Text Analysis with Mallet

Author

Table of Contents

[**Introduction** 3](#_Toc445938623)

[**Installing Mallet** 3](#_Toc445938624)

[**Testing Your Mallet Installation** 5](#_Toc445938625)

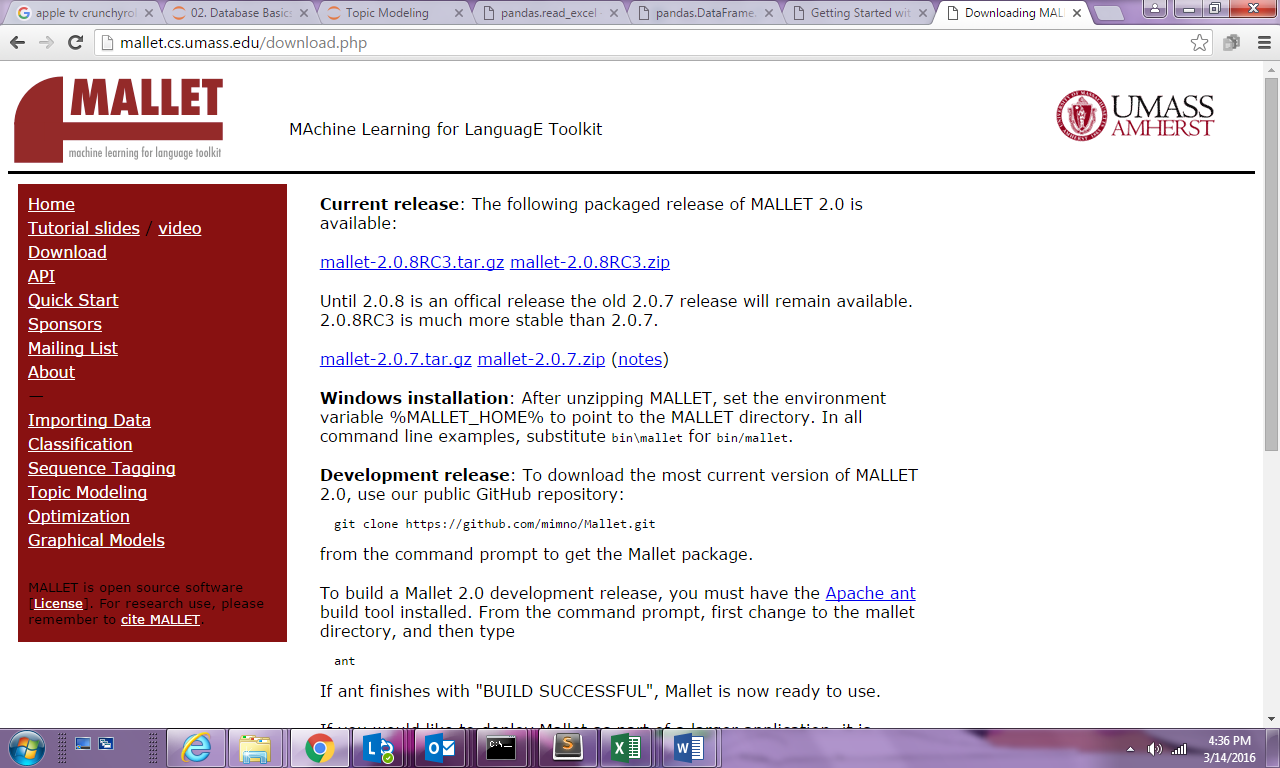
# **Introduction**

Text analysis is the process of applying analytical techniques to the content of natural langrage and text to collect information. Text analysis is a valuable and growing field because much of the ‘data’ people share with each other is in the form of natural language. Sending an email to a college, writing a movie review, or delivering a speech are forms of data sharing via text. For example, an activist might write an option column for a local newspaper expressing his views on an issue. Social media outlets such as twitter allow users to share text data with their followers. Text analysis can be understood as the process of taking natural language and organizing its features in to be quantifiable. A researcher may use Twitter to look at the rise and fall of certain slang terms and their usage over time. Job postings might be reviewed to look at changes in an industry over time.

This chapter will focus on Topic Analysis. This analysis done on a collection of text items to find categories which will best separate the text items into groups. For example, topic in analysis on Job postings for a particular company might reveal different collections of desired skills. Topic Analysis is a wide and developing discipline and there are many instructional resources available online. In this chapter, we will be learning the basics of Topic Modeling by attempting to categorized journal abstracts into topic categories.

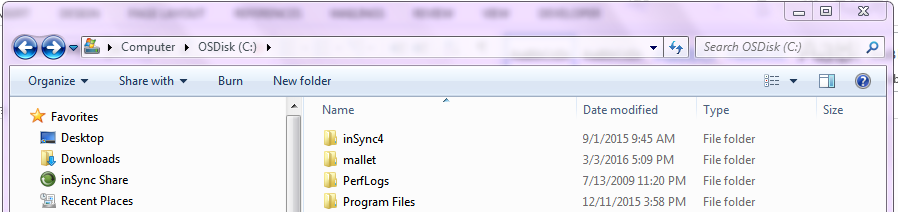
# **Installing Mallet**

For Topic Modeling in the chapter we will be using Mallet. Mallet is a Java based statistical learning package. You can download mallet from this site: <http://mallet.cs.umass.edu/index.php>.



**Download the zip file from** [**http://mallet.cs.umass.edu/index.php**](http://mallet.cs.umass.edu/index.php)

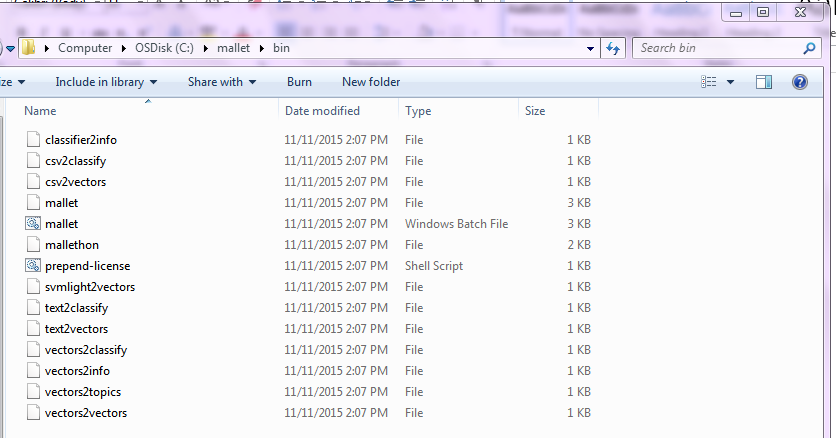
Extract the zip file to an easy to find location. It is suggested to place the folder directly under the main drive of your computer, for Example directly under the C drive.



**In the screenshot above I extracted the mallet folder under the C drive. The path the mallet folder is thus "C:\mallet"**

Using Mallet is not like using applications like Microsoft Word or MySQL Workbench. There is no user interface to open and use Mallet. Instead, Mallet is similar to a python module where we interact with Mallet and its features with code and commands. In this chapter’s notebook, it will be explained how to use python to communicate with mallet. But first, it is necessary to know where the batch file of mallet is located.

A batch file is essentially the main piece of code that allows us to access all of mallets features and abilities. By calling the batch file, we can then call mallet commands. Just like how python understands python commands, mallet understands mallet commands. From the main mallet folder, the batch file for mallet is located at "mallet\bin\mallet"



**The file highlighted above is the batch file. The batch file must be called before any mallet commands.**

The complete path to this batch file in this example is "C:\mallet\bin\mallet". To run a mallet command from the command line, you would type

> C:\mallet\bin\mallet <Mallet Command Here>

This method of calling mallet will be used frequently in the chapter

If you are interested in learning more about mallet and how to use mallet from the command line, this link is a great tutorial <http://programminghistorian.org/lessons/topic-modeling-and-mallet>

# **Testing Your Mallet Installation**

To test whether Mallet was successfully installed and that you have successfully identified the Mallet batch file, we can attempt to use the –help option on a Mallet command.

Open a command window and type the path to your mallet batch file, the mallet command “*import-dir*” and the option --*help*

For example,

> C:\mallet\bin\mallet import-dir –help

If this command returns the help documentation for the command “import-dir” then you have successfully installed mallet and identified the location of the batch file.