# Homework #3: Using and Understanding MALLET

HIST 318, Dr. McDaniel
Due Monday, February 3, 3 p.m.

## Introduction

For our January 31 class, you read several articles about using a method called "topic modeling" to "read" texts algorithmically. In this homework assignment, you will have a chance to use MALLET, a topic modeling software package, yourself and then write a reflection on your experience that applies what you have learned to our class project.

## Before You Begin

This assignment will require you to use the command line on your computer. I recommend that before you begin, you review some of the material on this that we covered in class on Friday.

If you have a Mac or Linux machine, the Command Line Bootcamp from the Scholars' Lab at the University of Virginia is a useful place to begin, and it is aimed at humanities students and scholars. If you have a Windows machine, here is a basic introduction to the DOC prompt.

Regardless of your machine, there are three main things you will need to be able to do in this assignment from the command line, so make sure you understand how to do each of them:

- See what directory you are currently in.
- Change directories.
- List the contents of the current directory.
- See inside the contents of a file.

You may also want to know how to clear your terminal screen if it becomes too crowded with text. You can do this with the command `cls` at the Windows command prompt and the command `clear` at the Unix/Mac command line. (Even after clearing the screen, you should be able to scroll up in your terminal windows to see what you've done in the past.)

## Objectives

1. To gain a basic familiarity with the command line.
2. To install and use MALLET with the sample data included in the package.
3. To reflect on the uses and limitations of topic modeling in historical research.
4. To gain experience and confidence in following a detailed tutorial for an unfamiliar tool.

## Requirements

There are both technical and non-technical requirements for this assignment, but the two parts are separable. I recommend that you attempt the technical part first since it will probably take longer, but if you get stuck, you should be able to answer the questions in the non-technical part before completing the techy stuff.

### Technical Requirements

Complete the tutorial on Getting Started with Topic Modeling and MALLET at the Programming Historian, which will show you how to install MALLET and then use it on the sample documents included with the package.

This requirement will be completed when you **tweet two screenshots of your work** to the course hashtag `#ricedh`. More specifically:

- One screenshot should, like Figure 8 in the tutorial, show the output of a `train-topics` command on the sample data set discussed in the tutorial, but should show that you generated **15 topics** instead of the default 10.
- One screenshot should, like Figure 10 in the tutorial, show a screenshot of the `tutorial_composition.txt` file generated by your 15-topic model opened in Excel. (If you don't have Excel installed on your computer, you can also satisfy this requirement by creating a GitHub Gist containing the contents of your `tutorial_composition.txt` file and tweeting the link to the Gist instead.)

If you are not familiar with how to take screenshots on your computer, do some Googling to find out the answer, or ask on Twitter for help. You will also need to learn how to post photos on Twitter.

### Non-Technical Requirements

After reading the Friday texts about topic modeling and trying out MALLET yourself, you should be able to figure out answers to the following two questions:

1. Suppose we wanted to create a topic model of the runaway slave ads we have collected on our Google Spreadsheet. What first steps would we have to take to get from our spreadsheet of permalinks to a `*.mallet` file that we could train topics on?
2. In his Mining the Dispatch project, Robert K. Nelson used MALLET to find articles that were likely to be fugitive slave ads in a large corpus of digitized newspapers. What feature(s) of the Portal to Texas History would have prevented us from using the same method to discover ads in the *Telegraph and Texas Register*? Be as specific and thorough as possible. (Here's a hint: do some searching for keywords in the *Telegraph and Texas Register* on the Portal, and notice what kinds of results you get back. Does the kind of result returned by a keyword search tell you something about

the way that the underlying text documents in the Portal are stored and separated from each other?)

Write up an email to me answering *both* of these questions. You should be able to answer them with just a few sentences in each case—no more than two good-sized paragraphs should do the job.

## Summary and Evaluation

Successful completion of this assignment will include:

- Two screenshots posted to Twitter to satisfy the technical requirements.
- An email to me answering the two non-technical questions.

Because this assignment has several, separable part, I will divide up the points for the assignment this way when evaluating your homework: two points for each screenshot, and three points for each answer in the email.

## Help! I'm Stuck!

There is a good possibility you'll encounter technical difficulties when doing this assignment. Don't fret or bang your head against the wall all weekend if you are getting an error message that is not mentioned in the tutorial, or if you are having trouble getting the same results shown in the tutorial. Instead, get help!

You can always take to Twitter if you need help. If you are getting error messages in your terminal that are longer than 140-characters or difficult to explain, you can also use a Gist, as you did in the first homework, to get help. Copy and paste the strange output of your terminal into a Gist, putting an explanation of what produced it in the Gist "description," and then tweet the URL to that Gist to our course hashtag to see if I or another student can help. (And remember, helping out other students is a way to score well on the Team Participation part of your grade.)

Remember, though, the academic integrity policies for the course. Do not get someone else to do the work for you and be sure to acknowledge any pointers or technical assistance you received—in this case by noting it in your email to me.