Appendix B: Topic Modeling with Lucerna data

This appendix describes how I transformed the .csv files generated by the SQL queries into individual files for topic modeling with MALLET.

From table to individual files

To automate the process of dividing one text file into many, I modified a php script written by Dr. Kalani Craig. The script 'explodes', or divides, the text file every time it sees a distinct string of numbers, letters, or spacing features. I changed Mac-specific language to PC and by modified the cue to explode the text to be five tabs ('\t\t\t\t\t'). After proofreading quotations from primary sources, I copied columns which contained the text's Lucerna ID, the event's Lucerna ID, the quotation for the primary source, and a column which had five tabs. Below is the text for the php:

```
?php
//php /path/to/wwwpublic/path/to/script.php arg1 arg2

//$Website = $argv[1];
//$Filename = $argv[2];

$string = file_get_contents("AllLucernaText.txt");
$Section = explode("\t\t\t\t\t", $string);
for ($I = 0; $I < count($Section); $I++)
{
        $Lines = explode("\t", $Section[$I]);
        $Title = array_shift($Lines);
//      $Title = array_shift($Lines);
//      file_put_contents ("LucernaTexts/".$I."_".preg_replace('~[\W\s]~', '_',
$Title).".txt",  $Lines[$I]);
        file_put_contents ("LucernaTexts/".$I."_".preg_replace('~[\W\s]~', '_',
$Title).".txt",  $Section[$I]);
}

?>
```

I saved this script as a php file using Sublime, a free text editor. Using the command terminal on my PC, I navigated to the folder with both the php script and the txt file that contained all the information from Lucerna. Once in this folder, I ran the php script. This generated a new folder called 'LucernaTexts' that contained all quotations of primary sources as individual files; the file names contained the event and text ID.

From individual files to topic model

For this study, I followed the directions for installing and using MALLET from the Programming Historian blog. The post provides detailed instructions for PC and Mac users. The following code is a lightly modified version from the blog; it loads the individual files from 'LucernaTexts' folder on my PC, MALLET's stopword list, and a customized stopword list (described below). For the sake of simplicity, I put the folder containing the texts from Lucerna in the same folder as the sample-data included in the MALLET download.

```
bin\mallet import-dir --input sample-data\LucernaTexts --output Lucerna.mallet --keep-sequence --remove-stopwords --extra-stopwords C:\Mallet\stoplists\lantern.txt
```

The code below runs the topic model. I added 'random-seed' so that I could control where the topic modeler started. This enabled me to experiment with different optimization intervals as well as the number of topics and words in a topic. Setting the random seed also meant that my results would be repeatable.

```
bin\mallet train-topics  --input Lucerna.mallet --random-seed 750 --num-top-words 8 --num-topics 25 --optimize-interval 10 --output-state sample-data\LucernaResults\versionthree.gz --output-doc-topics sample-data\LucernaResults\versionthree-doc.csv --output-8topic-keys sample-data\LucernaResults\versionthree-doc.txt
```

Stopword List

I developed a customized stop word list in order to prevent the most commonly appearing terms from appearing in the topics, the most obvious being ‘magic’ and ‘lantern’. Eliminating these terms from the topic model mitigated false correlations. I also excluded terms that related to specific organisations and types of events because this information was represented in other tables. These included

Organization Names: church, army, van, c. a, c.a., band, hope, Sabbath, school, union
Type of event: service, services, lecture, lectures, meeting, meetings
Location: mission, village
Titles: capt, captain, rev, reverend, mr, miss, mrs
Typographical features: ndash, Unlesbar
Time: monday, tuesday, wednesday, thursday, friday, saturday, sunday, morning, evening, night, pm, p.m, p.m.

There were several terms regarding location that I added, then removed from the stopword list: parish, chapel, hall, room, schoolroom. These terms turned out to be pivotal nodes in the topics. ‘Lecturer’ functioned similarly in topics relating to scientific lectures.

Reuniting the data

In order to study topics by organization and by location, I copied and pasted the text file that described the percentages of each topic in each document into a spreadsheet. In order to find the most representative topic in each document, I used the MAX function in Microsoft Excel. I then created a table using the VLOOKUP function that described events, their location, organisations associated with the event, quotes from primary sources, and that text’s top topic.