Text Analysis with Mallet

Installation and setup guide

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.

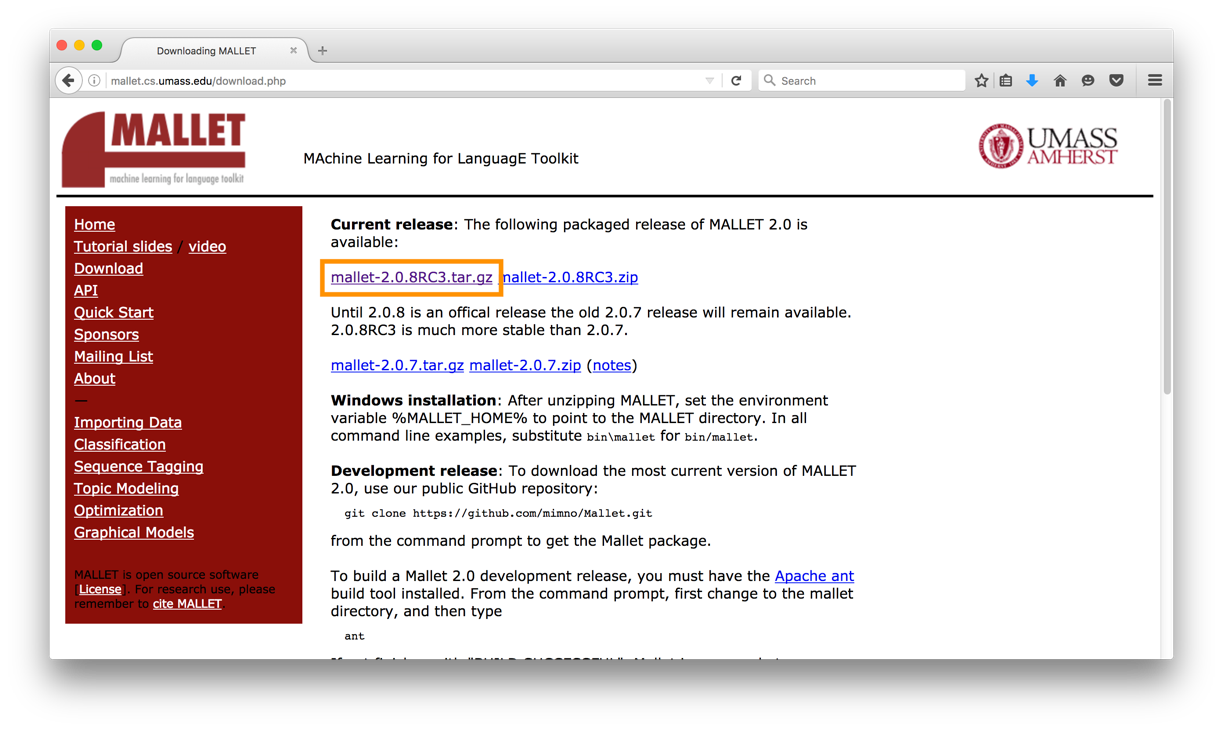
# **Installing Mallet**

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>.

## On Mac OS

### Install Mallet

Download the tar.gz file for the latest version of mallet (2.0.8RC3 at the time of this writing) from <http://mallet.cs.umass.edu/download.php>:



Next, we’ll uncompress and un-tar the downloaded tar.gz file. You can either just double-click the file in Finder or, in a command prompt in the folder where the tar.gz file was downloaded, enter the following command (using file name mallet-2.0.8RC3.tar.gz as an example):

tar –xvzf mallet-2.0.8RC3.tar.gz

This will result in a folder named after the version of the file (so for “mallet-2.0.8RC3.tar.gz”, there will be a folder named “mallet-2.0.8RC3”). Change this folder’s name to just “mallet”, and place it somewhere you can find it later – perhaps in your user directory so the path to the mallet folder would be “/Users/<your\_username>/mallet”.

### Install Java

From the Oracle Java download site, <http://www.oracle.com/technetwork/java/javase/downloads/index.html>, download the latest version of the Java Development Kit (JDK).

Double-click the DMG archive that you downloaded, then run the installer inside.

One the installer has completed, to verify that you have Java installed, open a command prompt and type “java -version”. You should see something like the following:

java version "1.8.0\_101"

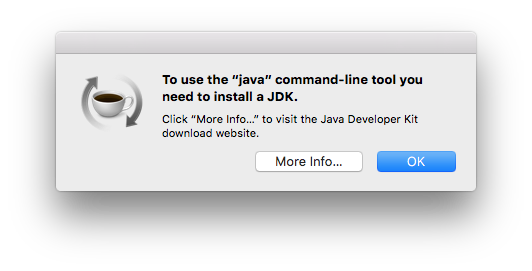
Java(TM) SE Runtime Environment (build 1.8.0\_101-b13)

Java HotSpot(TM) 64-Bit Server VM (build 25.101-b13, mixed mode)

If you see a line that says:

No Java runtime present, requesting install.

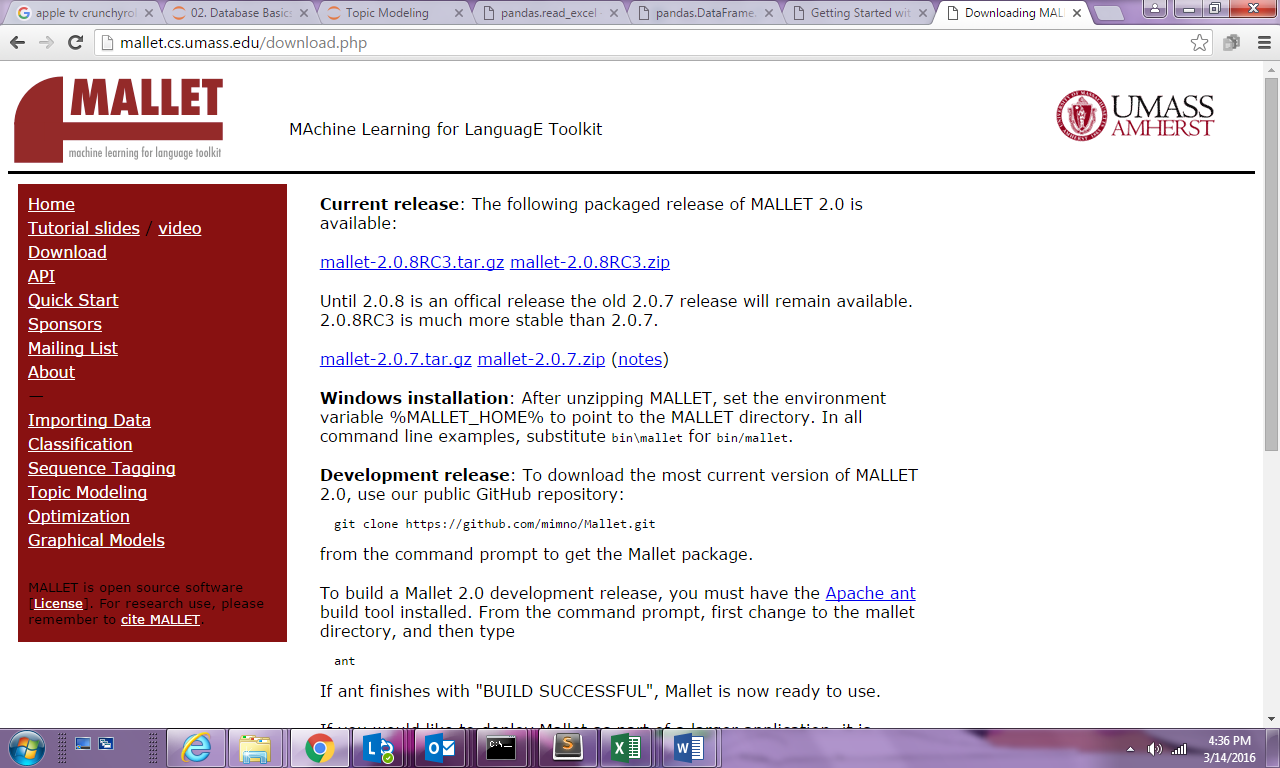
And a window like this:



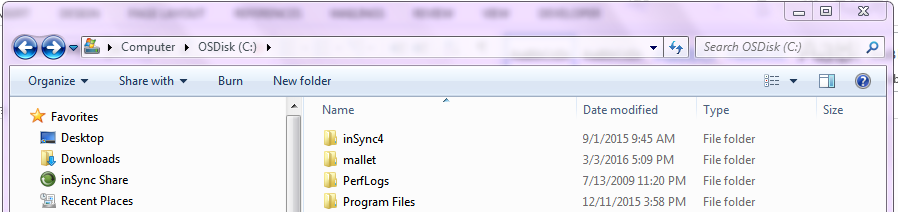
You don’t have java installed (and you don’t want to install the Java version the “More Info” button will offer you, from Apple – it is very old). Make sure that you’ve installed the latest Java Development Kit from Oracle, and not just a Java Runtime Environment (JRE).

## On Windows

**Download the zip file from** <http://mallet.cs.umass.edu/download.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"**

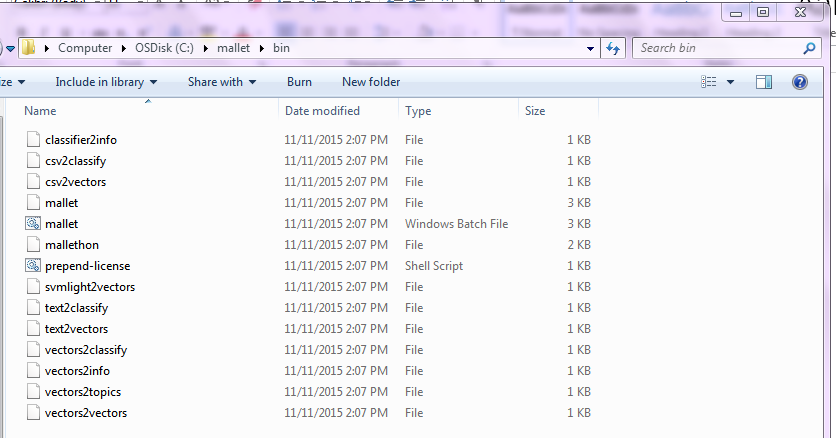
# Using Mallet

Using Mallet is not like using graphical user interface (GUI) applications like Microsoft Word or MySQL Workbench. Mallet has no user interface that one opens and uses. Instead, Mallet is a command line program. To use Mallet, you run the “mallet” executable in your computer’s command shell, and you pass the “mallet” command parameters to tell it what to do.

This chapter’s notebook explains how to use Python to run the “mallet” command without needing a command shell. In order for this to work, however, you’ll need to know the full path to the mallet command.

A command line program is a small program that can accept parameters and that uses the values of parameters passed to it to tell mallet what features the user is trying to make use of. Mallet has a series of actions it can perform. When we run the mallet command line program, we tell it which action we want it to perform, and details of how we want it to perform that action. It receives the parameters we pass it and converts those into correct calls to the requested action.

The “mallet” command line program is located inside the “mallet” folder you just downloaded and installed, in the “bin” folder. The relative path is “mallet/bin/mallet” on both Windows and Mac OS. The absolute path (needed for the notebook), is different depending on your operating system.



**The file highlighted above is the “mallet” command line program on windows.**

## On Mac OS

On Mac, if you installed mallet in your user’s home directory, the complete path to the “mallet” command line program would be:

/Users/<your\_username>/mallet/bin/mallet <mallet\_command>

So if your username is “jonathanmorgan”, the path to mallet would be:

/Users/jonathanmorgan/mallet/bin/mallet <mallet\_command>

## On Windows

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>

## Learning More

If you are interested in more detailed instructions on installing and using mallet (including ways to make the command easier to use), 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.

The mallet command itself is structured as such:

mallet <mallet\_command> <command\_parameters>

Mallet offers a number of commands, including “import-dir”, “train-topics”, and “infer-topics”. Each of these can then accept a number of different parameters to tell them how to do their work. This command syntax is discussed in greater detail in the workbook. To test, we will call the “import-dir” command, passing it the “--help” parameter to tell it we want to see help output (“--help” tells you all the available parameters for a given command – a good one to remember when using mallet).

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

For example, on windows, if you installed mallet in the root of the C drive:

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

Or on a mac, if you installed mallet in your user directory, and your username is “jonathanmorgan”:

$ /Users/jonathanmorgan/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. If not, make sure you got the path to the mallet program correct, and that you didn’t misspell the command or the parameter. If you see an error message similar to “file not found”, you likely have the wrong path. If you get an error about not being able to find “java”, you likely don’t have the Java Development Kit (JDK) installed.