

Installing MALLET

There are many tools one could use to create topic models, but at the time of this writing (summer 2012) the simplest tool to run your text through is called MALLET. [MALLET](#) uses an implementation of [Gibbs sampling](#), a statistical technique meant to quickly construct a sample distribution, to create its topic models. MALLET requires using the command line – we'll talk about that more in a moment, although you typically use the same few commands over and over.

While there is currently a preview release of MALLET 2.0.8 available, this lesson uses the official release of MALLET 2.0.7. If you are following along with our instructions, please be sure to download the correct version.

The installation instructions are different for Windows and Mac. Follow the instructions appropriate for you below:



Windows Instructions

1. Go to the [MALLET](#) project page, and [download MALLET](#) . (As of this writing, remember, we are working with version 2.0.7.)
2. You will also need the [Java developer's kit](#) – that is, not the regular Java that's on every computer, but the one that lets you program things. Install this on your computer.
3. Unzip MALLET into your `c:` directory . This is important: it cannot be anywhere else. You will then have a directory called `c:\mallet-2.0.7` or similar. For simplicity's sake, rename this directory just `mallet` .
4. MALLET uses an *environment variable* to tell the computer where to find all the various components of its processes when it is running. It's rather like a shortcut for the program. A programmer cannot know exactly where every user will install a program, so the programmer creates a variable in the code that will always stand in for that location. We tell the computer, once, where that location is by setting the

environment variable. If you moved the program to a new location, you'd have to change the variable.

To create an environment variable in Windows 7, click on your Start Menu -> Control Panel -> System -> Advanced System Settings (Figures 1,2,3). Click new and type `MALLET_HOME` in the variable name box. It must be like this – all caps, with an underscore – since that is the shortcut that the programmer built into the program and all of its subroutines. Then type the exact path (location) of where you unzipped MALLET in the variable value, e.g., `c:\mallet`.

To see if you have been successful, please read on to the next section.

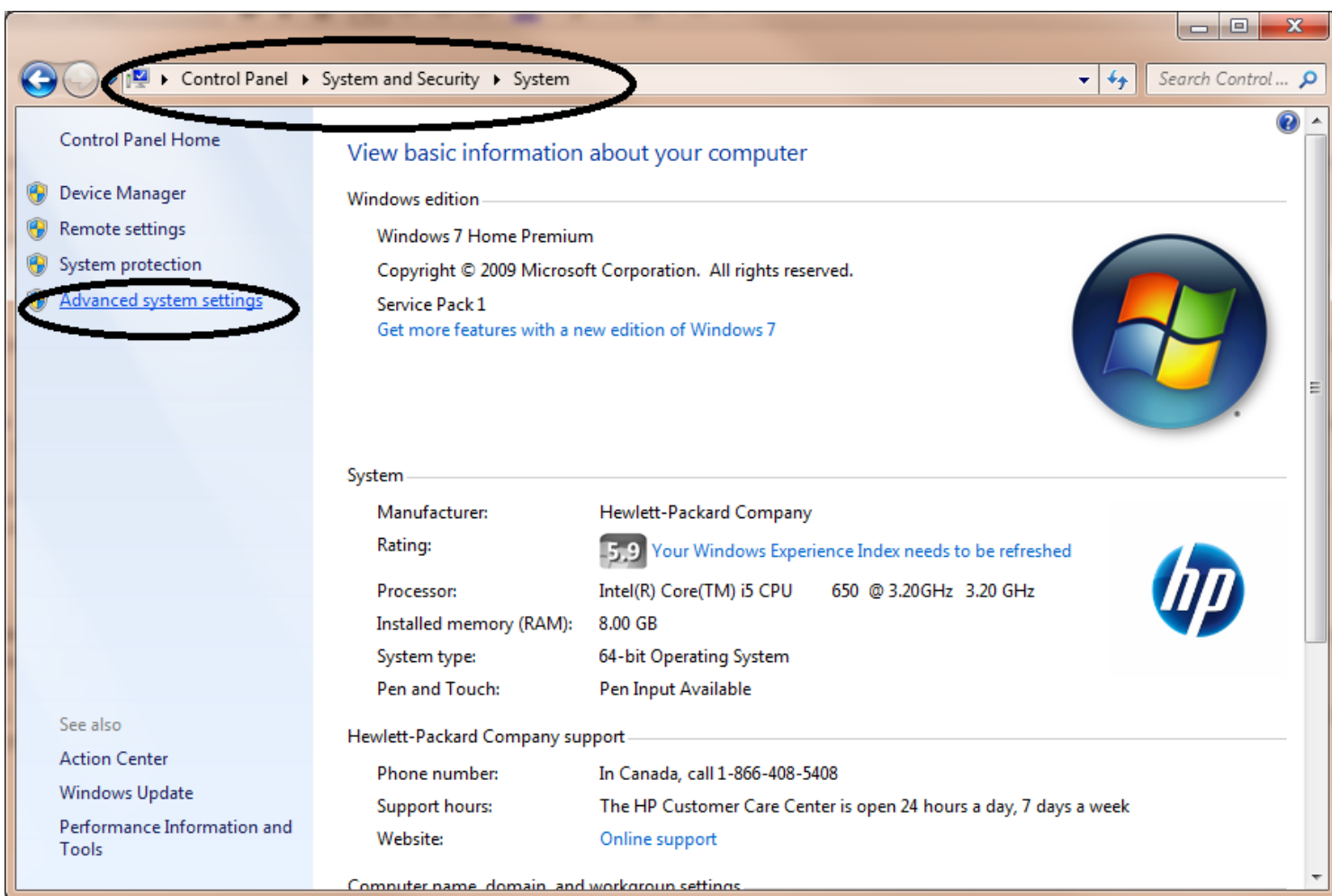


Figure 1: Advanced System Settings on Windows

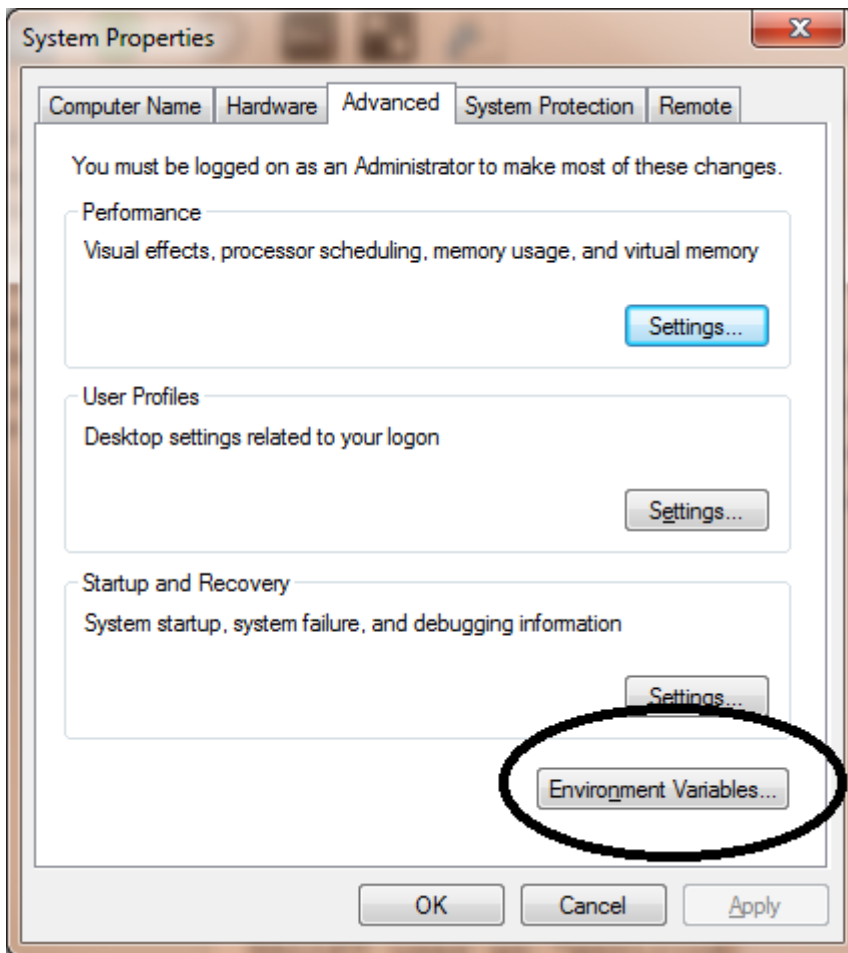


Figure 2: Environment Variables Location

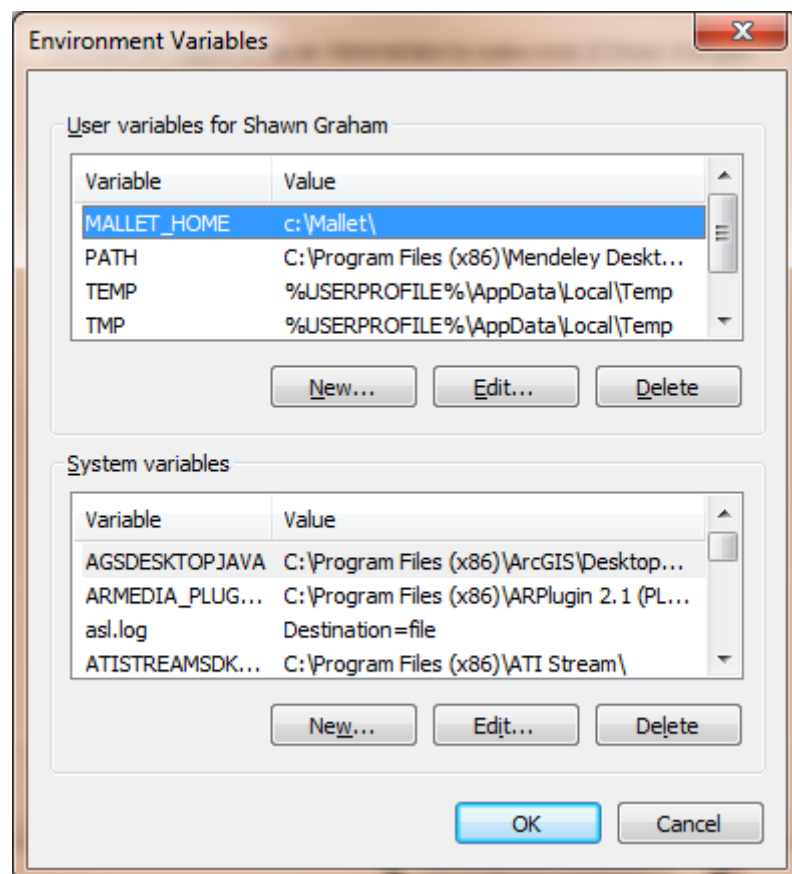


Figure 3: Environment Variable

Running MALLET using the Command Line

MALLET is run from the command line, also known as *Command Prompt* (Figure 4). If you remember MS-DOS, or have ever played with a Unix computer Terminal, this will be familiar. The command line is where you can type commands directly, rather than clicking on icons and menus.

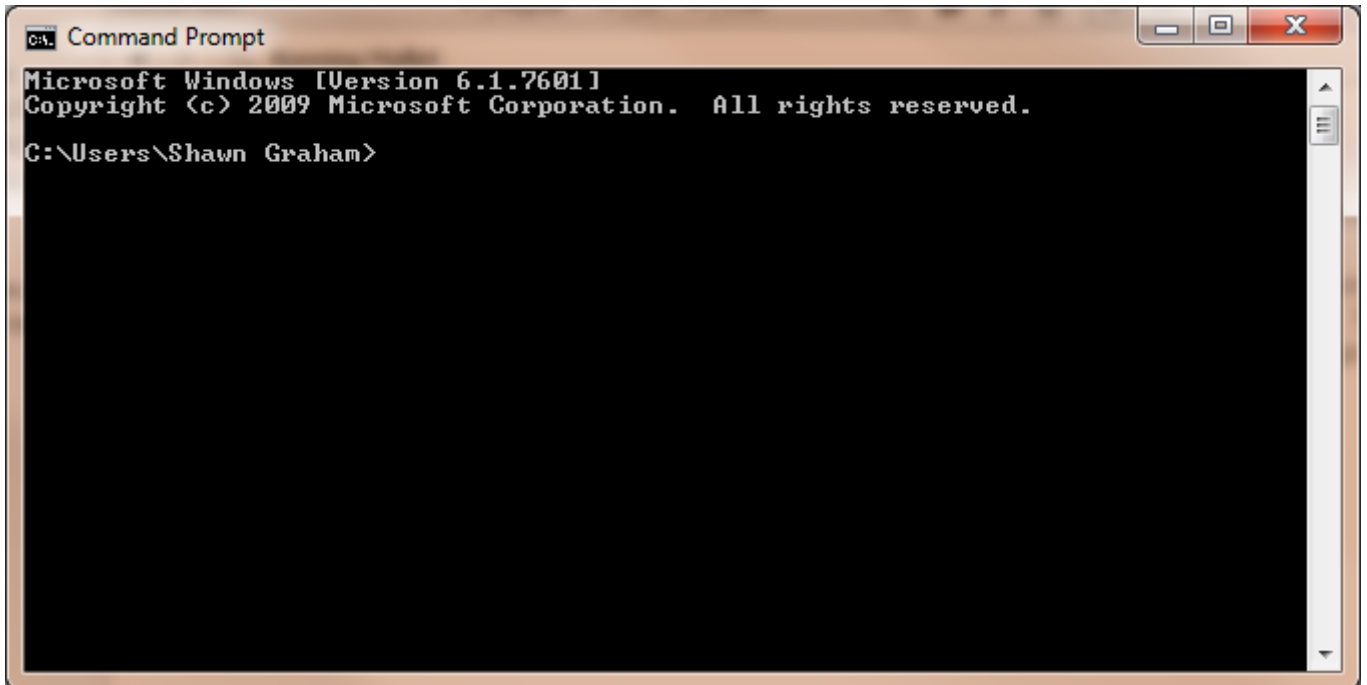


Figure 4: Command Prompt on Windows

1. Click on your Start Menu -> All Programs -> Accessories -> Command Prompt .\ You'll get the command prompt window, which will have a cursor at `c:\user\user>` (or similar; see Figure 4).
2. Type `cd ..` (That is: cd-space-period-period) to *change directory*. Keep doing this until you're at the `c:\` . (as in Figure 5)

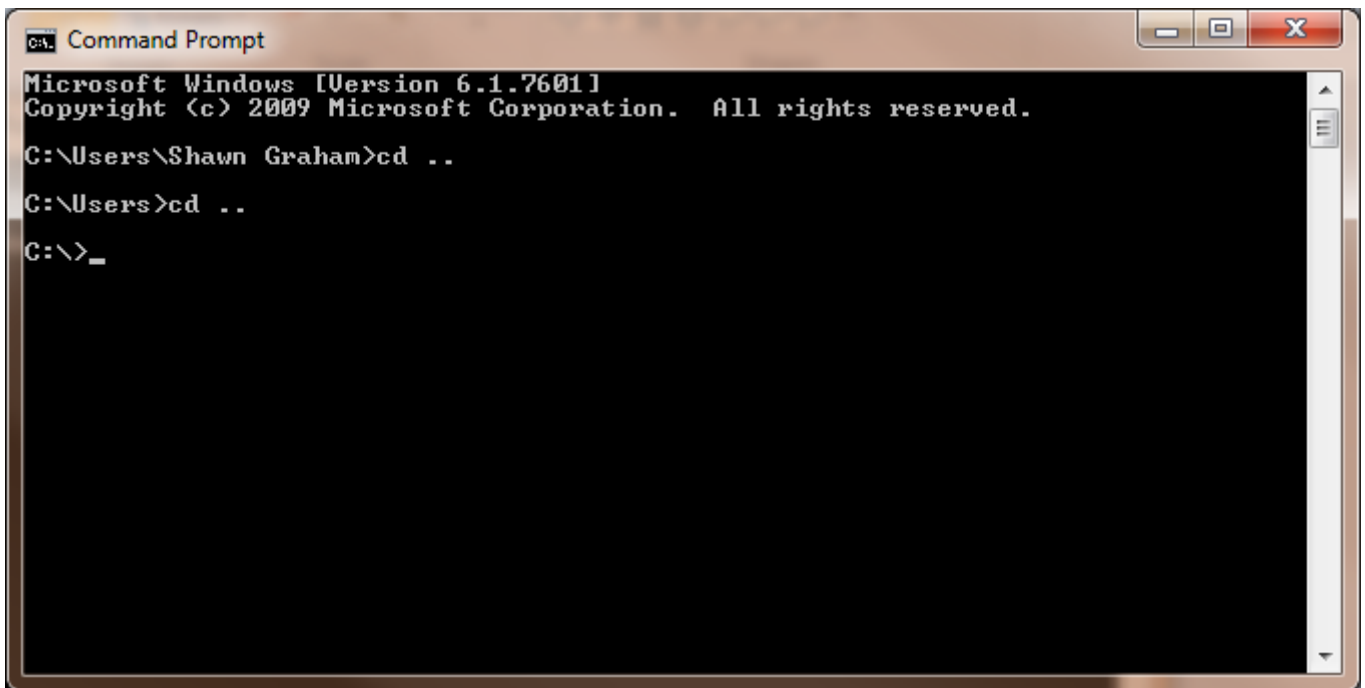


Figure 5: Navigating to the C:\ Directory in Command Prompt

1. Then type `cd mallet` and you are in the MALLET directory. Anything you type in the command prompt window is a *command*. There are commands like `cd` (change directory) and `dir` (list directory contents) that the computer understands. You have to tell the computer explicitly that 'this is a MALLET command' when you want to use MALLET. You do this by telling the computer to grab its instructions from the MALLET *bin*, a subfolder in MALLET that contains the core operating routines.
2. Type `bin\mallet` as in Figure 6. If all has gone well, you should be presented with a list of MALLET commands – congratulations! If you get an error message, check your typing. Did you use the wrong slash? Did you set up the environment variable correctly? Is MALLET located at `C:\mallet` ?

```
Command Prompt
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\Users\Shawn Graham>cd ..
C:\Users>cd ..
C:\>cd mallet
C:\Mallet>bin\mallet
Mallet 2.0 commands:
import-dir      load the contents of a directory into mallet instances (one
per file)
import-file     load a single file into mallet instances (one per line)
train-classifier train a classifier from Mallet data files
train-topics    train a topic model from Mallet data files
infer-topics    use a trained topic model to infer topics for new documents
estimate-topics estimate the probability of new documents given a trained mo
del
hlda            train a topic model using Hierarchical LDA
prune           remove features based on frequency or information gain
split           divide data into testing, training, and validation portions
Include --help with any option for more information

C:\Mallet>
```

Figure 6: Command Prompt MALLET Installed

You are now ready to skip ahead to the next section.



Mac Instructions

Many of the instructions for OS X installation are similar to Windows, with a few differences. In fact, it is a bit easier.

1. Download and [install MALLET 2.0.7 \(mallet-2.0.7.tar.gazas of Summer 2012\)](#).
2. Download the [Java Development Kit](#).

Unzip MALLET into a directory on your system (for ease of following along with this tutorial, your `/user/` directory works but anywhere is okay). Once it is unzipped, open up your Terminal window (in the `Applications` directory in your Finder. Navigate to the directory where you unzipped MALLET using

the Terminal (it will be `mallet-2.0.7`). If you unzipped it into your `/user/` directory as was suggested in this lesson, you can navigate to the correct directory by typing `cd mallet-2.0.7`). `cd` is short for “change directory” when working in the Terminal.

The same command will suffice to run commands from this directory, except you need to append `./` (period-slash) before each command. This needs to be done before all MALLET commands when working on a Mac.

Going forward, the commands for MALLET on a Mac will be nearly identical to those on Windows, except for the direction of slashes (there are a few other minor differences that will be noted when they arise). If on Windows a command would be `\bin\mallet` , on a Mac you would instead type:

```
./bin/mallet
```

A list of commands should appear. If it does, congratulations – you’ve installed it correctly!

Typing in MALLET Commands

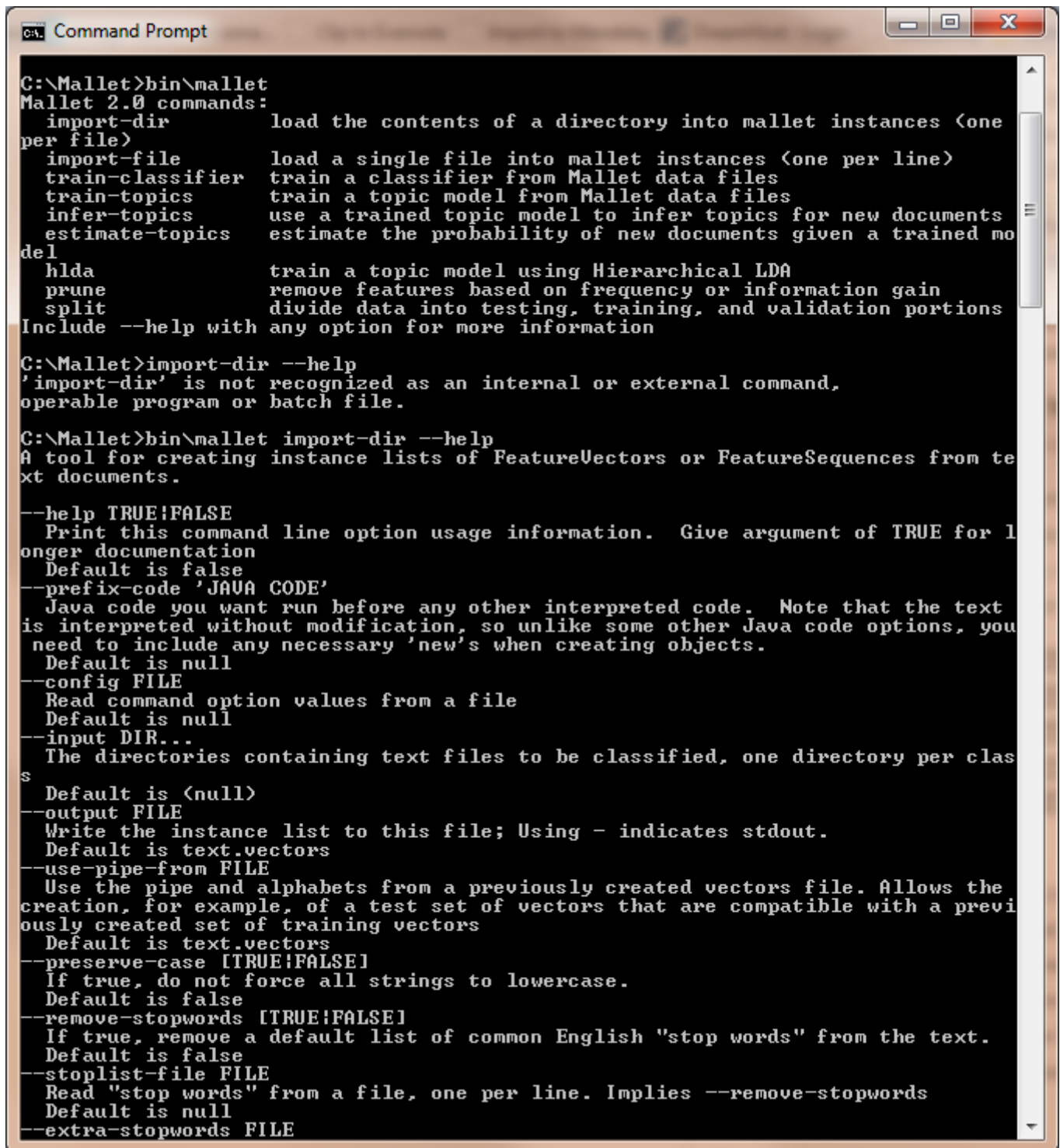
Now that you have MALLET installed, it is time to learn what commands are available to use with the program. There are nine MALLET commands you can use (see Figure 6 above). Sometimes you can combine multiple instructions. At the Command Prompt or Terminal (depending on your operating system), try typing:

```
import-dir --help
```

You are presented with the error message that `import-dir` is not recognized as an internal or external command, operable program, or batch file. This is because we forgot to tell the computer to look in the MALLET `bin` for it. Try again, with

```
bin\mallet import-dir --help
```

Remember, the direction of the slash matters (See Figure 7, which provides an entire transcript of what we have done so far in the tutorial). We checked to see that we had installed MALLET by typing in `bin\mallet` . We then made the mistake with `import-dir` a few lines further down. After that, we successfully called up the help file, which told us what `import-dir` does, and it listed all of the potential *parameters* you can set for this tool.



```

C:\Mallet>bin\mallet
Mallet 2.0 commands:
import-dir      load the contents of a directory into mallet instances (one
per file)
import-file     load a single file into mallet instances (one per line)
train-classifier train a classifier from Mallet data files
train-topics    train a topic model from Mallet data files
infer-topics    use a trained topic model to infer topics for new documents
estimate-topics estimate the probability of new documents given a trained mo
del
hlda           train a topic model using Hierarchical LDA
prune          remove features based on frequency or information gain
split         divide data into testing, training, and validation portions
Include --help with any option for more information

C:\Mallet>import-dir --help
'import-dir' is not recognized as an internal or external command,
operable program or batch file.

C:\Mallet>bin\mallet import-dir --help
A tool for creating instance lists of FeatureVectors or FeatureSequences from te
xt documents.

--help TRUE|FALSE
Print this command line option usage information. Give argument of TRUE for l
onger documentation
Default is false
--prefix-code 'JAVA CODE'
Java code you want run before any other interpreted code. Note that the text
is interpreted without modification, so unlike some other Java code options, you
need to include any necessary 'new's when creating objects.
Default is null
--config FILE
Read command option values from a file
Default is null
--input DIR...
The directories containing text files to be classified, one directory per clas
s
Default is (null)
--output FILE
Write the instance list to this file; Using - indicates stdout.
Default is text.vectors
--use-pipe-from FILE
Use the pipe and alphabets from a previously created vectors file. Allows the
creation, for example, of a test set of vectors that are compatible with a previ
ously created set of training vectors
Default is text.vectors
--preserve-case [TRUE|FALSE]
If true, do not force all strings to lowercase.
Default is false
--remove-stopwords [TRUE|FALSE]
If true, remove a default list of common English "stop words" from the text.
Default is false
--stoplist-file FILE
Read "stop words" from a file, one per line. Implies --remove-stopwords
Default is null
--extra-stopwords FILE

```

Figure 7: The Help Menu in MALLET

Note: there is a difference in MALLET commands between a single hyphen and a double hyphen. A single hyphen is simply part of the name; it replaces a space (e.g., `import-dir` rather than `import dir`), since spaces offset multiple commands or parameters. These parameters let us tweak the file that is created when we import our texts into MALLET. A double hyphen (as with `-help` above) modifies, adds a sub-command, or specifies some sort of parameter to the command.

For Windows users, if you got the error *'exception in thread "main" java.lang.NoClassDefFoundError:'* it might be because you installed MALLET somewhere other than in the `c:\` directory. For instance, installing MALLET

at `C:\Program Files\mallet` will produce this error message. The second thing to check is that your environment variable is set correctly. In either of these cases, check the Windows installation instructions and double check that you followed them properly.