Topic Modeling

Topic Modeling

A topic-modeling tool takes a text corpus and looks for patterns in the use of words. For large amounts of text, topic modeling provides a quick way to get “the lay of the land”, to get a sense of what the corpus is all about. This “**distant reading**” of a corpus is not a substitute for “**close reading**”, but it is a good start, like statistics’ EDA (Exploratory Data Analysis).

Topic models are computer programs that extract topics from texts. A **topic**, for these computer programs, is a list of words that occur in statistically meaningful ways. A **text** can be anything: a novel, a university mission statement, a newspaper editorial, an email, a blog post, a book chapter, a journal article, a diary entry. This text is unstructured, i.e., it does not contain any computer-readable annotations (tags) that tell the computer the semantic meaning of the words in the text.

There are many different topic-modeling programs available; Mallet, written in Java, is one of the best known ones, albeit its development has now been abandoned by its developers, for a different approach. Gensim is a Python-based topic-modeling tool.

Topic Modeling and LDA are often cited together. But LDA is a special case of topic modeling created by David Blei et al. in 2002. Among the many topic modeling approaches, LDA is by far the most popular. The myriad variations of topic modeling have resulted in an alphabet soup of techniques and programs to implement them that might be confusing or overwhelming to the uninitiated; ignore them for now. They all work in much the same way. Both MALLET and Gensim use LDA.

Mallet also uses an implementation of Gibbs sampling, a statistical technique meant to quickly construct a sample distribution, to create its topic models.

Your Corpus

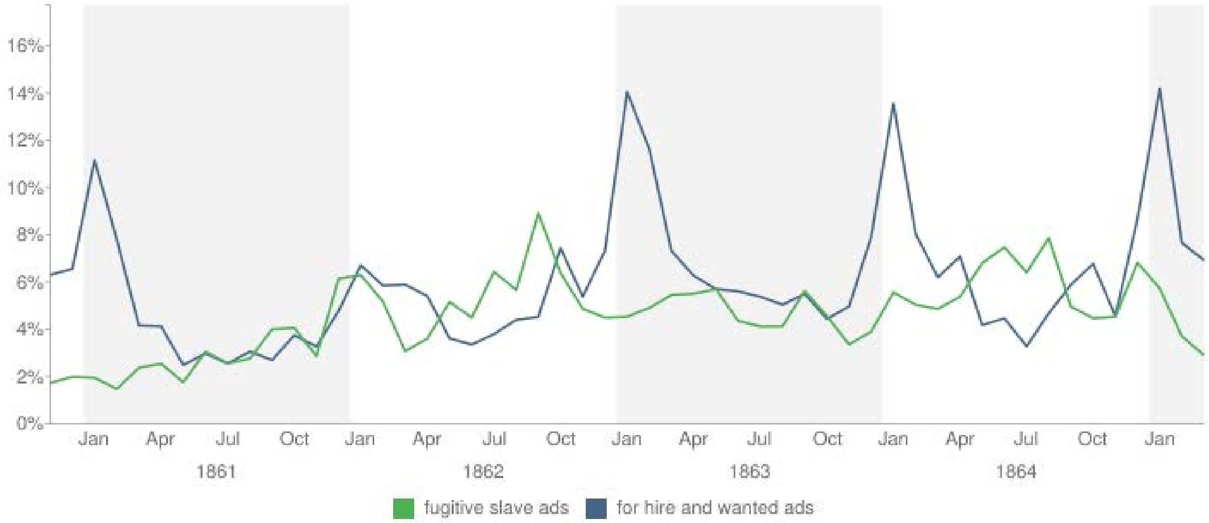
You want to put everything you wish to topic model into a single folder. This folder needs NOT be inside the Mallet directory (e.g., c:\mallet\mydata). But, wherever you put it, the full directory path of your corpus data should **NOT contain any blanks** or Mallet will bomb.

Your texts should be in **.txt format** (that is, you create them with Notepad, or in Word choose Save As -> MS Dos text).

What should your text files contain? It all depends upon the kind of analyses you want to carry out. Do you want to explore topics at a paragraph by paragraph level? Then each txt file should contain one paragraph. Things like page numbers or other identifiers can be indicated in the name you give the file, e.g., pg32_paragraph1.txt. If you are working with a diary, each text file might be a single entry, e.g., april_25_1887.txt. **Note that when naming folders or files, do not leave spaces in the name. Instead use underscores to represent spaces**. If the texts that you are interested in are on the web, you might be able to automate this process.

Chronologically-ordered Corpus

With a corpus of text files arranged in chronological order (e.g., 1.txt is earlier than 2.txt), then you can graph this output in your spreadsheet program, and begin to see changes over time (see Robert Nelson *Mining the Dispatch*, http://dsl.richmond.edu/dispatch/).



Software Options

R

R provides a topic-modeling algorithm (Structural Topic Modeling) that takes advantage of a date embedded in the filename as metadata (e.g., The New York imes_8-9-1999) to construct dynamic topics that may change overtime. Structural Topic Modeling.

Stanford Topic Modeling Toolbox

The Stanford Topic Modeling Toolbox (TMT) (https://nlp.stanford.edu/software/tmt/tmt-0.4/) brings topic modeling tools to social scientists and others who wish to perform analysis on datasets that have a substantial textual component. The toolbox features that ability to:

Stanford Topic Modeling Toolbox

Load a TMT script into a new tab using the File -> Open script.

Copyright (c) 2009. The Board of Trustees of
The Leland Stanford Junior University. All Rights Reserved.

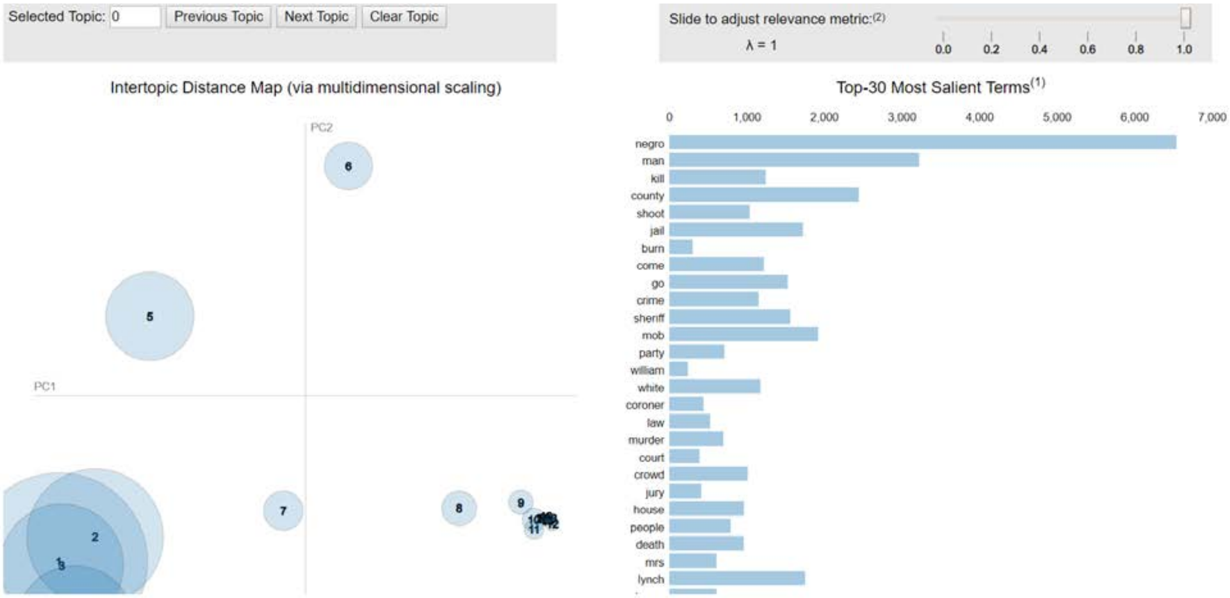
This is the current release.

- Import and manipulate text from cells in Excel and other spreadsheets.
- Train topic models (LDA, Labeled LDA, and PLDA *new*) to create summaries of the text.
- Select parameters (such as the number of topics) via a data-driven process.
- Generate rich Excel-compatible outputs for tracking word usage across topics, time, and other groupings of data.

Like Mallet, “TMT was written during 2009-10 in what is now a very old version of Scala, using a linear algebra library that is also no longer developed or maintained.”

Gensim

Gensim is a Python-based topic-modeling LDA algorithm (https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/). Gensim will display dynamic topic results as an html file. Since it is not easy to install Gensim, we created a Python algorithm for you.



Mallet

For an easy tutorial on Mallet, see the online document by Shawn Graham, Scott Weingart, and Ian Milligan http://programminghistorian.org/lessons/topic-modeling-and-mallet from which some of this documentation was taken.

Mallet

Mallet Installation

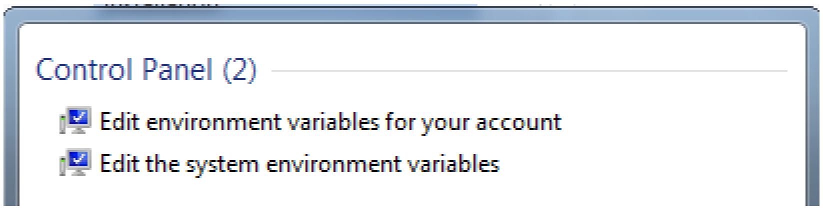
The PC-ACE setup comes with Mallet. It will be automatically setup for you. But, Mallet can be downloaded from http://mallet.cs.umass.edu/download.php.

Wherever you install Mallet, remember that Mallet will bomb if the full installation file path contains blanks. Thus, C:\Program files\Mallet will cause Mallet to fail. But not C:\Mallet_Installation\Mallet.

The same is true for the directory where the TXT corpus files are stored. You do not need to have these files under the Mallet subdirectory. But wherever you store your TXT files, do not leave any blanks in the full directory path of your corpus files. Thus, C:\My Text Corpus\ will cause Mallet to fail. But not C:\MyTextCorpus\ or C:\My_Text_Corpus\

MALLET installation also requires modifying an environment variable (essentially, setting up a short-cut so that your computer always knows where to find the MALLET program) and working with the command line (i.e., by typing in commands manually, rather than clicking on icons or menus). From PC-ACE, these line commands are carried out automatically by the Mallet routine.

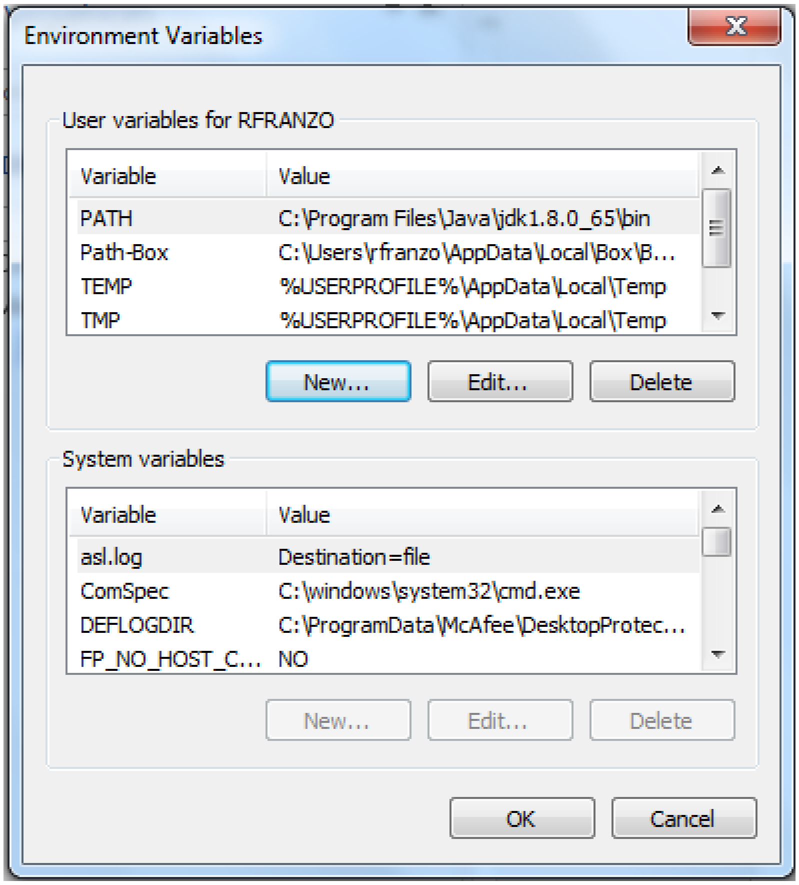
To modify environment variable, goto Start and enter environment in the search bar.



Control Panel (2)

- Edit environment variables for your account
- Edit the system environment variables

Click on the first entry and click on NEW.



Environment Variables

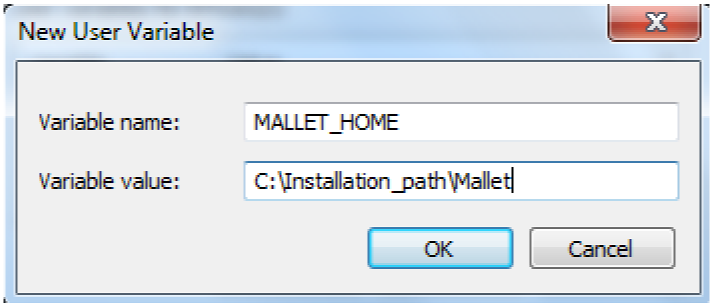
User variables for RFRANZO

Variable	Value
PATH	C:\Program Files\Java\jdk1.8.0_65\bin
Path-Box	C:\Users\rfranzo\AppData\Local\Box\B...
TEMP	%USERPROFILE%\AppData\Local\Temp
TMP	%USERPROFILE%\AppData\Local\Temp

System variables

Variable	Value
asl.log	Destination=file
ComSpec	C:\windows\system32\cmd.exe
DEFLOGDIR	C:\ProgramData\McAfee\DesktopProtec...
FP_NO_HOST_C...	NO

Enter MALLET_HOME under Variable name and the full Mallet installation directory path (e.g., C:\Installation_path\Mallet) under Variable value. Then click OK. And OK again in the previous form.



New User Variable

Variable name: MALLET_HOME

Variable value: C:\Installation_path\Mallet

Mallet Options

You can specify a number of different options in Mallet. PC-ACE, by and large, will run Mallet with default options. To set non-default options, please run Mallet from the command promt, setting the options you want.

Number of Topics

How do you know the number of topics to search for? Is there a natural number of topics? What we have found is that one has to run the train-topics with varying numbers of topics to see how the composition file breaks down. If we end up with the majority of our original texts all in a very limited number of topics, then we take that as a signal that we need to increase the number of topics; the settings were too coarse. There are computational ways of searching for this, including using MALLETs hlda command, but for the reader of this tutorial, it is probably just quicker to cycle through a number of iterations (but for more see Griffiths, T. L., & Steyvers, M. (2004). “Finding scientific topics”. *Proceedings of the National Academy of Science*, 101, 5228-5235).

Optimize Topics Interval

In general, including **–optimize-interval** leads to better identification of topics.

hlda Parameter

The hlda parameter may help you identify the “best” number of topics in your corpus.

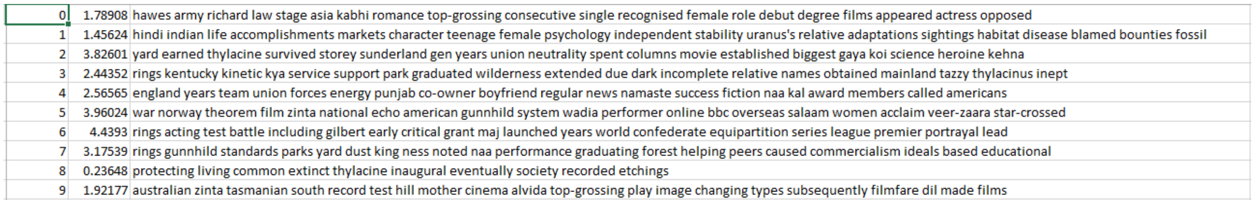
Dirichlet Parameter for the Topic

What Mallet Output Looks Like

It is important to note that MALLET includes an element of randomness, so the keyword lists will look different every time the program is run, even if on the same dataset.

Keys File

Using the sample-data EN in Mallet, this is what the keys file output looks like.



0	1.78908	hawes army richard law stage asia kabhi romance top-grossing consecutive single recognised female role debut degree films appeared actress opposed
1	1.45624	hindi indian life accomplishments markets character teenage female psychology independent stability uranus's relative adaptations sightings habitat disease blamed bounties fossil
2	3.82601	yard earned thylacine survived storey sunderland gen years union neutrality spent columns movie established biggest gaya koi science heroine kehna
3	2.44352	rings kentucky kinetic kya service support park graduated wilderness extended due dark incomplete relative names obtained mainland tazzy thylacinus inept
4	2.56565	england years team union forces energy punjab co-owner boyfriend regular news namaste success fiction naa kal award members called americans
5	3.96024	war norway theorem film zinta national echo american gunnhild system wadia performer online bbc overseas salaam women acclaim veer-zaara star-crossed
6	4.4393	rings acting test battle including gilbert early critical grant maj launched years world confederate equipartition series league premier portrayal lead
7	3.17539	rings gunnhild standards parks yard dust king ness noted naa performance graduating forest helping peers caused commercialism ideals based educational
8	0.23648	protecting living common extinct thylacine inaugural eventually society recorded etchings
9	1.92177	australian zinta tasmanian south record test hill mother cinema alvida top-grossing play image changing types subsequently filmfare dil made films

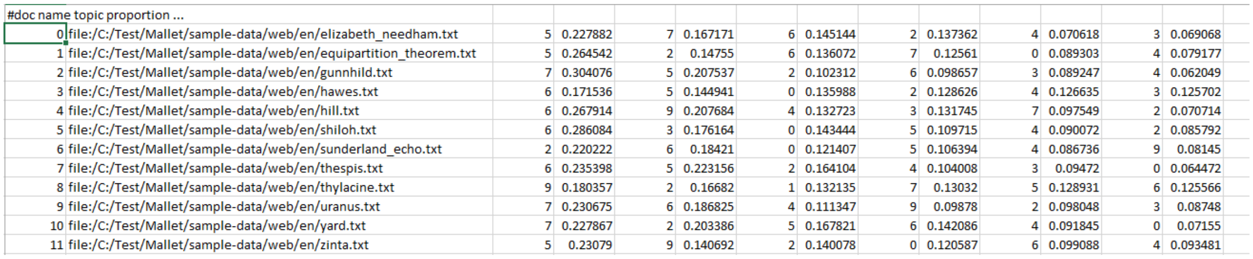
Column 1: contains the topic number, as many as specified in the parameter "Number of Topics" (topic 0, 1, 2, …);

Column 2: gives an indication of the weight of that topic across all the documents analyzed; it is the Dirichlet parameter for the topic; if this parameter is not specified, the values in the second columns will all be the same;

Column 3: The list of words in column 3 are key words that belong to each topic, one topic per line.

Composition File

The composition file has as many lines as documents imported (one document per line) and several columns, column 1, document number, column 2, document name and directory path, column 3, the topic number corresponding to the number in column 1 in the keys file followed by a column of numbers with the proportion of words in the document corresponding to that topic. Pairs of columns follow on each row for each computed topic: topic, proportion, topic, proportion, topic, proportion, topic, proportion, …



#doc name topic proportion ...													
0	file:/C:/Test/Mallet/sample-data/web/en/elizabeth_needham.txt	5	0.227882	7	0.167171	6	0.145144	2	0.137362	4	0.070618	3	0.069068
1	file:/C:/Test/Mallet/sample-data/web/en/equipartition_theorem.txt	5	0.264542	2	0.14755	6	0.136072	7	0.12561	0	0.089303	4	0.079177
2	file:/C:/Test/Mallet/sample-data/web/en/gunnhild.txt	7	0.304076	5	0.207537	2	0.102312	6	0.098657	3	0.089247	4	0.062049
3	file:/C:/Test/Mallet/sample-data/web/en/hawes.txt	6	0.171536	5	0.144941	0	0.135988	2	0.128626	4	0.126635	3	0.125702
4	file:/C:/Test/Mallet/sample-data/web/en/hill.txt	6	0.267914	9	0.207684	4	0.132723	3	0.131745	7	0.097549	2	0.070714
5	file:/C:/Test/Mallet/sample-data/web/en/shiloh.txt	6	0.286084	3	0.176164	0	0.143444	5	0.109715	4	0.090072	2	0.085792
6	file:/C:/Test/Mallet/sample-data/web/en/sunderland_echo.txt	2	0.220222	6	0.18421	0	0.121407	5	0.106394	4	0.086736	9	0.08145
7	file:/C:/Test/Mallet/sample-data/web/en/thespis.txt	6	0.235398	5	0.223156	2	0.164104	4	0.104008	3	0.09472	0	0.064472
8	file:/C:/Test/Mallet/sample-data/web/en/thylacine.txt	9	0.180357	2	0.16682	1	0.132135	7	0.13032	5	0.128931	6	0.125566
9	file:/C:/Test/Mallet/sample-data/web/en/uranus.txt	7	0.230675	6	0.186825	4	0.111347	9	0.09878	2	0.098048	3	0.08748
10	file:/C:/Test/Mallet/sample-data/web/en/yard.txt	7	0.227867	2	0.203386	5	0.167821	6	0.142086	4	0.091845	0	0.07155
11	file:/C:/Test/Mallet/sample-data/web/en/zinta.txt	5	0.23079	9	0.140692	2	0.140078	0	0.120587	6	0.099088	4	0.093481

Document # 0 (the first document loaded into MALLET), elizabeth_needham.txt has topic 5 as its principal topic, with 22.7%; topic 7 with 16.7%, topic 6 at 14.5% and so on in decreasing weight; equipartition_theorem.txt (document # 1) and zinta.txt (document # 11) also have topic 5 as their largest topic, at 26.4% and 23% respectively. The topic model suggests a connection between these three documents that you might not at first have suspected.