Project Charter: Shoah Foundation

Krishna Bhogaonker

IPAM-RIPS

Revision History

Revision	Date	Author(s)	Description
0.99	06-03-2015	krishnab	created
1.0	06-05-2015	krishnab	Updated based upon comments from Christian Ratsch, Russ Calfisch, and Tim Tangherlini. Executive summary expanded to include topic models beyond LDA. Added additional student to list of team members.sprin

Contents

1 EXECUTIVE SUMMARY

When a scholar steps through the doors of the Shoah archive, he/she expects the archivist and computer search tools to help them locate artifacts relevant to their specific search query. However, defining which artifacts are relevant to a query and the priority of those relevant artifacts are among the key challenges that collections such as the Shoah Foundation face.

However, archivists face challenges on another front as well. As their archives grow in size, defining relevance becomes even more difficult. Not only are archivists asked to find the needle in the proverbial haystack, they are asked to find a needle in a haystack that is growing all around them. Indeed, as the archive's collection continues to grow year upon year, assigning relevance and prioritizing search results become more important even as these tasks become more difficult.

Existing archive search methods rely on keyword searches and keyword correlations to assign relevance to search results. Keyword methods, however, can be imprecise, especially as the scale of the archive grows. As an archive grows, the number of possible keywords and the number of artifacts with that keyword also grows. To meet this challenge search tools must extract more information from the query phrasing and more intelligently assign relevance based upon more than just keyword associations.

The 2015 IPAM-RIPS(Research in Industrial Projects) team will develop a novel text analysis tool to assist Shoah archivists better identify relevant artifacts/interviews, given a particular research query. Using machine learning methods, the team will identify sets of "latent" or undiscovered relationships between archived interviews. Then using tools such as Latent Dirichlet allocation, Non-negative matrix factorization, etc., the RIPS team's tool will assign each interview to the top k topics based upon the content of the observed video "tags." Assignment of these latent topics to archive artifacts enables search methods to incorporate more of the artifact's semantic information into the assignment of relevance. The final results should produce a more accurate and better prioritized set of search results given a query.

Not only are archivists asked to find the needle in the proverbial haystack, they are asked to find a needle in a haystack that is growing all around them.

2 PROJECT PURPOSE

The purpose of this project is to create a search tool for the Shoah archive that improves the identification of relevant artifacts given a search query. Archivists face the daily challenge of indexing their collections so that users may search and retrieve appropriate documents and artifacts related to some particular research query. As an archive grows larger, the existing index methods may clash with new records because the new records may not fit well within the existing indexing scheme–thus encouraging a revision to archive-wide indexes. Reindexing is no easy task. In lieu of reindexing challenges, archivists have difficulty imposing complex indexing structures that could improve the appropriateness of query results but incur substantial costs to recode. The tool that the RIPS team develops will reduce the complexity/cost of reindexing while simultaneously improving the accuracy of search results.

2.1 BUSINESS NEED/CASE

According to the Shoah Foundations's website, their priorities include:

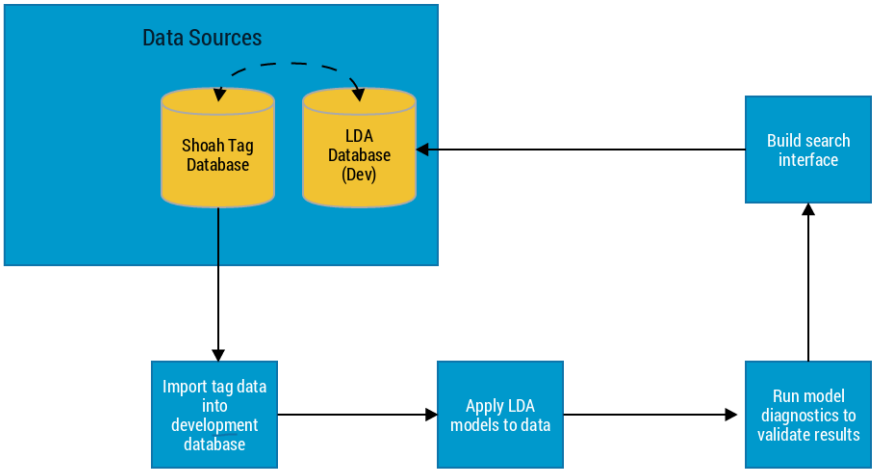
1. Making the archive a compelling voice for education and action
2. Developing content with consequence
3. Teaching the World through testimony
4. Sharing testimonies through global access

The proposed project applies directly to the second objective. This project will enable scholars to examine the existing interview archive in a novel way. Using statistical tools, the project team will use archive-wide trends to identify "hidden" or latent topics within the segment tags captured from individual testimony. These latent topics provide a new educational opportunity for scholars to examine archive wide relationships.

3 PROJECT DESCRIPTION

3.1 PROJECT OBJECTIVES AND SUCCESS CRITERIA

Using techniques developed by David Blei (Columbia) and John Lafferty (Carnegie Mellon), among others, the team will apply Latent Dirichlet Allocation(LDA) topic models to the Shoah interview tag database. Of course LDA is only one possible tool for topic analysis, and the RIPS team will explore multiple machine learning approaches to extract topics from interview artifacts. The outline of project activities is documented in Figure 1 below.



Project Activity

FIGURE 1

The success of the project depends upon the ability to generate novel search results compared to existing search results using keywords or shadow keywords.

3.2 REQUIREMENTS

We require the following support to execute this project.

1. Read access to video library tag database, either through network access or local copy.
2. Access to Shoah technical team to help resolve questions or access issues.
3. A sample of test users to evaluate the search interface and results. These testers may be the Shoah technical staff.

3.3 CONSTRAINTS

Because this project depends upon a large dataset, we cannot estimate just how novel the resulting topic clusters will be. While we expect some overlap between the existing keyword and shadow-keyword searches, we cannot *ex ante* predict just how great the differences will be.

in response to this constraint, we will provide the Shoah technical team with instructions on how to re-generate the database after tuning some of the topic model parameters. This training will allow the Shoah team to update the latent topic search methods as new interviews and documents are added to their collection.

4 RISKS

The table below documents anticipated project risks and potential mitigation strategies.

Risk	Description	Mitigation
Model tuning duration	The model tuning process may take longer than expected due to the unfamiliarity with the dataset and the large number of adjustable parameters.	Work with users in advance to obtain a good set of use-cases. Develop a small training environment to gain familiarity with the data and models.
Technical challenges associated with managing large dataset	The size of the Shoah dataset may require special tools to query. Learning these tools or incorporating them into the workflow may require extra time.	Set up test database early on, even if with a subsample of the data. Also gather information on the existing Shoah database structure and tables.

5 PROJECT DELIVERABLES

1. Create an latent topic database that contains the list of latent topics and a table that indicates the latent topics associated with each interview.

2. Instructions on how to regenerate the database if tuning is required post-project. Appropriate scripts will be provided as well.

6 PROJECT TEAM

Name	Role	Contact Info
Krishna Bhogaonker	Project Mentor	uc.krishnab@gmail.com
Adam Foster	Technical Team Member	aef39@cam.ac.uk
Hangjian Li	Technical Team Member	lihangjian123@g.ucla.edu
Megan Shearer	Technical Team Member	megshearer@email.arizona.edu
Georg Maierhofer	Technical Team Member	gam37@cam.ac.uk