Visualizations for Mental Health Topic Models

by

Ge (Jackie) Chen

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2014

Author .....
Department of Electrical Engineering and Computer Science
May 9, 2014

Certified by .....
Henry A. Lieberman
Principal Research Scientist
Thesis Supervisor

Accepted by .....
Prof. Albert R. Meyer
Chairman, Masters of Engineering Thesis Committee

Visualizations for Mental Health Topic Models

by

Ge (Jackie) Chen

Submitted to the Department of Electrical Engineering and Computer Science
on May 9, 2014, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

In this thesis, I designed and implemented a compiler which performs optimizations that reduce the number of low-level floating point operations necessary for a specific task; this involves the optimization of chains of floating point operations as well as the implementation of a "fixed" point data type that allows some floating point operations to simulated with integer arithmetic. The source language of the compiler is a subset of C, and the destination language is assembly language for a micro-floating point CPU. An instruction-level simulator of the CPU was written to allow testing of the code. A series of test pieces of codes was compiled, both with and without optimization, to determine how effective these optimizations were.

Thesis Supervisor: Henry A. Lieberman
Title: Principal Research Scientist

Acknowledgments

This is the acknowledgements section. You should replace this with your own acknowledgements.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Crisis Text Line (CTL) is an organization that provides counseling services to young people in crisis through texting. The goal of this thesis is to supply CTL counselors with assistive tools in order to offer their clients the best possible service. This section explains the motivation behind this research project, the problems we want to solve, the approaches to implement and evaluate, and the contributions made.

1.1 Motivations for Mental Health Visualizations

The main motivating factor for this thesis is to help people with mental health crises. Many people suffer from depression, suicidal thoughts, and emotional stress every day. Crisis Text Line provides an outlet for clients to discuss their issues and ask for support. However, there is a shortage of counselors compared to the countless number of people seeking aid. Each counselor may have to manage various conversations continuously for several hours. Many of these counselors are also volunteers who undergo a short period of training as preparation. These constraints motivate us to maximize the amount of time and brainpower counselors spend on client support. Time is a particularly critical factor in mental health situations because clients may be at risk of suicide or physical harm.

Fortunately, the unique thing about a texting hotline is that the use of written communication allows computer programs to analyze and extract meaningful data

from the text. Topic modeling is a machine learning technique that discovers abstract topics occurring in a set of documents. It can be useful for summarizing large amounts of text. Counselors may benefit from summaries of their various conversations with clients, but topic modeling is a complex and advanced artificial intelligence concept. Therefore, we would like to provide counselors with an easy method of understanding the data through visualizations. Visualization is a powerful approach for presenting data that most people are familiar with.

1.2 Problem Definition

As mentioned in the motivation section, some main difficulties with Crisis Text Line are that there are not enough counselors to talk with all of the clients in crisis, the counselors are usually context switching between multiple different client texts at a time, and they need to be properly trained. Since it is more difficult to control external factors such as the number of available counselors, our approach to these issues consist of tackling four specific problems:

1. Reduce the amount of time counselors spend not talking to clients.
2. Reduce the cognitive load of counselors so they can concentrate on clients.
3. Improve the quality of counselor training.
4. Improve the quality of client service.

1.3 Hypotheses

We believe a variety of topic model visualizations will offer assistance in solving the proposed problems. Using topic models, mental health conversations may be summarized by a combination of topics, such as job-related issues, family troubles, relationship difficulties, or self-injurious behavior, to name a few. Topic models also provide indexing information, which tells us where each specific topic can be found in the conversation text.

1.3.1 Context Switching

As the medium of texting usually involves gaps in response time during a conversation, counselors often switch context between talking to different clients. When a counselor returns to a previous conversation after a client response, he or she may have to spend time recalling what that particular conversation was about. However, if the counselor was given a visual summary of the conversation, with the option of quickly reading through chat details, less time may be spent recognizing the conversation topics. This approach can minimize both the time a counselor spends not talking to a client and the cognitive load that context switching has on the counselor.

1.3.2 Shift Changes

Counselors usually handle incoming client texts in shifts. A shift change may occur in the middle of conversations, in which case the leaving counselor gives the incoming counselor a brief summary of the talking points so he or she can take over. However, this summary is general and transient, and the incoming counselor would have to take time scanning through the existing conversation text for details. We suspect that a permanent visual summary computed using topic models would be more helpful for the incoming counselor. The visualizations can provide different levels of detail depending on what the counselor needs to know about the conversation history. Visual indexing can quickly point him or her to the parts of the conversation related to a certain topic. This technique minimizes the amount of time necessary to search through the text for details and potentially improves the quality of client service by better preparing the new counselor for the interaction.

1.3.3 Counselor Training

One potentially significant contributor to quality counseling, especially for new counselors, is the initial training that they receive. As many counselors are simply volunteers with little professional background in mental health, it can be important to show them the right way to respond to certain situations or test how they would reply

to a real client message. I think topic models are a great resource for finding real and relevant training examples. Managers or supervisors can utilize the visual indexing aspect of the topic model visualizations to find helpful instances of client messages related to specific topics. Then a counselor can be easily trained for a specific area, such as crises involving jobs or family. This method may enhance the quality of counselor training and therefore client service, while decreasing the amount of time spent searching for or fabricating relevant samples.

1.3.4 Conversation Trends

As previously mentioned, counselors must keep track of multiple conversations at a time. These conversations may also contain gaps of time due to the use of texting. In order to aid the counselor's memory, we believe that displaying topic trends over time for each conversation could be useful. Showing trends, including where topics appear in the course of a conversation and how they accumulate, may potentially improve the quality of client service. A chart of topic trends could alert the counselor to important focus points. For example, if the topic of self-injurious behavior is on the rise, the counselor might want to react in a certain way to prevent escalation of injury. Conversation trends may also be useful for organization leaders to detect patterns that might be of use in supporting clients.

1.4 Contributions

Based on a topic model developed from a collection of real mental health conversations, I designed and implemented a website prototype for Crisis Text Line with four visualizations. These visualizations were designed based on four different levels of granularity, so the counselor can choose the amount of detail he or she wants.

The **Topic List** visualization lists the topics discovered in a conversation that are above a certain threshold. Topics are ordered from highest to lowest percentage detected in the conversation. This visualization is a quick, glance-able summary of the conversation topics.

The **Donut Chart** visualization adds a small level of detail by displaying the topic proportions in a pie chart variation to show the parts of the whole relationship. User interaction by hovering over the chart or the legend provides the topic percentages for quantitative information.

The **Line Chart** visualization reaches finer granularity by revealing topic proportions at the message level, where each client message in the conversation is analyzed for topics. A line exists for each topic above a certain threshold that shows the trend of that topic throughout the conversation timeline. There is also the option of viewing the accumulation of topic proportions across the conversation. When the user clicks on a topic, points are displayed to reveal the client messages in the conversation that contain the topic.

The **Scatter Plot** visualization is the deepest detail level, allowing the user to click on the topic instances that occur throughout the conversation. The conversation text then automatically scrolls to the appropriate message. The size of the scatter plot points represent the proportion of that topic in the corresponding message.

1.5 Thesis Outline

Chapter two presents related work, consisting of topic models, visualizations of topic models for other fields, and general visualization techniques.

Chapter three discusses the design of the four visualizations contributed in this this thesis: a topic list, a donut chart, a multi-series line chart, and a scatter plot.

Chapter four explains how the system was implemented and lists the existing technologies that were used.

Chapter five evaluates the visualizations based on the user test results.

Chapter six explores ideas for future work, some of which could not be completed due to time, resource, and technological constraints.

Finally, chapter seven discusses the main contributions presented in this thesis.

Chapter 2

Related Work and Scope

In this section, we first summarize relevant research presented in three categories: topic modeling, topic model visualizations, and general visualization techniques. We then provide the scope and limitations of this thesis project.

2.1 Topic Modeling

Probabilistic topic models [1] are algorithms that aim to extract the main themes from a large collection of documents. These algorithms use statistics to analyze the words in each document's text and organize them into topics. Topic modeling can be used to aid summarization and information retrieval for various types of data without the need for humans to manually annotate a large amount of text.

The simplest topic model is *latent Dirichlet allocation* (LDA) [1]. LDA uses a statistical process to discover the topics in a corpus of documents. A *topic* is formally defined as a distribution over a fixed vocabulary. For example, a *genetics* topic should have the words *genetics* and *genes* with high probability. LDA consists of reverse-engineering an imaginary generative process. This process begins by taking a random distribution over topics. Each word for each document is then generated by randomly choosing a topic from the distribution over topics and randomly choosing a word from that topic's distribution over words. We refer to the topics, the per-document topic distributions, and the per-document per-word topic assignments as

the topic structure. This generative process must be reverse-engineered because the words in the documents are observed, while the hidden topic structure that most likely generated the words must be inferred.

The topic modeling algorithm used for the visualizations in this thesis is contributed by Karthik Dinakar. His approach is similar to the algorithm developed for a previous story-matching project [5] that involves mental health topic models, which is a very new research area. First, LDA is applied to the set of documents. The output is topics, in the form of word clusters, and a distribution over the topics for each document. Each word cluster is then analyzed by a human and interpreted as a theme if possible. This process iterates with an increasing number of desired topics until a satisfactory collection of themes have been extracted from the documents. Each document has a distribution over the themes. The topic modeling used in this thesis is slightly different in two ways. First, the documents are conversations between a client and a counselor, so only the words in the client messages are analyzed. Second, each client message also has a distribution over themes, so the conversations can be broken down further and annotated with themes.

We will not discuss the specifics of topic modeling at the deeper level of probability and statistics because this thesis is concentrated on visualization. The purpose of this overview is to familiarize the reader with the concept of topic modeling, focusing on how it is used to extract a set of topics from a document corpus and annotate documents with topics based on the document words.

2.2 Visualizing Topic Models

Numerous research projects in topic model visualization revolve around visualizing documents to show relationships or similarities based on their latent topics. *Probabilistic Latent Semantic Visualization* (PLSV) [7] is a topic model approach to visualizing documents and topics as coordinate points in a visualization space. The distances between documents and topics are based on the topic distribution of a document. *Topic maps* [9] and *Exemplar-based Visualization* (EV) [3] provide similar

graphs of a large collection of documents, with document points color-coded by their dominant topic. The Stanford Dissertation Browser [4] is also a notable visualization developed to evaluate word and topic similarities between the Ph.D. theses of different departments over time. The general purpose of these visualizations is to show documents with similar topics in clustered areas for a global overview of the corpus.

Now we turn to a few systems that are more relevant to our research in terms of their goals, end-users, or visual design. We are focused on summarizing individual documents using topics, revealing topic trends of a document over time, and indexing topics within document text using simple visualizations for non-technical users. Our developed visualizations were inspired by different aspects of these projects.

The Wikipedia navigator [2] was specifically designed to summarize the corpus and show relationships between textual content and topics for non-technical users. Three straightforward visualizations were produced: an overview page that lists the set of topics associated with all documents, a topic page that displays associated words as well as related document and topic links, and a document page showing the content in addition to related document links and a pie chart of related topics. These visuals allow the user to be completely unaware of the underlying LDA topic models.

The interactive visual text analysis tool TIARA [8] summarizes a corpus over time using a stream graph with topic layers and distributed keywords. ThemeRiver [6] provides the same type of graph without keywords. The height of the topic layer areas illustrate the strength of each topic at a certain point in time. Although I personally find stream graphs difficult to comprehend, these visualizations show that area or line charts can be useful for expressing topic trends over time.

Finally, Termite [**?**] is a visual analysis tool for evaluating the quality of topic models. The main visualization of this tool is a term-topic matrix that can be described as a scatter plot of words for each topic, with the size of each point proportional to the word frequency for that topic. Clicking on a topic in this matrix shows its representative documents and a one-dimensional plot of where topical terms can be found within each document. These simple designs seem effective for visually indexing topics in each document.

2.3 Visualization Techniques

2.4 Scope and Limitations

Chapter 3

Visualization Design

3.1 Topic List

3.2 Donut Chart

3.3 Line Chart

3.4 Scatter Plot

Chapter 4

System Implementation

4.1 Front-End Website

4.2 Back-End Server and Database

4.3 Topic Model

4.4 Visualizations

4.5 Texting Integration

Chapter 5

Project Evaluation

Chapter 6

Future Work

6.1 Currently Possible Features

6.2 Additional Resources

6.3 Advanced Topic Models

Chapter 7

Conclusion

Appendix A

Tables

Table A.1: Armadillos

Armadillos	are
our	friends

Appendix B

Figures

Figure B-1: Armadillo slaying lawyer.

Figure B-2: Armadillo eradicating national debt.

Bibliography

[1] David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, April 2012.

[2] Allison J. B. Chaney and David M. Blei. Visualizing topic models. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, pages 419–422, 2012.

[3] Yanhua Chen, Lijun Wang, Ming Dong, and Jing Hua. Exemplar-based visualization of large document corpus. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1161–1168, 2009.

[4] Jason Chuang, Christopher D. Manning, and Jeffrey Heer. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '12, pages 74–77, New York, NY, USA, 2012. ACM.

[5] Jason Chuang, Daniel Ramage, Christopher Manning, and Jeffrey Heer. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 443–452, New York, NY, USA, 2012. ACM.

[6] Karthik Dinakar, Birago Jones, Henry Lieberman, Rosalind Picard, Carolyn Rose, Matthew Thoman, and Roi Reichart. You too?! mixed-initiative lda story matching to help teens in distress. *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[7] Susan Havre, Elizabeth Hetzler, Paul Whitney, and Lucy Nowell. Themeriver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):9–20, 2002.

[8] Tomoharu Iwata, Takeshi Yamada, and Naonori Ueda. Probabilistic latent semantic visualization: Topic model for visualizing documents. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 363–371, New York, NY, USA, 2008. ACM.

[9] Shixia Liu, Michelle X. Zhou, Shimei Pan, Weihong Qian, Weijia Cai, and Xiaoxiao Lian. Interactive, topic-based visual text summarization and analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 543–552, New York, NY, USA, 2009. ACM.

[10] David Newman, Timothy Baldwin, Lawrence Cavedon, Eric Huang, Sarvnaz Karimi, David Martinez, Falk Scholer, and Justin Zobel. Visualizing search results and document collections using topic maps. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(2):169–175, 2010.