a potentially significant underlying theme or scenario contained in the email dataset.

The dataset consisted of 1,072 climate research electronic mail messages, involving 271 different authors and covering the time period between March 1996 and November 2009. Parsing the dataset yielded a dictionary of 11,829 terms, each term having to appeared in at least two different messages and at least twice across the collection in order to be included in the dictionary. Using NTF and FutureLens, it was possible to locate several interesting documents, such as the one shown in Figure 8 [13]. While the article shown here does not prove or disprove the charges of unethical behavior, it goes show one of the East Anglia climate researchers using the phrase “beat the crap out of him” in reference to a more skeptical scientist.



Filename: 09-Oct-2009-1043.txt

Date: October 9, 2009

Mon Oct 09 00:00:00 2009 From: Ben Santer santer1@xxxxxxxxx.xxx To: P.Jones@xxxxxxxxx.xxx Subject: Re: CEI formal petition to derail EPA GHG endangerment finding with charge that destruction of CRU raw data undermines integrity of global temperature record Date: Fri, 09 Oct 2009 11:07:xxx xxxx xxxx Reply-to: santer1@xxxxxxxxx.xxx x-flowed Dear Phil, I've known Rick Piltz for many years. He's a good guy. I believe he used to work with Mike MacCracken at the U.S. Global Change Research Program. I'm really sorry that you have to go through all this stuff, Phil. Next time I see Pat Michaels at a scientific meeting, I'll be tempted to beat the crap out of him. Very tempted. I'll help you to deal with Michaels and the CEI in any way that I can. The only reason these guys are going after you is because your work is of crucial importance - it changed the way the world thinks about human effects on climate. Your work mattered in the 1980s, and it matters now. With best wishes, Ben P.Jones@xxxxxxxxx.xxx wrote: Ben, Thanks for backing me up with whoever Rick is. I forwarded the message to Rick. So if you want to add anything else feel free to do so. We have more stations going into the latest CRU data than we did in the 1980s. In Lecce next week for 2 days at a GKSS summer school led by Hans VS! Cheers Phil Dear Rick, I am prepared to help in any way that I can. As I see it, there are two key issues here. First, the CEI and Pat Michaels are arguing that Phil Jones and colleagues at the Climatic Research Unit (CRU) willfully and intentionally "destroyed" some of the raw surface temperature data used in the construction of the gridded surface temperature datasets. Second, the CEI and Pat Michaels contend that the CRU surface temperature datasets provided the sole basis for IPCC "discernible human influence" conclusions. Both of these arguments are factually incorrect. First, there was no intentional destruction of the primary source data. I am sure that, over 20 years ago, Phil could not have foreseen that the raw station data might be the subject of legal proceedings by the CEI and Pat Michaels. Raw data were NOT secretly destroyed to avoid efforts by other scientists to replicate the CRU and Hadley Centre-based estimates of global-scale changes in near-surface temperature. In fact, a key point here is that other groups (primarily at NCDC and at GISS, but also in Russia) WERE able to replicate the major findings of the CRU and Hadley Centre groups. The NCDC and GISS groups performed this replication completely independently. They made different choices in the complex process of choosing input data, adjusting raw station data for

Figure 8: NTF output Feature 5 corresponds to discussions that contained insults directed at scientists skeptical of human influences on global warming. A collection of terms (Tree, Ring) *has been created here in order to direct FutureLens to retrieve information pertaining to tree ring climate data (which has on occasion been contradictory to the human-influenced climate change theories). Additionally, selecting the terms* Explain *and* Influence *leads us to an email in which a climate researcher threatens to "beat the crap out of" a skeptic (text highlighted in blue in the window on the right of the figure).*

Chapter 3

Implementation of Portable and User-Friendly NTF in a Visual Analysis Environment

One of the main goals of this dissertation is to create a single, unified, user-friendly textual dataset analysis environment. Significant research has been conducted in the fields of text mining, visual analytics, and sentiment tracking. However, no substantial attempts to integrate techniques from these vastly different fields into a single, convenient, highly usable text analysis environment have been documented. Therefore, one of the goals of this dissertation is to evaluate the potential effectiveness of the combined approach to text analysis, which heretofore has remained a largely unexplored research area. The results of this evaluation are described more fully in Chapter 5.

3.1 Overall Design Goals for Text Analysis Environment

3.1.1 User-Friendliness

First and foremost, the integrated analysis approach should be highly user-friendly. Ideally, a user without a great deal of technical experience or knowledge should be able to utilize the analysis software without much time having to be spent on training. Specific knowledge pertaining to data mining or visual analytics should be unnecessary. It is therefore imperative to conceal the underlying nonnegative tensor factorization process, while providing the user with a clear set of controls that would allow him or her to influence the NTF process in ways that may facilitate data analysis.

3.1.2 Portability, Flexibility, Cost of Use

The second major design goal for the integrated analysis environment is portability and flexibility. A significant amount of NTF-related work has in the past been performed using Matlab®. The Matlab® Tensor Toolbox that was created at Sandia National Laboratory is a great example of such work [26]. However, experience suggests that even though Matlab® is a powerful programming environment for scientific applications, code written in it does not transition well into general usage. Since one of the main goals is to create a highly usable textual data analysis tool, it is therefore critical to write it in languages that are more portable and flexible than Matlab®. For instance, Python and its NumPy/Pylab libraries have been proven on many occasions to be effective alternatives to Matlab®. Python has the additional advantage of being freely available to programmers and users, and completely cross-platform. For the visualization portion of the analysis environment, a Java-based graphical post-processing tool (FutureLens) has been previously shown to be helpful to the text analysis process. Java, being a cross-platform language, is a good choice for accomplishing the portability/flexibility goal.

3.1.3 Speed and Efficiency for Real-time User Analysis

The third major goal is to make the analysis process efficient and as scalable as possible. Because of the very nature of the field of data mining, scalability is always a great challenge. One of the major design goals for this software therefore will be to make the approach as efficient and scalable as possible. A significant area of exploration will be Python's efficiency in performing NTF decompositions on large datasets. According to some sources, Python’s NumPy library yields computational performance comparable to Matlab® [12]. One of the goals will be to attempt to make the performance of the portable (Python-based) analysis environment to match (or at least approach) that of the older Matlab®-based methodology.

3.1.4 Automation and User Input Necessity

This approach requires the integration of several vastly different elements. Figure 9 summarizes the overall design of the analysis environment. It is important to note that while many of the steps of the analysis described below include a high degree of automation, input from a human analyst is necessary at many of the stages.

The first stage during which human input is essential is the selection of the document collection. This clearly depends on the goals and interests of the analyst, and thus it would be impossible to remove the human element from this stage. The user also must provide a set of timestamps corresponding to the elements of the document collection. Many databases include such information as part of the metadata. Alternatively, a parsing script may be utilized to extract temporal information directly from the documents. A dictionary of terms found in the data collection is also required. While the environment does not provide a feature for generating one, virtually every text parsing software includes such a feature and thus generating a dictionary file should be a straightforward operation.

The user’s input is also required in creating a list of entities. In the current scheme, an entity is defined as a potentially significant, but relatively rare term or phrase. A person’s first and last name may be a good example of a phrase that an analyst may wish to define as an entity. This process is highly subjective and depends entirely upon the analyst’s goals – thus requiring human input.

Once the user has created the elements described above, the rest of the process up to final results analysis is completely automated. As shown in Figure 9, The analysis environment is capable of tagging the user defined entities and timestamps within the elements of the dataset using SGML-style tags, then generating a plain-text file describing a tensor, and applying the NTF algorithm to the tensor. If the analyst desires to adjust term weights within the tensor, additional human input in the form of a user-created weights file is required. The reason for this is similar to the reason for requiring user input in generating a list of entities – the environment is flexible and capable of working with a wide variety of analysis goals, but this means that input from the analyst must necessarily be required.

Upon completion of the NTF algorithm, the user has the option of launching FutureLens directly from the analysis environment. The user may then load a set of pre-defined category description files. Creation of these may require input from the analyst, although in many cases, these descriptions may have been created already (e.g., there may exist an agreed-upon definition of what constitutes "unrest" and what terms describe the concept). Once the user has loaded the category descriptions, FutureLens will automatically label and color-code each loaded NTF output group.

All of these steps are summarized by Figure 9, and described in much greater detail in Chapters 4 and 5.

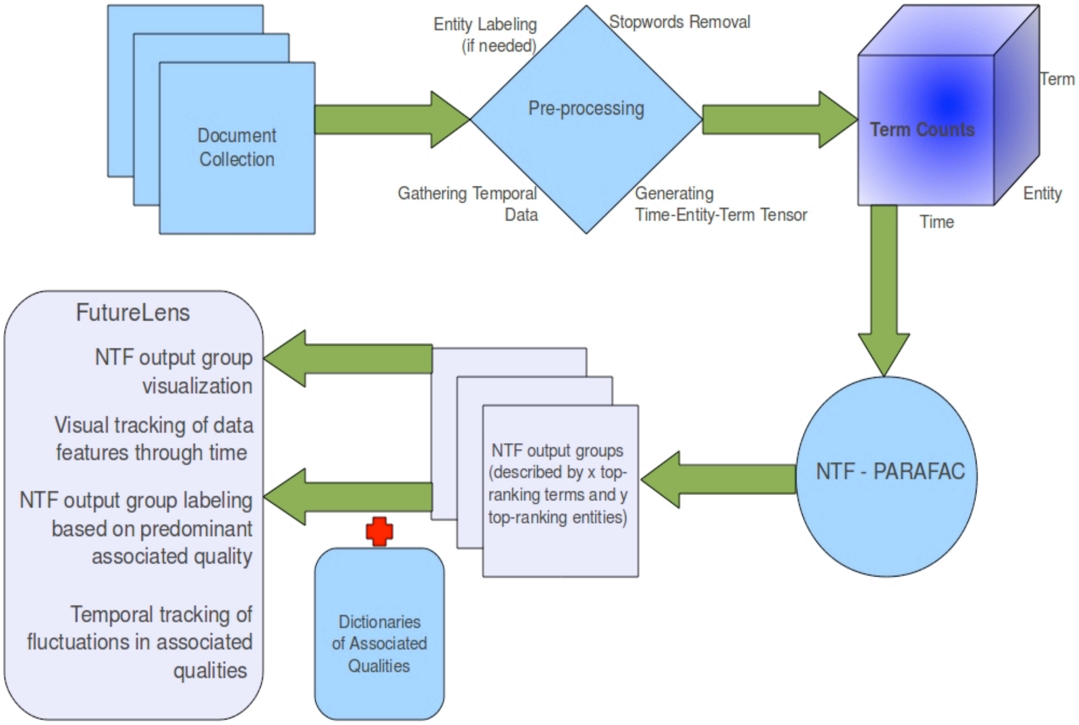

Document Collection
Entity Labeling (if needed)
Stopwords Removal
Pre-processing
Gathering Temporal Data
Generating Time-Entity-Term Tensor
Term Counts
Term
Entity
Time
NTF - PARAFAC
NTF output groups (described by x top-ranking terms and y top-ranking entities)
Dictionaries of Associated Qualities
FutureLens
NTF output group visualization
Visual tracking of data features through time
NTF output group labeling based on predominant associated quality
Temporal tracking of fluctuations in associated qualities

Figure 9: Proposed design of the text analysis environment. This design will allow the user to easily perform a number of operations upon a textual input dataset, such as entity tagging, timestamp insertion, and tensor term weight adjustment. The environment will also allow the user to easily execute the NTF algorithm, and analyze the results. Among the most important aids in results analysis is the environment's capability to automatically label the resulting NTF output features in accordance with a user-defined categorization scheme.

3.2 Nonnegative Tensor Decomposition Improvements

One may argue that the single most important stage of the integrated analysis process is the nonnegative tensor factorization. Other stages may be viewed as pre-processing data for NTF, or post-processing and visualizing the results of NTF. While the other stages are clearly important, NTF may be said to be the most significant step. The conversion of the PARAFAC code from Matlab® to Python was therefore quite important to the development of this analysis environtment. The language conversion of this critical step allowed for the implementation of additions described in the sections below.

3.2.1 Python Conversion

In addition to being very helpful to the process of integrating more pre-processing and post-processing features into the analysis environment, Python has several important advantages. First, Python is widely accepted to be one of most programmer-friendly, writeable and readable programming languages in existence today. However, Python far more than a vacuous, "toy" language. It includes many powerful features, such as full support of object-oriented programming, multithreading, graphical user interface building tools, and a variety of data structures. Additionally, basic Python has been greatly extended in recent years specifically for scientific computing. Packages such as Numpy [19] and SciPy [20] significantly expand the language's capabilities for scientific applications. This also illustrates another major advantage of Python, namely that it is fully extensible, freely available, and cross-platform [22].

3.2.2 Additional Dimension Creation through Entity Tagging

Giving the user an ability to create an additional tensor dimension through tagging a subset of significant terms ("entities") is one of the major NTF improvements included in the integrated analysis environment. This is distinct from the trust measures described in the subsequent section, because relative significance in the case of entities is the result of their type, rather than of the nature of the specific terms. For example, *Person*-type entities could include all the people's names found in the dataset. *Location*-type entities could include a wide variety of geographical labels: city names, state/province names, countries, mountain ranges, lakes, etc. In other words, a user could emphasize an entire group of terms (created because of common type), without having to consider each individual term's potential significance.

3.2.3 Significance or Trust Measure Integration into NTF

Under some circumstances, it could be greatly helpful to the analysis process for the environment to include an integrated significance or trust measures capability. It is possible, indeed likely, that a knowledgeable user will have access to potentially important information which normally would be inaccessible to the NTF algorithm. In other words, different elements of the data may have different levels of significance to the user because of the user's prior knowledge about the data. Alternatively, this may be viewed as a trustworthiness issue—meaning, for example, that the user may consider certain sources as inherently worthy of trust, while others may be entirely untrustworthy in the user's mind. The Python NTF implementation includes the ability to alter the tensor values in accordance with a user-supplied trust list. The trust list is simply a list of terms and corresponding weights. Terms that are more worthy of consideration may be assigned a higher weight by the user, while some other terms may be assigned a lower weight. The NTF-PARAFAC approach then integrates these significance/trust measures into the factorization process. Incorporation of different term weighting schemes could also be included as part of this user-influenced NTF approach. The integrated analysis environment provides the user with significance/trust controls that do not requiring the user to be exposed to the underlying NTF code.

3.3 Visualization of NTF Results

Visualization of the NTF output results is accomplished via the use of a Java-based software tool called FutureLens. In addition to allowing the user to visually analyze the output of the NTF code, FutureLens also connects the tensor factorization output features to the original dataset. FutureLens is capable of displaying components of the original text corpus, and of constructing a timeline (assuming the elements contain the necessary temporal information in the appropriate format—i.e., an SGML-format date tag). The integrated analyis environment provides the user with simple, easy-to-use tools that facilitate the process of altering the format of the original dataset to be fully-readable for FutureLens. After loading the dataset and the NTF output groups into FutureLens, the user may add elements of the NTF output group files (terms or entities) to the FutureLens display. FutureLens then locates these elements within the data and constructs plots of their occurrence over time. Additionally, the occurrences are color-coded within the original data itself. FutureLens also provides the user with a capability to create and search for collections of terms (where term order and adjacency are unimportant), and phrases (where order and adjacency matter) [9].

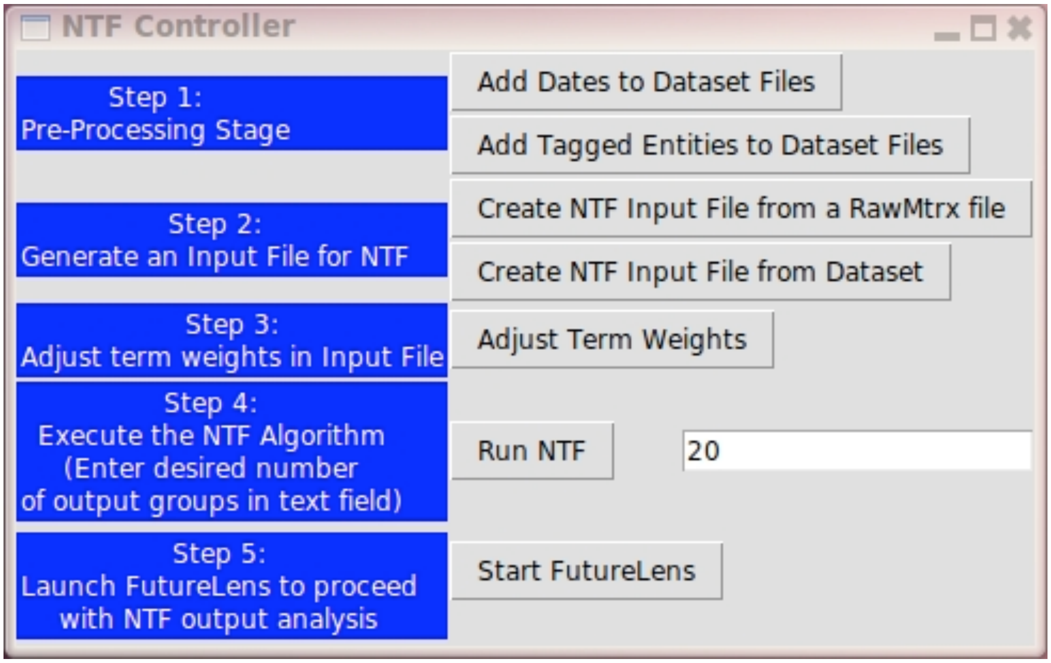
3.4 Sentiment Tracking and Automatic NTF Output Labeling

The idea of automatic group labeling was inspired by research into visual sentiment tracking. Sentiment tracking is a highly promising technique that has great potential for providing insight during analysis of textual data. Various techniques exist, for example, one common approach utilizes synonym connections between adjectives within the dataset and certain key emotion descriptors (e.g., “angry”, “joyful”, etc.) [5]. A different approach to sentiment tracking utilizes pre-defined dictionaries describing a particular sentiment. This approach may be considered somewhat more flexible, because it allows user input while creating sentiment descriptions, rather than relying upon a pre-defined thesaurus. It also allows parts of speech other than adjectives to be considered, which may be highly useful in certain cases [4]. The NTF output group labeling by category is based upon the latter approach, since it allows the users to create descriptions of completely arbitrary categories that do not have to correspond to any one particular emotion or sentiment.

Chapter 4

Integrated Environment Capabilities and Input Format Requirements

The sections in this chapter describe the specific capabilities of the analysis environment. The required input formats are also described here, including examples. The sections also contain information about the format of the output produced by each of the functions that constitute the environment. It is important to note that the functions may be used in sequence, starting with Step 1, however this is not strictly required in order for the environment to function. For example, if the user already had an input file describing a tensor in the appropriate format, he or she could simply go straight to the non-negative factorization step without having to re-generate the tensor description file. Furthermore, certain steps may be skipped entirely. For instance, there is no requirement that a dataset must contain tagged entities. Figure 10 below shows the graphical user interface of the analysis environment. The features shown in this figure are described in the sections below.



Step 1:
Pre-Processing Stage

Add Dates to Dataset Files
Add Tagged Entities to Dataset Files

Step 2:
Generate an Input File for NTF

Create NTF Input File from a RawMtrx file
Create NTF Input File from Dataset

Step 3:
Adjust term weights in Input File

Step 4:
Execute the NTF Algorithm
(Enter desired number
of output groups in text field)

20

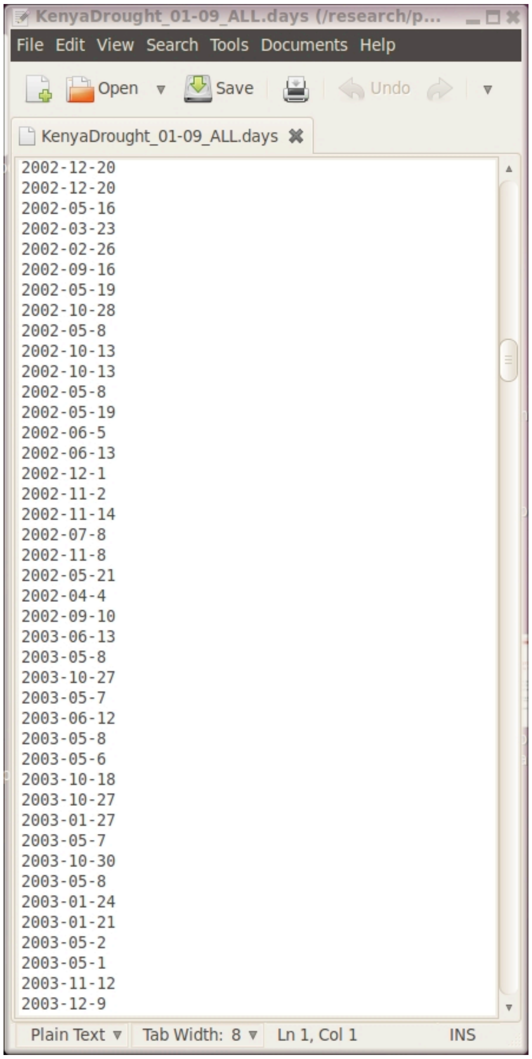
Step 5:
Launch FutureLens to proceed
with NTF output analysis

Figure 10: The graphical user interface of the analysis environment. Some of the steps shown here are optional, although they do enhance the analysis process and aid knowledge discovery.

4.1 Pre-processing: Inserting Dates into the Dataset

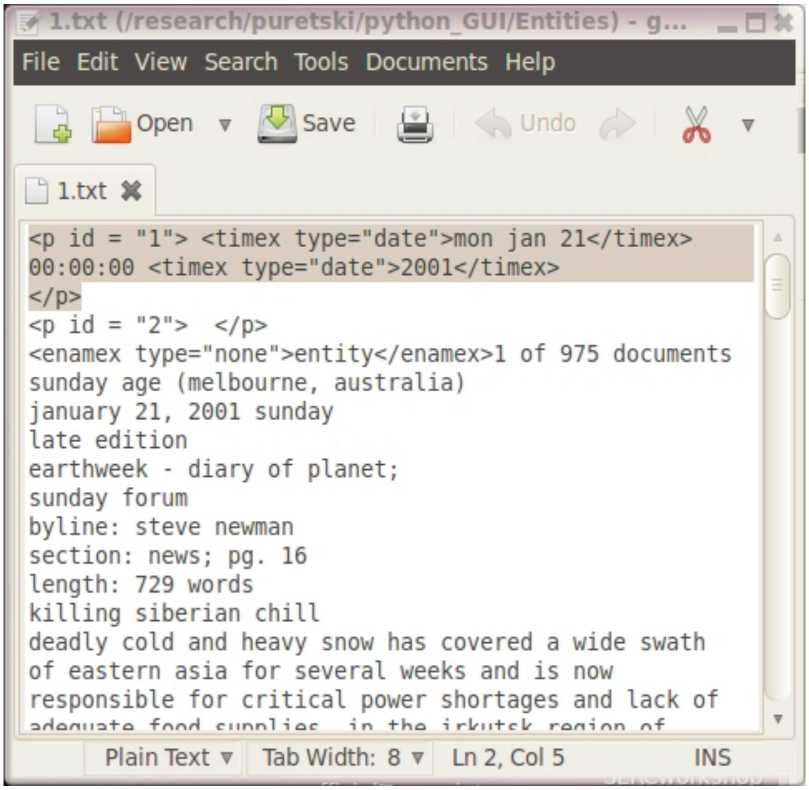
Many of the capabilities of the analysis environment depend upon the user being able to provide temporal information relating to the dataset. While it is possible to utilize the environment for potentially effective knowledge discovery even without such information, one would not be able to take advantage of its full capability in that case. If the temporal information is not already encoded in the dataset, the user may do so using the analysis environment.

In order to utilize this feature, a file containing all of the dates corresponding to the elements of the dataset is necessary. The dates should appear one per line, and in the following format: *yyyy-mm-dd*. The ordering of the files in the dataset and the order of the dates in this input file should match. Figure 11 shows a sample dates list file, while Figure 12 shows the results of utilizing this feature: one of the data set files with the added SGML-format date tag.


KenyaDrought_01-09_ALL.days (/research/p...
File Edit View Search Tools Documents Help
Open
Save
Undo
KenyaDrought_01-09_ALL.days
Plain Text
Tab Width: 8
Ln 1, Col 1
INS

```
2002-12-20
2002-12-20
2002-05-16
2002-03-23
2002-02-26
2002-09-16
2002-05-19
2002-10-28
2002-05-8
2002-10-13
2002-10-13
2002-05-8
2002-05-19
2002-06-5
2002-06-13
2002-12-1
2002-11-2
2002-11-14
2002-07-8
2002-11-8
2002-05-21
2002-04-4
2002-09-10
2003-06-13
2003-05-8
2003-10-27
2003-05-7
2003-06-12
2003-05-8
2003-05-6
2003-10-18
2003-10-27
2003-01-27
2003-05-7
2003-10-30
2003-05-8
2003-01-24
2003-01-21
2003-05-2
2003-05-1
2003-11-12
2003-12-9
```

Figure 11: A small portion of a temporal information file. The analysis environment provides the user with a feature that can be used to extend the dataset with temporal information by adding SGML-format date tags to the data files.

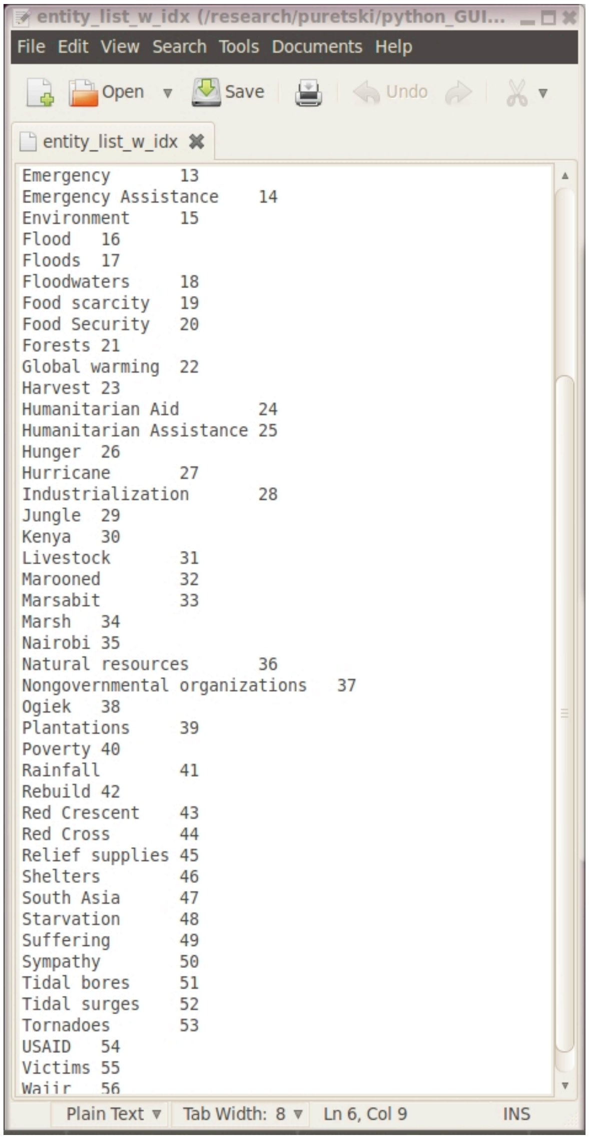

1.txt (/research/puretski/python_GUI/Entities) - g...
File Edit View Search Tools Documents Help
Open
Save
Undo
1.txt
Plain Text
Tab Width: 8
Ln 2, Col 5
INS

```
<p id = "1"> <timex type="date">mon jan 21</timex>
00:00:00 <timex type="date">2001</timex>
</p>
<p id = "2">  </p>
<enamex type="none">entity</enamex>1 of 975 documents
sunday age (melbourne, australia)
january 21, 2001 sunday
late edition
earthweek - diary of planet;
sunday forum
byline: steve newman
section: news; pg. 16
length: 729 words
killing siberian chill
deadly cold and heavy snow has covered a wide swath
of eastern asia for several weeks and is now
responsible for critical power shortages and lack of
adequate food supplies, in the irkutsk region of
```

Figure 12: This figure shows the results of utilizing the date-insertion feature of the analysis environment. Shown here is one of the dataset files, now containing an SGML-format date tag that was inserted by using this feature and based upon a user-provided temporal information file (shown in Fig. 10).

4.2 Pre-processing: Entity Tagging

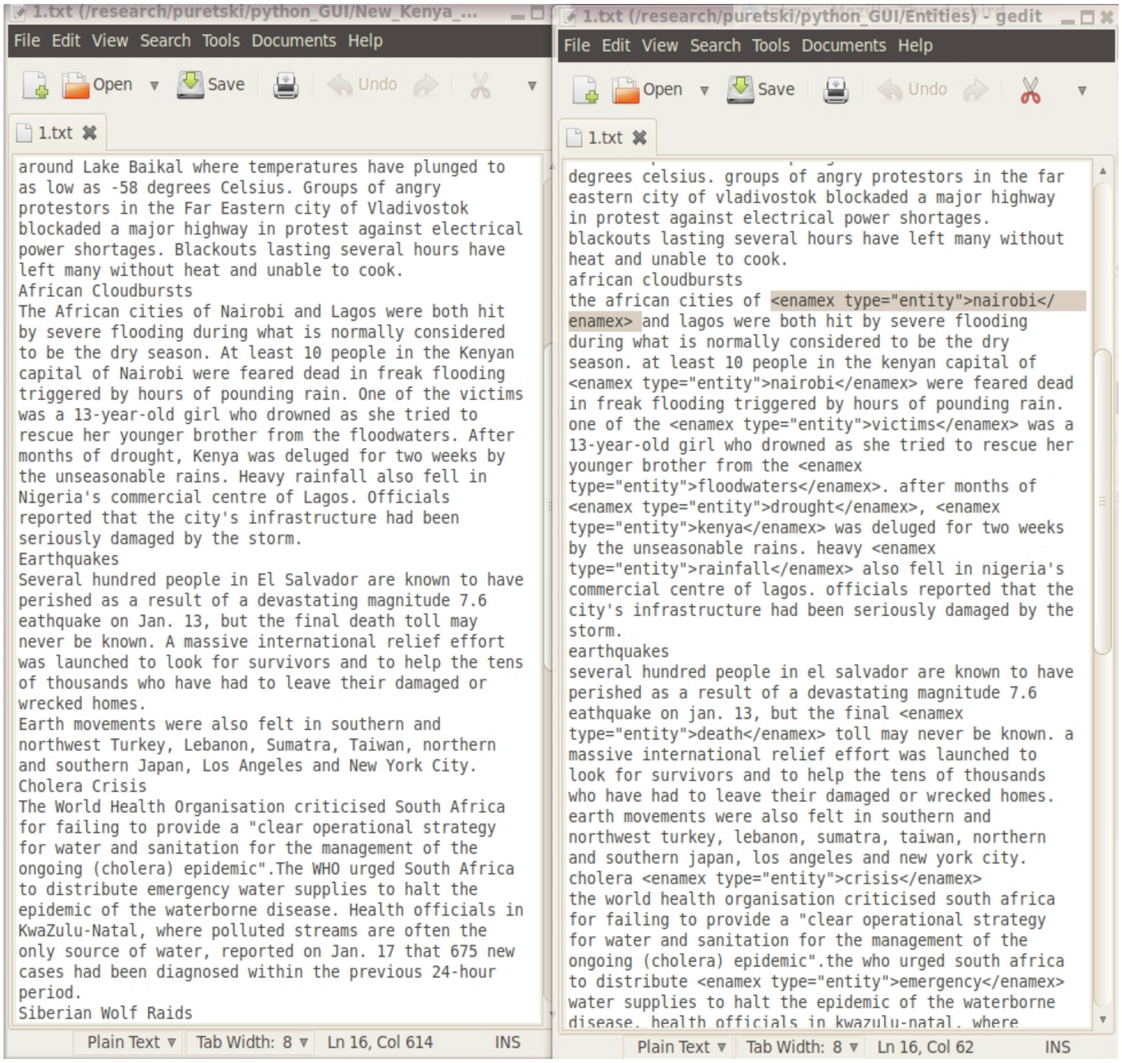
As stated in Section 3.2.2, including the ability to create an additional tensor dimension through tagging a subset of significant terms ("entities") is one of the major NTF improvements included in the integrated analysis environment. In order to take advantage of this feature, the user needs to provide a dictionary of entities. This file can contain whatever terms or phrases the user considers to be "entities" in the context of that particular study. For example, the file could contain all of the personal and geographical names contained in the dataset, or just some subset that is of interest to the user. The entity dictionary file should contain one entry per line, and the format should be as follows: *entity<tab>index*. The indexes should be unique integers, but other than this rule, the index creation is completely up to the user. The indexes do not have to be consecutive, nor do they have to be begin with "1". However, a simple way to generate these indexes is to create a spreadsheet with an entity column and an adjoining index column (a list of consecutive integers is very easy to generate in most spreadsheet software). The user would then be able to export this file in a separated-values format, using the <*tab*> character as the separator. Figure 13 shows a sample indexed entity list.


entity_list_w_idx (/research/puretski/python_GUI...
File Edit View Search Tools Documents Help
Open
Save
Undo
entity_list_w_idx
Plain Text
Tab Width: 8
Ln 6, Col 9
INS

```
Emergency       13
Emergency Assistance    14
Environment     15
Flood   16
Floods  17
Floodwaters     18
Food scarcity   19
Food Security   20
Forests 21
Global warming  22
Harvest 23
Humanitarian Aid        24
Humanitarian Assistance 25
Hunger  26
Hurricane       27
Industrialization       28
Jungle  29
Kenya   30
Livestock       31
Marooned        32
Marsabit        33
Marsh   34
Nairobi 35
Natural resources       36
Nongovernmental organizations   37
Ogiek   38
Plantations     39
Poverty 40
Rainfall        41
Rebuild 42
Red Crescent    43
Red Cross       44
Relief supplies 45
Shelters        46
South Asia      47
Starvation      48
Suffering       49
Sympathy        50
Tidal bores     51
Tidal surges    52
Tornadoes       53
USAID   54
Victims 55
Wajir   56
```

Figure 13: A sample indexed entity list may be used to add SGML-tagged entities to the dataset. Doing so allows the subsequent analysis environment steps to construct an addition (entity-based) dimension for the tensor. Adding this dimension can greatly enhance analysis and aid knowledge discovery.

Upon utilizing this feature of the analysis environment, a copy of the original dataset containing entity tags will be created. The following is an example of how an SGML entity tag will appear in the data, for an entity *John Brown*: *<enamex type =”entity”>John Brown</enamex>*. Figure 14 shows a comparison of the two styles of the same dataset file, one that does not include SGML-style tagged entities, and the other one that does.


1.txt (/research/puretski/python_GUI/New_Kenya_...
File Edit View Search Tools Documents Help
Open
Save
Undo
1.txt
Plain Text
Tab Width: 8
Ln 16, Col 614
INS
1.txt (/research/puretski/python_GUI/Entities) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
1.txt
Plain Text
Tab Width: 8
Ln 16, Col 62
INS

```
around Lake Baikal where temperatures have plunged to
as low as -58 degrees Celsius. Groups of angry
protestors in the Far Eastern city of Vladivostok
blockaded a major highway in protest against electrical
power shortages. Blackouts lasting several hours have
left many without heat and unable to cook.
African Cloudbursts
The African cities of Nairobi and Lagos were both hit
by severe flooding during what is normally considered
to be the dry season. At least 10 people in the Kenyan
capital of Nairobi were feared dead in freak flooding
triggered by hours of pounding rain. One of the victims
was a 13-year-old girl who drowned as she tried to
rescue her younger brother from the floodwaters. After
months of drought, Kenya was deluged for two weeks by
the unseasonable rains. Heavy rainfall also fell in
Nigeria's commercial centre of Lagos. Officials
reported that the city's infrastructure had been
seriously damaged by the storm.
Earthquakes
Several hundred people in El Salvador are known to have
perished as a result of a devastating magnitude 7.6
eathquake on Jan. 13, but the final death toll may
never be known. A massive international relief effort
was launched to look for survivors and to help the tens
of thousands who have had to leave their damaged or
wrecked homes.
Earth movements were also felt in southern and
northwest Turkey, Lebanon, Sumatra, Taiwan, northern
and southern Japan, Los Angeles and New York City.
Cholera Crisis
The World Health Organisation criticised South Africa
for failing to provide a "clear operational strategy
for water and sanitation for the management of the
ongoing (cholera) epidemic".The WHO urged South Africa
to distribute emergency water supplies to halt the
epidemic of the waterborne disease. Health officials in
KwaZulu-Natal, where polluted streams are often the
only source of water, reported on Jan. 17 that 675 new
cases had been diagnosed within the previous 24-hour
period.
Siberian Wolf Raids
```

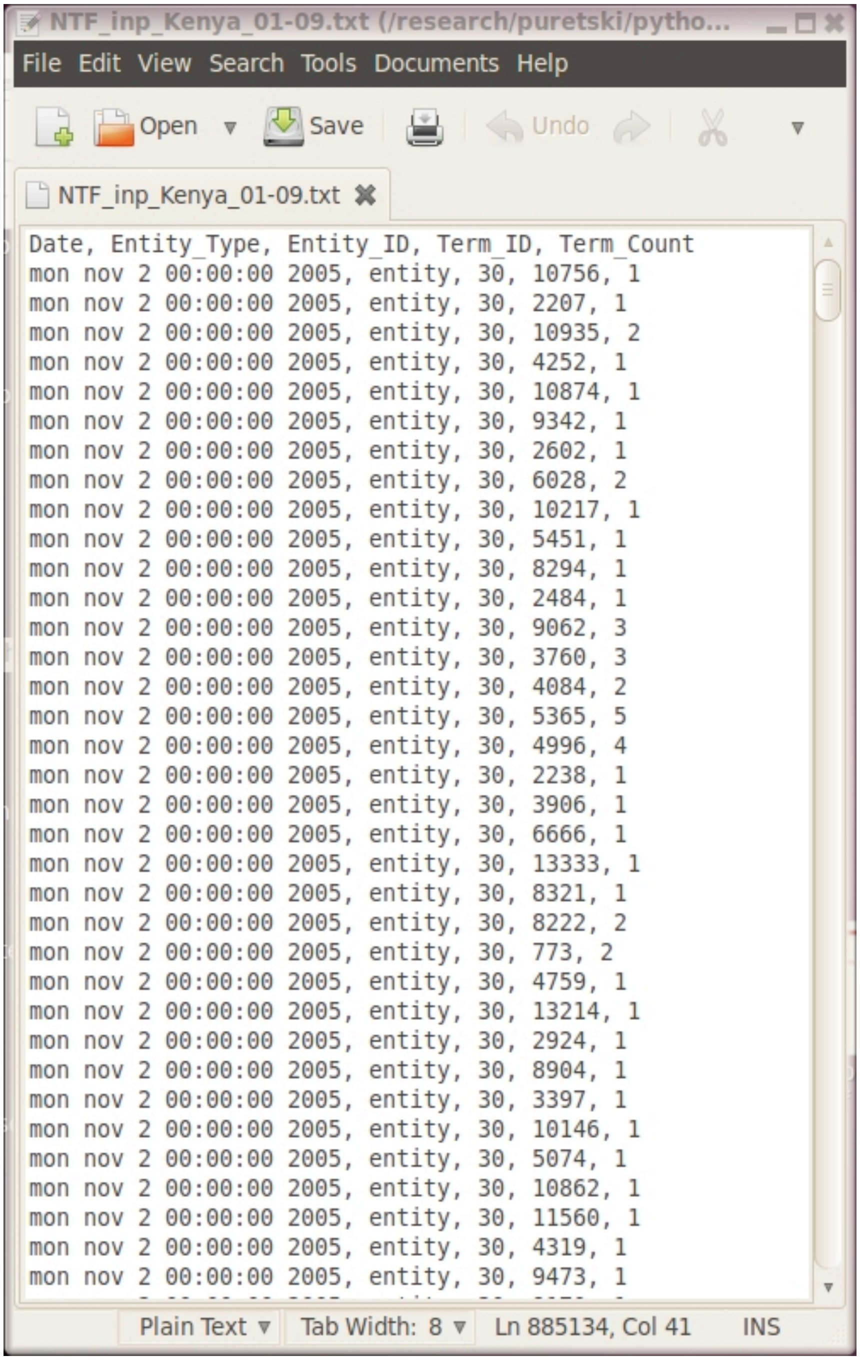
```
degrees celsius. groups of angry protestors in the far
eastern city of vladivostok blockaded a major highway
in protest against electrical power shortages.
blackouts lasting several hours have left many without
heat and unable to cook.
african cloudbursts
the african cities of <enamex type="entity">nairobi</
enamex> and lagos were both hit by severe flooding
during what is normally considered to be the dry
season. at least 10 people in the kenyan capital of
<enamex type="entity">nairobi</enamex> were feared dead
in freak flooding triggered by hours of pounding rain.
one of the <enamex type="entity">victims</enamex> was a
13-year-old girl who drowned as she tried to rescue her
younger brother from the <enamex
type="entity">floodwaters</enamex>. after months of
<enamex type="entity">drought</enamex>, <enamex
type="entity">kenya</enamex> was deluged for two weeks
by the unseasonable rains. heavy <enamex
type="entity">rainfall</enamex> also fell in nigeria's
commercial centre of lagos. officials reported that the
city's infrastructure had been seriously damaged by the
storm.
earthquakes
several hundred people in el salvador are known to have
perished as a result of a devastating magnitude 7.6
eathquake on jan. 13, but the final <enamex
type="entity">death</enamex> toll may never be known. a
massive international relief effort was launched to
look for survivors and to help the tens of thousands
who have had to leave their damaged or wrecked homes.
earth movements were also felt in southern and
northwest turkey, lebanon, sumatra, taiwan, northern
and southern japan, los angeles and new york city.
cholera <enamex type="entity">crisis</enamex>
the world health organisation criticised south africa
for failing to provide a "clear operational strategy
for water and sanitation for the management of the
ongoing (cholera) epidemic".the who urged south africa
to distribute <enamex type="entity">emergency</enamex>
water supplies to halt the epidemic of the waterborne
disease. health officials in kwazulu-natal, where
```

Figure 14: A side-by-side comparison of the two input dataset file styles. On the left, the original dataset file containg no tagged entities. SGML-style entity tags have been added to the file on the right using the entity tagging feature of the analysis environment. A newly-tagged entity, Nairobi *is highlighted in the window on the right.*

4.3 Generating an NTF Input File Based on a Dataset

The analysis environment is capable of generating an NTF input file based on a dataset regardless of whether it contains tagged entities. To utilize this function, the user simply needs to select a directory containing all of the dataset's files. Each file should contain only one element of the dataset (an article, paper, email message, chat log, or anything else that for the purposes of that particularly study may be considered "an element of the dataset").

The software will then request an output file name from the user, and proceed to create the file. Once it has been created, the user may proceed to the next step and run the NTF algorithm using it as input. A typical line of NTF input file will have the following format: *Date, Entity Type, Entity ID, Term ID, Term Count.* The Entity Type field is not part of the actual NTF process, it is used mostly for human validation of correctness. The term counts are not weighted or scaled at this point in the analysis process, although the feature described in the next section allows the user to adjust some/all of these. Figure 15 shows a small portion of a sample NTF output file.


NTF_inp_Kenya_01-09.txt (/research/puretski/pytho...
File Edit View Search Tools Documents Help
Open
Save
Undo
NTF_inp_Kenya_01-09.txt
Plain Text
Tab Width: 8
Ln 885134, Col 41
INS

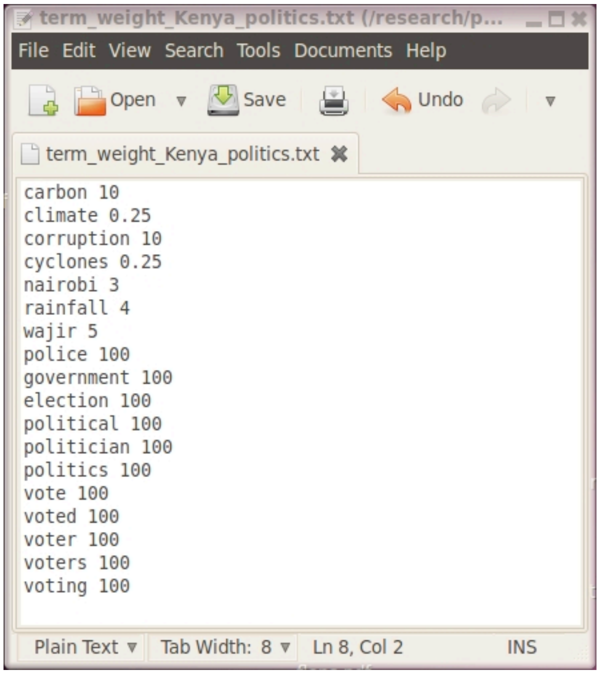
```
Date, Entity_Type, Entity_ID, Term_ID, Term_Count
mon nov 2 00:00:00 2005, entity, 30, 10756, 1
mon nov 2 00:00:00 2005, entity, 30, 2207, 1
mon nov 2 00:00:00 2005, entity, 30, 10935, 2
mon nov 2 00:00:00 2005, entity, 30, 4252, 1
mon nov 2 00:00:00 2005, entity, 30, 10874, 1
mon nov 2 00:00:00 2005, entity, 30, 9342, 1
mon nov 2 00:00:00 2005, entity, 30, 2602, 1
mon nov 2 00:00:00 2005, entity, 30, 6028, 2
mon nov 2 00:00:00 2005, entity, 30, 10217, 1
mon nov 2 00:00:00 2005, entity, 30, 5451, 1
mon nov 2 00:00:00 2005, entity, 30, 8294, 1
mon nov 2 00:00:00 2005, entity, 30, 2484, 1
mon nov 2 00:00:00 2005, entity, 30, 9062, 3
mon nov 2 00:00:00 2005, entity, 30, 3760, 3
mon nov 2 00:00:00 2005, entity, 30, 4084, 2
mon nov 2 00:00:00 2005, entity, 30, 5365, 5
mon nov 2 00:00:00 2005, entity, 30, 4996, 4
mon nov 2 00:00:00 2005, entity, 30, 2238, 1
mon nov 2 00:00:00 2005, entity, 30, 3906, 1
mon nov 2 00:00:00 2005, entity, 30, 6666, 1
mon nov 2 00:00:00 2005, entity, 30, 13333, 1
mon nov 2 00:00:00 2005, entity, 30, 8321, 1
mon nov 2 00:00:00 2005, entity, 30, 8222, 2
mon nov 2 00:00:00 2005, entity, 30, 773, 2
mon nov 2 00:00:00 2005, entity, 30, 4759, 1
mon nov 2 00:00:00 2005, entity, 30, 13214, 1
mon nov 2 00:00:00 2005, entity, 30, 2924, 1
mon nov 2 00:00:00 2005, entity, 30, 8904, 1
mon nov 2 00:00:00 2005, entity, 30, 3397, 1
mon nov 2 00:00:00 2005, entity, 30, 10146, 1
mon nov 2 00:00:00 2005, entity, 30, 5074, 1
mon nov 2 00:00:00 2005, entity, 30, 10862, 1
mon nov 2 00:00:00 2005, entity, 30, 11560, 1
mon nov 2 00:00:00 2005, entity, 30, 4319, 1
mon nov 2 00:00:00 2005, entity, 30, 9473, 1
```

Figure 15: A small portion of an NTF input file. Other features of the analysis environment construct allow the user to adjust term weight, construct a tensor, and run the NTF algorithm, using this file as input.

4.4 Adjusting Term Weights in an NTF Input File

One of the key features of the analysis environment is the ability to incorporate additional externally sourced information into the dataset-based tensor. There is a wide variety of possible motivations and reasons for doing so, and a variety of sources for the additional information. For instance, a user could have prior knowledge regarding trustworthiness of certain terms (which might correspond to person or organization names). The user could then utilize this feature to quickly and easily decrease the weight of these terms within the tensor, lessening their impact on the overall analysis results.

Alternatively, it is possible that in the course of a particular study, the analyst may develop a strong interest in a particular subset of terms. Frequently, these subsets are based on some well-established ontology — a domain-specific collection of knowledge, describing various concepts and relationships between them. For example, if the user is particularly interested in events related to unrest in a particular geographical area, the user can quickly and easily generate a number of term weights files based on an unrest-related ontology. Each file could contain a different interest model. The user could then utilize this feature of the analysis environment to create a number of different tensor models, which would differ only in which term weights had been adjusted. In order to utilize this feature, a term weights file is necessary. A sample file is shown in Figure 16.


term_weight_Kenya_politics.txt (/research/p...
File Edit View Search Tools Documents Help
Open
Save
Undo
term_weight_Kenya_politics.txt
Plain Text
Tab Width: 8
Ln 8, Col 2
INS

```
carbon 10
climate 0.25
corruption 10
cyclones 0.25
nairobi 3
rainfall 4
wajir 5
police 100
government 100
election 100
political 100
politician 100
politics 100
vote 100
voted 100
voter 100
voters 100
voting 100
```

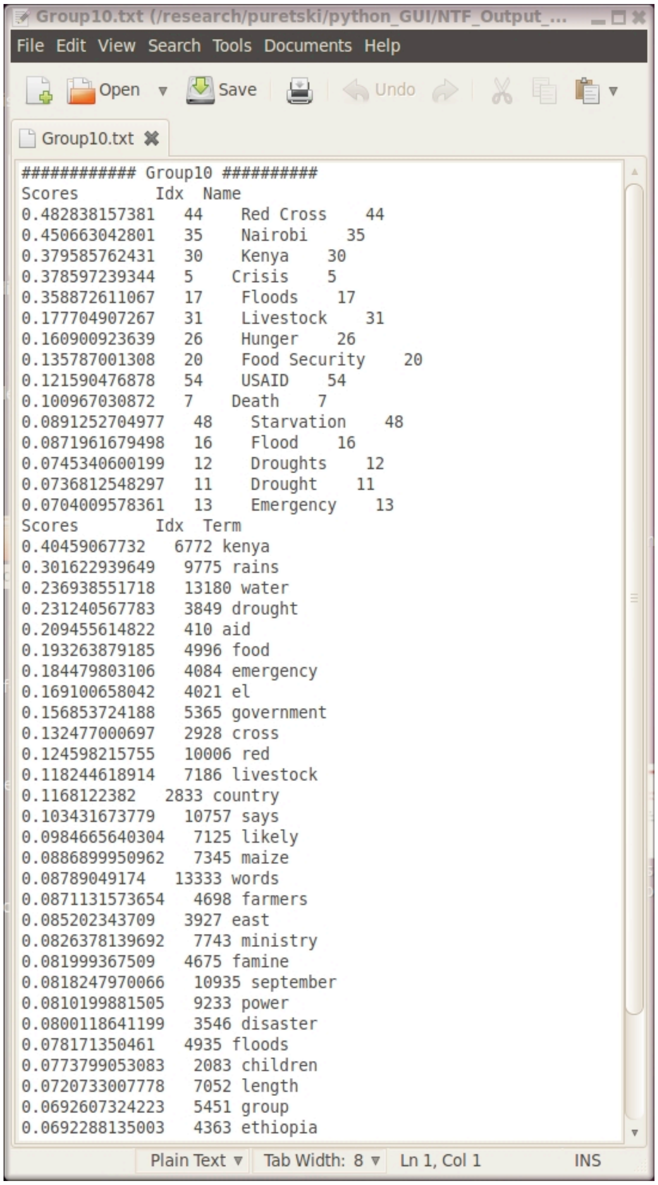
Figure 16: A sample user-created term weights file that can be used to selectively manipulate the term weights in the tensor. Reasons for utilizing this feature may vary, and it is an optional part of the analysis process that has a potential to enhance the analysis process by allowing the user to generate and evaluate different models.

4.5 Executing the NTF Algorithm as Implemented in Python

All of the features described in the previous sections of this chapter are meant to facilitate the execution of the nonnegative tensor factorization algorithm. While the features of analysis environment described there are important and enhance the potential effectiveness of the environment as it relates to knowledge discovery, the NTF step is by far the most significant. In order to utilize this feature, the user will need to provide an NTF input file (Section 4.3). Whether this file contains tagged entities or not does not matter, however it should be noted that the inclusion of tagged entities may greatly enhance the analysis process. The additional dimension that can be constructed based on the tagged entities may allow for the establishment of connections that would not have otherwise been revealed.

The user may choose the number of desired NTF output features by entering that number into the text field shown in Figure 10, which appears at the beginning of this chapter. The NTF algorithm will then attempt to create that number of output groups, each described in a separate file and labeled *GroupX.txt*, where *X* is the arbitrarily assigned group number. It should be noted that the group number does not carry any significance. For example, *Group1.txt* does not necessarily describe a feature of the data that is more interesting or important than that described by *Group20.txt*. This is in large part due to the highly subjective and context-dependent nature of concepts such as "interesting" and "important". These concepts depend on the nature and the context of the analysis, the nature of the dataset and the problem, as well as the user's personal opinions and biases. It is impossible to quantify all of these highly subjective and unstable variables to incorporate them into a deterministic computer algorithm.

When entities are included in the dataset, each NTF output group file includes a list of top 15 most relevant entities and top 35 most relevant terms. The entities and terms are ranked in accordance with an internally generated relevance score. The score attempts to quantify the term's relative importance to this particular feature. As shown in Figure 17, both the terms and the entities are listed in descending order of importance in an NTF output group file. However, it is again important to remember that this quantification is just an attempt at reflecting subjective, human judgment, and may not reflect the opinions of a human analyst precisely.


Group10.txt (/research/puretski/python_GUI/NTF_Output_...
File Edit View Search Tools Documents Help
Open
Save
Undo
Group10.txt
Plain Text
Tab Width: 8
Ln 1, Col 1
INS

```
############ Group10 ##########
Scores        Idx  Name
0.482838157381   44    Red Cross    44
0.450663042801   35    Nairobi    35
0.379585762431   30    Kenya    30
0.378597239344   5    Crisis    5
0.358872611067   17    Floods    17
0.177704907267   31    Livestock    31
0.160900923639   26    Hunger    26
0.135787001308   20    Food Security    20
0.121590476878   54    USAID    54
0.100967030872   7    Death    7
0.0891252704977   48    Starvation    48
0.0871961679498   16    Flood    16
0.0745340600199   12    Droughts    12
0.0736812548297   11    Drought    11
0.0704009578361   13    Emergency    13
Scores        Idx  Term
0.40459067732   6772 kenya
0.301622939649   9775 rains
0.236938551718   13180 water
0.231240567783   3849 drought
0.209455614822   410 aid
0.193263879185   4996 food
0.184479803106   4084 emergency
0.169100658042   4021 el
0.156853724188   5365 government
0.132477000697   2928 cross
0.124598215755   10006 red
0.118244618914   7186 livestock
0.1168122382   2833 country
0.103431673779   10757 says
0.0984665640304   7125 likely
0.0886899950962   7345 maize
0.08789049174   13333 words
0.0871131573654   4698 farmers
0.085202343709   3927 east
0.0826378139692   7743 ministry
0.081999367509   4675 famine
0.0818247970066   10935 september
0.0810199881505   9233 power
0.0800118641199   3546 disaster
0.078171350461   4935 floods
0.0773799053083   2083 children
0.0720733007778   7052 length
0.0692607324223   5451 group
0.0692288135003   4363 ethiopia
```

Figure 17: A sample NTF output file. This file was generated by the Python-based analysis environment using the NTF-PARAFAC algorithm. The algorithm was applied to a dataset of news articles about Kenya, covering the years of 2001-2009. As can be seen in this figure, this NTF output feature describes a drought-related theme in the dataset. Terms such as rains, water, drought, emergency, *and* aid *appear near the top of the terms list.*

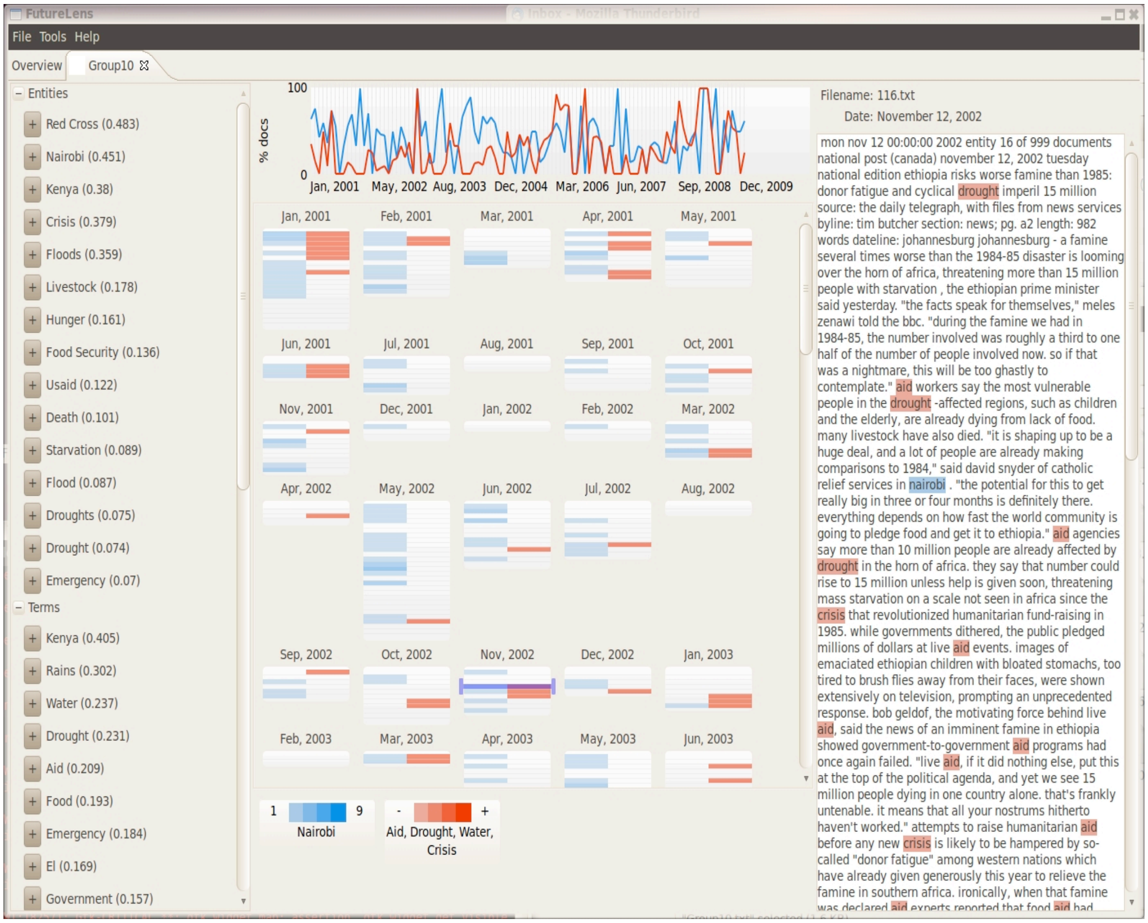
The core of the Python NTF implementation integrated into the analysis environment was created by Papa Diaw in the Summer of 2010. The implementation was created as part of his work to obtain a Master of Science in Computer Science degree at the University of Tennessee, Knoxville. The project used a Matlab® NTF implementation of the PARAFAC algorithm as a starting point [3]. The free and open-source Python implementation provides an alternative to the commercial Matlab® version, in the hope that this will allow for much greater distribution, use, modification, and improvement of this approach to knowledge discovery. Python as a language has a number of other advantages, namely that it is object-oriented, extensible, open source and freely available, and extremely easy to read and write. Python is capable of running on essentially any of the major operating systems, including various flavors of Unix and Linux, as well as all modern versions of Windows and Apple's OS X. As mentioned previously, it is free and its complete source code is available in its distributions [22].

The openness and easy to use qualities of Python have lead to the development of a number of freely distributed and open source libraries that are highly useful in Scientific Computing. These include such well-known and popular packages as Numpy and Scipy [19, 20]. Numpy in particular was instrumental to this project, as the factorization code uses a number of Numpy-based structures and functions.

4.6 Continuing Analysis with FutureLens

As demonstrated in Figure 17 in Section 4.5, the output of the NTF algorithm is simply a series of lists of terms, each list describing some feature of the dataset. Further human analysis and knowledge discovery may be difficult to accomplish based on nothing more than a list of terms. This was the motivation for the creation of the visual NTF output analysis tool called FutureLens [9]. This dissertation integrates FutureLens into the overall analysis environment, and adds an important automated NTF output labeling feature that greatly enhances and speeds up the analysis process.

FutureLens allows the user to import the output of the NTF algorithm and analyze it further, while connecting it back to the original dataset. The user has the option of loading any number of NTF output groups at the same time, and in any combination. Each group is allocated its own separate tab in the graphical user interface. The button labeled with a "+" symbol that appears to the left of each term may be used to add that term to the main FutureLens display. Once a term has been added, FutureLens will plot that term's temporal distribution summary in the top-center display panel. This allows the user to get a quick impression of how the term is used throughout the dataset, perhaps taking note of peak usage times. FutureLens also locates and color-codes the term within the dataset's document space. This is shown in the central display panel, where every line segment is clickable and corresponds to a single document within the dataset. If the user clicks on one of these line segments, the corresponding document will be displayed in the panel on the right. The displayed document will include the selected terms, highlighted and color-coded in accordance with the color legend displayed in the bottom-enter frame. Figure 18 demonstrates all of the features described above.



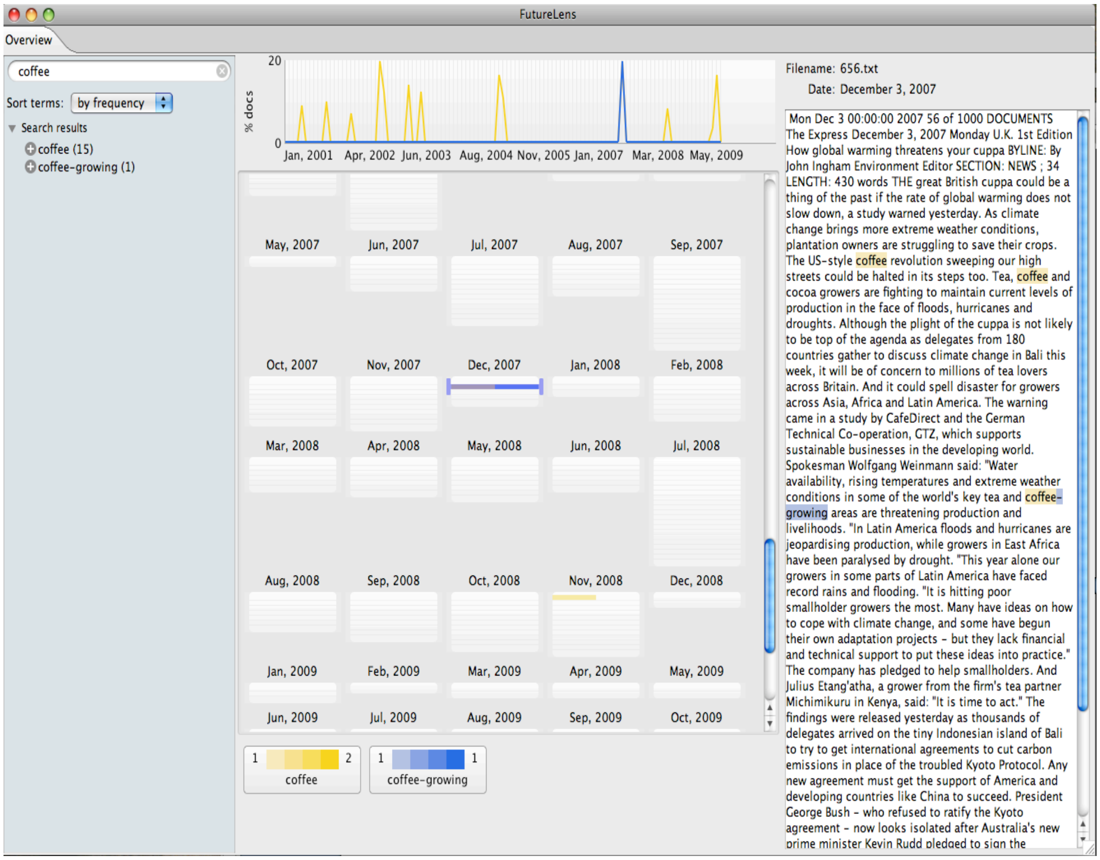
Filename: 116.txt

Date: November 12, 2002

mon nov 12 00:00:00 2002 entity 16 of 999 documents national post (canada) november 12, 2002 tuesday national edition ethiopia risks worse famine than 1985: donor fatigue and cyclical drought imperil 15 million source: the daily telegraph, with files from news services byline: tim butcher section: news; pg. a2 length: 982 words dateline: johannesburg johannesburg - a famine several times worse than the 1984-85 disaster is looming over the horn of africa, threatening more than 15 million people with starvation , the ethiopian prime minister said yesterday. "the facts speak for themselves," meles zenawi told the bbc. "during the famine we had in 1984-85, the number involved was roughly a third to one half of the number of people involved now. so if that was a nightmare, this will be too ghastly to contemplate." aid workers say the most vulnerable people in the drought -affected regions, such as children and the elderly, are already dying from lack of food. many livestock have also died. "it is shaping up to be a huge deal, and a lot of people are already making comparisons to 1984," said david snyder of catholic relief services in nairobi . "the potential for this to get really big in three or four months is definitely there. everything depends on how fast the world community is going to pledge food and get it to ethiopia." aid agencies say more than 10 million people are already affected by drought in the horn of africa. they say that number could rise to 15 million unless help is given soon, threatening mass starvation on a scale not seen in africa since the crisis that revolutionized humanitarian fund-raising in 1985. while governments dithered, the public pledged millions of dollars at live aid events. images of emaciated ethiopian children with bloated stomachs, too tired to brush flies away from their faces, were shown extensively on television, prompting an unprecedented response. bob geldof, the motivating force behind live aid, said the news of an imminent famine in ethiopia showed government-to-government aid programs had once again failed. "live aid, if it did nothing else, put this at the top of the political agenda, and yet we see 15 million people dying in one country alone. that's frankly untenable. it means that all your nostrums hitherto haven't worked." attempts to raise humanitarian aid before any new crisis is likely to be hampered by so-called "donor fatigue" among western nations which have already given generously this year to relieve the famine in southern africa. ironically, when that famine was declared aid experts reported that food aid had

Figure 18: FutureLens allows the user to analyze NTF output results in depth by tracking the constituent NTF group terms through the dataset.

It is important to note that FutureLens may be highty useful as a text analysis tool even without NTF output results, since it functions quite effectively as stand-alone software. For instance, the user has the ability to load a dataset into FutureLens independently of NTF output groups. Once a dataset is loaded, the user may search for particular terms and track their occurrence temporally through the dataset (if the dataset contains SGML-style date tags, which can be added using the feature of the analysis environment described in Section 4.1). It is also possible to display all of the terms contained within the dataset (excluding the ones on a user-defined stop words list), sorted either alphabetically or by frequency. FutureLens displays the terms thirty at a time, providing the user with “Next Page” and “Previous Page” buttons. These features of FutureLens are demonstrated in Figures 19 and 20 below.



coffee

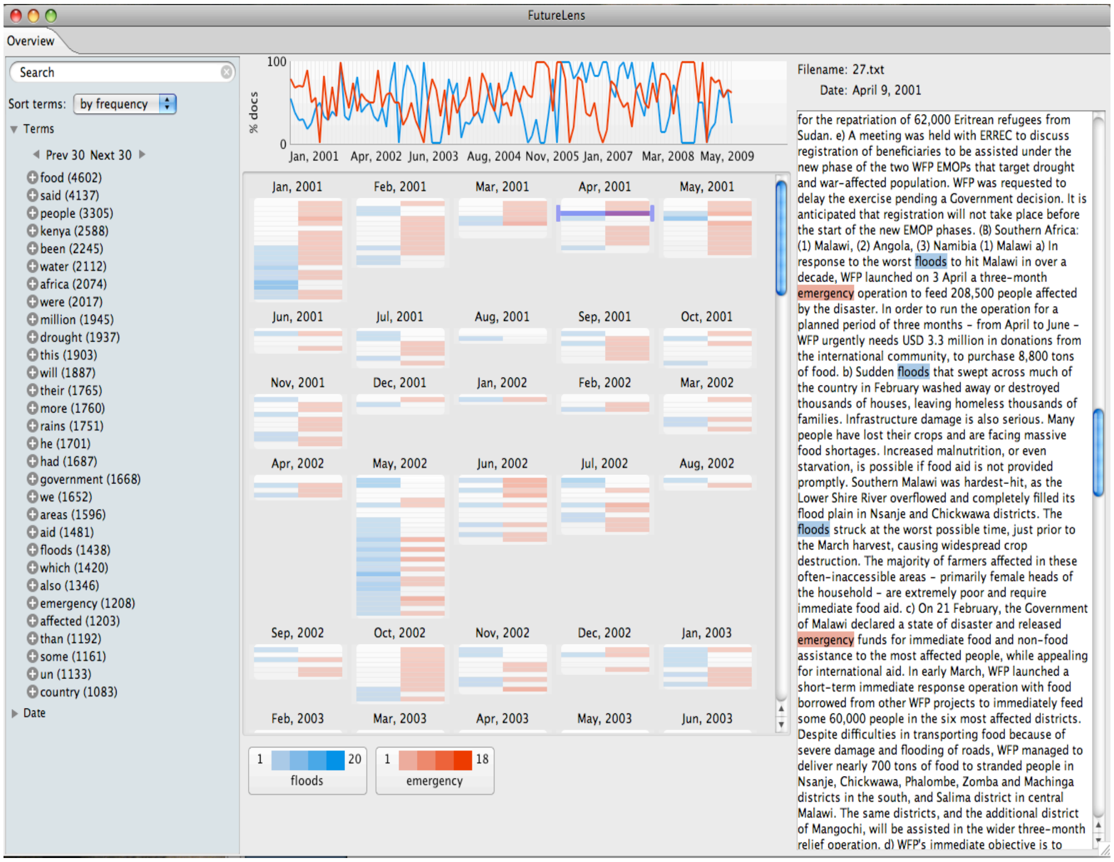
Sort terms: by frequency

Filename: 656.txt

Date: December 3, 2007

Mon Dec 3 00:00:00 2007 56 of 1000 DOCUMENTS The Express December 3, 2007 Monday U.K. 1st Edition How global warming threatens your cuppa BYLINE: By John Ingham Environment Editor SECTION: NEWS ; 34 LENGTH: 430 words THE great British cuppa could be a thing of the past if the rate of global warming does not slow down, a study warned yesterday. As climate change brings more extreme weather conditions, plantation owners are struggling to save their crops. The US-style coffee revolution sweeping our high streets could be halted in its steps too. Tea, coffee and cocoa growers are fighting to maintain current levels of production in the face of floods, hurricanes and droughts. Although the plight of the cuppa is not likely to be top of the agenda as delegates from 180 countries gather to discuss climate change in Bali this week, it will be of concern to millions of tea lovers across Britain. And it could spell disaster for growers across Asia, Africa and Latin America. The warning came in a study by CafeDirect and the German Technical Co-operation, GTZ, which supports sustainable businesses in the developing world. Spokesman Wolfgang Weinmann said: "Water availability, rising temperatures and extreme weather conditions in some of the world's key tea and coffee-growing areas are threatening production and livelihoods. "In Latin America floods and hurricanes are jeopardising production, while growers in East Africa have been paralysed by drought. "This year alone our growers in some parts of Latin America have faced record rains and flooding. "It is hitting poor smallholder growers the most. Many have ideas on how to cope with climate change, and some have begun their own adaptation projects - but they lack financial and technical support to put these ideas into practice." The company has pledged to help smallholders. And Julius Etang'atha, a grower from the firm's tea partner Michimikuru in Kenya, said: "It is time to act." The findings were released yesterday as thousands of delegates arrived on the tiny Indonesian island of Bali to try to get international agreements to cut carbon emissions in place of the troubled Kyoto Protocol. Any new agreement must get the support of America and developing countries like China to succeed. President George Bush - who refused to ratify the Kyoto agreement - now looks isolated after Australia's new prime minister Kevin Rudd pledged to sign the

Figure 19: FutureLens may be used as a robust text search tool. Here, the user's search for the term coffee *provides two results:* coffee *and* coffee-growing. *Selecting both terms allows the user to quickly visualize the terms' distribution in the dataset. The user is also able to quickly access the dataset elements containing these terms, which provides much-needed context.*



Sort terms: by frequency

- food (4602)
- said (4137)
- people (3305)
- kenya (2588)
- been (2245)
- water (2112)
- africa (2074)
- were (2017)
- million (1945)
- drought (1937)
- this (1903)
- will (1887)
- their (1765)
- more (1760)
- rains (1751)
- he (1701)
- had (1687)
- government (1668)
- we (1652)
- areas (1596)
- aid (1481)
- floods (1438)
- which (1420)
- also (1346)
- emergency (1208)
- affected (1203)
- than (1192)
- some (1161)
- un (1133)
- country (1083)

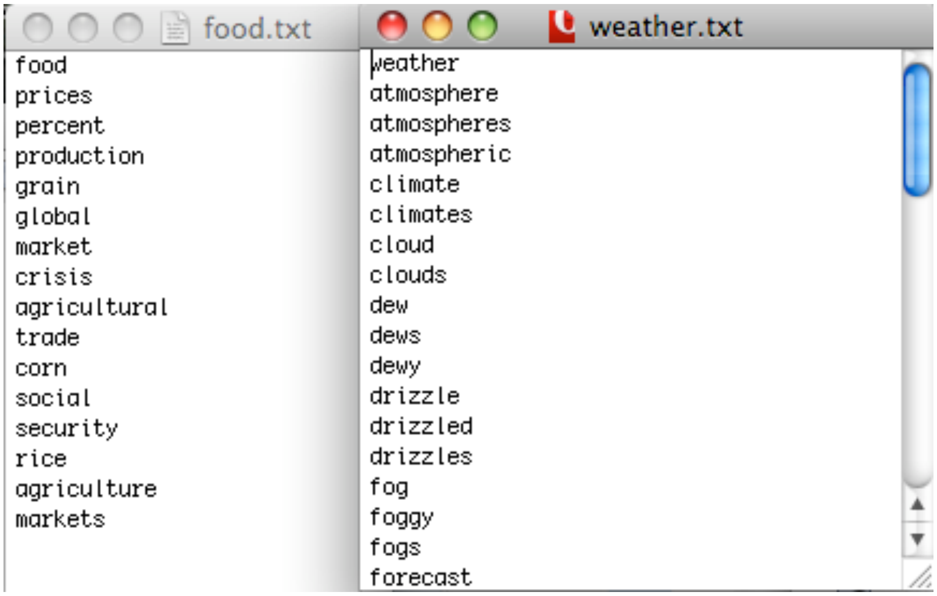
Filename: 27.txt

Date: April 9, 2001

for the repatriation of 62,000 Eritrean refugees from Sudan. e) A meeting was held with ERREC to discuss registration of beneficiaries to be assisted under the new phase of the two WFP EMOPs that target drought and war-affected population. WFP was requested to delay the exercise pending a Government decision. It is anticipated that registration will not take place before the start of the new EMOP phases. (B) Southern Africa: (1) Malawi, (2) Angola, (3) Namibia (1) Malawi a) In response to the worst floods to hit Malawi in over a decade, WFP launched on 3 April a three-month emergency operation to feed 208,500 people affected by the disaster. In order to run the operation for a planned period of three months – from April to June – WFP urgently needs USD 3.3 million in donations from the international community, to purchase 8,800 tons of food. b) Sudden floods that swept across much of the country in February washed away or destroyed thousands of houses, leaving homeless thousands of families. Infrastructure damage is also serious. Many people have lost their crops and are facing massive food shortages. Increased malnutrition, or even starvation, is possible if food aid is not provided promptly. Southern Malawi was hardest-hit, as the Lower Shire River overflowed and completely filled its flood plain in Nsanje and Chickwawa districts. The floods struck at the worst possible time, just prior to the March harvest, causing widespread crop destruction. The majority of farmers affected in these often-inaccessible areas – primarily female heads of the household – are extremely poor and require immediate food aid. c) On 21 February, the Government of Malawi declared a state of disaster and released emergency funds for immediate food and non-food assistance to the most affected people, while appealing for international aid. In early March, WFP launched a short-term immediate response operation with food borrowed from other WFP projects to immediately feed some 60,000 people in the six most affected districts. Despite difficulties in transporting food because of severe damage and flooding of roads, WFP managed to deliver nearly 700 tons of food to stranded people in Nsanje, Chickwawa, Phalombe, Zomba and Machinga districts in the south, and Salima district in central Malawi. The same districts, and the additional district of Mangochi, will be assisted in the wider three-month relief operation. d) WFP's immediate objective is to

Figure 20: FutureLens allows the user to diplay the entire list of terms contained in the dataset, sorted either by frequency of alphabetically. As shown in this figure, the terms are displayed 30 per page.

Automated NTF output labeling is a significant addition to FutureLens that was made as part of its integration into the analysis environment. Automated NTF group labeling has the ability to speed up the analysis process by allowing the user to quickly focus attention of most relevant groups. Naturally, relevance and relative importance are highly subjective and depend on the exact nature of the user's particular research study. It is therefore highly beneficial to allow easily customizable, plain-text files to serve as category descriptors. The format of these files is extremely straightforward, as shown in Figure 21.

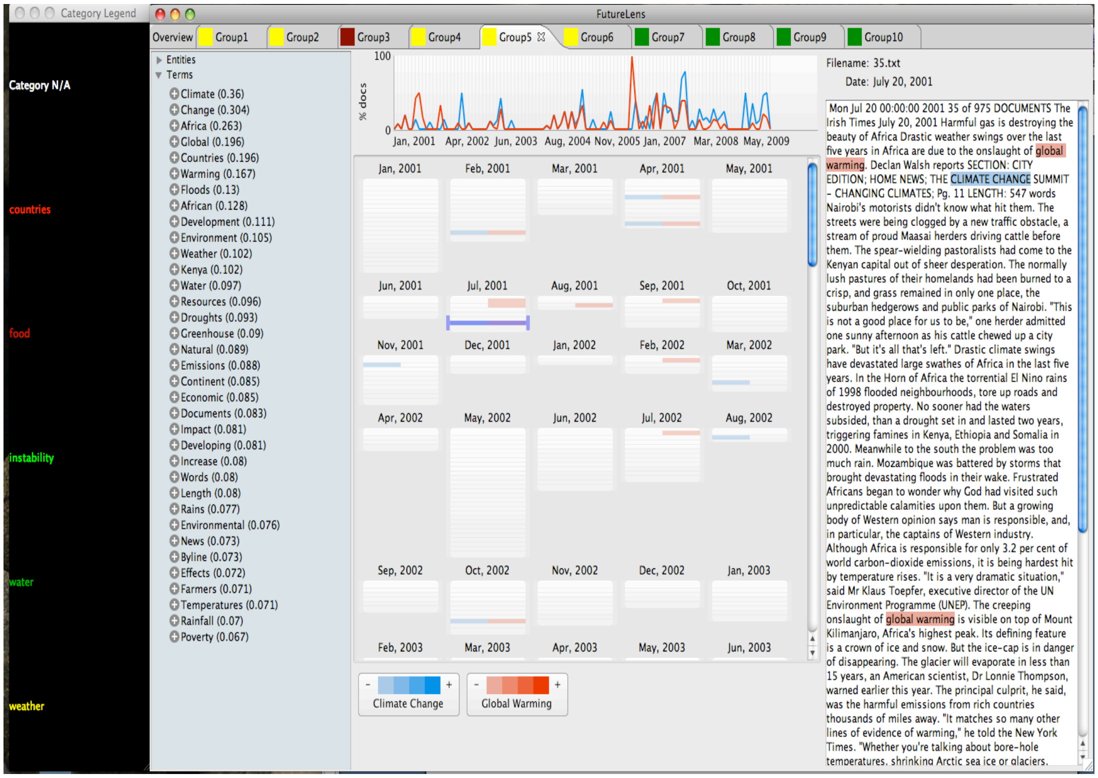


```
food
prices
percent
production
grain
global
market
crisis
agricultural
trade
corn
social
security
rice
agriculture
markets
```

```
weather
atmosphere
atmospheres
atmospheric
climate
climates
cloud
clouds
dew
dews
dewy
drizzle
drizzled
drizzles
fog
foggy
fogs
forecast
```

Figure 21: Sample category description files that are required to use FutureLens's automated NTF output labeling feature. The first term in a file is used as the category label. The number of terms in each file may be different--there is no required minimum number of a maximum limit.

The category descriptor files can be very easily created and/or modified by the user, in accordance with the exact nature of the goals and desired focus of each particular study or model. Any number of categories is possible, but experience has shown that it is generally more helpful to keep the number relatively small. After the categories have been loaded, FutureLens compares the terms constituting each NTF output group with the terms found in the category descriptor files. The category with the highest number of matches becomes the label for that NTF group. Figures 22 and 23 demonstrate how this feature may be highly useful to furthering text analysis. In this example, the user can immediately see that of the ten NTF output groups loaded into FutureLens, five have been labeled as belonging to the *weather* category (light yellow), four have been labeled under the *water* category (dark green), and one has been labeled as belonging to the *food* category (dark red). It should be noted that the category labels also appear as a tool-tip if the user places the mouse cursor over GUI tab containing the NTF group file name. This may be helpful for color-blind users, and users who have closed the legend window that appears on the left in Figures 22 and 23.

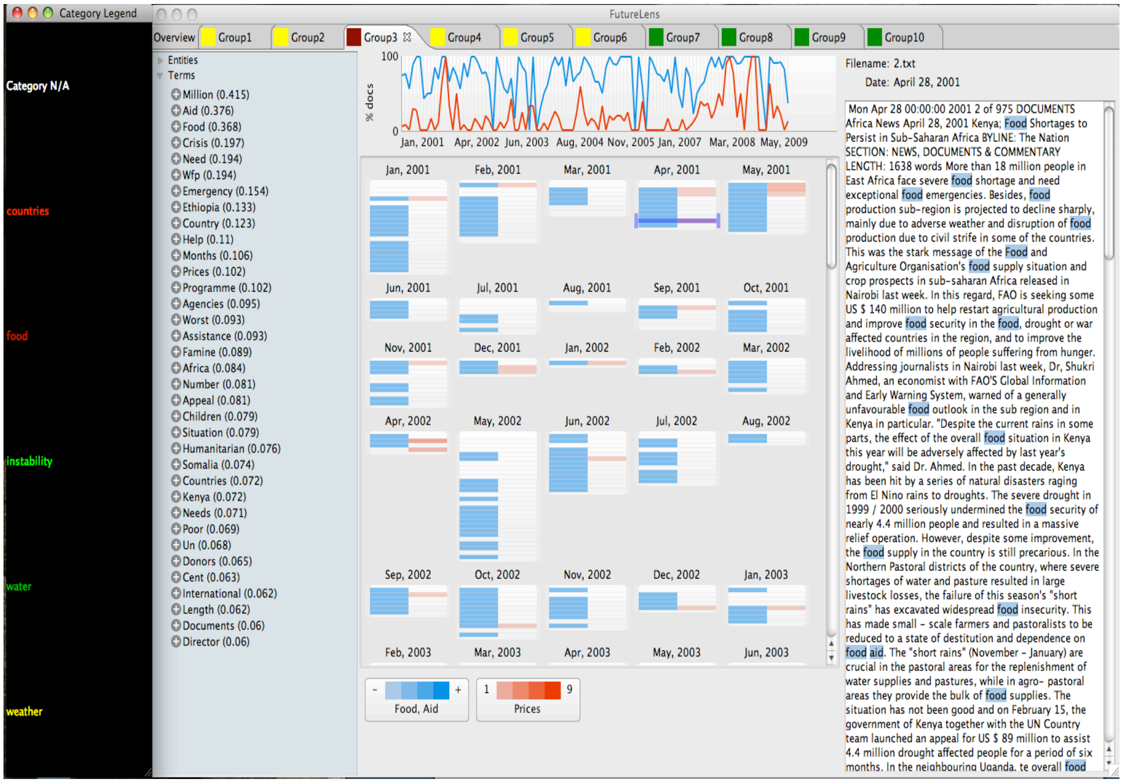


Entities
Terms
Climate (0.36)
Change (0.304)
Africa (0.263)
Global (0.196)
Countries (0.196)
Warming (0.167)
Floods (0.13)
African (0.128)
Development (0.111)
Environment (0.105)
Weather (0.102)
Kenya (0.102)
Water (0.097)
Resources (0.096)
Droughts (0.093)
Greenhouse (0.09)
Natural (0.089)
Emissions (0.088)
Continent (0.085)
Economic (0.085)
Documents (0.083)
Impact (0.081)
Developing (0.081)
Increase (0.08)
Words (0.08)
Length (0.08)
Rains (0.077)
Environmental (0.076)
News (0.073)
Byline (0.073)
Effects (0.072)
Farmers (0.071)
Temperatures (0.071)
Rainfall (0.07)
Poverty (0.067)

Filename: 35.txt
Date: July 20, 2001

Mon Jul 20 00:00:00 2001 35 of 975 DOCUMENTS The Irish Times July 20, 2001 Harmful gas is destroying the beauty of Africa Drastic weather swings over the last five years in Africa are due to the onslaught of global warming. Declan Walsh reports SECTION: CITY EDITION; HOME NEWS; THE CLIMATE CHANGE SUMMIT - CHANGING CLIMATES; Pg. 11 LENGTH: 547 words Nairobi's motorists didn't know what hit them. The streets were being clogged by a new traffic obstacle, a stream of proud Maasai herders driving cattle before them. The spear-wielding pastoralists had come to the Kenyan capital out of sheer desperation. The normally lush pastures of their homelands had been burned to a crisp, and grass remained in only one place, the suburban hedgerows and public parks of Nairobi. "This is not a good place for us to be," one herder admitted one sunny afternoon as his cattle chewed up a city park. "But it's all that's left." Drastic climate swings have devastated large swathes of Africa in the last five years. In the Horn of Africa the torrential El Nino rains of 1998 flooded neighbourhoods, tore up roads and destroyed property. No sooner had the waters subsided, than a drought set in and lasted two years, triggering famines in Kenya, Ethiopia and Somalia in 2000. Meanwhile to the south the problem was too much rain. Mozambique was battered by storms that brought devastating floods in their wake. Frustrated Africans began to wonder why God had visited such unpredictable calamities upon them. But a growing body of Western opinion says man is responsible, and, in particular, the captains of Western industry. Although Africa is responsible for only 3.2 per cent of world carbon-dioxide emissions, it is being hardest hit by temperature rises. "It is a very dramatic situation," said Mr Klaus Toepfer, executive director of the UN Environment Programme (UNEP). The creeping onslaught of global warming is visible on top of Mount Kilimanjaro, Africa's highest peak. Its defining feature is a crown of ice and snow. But the ice-cap is in danger of disappearing. The glacier will evaporate in less than 15 years, an American scientist, Dr Lonnie Thompson, warned earlier this year. The principal culprit, he said, was the harmful emissions from rich countries thousands of miles away. "It matches so many other lines of evidence of warming," he told the New York Times. "Whether you're talking about bore-hole temperatures, shrinking Arctic sea ice or glaciers,

Figure 22: The automated NTF results labeling feature can be extremely helpful by directing the user's attention to particular NTF output files. One of the weather-*themed groups is selected here. The user was able to quickly and easily construct phrases* climate change *and* global warming *using the terms contained in this group.*



Filename: 2.txt

Date: April 28, 2001

Mon Apr 28 00:00:00 2001 2 of 975 DOCUMENTS Africa News April 28, 2001 Kenya; Food Shortages to Persist in Sub-Saharan Africa BYLINE: The Nation SECTION: NEWS, DOCUMENTS & COMMENTARY LENGTH: 1638 words More than 18 million people in East Africa face severe food shortage and need exceptional food emergencies. Besides, food production sub-region is projected to decline sharply, mainly due to adverse weather and disruption of food production due to civil strife in some of the countries. This was the stark message of the Food and Agriculture Organisation's food supply situation and crop prospects in sub-saharan Africa released in Nairobi last week. In this regard, FAO is seeking some US $ 140 million to help restart agricultural production and improve food security in the food, drought or war affected countries in the region, and to improve the livelihood of millions of people suffering from hunger. Addressing journalists in Nairobi last week, Dr, Shukri Ahmed, an economist with FAO'S Global Information and Early Warning System, warned of a generally unfavourable food outlook in the sub region and in Kenya in particular. "Despite the current rains in some parts, the effect of the overall food situation in Kenya this year will be adversely affected by last year's drought," said Dr. Ahmed. In the past decade, Kenya has been hit by a series of natural disasters raging from El Nino rains to droughts. The severe drought in 1999 / 2000 seriously undermined the food security of nearly 4.4 million people and resulted in a massive relief operation. However, despite some improvement, the food supply in the country is still precarious. In the Northern Pastoral districts of the country, where severe shortages of water and pasture resulted in large livestock losses, the failure of this season's "short rains" has excavated widespread food insecurity. This has made small - scale farmers and pastoralists to be reduced to a state of destitution and dependence on food aid. The "short rains" (November - January) are crucial in the pastoral areas for the replenishment of water supplies and pastures, while in agro- pastoral areas they provide the bulk of food supplies. The situation has not been good and on February 15, the government of Kenya together with the UN Country team launched an appeal for US $ 89 million to assist 4.4 million drought affected people for a period of six months. In the neighbouring Uganda, te overall food

Figure 23: A further demonstration of the automated labeling feature in action. In this example, the user is more interested in the only food*-themed group contained among the NTF results. The list of terms describing this group is clearly very different from those found in Figure 22. The user was able to construct a terms collection* (food, aid)*, in addition to selecting the term* prices.

As discussed in this chapter, the integrated analysis environment provides the analyst with a number of significant features, ranging from data pre-processing, to NTF execution, to deeper, post-processing NTF results analysis. A number of experimental datasets have been used to demonstrate the potential of this approach. The studies varied greatly in the source material type (e.g., news articles, email messages), and the study goals. Chapter 5 continues the discussion of experimental uses of the analysis environment in greater detail.

Chapter 5

Practical Knowledge Discovery with NTF

This dissertation developed a potentially effective integrated analysis environment that includes robust pre-processing, processing, and post-processing capabilities. The approach has been shown to be effective using a pre-defined scenario discovery goal of the IEEE VAST-2007 contest (as described in Chapter 2). It is difficult to state the effectiveness of any approach to knowledge discovery on real-world data with a set of relatively open-ended analysis goals. In this context, the phrase "potentially effective" is more reasonable. The effectiveness depends in large part on the specific goals of the user. This chapter describes how the integrated analysis environment is potentially effective in facilitating knowledge discovery, given a large collection of textual data as input.

5.1 Datasets and Methodologies

A number of datasets have been gathered and utilized for the purposes of demonstrating the potential of the integrated NTF/Visualization text analysis approach. These include the IEEE VAST-2007 fictitious news articles dataset, the climategate email dataset, the Voices Heard Media (VHM) chat-log dataset, and a number of datasets consisting of real news articles gathered in collaboration with Information International Associates, Inc. (IIA). The datasets were used throughout the development process, over the course of the last three years. The following subsections describe the datasets more fully.

5.1.1 IEEE VAST-2007 Dataset

The largest in terms of the resulting tensor dimensions is the IEEE VAST-2007 dataset. It consists of 1,455 text files corresponding to fictitious news articles. The articles include SGML-tagged entities in four categories: *person, location, organization and money*. Parsing this dataset produces a dictionary of 12,121 terms and 7141 entities [1]. Experiments involving this dataset focus on scenario extraction and knowledge discovery.

The IEEE VAST-2007 dataset was the original motivator for the creation of FutureLens, which in turn spurred the creation of the integrated analysis environment and this dissertation. The original FutureLens that was created specifically for the analysis of the NTF output results from this data performed admirably and proved significantly more effective than the purely text and statistics-based approach [9].

The original FutureLens created specifically for the analysis of NTF results from this data lacked many of the features included in the version that has been integrated into the analysis environment. Perhaps most importantly, an automated NTF group classification feature has been added due to experience that would be acquired in the course of future projects. Also quite importantly, the user has been provided with the ability to manipulated component terms, creating collections (where term adjacency is unimportant) and phrases (where term adjacency and term order both matter).

5.1.2 Climategate Email Dataset

The climategate email dataset is somewhat smaller, and contained far fewer "entities" (in this case, email authors and recipients are the only entities). The climategate dataset consists of 1,072 messages, and the parsing results in a dictionary of 11,829 terms and 271 entities (authors/recipients) [13]. While human input to scenario extraction and knowledge discovery is certainly valuable in all these experiments, automated NTF output group labeling gained in importance here due to the highly charged/emotional nature of some of the discussions involved in these email exchanges. For instance, certain authors were fond of using insulting language towards other individuals and even threatening (perhaps jokingly) physical violence. Automated labeling of "angry" NTF groups greatly facilitated the analysis process and aided knowledge discovery by drawing the analysts attention to these NTF output groups.

The work with this dataset resulted in a number of interesting conclusions. It was revealed that the widely reported incidents, such as use of threatening languages and discussions about data manipulation did indeed exist in the dataset. It was however also revealed that such incidents were relatively few, and the bulk of the themes found in the dataset had nothing to do with anything particularly sinister (or indeed, with anything of much interest to anyone who is not a climate researcher) [13].

5.1.3 Voices Heard Media (VHM) Dataset

The VHM data consists of many relatively short questions or statements that were submitted to VHM by viewers of one of the affiliated television shows. This dataset is the largest in terms of the total number of elements (3,257). However, these elements correspond to very short "chat"-style postings from viewers responding and reacting to a television show in real time. The vocabulary is, therefore, considerably less diverse. This means that the data in this set of experiments is markedly different from all other data used in these experiments. The brief nature of the textual elements of the data set makes automated NTF output classification all the more valuable.

5.1.4 Information International Associates, Inc. (IIA) Datasets

In the course of collaboration with Information International Associates, Inc. (IIA) under an SBIR contract[1], several extensive textual datasets were obtained. Most of these

[1] Small Business Innovation Research (SBIR) Phase I award entitled "Weather/Climate Variability Impact on Energy, Water and Food Resources and Implications for Regional Stability". Topic Number OSD09-HS1. Contract Number W913E5-10-C-0012.

datasets were country-specific. Some covered a span of multiple years, even decades, while others focused on a specific year. The total number of documents ranged from 818 to 5097. The average document length ranged from 137 to 848 terms per document. These sets were used to explore how FutureLens might be used in a predictive capacity. The goal was to determine whether a spike in discussions of certain subjects in the news media may be used to predict a higher likelihood of negative developments occurring in the near future. The IIA datasets are summarized in the table below.

Table 1. IIA Dataset Summary. Sets obtained using Factiva are marked with (F) in the second column, while those obtained using LexisNexis are marked with (L).

	Country	Timeframe	Number of Documents	Average Document Length (# of terms)
1.	Bangladesh (F)	2009	1000	848
2.	Bangladesh (F)	1972-1976	818	137
3.	Kenya (L)	2001	971	738
4.	Kenya (L)	2001-2009	900	696
5.	Somalia (L&F)	1980-1989	5097	791
6.	Somalia (L&F)	1970-1979	2653	450
7.	Somalia (L&F)	1970-2009	8983	685

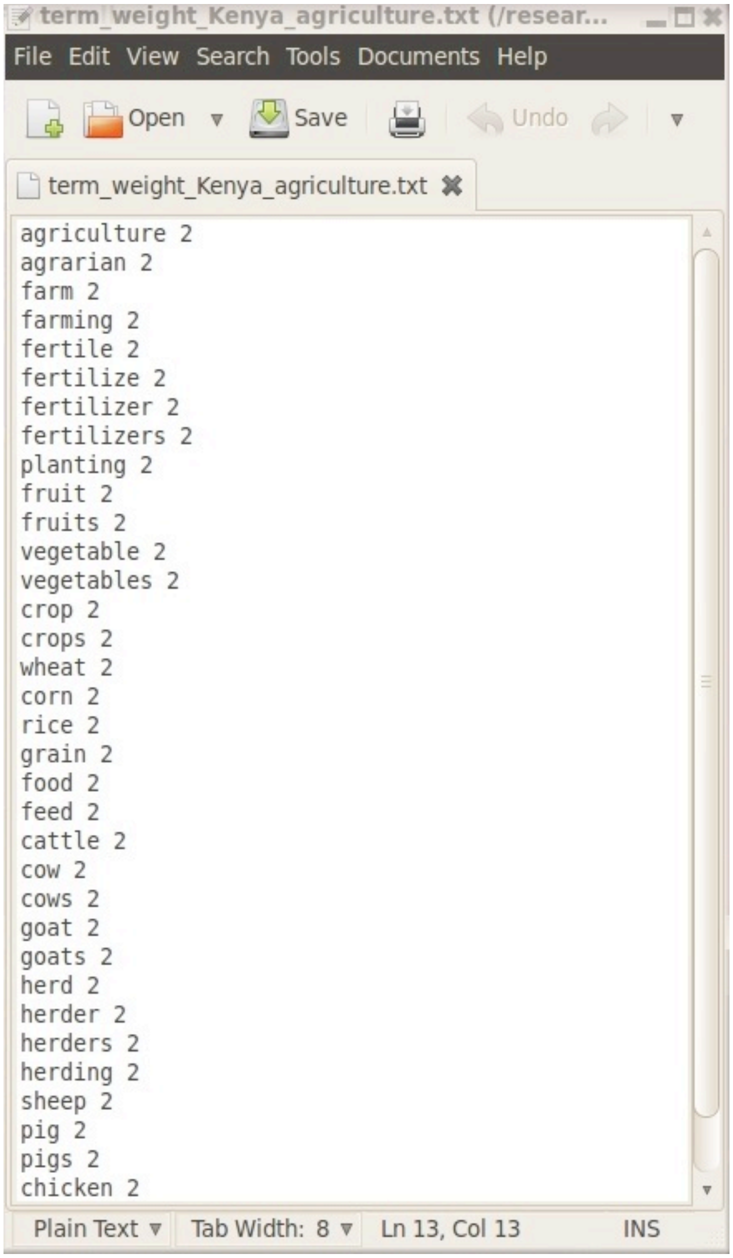
The IIA datasets and collaboration work had been instrumental in the development and testing of several of the features of the integrated analysis environment. Namely, the date insertion and entity tagging features were motivated and necessitated by the IIA collaboration. These features were used extensively in the course of that project, and each of the datasets listed in Table 1 had been used as input to utilize them. The results were deemed to be effective, helping to facilitate the creation of the final report under the SBIR contract. As of this writing, IIA is involved in planning for Phase II work that will advance this line of research further.

5.2 Evidence of Integrated Analysis Environment Effectiveness

The two usage examples described in this section demonstrate the potential effectiveness of the integrated analysis environment and its potential for knowledge discovery. The first example focuses on demonstrating the potential effectiveness of adjusting term weights as it applies to knowledge discovery. This example utilizes the Kenya 2001-2009 IIA dataset (#4 in Table 1 above). The second example shows the potential of the automated category labeling feature, and uses the Bangladesh 1972-1976 dataset (#2 in Table 1 above).

5.2.1 Effect of Tensor Weights Adjustment on Analysis

The Kenya 2001-2009 IIA dataset is fascinating in many regards, as it includes a number of greatly varied themes that appear and change in prominence over the dataset's decade-long time span. It is easy to imagine an analyst with a significant amount of prior knowledge about the dataset, and a desire to focus on a particular theme. In fact, this was typically the case in the course of the actual IIA collaboration and such situations are realistic scenarios. For the purpose of this example, the hypothetical analyst is interested in agriculture- and animal husbandry-related features of the dataset, as revealed through nonnegative tensor factorization. The first step in focusing the NTF algorithm on the themes of interest is the creation of a term weights adjustment file, shown in Figure 24.

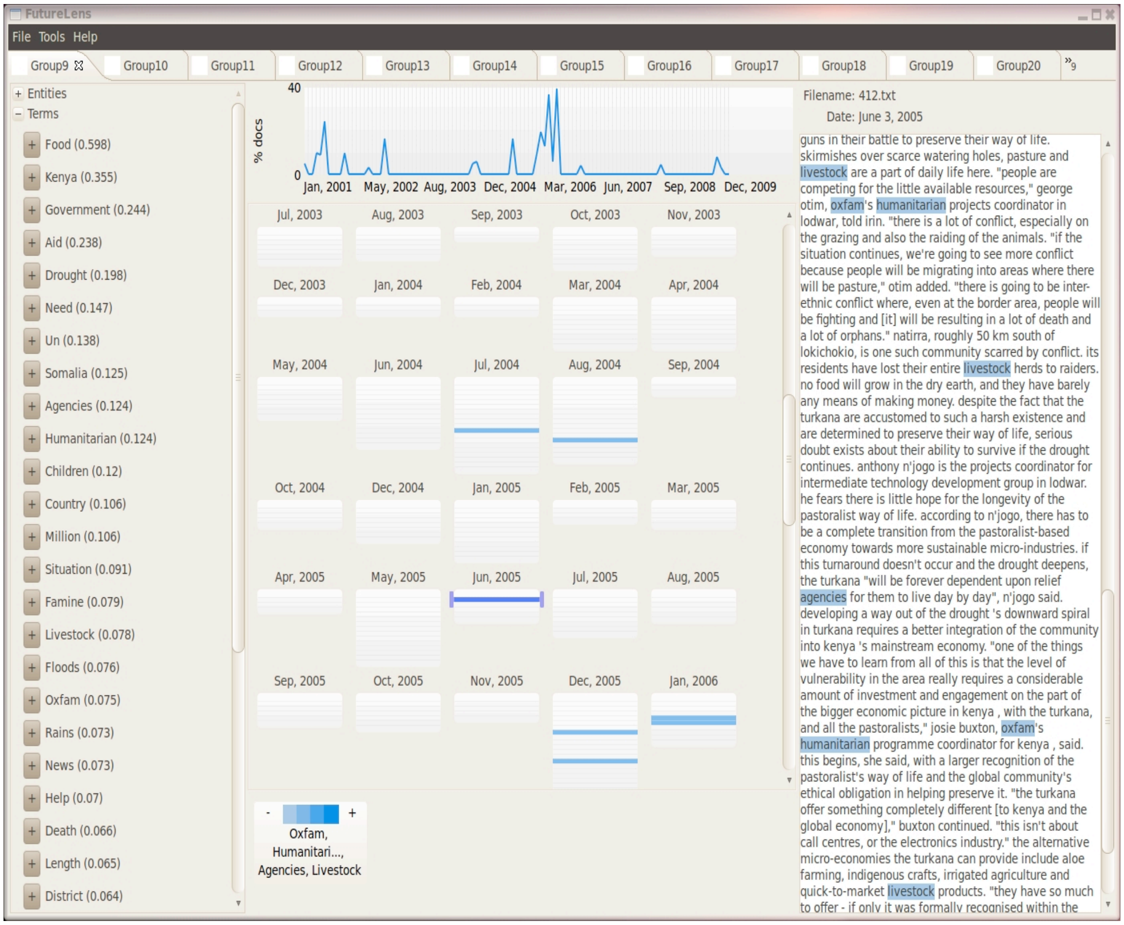

term_weight_Kenya_agriculture.txt (/resear...
File Edit View Search Tools Documents Help
Open
Save
Undo
term_weight_Kenya_agriculture.txt
Plain Text
Tab Width: 8
Ln 13, Col 13
INS

```
agriculture 2
agrarian 2
farm 2
farming 2
fertile 2
fertilize 2
fertilizer 2
fertilizers 2
planting 2
fruit 2
fruits 2
vegetable 2
vegetables 2
crop 2
crops 2
wheat 2
corn 2
rice 2
grain 2
food 2
feed 2
cattle 2
cow 2
cows 2
goat 2
goats 2
herd 2
herder 2
herders 2
herding 2
sheep 2
pig 2
pigs 2
chicken 2
```

Figure 24: A term weights adjustment file that was used to direct the NTF algorithm towards agriculture-related themes.

As shown in Figure 24, the model simply doubled the original term frequencies of certain agriculture-related terms. The integrated analysis environment was then utilized to created a modified tensor input file that reflected the adjusted weights. Referring back to Equation 1 in Section 2.1, it should be noted that only one of the three axes is affected by this adjustment (the terms axis). The NTF algorithm was subsequently applied to the newly generated modified input file, and the resulting NTF output groups were loaded into FutureLens for further visual analysis. The impact of the weight adjustment was readily apparent, as now every single output group featured at least some agriculture-related language. Interestingly, the NTF algorithm reveals many agricultural terms that were not included in the original weights file, such as *livestock*. Browsing through the groups, the analyst might note that Group 9 includes the name of a reputable charitable organization, Oxfam, as well as terms such as *humanitarian*, which also relate to charitable work. As shown in Figures 25 and 26, pursuing this line of analysis further yields interesting results and leads to discovery of (potentially) new knowledge.

Figure 25 shows a significant spike in the user-created term collection (*Oxfam, Humanitarian, Agencies, Livestock*), that occurs starting in mid-2005 and levels off by mid-2006. Selecting one of the color-coded (blue) bars in the June 2005 box in the central panel causes the corresponding article to be displayed in the panel on the right. Here, the user quickly learns about a recent spike in conflict over limited resources and grazing rights in Kenya's Rift Valley, partly caused by a recent drought's wiping out of 70 percent of the livestock in the Turkana province.



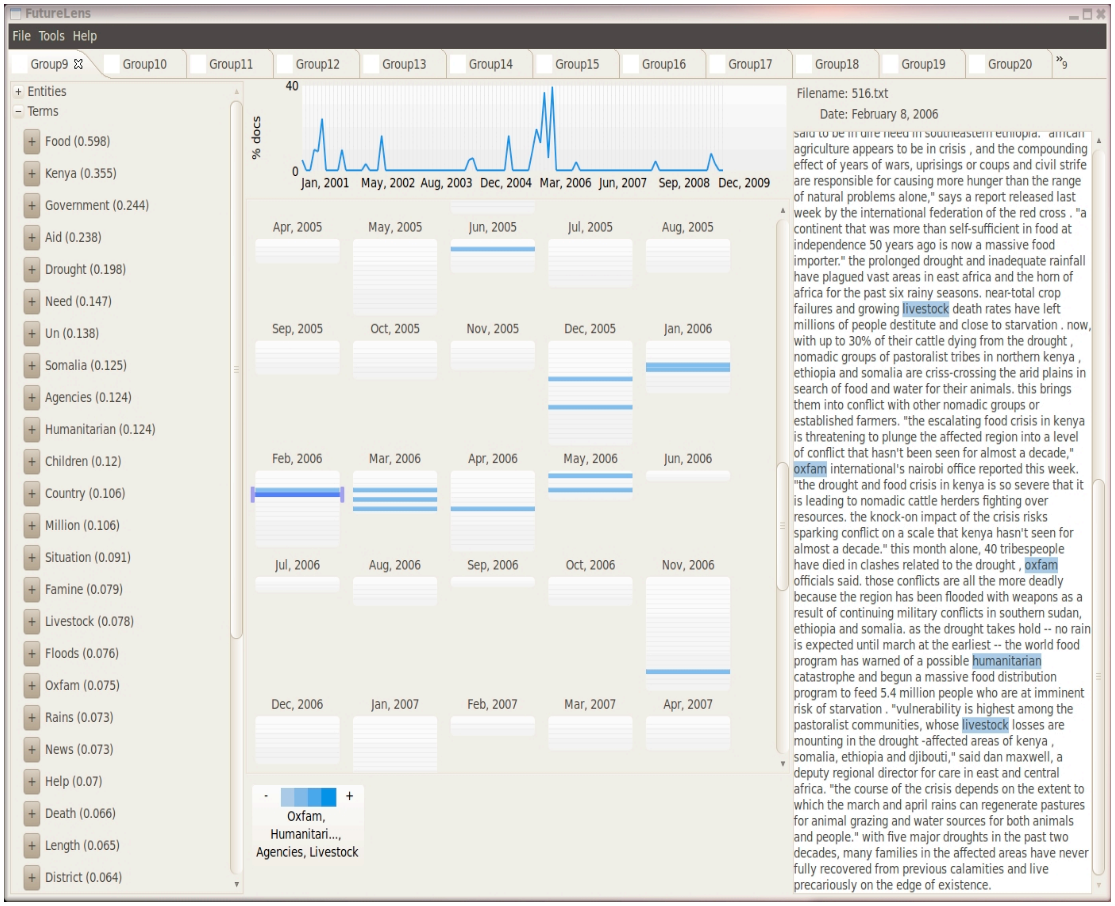
Filename: 412.txt

Date: June 3, 2005

guns in their battle to preserve their way of life. skirmishes over scarce watering holes, pasture and livestock are a part of daily life here. "people are competing for the little available resources," george otim, oxfam's humanitarian projects coordinator in lodwar, told irin. "there is a lot of conflict, especially on the grazing and also the raiding of the animals. "if the situation continues, we're going to see more conflict because people will be migrating into areas where there will be pasture," otim added. "there is going to be inter-ethnic conflict where, even at the border area, people will be fighting and [it] will be resulting in a lot of death and a lot of orphans." natirra, roughly 50 km south of lokichokio, is one such community scarred by conflict. its residents have lost their entire livestock herds to raiders. no food will grow in the dry earth, and they have barely any means of making money. despite the fact that the turkana are accustomed to such a harsh existence and are determined to preserve their way of life, serious doubt exists about their ability to survive if the drought continues. anthony n'jogo is the projects coordinator for intermediate technology development group in lodwar. he fears there is little hope for the longevity of the pastoralist way of life. according to n'jogo, there has to be a complete transition from the pastoralist-based economy towards more sustainable micro-industries. if this turnaround doesn't occur and the drought deepens, the turkana "will be forever dependent upon relief agencies for them to live day by day", n'jogo said. developing a way out of the drought 's downward spiral in turkana requires a better integration of the community into kenya 's mainstream economy. "one of the things we have to learn from all of this is that the level of vulnerability in the area really requires a considerable amount of investment and engagement on the part of the bigger economic picture in kenya , with the turkana, and all the pastoralists," josie buxton, oxfam's humanitarian programme coordinator for kenya , said. this begins, she said, with a larger recognition of the pastoralist's way of life and the global community's ethical obligation in helping preserve it. "the turkana offer something completely different [to kenya and the global economy]," buxton continued. "this isn't about call centres, or the electronics industry." the alternative micro-economies the turkana can provide include aloe farming, indigenous crafts, irrigated agriculture and quick-to-market livestock products. "they have so much to offer - if only it was formally recognised within the

Figure 25: After adjusting the NTF algorithm to have an agriculture focus, the user may utilize FutureLens for further visual analysis of the NTF results. Shown here, the discovery of the impact of a 2004-2005 drought on Kenyan agriculture and the corresponding social unrest it caused.

The dataset, however, includes news articles from 2001 through 2009, and the peak in the selected term group levels off in mid-2006. It may be interesting to track this collection further temporally, in order to attempt to determine why its importance decreased towards the end of this time period. Taking a look at a strong February 2006 spike in this collection's frequency, one may note that matters have in fact gotten worse at this time. The article shown in Figure 26 discusses escalating and increasingly violent conflict, made even worse by the fact that the region is "flooded" with weapons due to continuing military conflict in neighboring Sudan. This dire description of the situation makes the subsequent leveling off all the more mysterious.



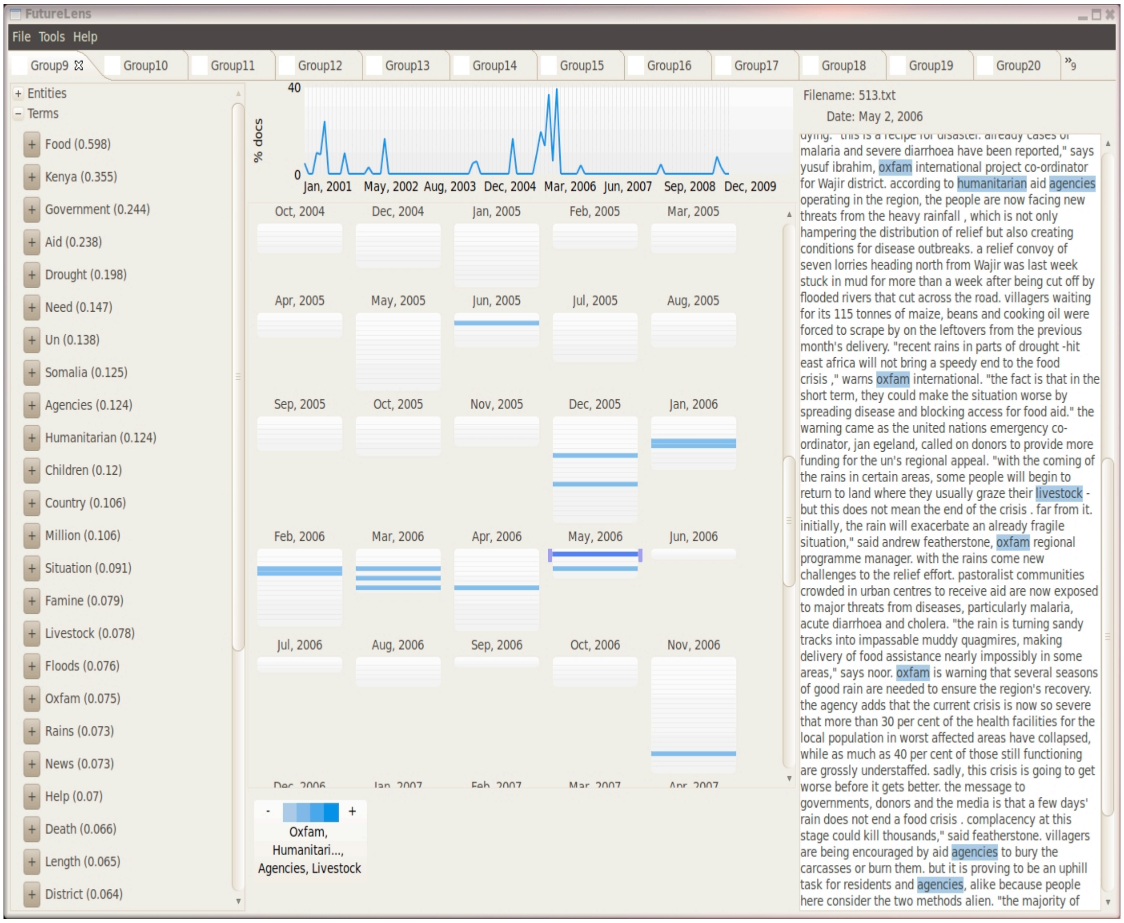
Filename: 516.txt

Date: February 8, 2006

said to be in dire need in southeastern ethiopia. "african agriculture appears to be in crisis , and the compounding effect of years of wars, uprisings or coups and civil strife are responsible for causing more hunger than the range of natural problems alone," says a report released last week by the international federation of the red cross . "a continent that was more than self-sufficient in food at independence 50 years ago is now a massive food importer." the prolonged drought and inadequate rainfall have plagued vast areas in east africa and the horn of africa for the past six rainy seasons. near-total crop failures and growing livestock death rates have left millions of people destitute and close to starvation . now, with up to 30% of their cattle dying from the drought , nomadic groups of pastoralist tribes in northern kenya , ethiopia and somalia are criss-crossing the arid plains in search of food and water for their animals. this brings them into conflict with other nomadic groups or established farmers. "the escalating food crisis in kenya is threatening to plunge the affected region into a level of conflict that hasn't been seen for almost a decade," oxfam international's nairobi office reported this week. "the drought and food crisis in kenya is so severe that it is leading to nomadic cattle herders fighting over resources. the knock-on impact of the crisis risks sparking conflict on a scale that kenya hasn't seen for almost a decade." this month alone, 40 tribespeople have died in clashes related to the drought , oxfam officials said. those conflicts are all the more deadly because the region has been flooded with weapons as a result of continuing military conflicts in southern sudan, ethiopia and somalia. as the drought takes hold -- no rain is expected until march at the earliest -- the world food program has warned of a possible humanitarian catastrophe and begun a massive food distribution program to feed 5.4 million people who are at imminent risk of starvation . "vulnerability is highest among the pastoralist communities, whose livestock losses are mounting in the drought -affected areas of kenya , somalia, ethiopia and djibouti," said dan maxwell, a deputy regional director for care in east and central africa. "the course of the crisis depends on the extent to which the march and april rains can regenerate pastures for animal grazing and water sources for both animals and people." with five major droughts in the past two decades, many families in the affected areas have never fully recovered from previous calamities and live precariously on the edge of existence.

Figure 26: The situation in Kenya's Rift Valley seems to have become even more dangerous by February of 2006. The articles corresponding to the spike in the selected term collection described a region "flooded" with weapons and on the brink of an outbreak of major violent conflict. This makes the subsequent leveling off in the frequency of this collection all the more mysterious.

To explore this mystery further, the user simply has to continue tracking the term collection temporally through the dataset, reading only a very small portion of the articles contained in the entire dataset. This is has the potential to greatly increase analyst efficiency, saving significant time and resources. The subsequent months' articles that were revealed by continued tracking of this term collection show the causes of the eventual sudden leveling off that indicates that the conflicts described in the previous articles may have been resolved. As shown in Figure27, the growing conflict was alleviated by a significant amount of rainfall that occurred in April and May of 2006 in this area of Kenya. The rainfall amount was in fact so great, that it even caused some additional danger through a risk of flooding. However, it did eventually stabilize the situation in the area by eliminating the drought. While the crisis had not been completely resolved, positive trends had began to emerge and cattle herders had began to return to previously-abandoned land.



Filename: 513.txt

Date: May 2, 2006

dying." this is a recipe for disaster. already cases of malaria and severe diarrhoea have been reported," says yusuf ibrahim, oxfam international project co-ordinator for Wajir district. according to humanitarian aid agencies operating in the region, the people are now facing new threats from the heavy rainfall , which is not only hampering the distribution of relief but also creating conditions for disease outbreaks. a relief convoy of seven lorries heading north from Wajir was last week stuck in mud for more than a week after being cut off by flooded rivers that cut across the road. villagers waiting for its 115 tonnes of maize, beans and cooking oil were forced to scrape by on the leftovers from the previous month's delivery. "recent rains in parts of drought -hit east africa will not bring a speedy end to the food crisis ," warns oxfam international. "the fact is that in the short term, they could make the situation worse by spreading disease and blocking access for food aid." the warning came as the united nations emergency co-ordinator, jan egeland, called on donors to provide more funding for the un's regional appeal. "with the coming of the rains in certain areas, some people will begin to return to land where they usually graze their livestock - but this does not mean the end of the crisis . far from it. initially, the rain will exacerbate an already fragile situation," said andrew featherstone, oxfam regional programme manager. with the rains come new challenges to the relief effort. pastoralist communities crowded in urban centres to receive aid are now exposed to major threats from diseases, particularly malaria, acute diarrhoea and cholera. "the rain is turning sandy tracks into impassable muddy quagmires, making delivery of food assistance nearly impossibly in some areas," says noor. oxfam is warning that several seasons of good rain are needed to ensure the region's recovery. the agency adds that the current crisis is now so severe that more than 30 per cent of the health facilities for the local population in worst affected areas have collapsed, while as much as 40 per cent of those still functioning are grossly understaffed. sadly, this crisis is going to get worse before it gets better. the message to governments, donors and the media is that a few days' rain does not end a food crisis . complacency at this stage could kill thousands," said featherstone. villagers are being encouraged by aid agencies to bury the carcasses or burn them. but it is proving to be an uphill task for residents and agencies, alike because people here consider the two methods alien. "the majority of

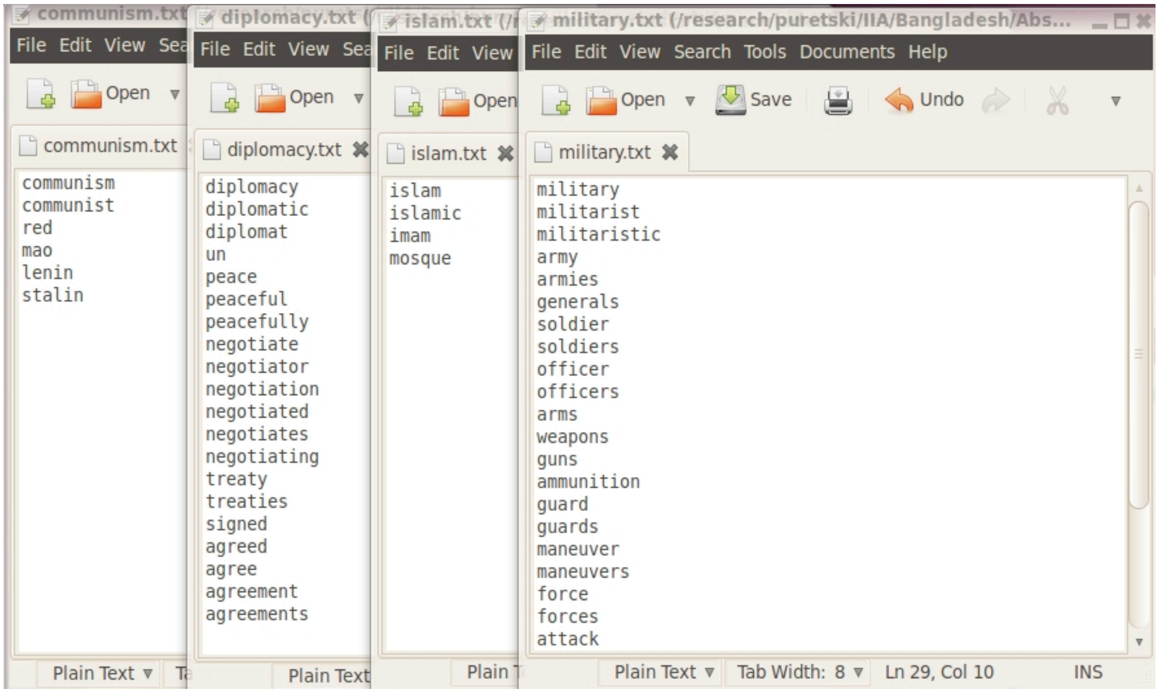
Figure 27: Continuing to track the term collection further through the dataset reveals that the dangerous situation described in Figures 25 and 26 had been resolved largely due to a high amount of rainfall that occured in April and May of 2006.

Thus, the use of a number of different features of the integrated analysis development environment has lead to significant knowledge discovery. Even an analyst who is completely new to this environment, having gone through the process described above, could learn a number of important pieces of information in just an hour or two. First, an agriculture-themed initial exploration had revealed serious and potentially critically important agriculture-based conflicts in the region of interest. Second, tracking the evolution of these conflicts through the dataset had revealed that these conflicts are by no means fully resolved. Even though they were alleviated before turning strongly violent, the alleviation was essentially just a lucky, weather-related break. The underlying risk factors and dangers, such as the “flood” of weapons and competition for scarce resources remain. And thus one might conclude that the situation in this region remains dangerous, though perhaps not immediately so.

5.2.2 Effect of Automated NTF Output Labeling on Analysis

The integrated analysis environment’s automated NTF output labeling capability is one of its most important features. As will be shown in this section, it can enormously improve an analyst’s efficiency by providing a quick automatic ability to sort NTF results in accordance with analyst-defined categories of interest.

For this example, the Bangladesh 1972-1976 dataset was processed using the analysis environment. As the first step, several category descriptor files were created. These categories represent realistic potential areas of interest to someone involved in research on 1970s South East Asia. However, for the purposes of this example, let us assume that the analyst is most interested in developments pertaining to Islam. The category described by the files shown in Figure 28, include *Communism, Diplomacy, Islam,* and *Military*.


communism.txt
diplomacy.txt (
islam.txt (/r
military.txt (/research/puretski/IIA/Bangladesh/Abs...
File Edit View Search Tools Documents Help
Open
Save
Undo
Plain Text
Tab Width: 8
Ln 29, Col 10
INS

communism.txt

```
communism
communist
red
mao
lenin
stalin
```

diplomacy.txt

```
diplomacy
diplomatic
diplomat
un
peace
peaceful
peacefully
negotiate
negotiator
negotiation
negotiated
negotiates
negotiating
treaty
treaties
signed
agreed
agree
agreement
agreements
```

islam.txt

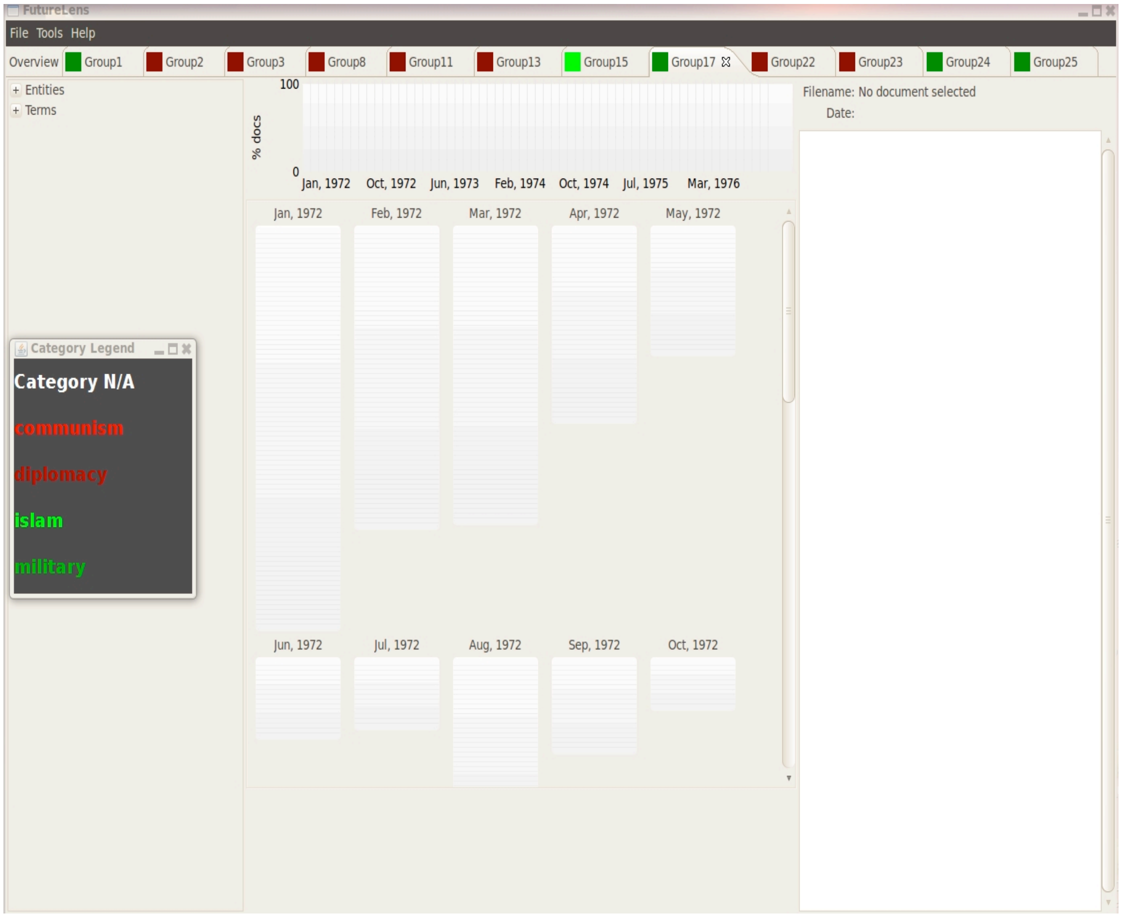
```
islam
islamic
imam
mosque
```

military.txt

```
military
militarist
militaristic
army
armies
generals
soldier
soldiers
officer
officers
arms
weapons
guns
ammunition
guard
guards
maneuver
maneuvers
force
forces
attack
```

Figure 28: A realistic set of categories that someone involved in research on 1970s South East Asia could potentially find interesting.

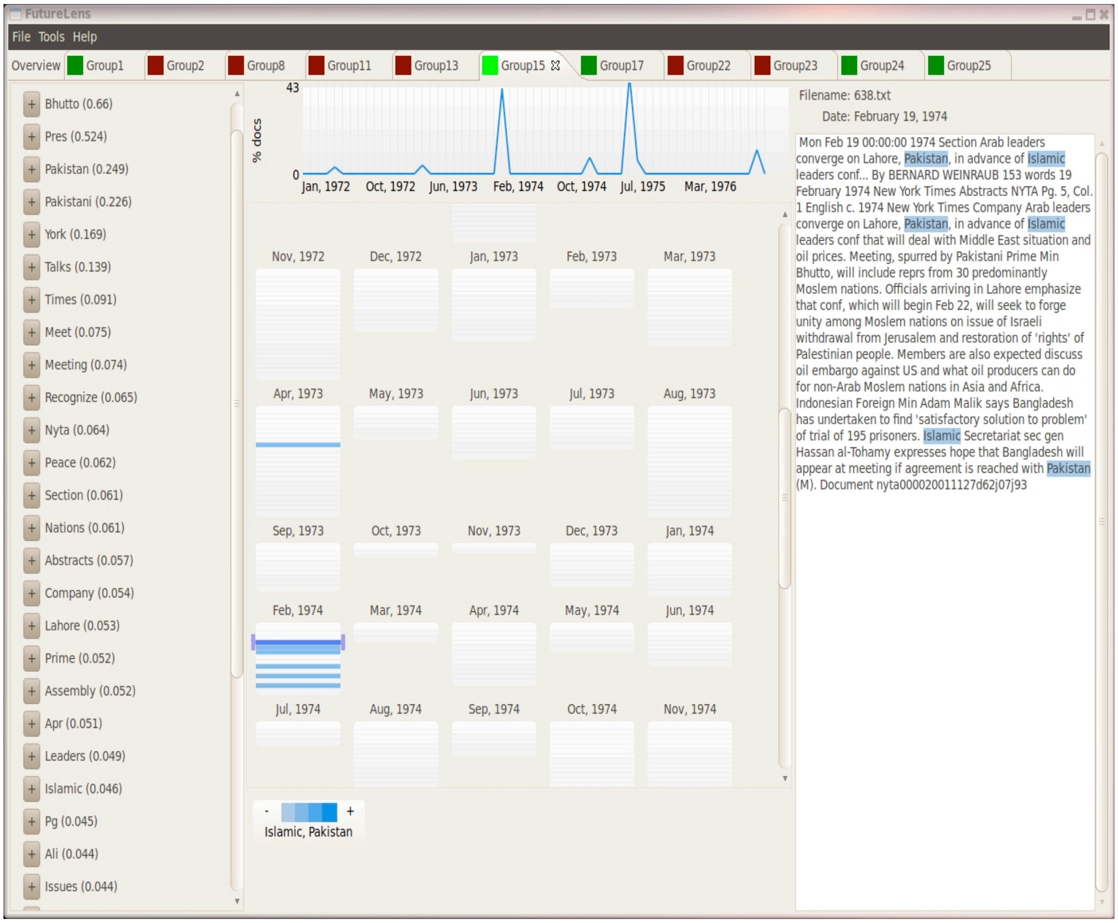
Following the creation of these category descriptors and the previously described process of execution of the NTF algorithm to generate NTF output group files, the user may utilize FutureLens’s automated group labeling feature. Without the automated labeling feature, the analyst must focus in great detail on every single one of the NTF output groups (25 total, for this example). This could take a considerable amount of time, and the process would be prone to human error. Using the automated NTF group labeling feature of the analysis environment, however, takes just a few second. The results are shown in Figure 29, where those groups that did not fit into any one of the four categories of interest have already been closed. Of the labeled groups, one fit into the *Islam* category, four were labeled as *Military*-related, ten had a *Diplomacy* theme, while the rest did not fit into any of the categories created by the user. There were no *Communism*-labeled groups in this set.



Filename: No document selected
Date:

Figure 29: NTF output groups have been automatically labeled in accordance with the categories loaded by the user (shown in the legend window on the right of the figured).

As one may recall, the hypothetical analyst in this scenario is most interested in developments pertaining to Islam. It just happens that only one of the NTF output features has been automatically labeled as belonging to the *Islam* category. This already provides the analyst with some important and potentially new knowledge, namely that Islam did not figure prominently into the news coming out of Bangladesh in the 1970s. Even more importantly, the analyst can save a great deal of time by focusing exclusively on just one of the twenty-five total NTF output groups. Shown in Figure 30, the analyst performs a detailed analysis of Group 15, labeled as belonging to the *Islam* category. Quickly revealed in the articles belonging to this category are Pakistan's efforts to improve its diplomatic position by strengthening ties with Islamic countries inside and outside of the South East Asia region.



Filename: 638.txt

Date: February 19, 1974

Mon Feb 19 00:00:00 1974 Section Arab leaders converge on Lahore, Pakistan, in advance of Islamic leaders conf... By BERNARD WEINRAUB 153 words 19 February 1974 New York Times Abstracts NYTA Pg. 5, Col. 1 English c. 1974 New York Times Company Arab leaders converge on Lahore, Pakistan, in advance of Islamic leaders conf that will deal with Middle East situation and oil prices. Meeting, spurred by Pakistani Prime Min Bhutto, will include reprs from 30 predominantly Moslem nations. Officials arriving in Lahore emphasize that conf, which will begin Feb 22, will seek to forge unity among Moslem nations on issue of Israeli withdrawal from Jerusalem and restoration of 'rights' of Palestinian people. Members are also expected discuss oil embargo against US and what oil producers can do for non-Arab Moslem nations in Asia and Africa. Indonesian Foreign Min Adam Malik says Bangladesh has undertaken to find 'satisfactory solution to problem' of trial of 195 prisoners. Islamic Secretariat sec gen Hassan al-Tohamy expresses hope that Bangladesh will appear at meeting if agreement is reached with Pakistan (M). Document nyta000020011127d62j07j93

Figure 30: The automated NTF group labeling feature allows the analyst to very quickly focus on the one most relevant group. Quickly revealed through deeper analysis of this group are Pakistan's efforts at diplomacy involving Islamic countries inside and outside of the South East Asia region.

The two usage examples described in Sections 5.2.1 and 5.2.2 only begin to describe the full capability of this approach to the analysis of textual data. The approach is extremely flexible and its capabilities are robust. However, it may be expanded in a number of interesting directions in the future. These are described in Chapter 6.

Chapter 6

Conclusions and Future Research Directions

The purpose of this dissertation has been on creating a text analysis environment that effectively integrates mathematical techniques (NTF) with visual post-processing tools. The integrated environment also provides effective pre-processing tools that allow the user to make various improvements to the analysis process, as well as to construct and evaluate a variety of models. The additions and improvements to the basic NTF-based analysis process have been made in both pre-processing and post-processing stages, with the goal of making the non-negative tensor factorization algorithm accessible to the casual user. The integrated analysis environment implementation presented in this dissertation allows the user to construct and modify the contents of the tensor, experiment with relative term weights and trust measures, and experiment with the total number of algorithm output features produced by NTF. Non-negative tensor factorization output feature production is closely integrated with a visual post-processing tool, FutureLens, which allows the user to perform in-depth analysis and has a great potential for discovery of interesting and novel patterns within a large collection of textual data. Section 6.1 summarizes the goals of this dissertation and discusses how the dissertation addressed each one. Sections 6.2 and 6.3 discuss potential future work.

6.1 Summary of Dissertation Goals

The purpose of this dissertation was to create a novel approach to text analysis, integrating nonnegative tensor factorization, tensor term weight adjustment, and visual factorization results post-processing into a single integrated text analysis environment. The following subsections summarize the specific goals that had to be met in order to create such an environment, and describe how this dissertation addresses each of them.

6.1.1 User-Friendliness

One of the primary goals was to create an integrated analysis approach that would be highly user-friendly. Ideally, a user without a great deal of technical experience or knowledge would be able to utilize the analysis software without much time having to be spent on training. Specific knowledge pertaining to data mining or visual analytics would be unnecessary. To this end, the environment conceals the underlying nonnegative tensor factorization process, while providing the user with a clear set of controls that would allow him or her to influence the NTF process in ways that may facilitate data analysis.

6.1.2 Portability, Flexibility, Cost of Use

The second major design goal was portability and flexibility. In order to create a highly usable textual data analysis tool, it was critical to utilize languages that are more portable and accessible than Matlab®. For this reason, Python and its NumPy/Pylab libraries have been utilized in this project. Python has the additional advantage of being freely available to programmers and users, and completely cross-platform. FutureLens has been added to the analysis environment for the visualization portion of the process. FutureLens is a Java-based graphical post-processing tool that has been shown to be helpful to the text analysis process, as discussed in detail throughout this dissertation. Java, being a cross-platform language, is a good choice for accomplishing the portability/flexibility goal.

6.1.3 Speed and Efficiency for Real-time User Analysis

The third major goal was to make the analysis process efficient and as scalable as possible. Because of the very nature of the field of data mining, scalability is always a great challenge. Making the approach as efficient and scalable as possible was one of the design goals for this dissertation. This meant attempting to make the performance of the portable (Python-based) analysis environment to match (or at least approach) that of the older Matlab®-based methodology.

As shown in Table 2, the speed of the Python NTF implementation does not quite approach that of the Matlab® version. The Python implementation still allows the user to perform analysis in real time, but further speed improvements would be greatly beneficial and constitute a good direction for future work.

Table 2. Performance comparison (averaged over 10 trials) between Matlab® and Python NTF implementations.

Dataset	Number of Documents	Avg. Document Length (terms)	Matlab® Execution Time (mins)	Python Execution Time (mins)
Kenya 2001-2009	900	696	4.54	17.15
VAST 2007	1455	391	3.95	16.13

6.1.4 Automation and User Input Necessity

Making partially automated processing a part of the analysis process has been another major goal of this dissertation. Automation manifests itself most significantly in the FutureLen's new capability to label NTF output groups in accordance with user-defined categories. Input from the user is required during various stages of the analysis process, which is described in detail in Section 3.1.4. Greater automation of the analysis process, perhaps through the use of artificial intelligence systems such as neural networks, may be an interesting future work direction.

There are many other potential future additions and improvements that can be made to the analysis environment. Perhaps the most significant of these, spatial tracking, is discussed in Section 6.2. It would also be interesting to apply these techniques to the field of bioinformatics. This is an interesting research challenge, since bioinformatics differs from news article related research in both the structure of its input data and the goals of its research studies. Section 6.3 discusses potential bioinformatics applications further.

6.2 Integrating Geographical Information into the Analysis Environment

While the integrated analysis environment already allows the user to track elements of the dataset temporally, currently there is no direct spatial tracking capability. Such a capability could prove to be extremely important for certain types of research projects. It could help an analyst to quickly and easily answer the *Where* question, in addition to the *When* question. The fact that a particular subject is being discussed at the same time could be very important. It is, however, potentially even more important to knowledge discovery to be able to demonstrate that the particular subject in question was being discussed in the same general geographic location at the same time. Alternatively, it could be just as important to be able to demonstrate that he subject in question was being discussed in vastly different location at the same time (or same location and different times, or different locations and different times). The exact nature of this analysis would, of course, depend on the particular circumstances and demands of each specific study.

Adding a geocoding capability to the analysis environment could facilitate the integration of spatial tracking ability into the overall process. Geocoding refers to the process of adding geographic descriptors, such as latitude and longitude coordinates to textual data. A large number of research studies have been conducted in this area, many with the goal of creating databases of geo-tagged data in a particular subject area. For example, a research team at the University of Berkley recently created a geocoded database of fatal and severe injury traffic accidents that occurred in the California between the years 1997 and 2006. The researchers concluded that the availability of this database to other researcher in the fields ranging from medical science to public health and safety could be beneficial. It’s clear how such geocoded data allows the study of potential connections between a wide variety of factors, environmental and social, and serious traffic accidents [23].

Virtually every area of study could benefit from incorporating geocoding information into the analysis process. The integrated analysis environment already offers a basic framework that would allow the addition of geocoding-related functionality into several stages of the analysis process. Perhaps the most significant amount of effort would be required for determining how to extract geographical information from textual data. For instance, when a study involves news articles, what matters more: the locations mentioned in the news article, the location of the reporter at the time of the article's creation, or the location of the publishing entity? Such questions would need to be answered before a geocoding implementation is even attempted. Adding to the challenge is the fact that the answer may depend and vary with the nature of each research study.

Once these matters have been resolved, however, further integration into the visual NTF results analysis stage may be relatively straightforward. Similarly to how temporal information is currently displayed in a top-level summary graph, spatial information could be summarized using color-coding of a simple geographical map that would be displayed in a separate panel. The color-coding capability already build into FutureLens would integrate well with such an approach.

6.3 Adjusting the Environment to Work with Bioinformatics Data

It may be interesting and perhaps extremely useful to apply the integrated analysis environment to the field of bioinformatics. Recent work in this field involving a related technique, nonnegative matrix factorization (NMF) has proven highly effective. The project, called FAUN attempts to facilitate research in medicine and genetics by providing tools that greatly increase efficiency of searching through medical literature. The FAUN environment is Web-based, and uses NMF to facilitate the discovery and classification of functional relationships among genes as discussed in medical research literature [24].

Expanding the approach to include NTF, using a Gene-by-Term-by-Expression tensor may reveal additional gene functional relationships using both biomedical literature and microarray data. Adding the spatial tracking capabilities to the analysis environment could have a great potential for this project.

Bibliography

1) B. Bader, A. Puretskiy, M. Berry, ”Scenario Discovery Using Nonnegative Tensor Factorization", In Proceedings of the Thirteenth Iberoamerican Congress on Pattern Recognition, J. Ruiz-Shulcloper and W.G. Kropatsch (Eds.): CIARP 2008, LNCS 5197, pp. 791-805. Springer-Verlag Berlin Heidelberg 2008.
2) L. De Lathauwer, “A Survey of Tensor Methods”, in Proc. of ISCAS 2009, Taipei, Taiwan, 2009.
3) T. Kolda, B. Bader, “Tensor Decompositions and Applications”. SIAM Review, Vol. 51, No. 3, 2009, pp. 455-500.
4) Zhe, X. & Boucouvalas, A.C., 2002. Text-to-Emotion Engine for Real Time Internet Communication. International Symposium on Communication Systems, Networks and DSPs, 15-17 July 2002, Staffordshire University, UK, pp 164-168.
5) SEASR. Sentiment Tracking from UIMA Data. http://seasr.org/documentation/uima-and-seasr/sentiment-tracking-from-uima-data/. Visited May 2010.
6) Harshman, R.A.: Foundations of the PARAFAC procedure: models and conditions for an “explanatory” multi-modal factor analysis. UCLA working papers in phonetics 16, 1-84 (1970), http://publish.uwo.ca/~harshman/wpppfac0.pdf
7) Scholtz, J., Plaisant, C., Grinstein, G.: IEEE VAST 2007 Contest (2007), http://www.cs.umd.edu/hcil/VASTcontest07
8) Carroll, J.D., Chang, J.J.: Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition. Psychometrika 35, 283-319 (1970).
9) Shutt, G.L., Puretskiy, A.A., Berry, M.W.: FutureLens: Software for Text Visualization and Tracking. Text Mining Workshop, Proceedings of the Ninth SIAM International Conference on Data Mining, Sparks, NV, April 30-May 2, 2009, ISBN: 978-0-898716-82-5.
10) “Hacked E-Mail Is New Fodder for Climate Dispute”, Andrew C. Revkin, The New York Times. November 20, 2009. http://www.nytimes.com/2009/11/21/science/earth/21climate.html?_r=1 (visited 5/3/2010).
11) East Anglia Emails Source. http://www.eastangliaemails.org. Visited 5/19/2010.
12) SciPy Python Performance Evaluation. http://www.scipy.org/PerformancePython (visited 5/25/2010).
13) Berry, M.W, Pjesivac, V., and Puretskiy, A.A. Nonnegative Tensor Factorization Models Detect and Track Questionable Behavior from Climate Research Electronic Mail. Submitted to UTK CUR Journal, 5/2010.
14) A.A. Puretskiy, G.L. Shutt, and M.W. Berry, ”Survey of Text Visualization Techniques,” in Text Mining: Applications and Theory, M.W. Berry and J. Kogan (Eds.), Wiley, Chichester, UK, pp. 107-127, 2010.
15) Jonathan Feinberg, “Wordle: Beautiful Word Clouds”. http://www.wordle.com, Visited July 2009.
16) Bradford Pailey, “TextArc”. http://www.textarc.org, Visited July 2009.

17) SEASR, “UIMA and SEASR”. http://seasr.org/documentation/uima-and-seasr/, Visited July 2009.
18) W.G. Parrot, “Emotions in social psychology: Essential readings”, pp. 1-19. Psychology Press, Philadelphia: 2001.
19) Numpy Documentation: http://docs.scipy.org/doc/numpy. Visited September 2010.
20) SciPy Documentation: http://docs.scipy.org. Visited September 2010.
21) Comon, P., “When tensor decomposition meets compressed sensing”. Ninth International Conference on Latent Variable Analysis and Signal Separation. September 27, 2010.
22) Diaw, P., “Sparse Tensor Decomposition Software”. Department of Electrical Engineering and Computer Science Report. July 16, 2010, University of Tennessee, Knoxville.
23) Bigham JM, Rice TM, Pande S, Lee J, Park SH, Gutierrez N, Ragland DR. Geocoding police collision report data from California: a comprehensive approach. International Journal of Health Geographics. 2009 Dec 29; 8:72.
24) Tjioe E, Berry MW, Homayouni R. Discovering gene functional relationships using FAUN (Feature Annotation Using Nonnegative matrix factorization). BMC Bioinformatics. 2010 Oct 7;11 Suppl 6:S14. PMID: 20946597
25) Hickman, Leo; Randerson (2009-11-20). "Climate sceptics claim leaked emails are evidence of collusion among scientists". Guardian.co.uk (London: The Guardian). http://www.guardian.co.uk/environment/2009/nov/20/climate-sceptics-hackers-leaked-emails. Visited 2010-07-27.
26) Brett W. Bader and Tamara G. Kolda, MATLAB Tensor Toolbox Version 2.4, http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox/, March 2010.

Vita

Andrey A. Puretskiy was born in Troitsk, Russia on January 14, 1982. He grew up in Troitsk, and attended that town's School #5 until the age of twelve. At twelve, Andrey moved to Knoxville, TN, where he would eventually graduate from Bearden High School. Andrey then attended the University of Tennessee, where he received his Bachelor's and Master's degrees in Computer Science. Andrey Puretskiy completed his doctorate in Computer Science at the University of Tennessee in December, 2010.