IST736 Text Mining
Tutorial on Topic Modeling with Mallet
Professor Bei Yu

First, go to the Mallet website (http://mallet.cs.umass.edu/) to download and install the toolkit. Consult with [1] for step-by-step instructions for installation.

I. Topic Modeling in Three Steps

Step 1: Data Preparation

- Save all documents to one folder
- Convert the input data to Mallet format
    - Run command "bin/mallet import-dir --input your_input_folder --output your_output.mallet --keep-sequence --remove-stopwords"
    - For example, let's use the sample data provided by Mallet, located at "sample-data/web/en", which includes 12 documents. The command is "bin/mallet import-dir --input sample-data/web/en --output sample.mallet --keep-sequence --remove-stopwords --gram-sizes 1,2"

More on data import:

- To see options for "import-dir", type command "bin/mallet import-dir --help", and you will see options like ngrams or converting to lowercase letters.
- You can create and use your own stoplist in Mallet, using the option "--extra-stopwords [stoplist_filename]". For example, " bin/mallet import-dir --input sample-data/web/en --output bei.mallet --keep-sequence --remove-stopwords --extra-stopwords my-stoplist.txt "
- Mallet can process non-English documents like Chinese using command like "bin/mallet import-dir --input someChineseDocs --output someChinese.mallet \ --keep-sequence --remove-stopwords --token-regex '[\p{L}\p{M}]+'
- See more details in the Mallet Tutorial http://mallet.cs.umass.edu/import.php

Step 2: Topic Modeling

Build Topic Model
- Run command "bin/mallet train-topics --input sample.mallet --num-topics 10 --optimize-interval 20 --output-state sample-topic-state.gz --output-topic-keys sample-keys.txt --output-doc-topics sample-topics.txt
- See more details in the Mallet Tutorial http://mallet.cs.umass.edu/topics.php

Step 3: Examine output files to explain the topic model

Output 1: the keyword file

```
/Users/byu/mallet-2.0.7>cat bei-keys.txt
0       5       time zinta role hindi indian acting sullivan gilbert thespis england mother grossi
1       5       battle union hawes confederate kentucky army grant gen tennessee career confederat
2       5       thylacine tasmanian back extinct survived states male devil found species related
3       5       including london record world ho films relative productions position landing leadi
4       5       gunnhild united norway life death king needham actors maj ended bragg virginia neu
5       5       years yard national wilderness actress parks park modern numerous kehna female sta
6       5       test cricket south australian war film century naa filmfare december northern laur
7       5       sunderland echo paper journalist performances creating daily edward areas east hil
8       5       system average equipartition theorem law energy tiger general newspaper fighting h
9       5       rings dust uranus number moons narrow uranian ring addition particles dark discove
```

If add option "--optimize-interval 20", the above column of all 5s will become topic weights, so this option usually gives better topics

Output 2: the topic distribution of a document

For each document, the probability that it contains each topic

```
/Users/byu/mallet-2.0.7>cat bei-topics.txt
#doc name topic proportion ...
0       file:/Users/byu/mallet-2.0.7/sample-data/web/en/elizabeth_needham.txt   4       0.17307692307692307     3       0
.09615384615384616      0       0.09615384615384616     2       0.08653846153846154     9       0.0673076923076923      8
.057692307692307696
1       file:/Users/byu/mallet-2.0.7/sample-data/web/en/equipartition_theorem.txt       8       0.5547945205479452      9
.0547945205479452       3       0.04794520547945205     0       0.04794520547945205     7       0.0410958904109589      4
.0410958904109589
2       file:/Users/byu/mallet-2.0.7/sample-data/web/en/gunnhild.txt    4       0.3157894736842105      7       0.1654135
7518796992481203        8       0.06766917293233082     6       0.06766917293233082     2       0.06766917293233082     3
.045112781954887216
3       file:/Users/byu/mallet-2.0.7/sample-data/web/en/hawes.txt       1       0.3161290322580645      4       0.1354838
967741935483871 2       0.07096774193548387     7       0.05806451612903226     3       0.05806451612903226     9       0
.03870967741935484
4       file:/Users/byu/mallet-2.0.7/sample-data/web/en/hill.txt        6       0.30201342281879195     3       0.1073825
9395973154362416        0       0.09395973154362416     7       0.087248322147651       1       0.06711409395973154     8
```

Output 3: the topic state for future reference

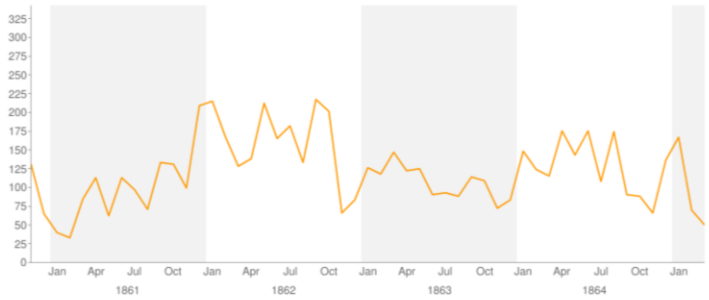
You can use the saved topic state to infer topics in new documents, just like using a trained classifier. See more details at http://mallet.cs.umass.edu/topics.php

II. Trend Analysis

If your data are in chronological order, we can visualize the trend of topic composition changes over time using each file's time stamp and topic distribution.

The topic trend may be calculated in different ways. For example, [2] calculated the trend of a topic by counting all documents with their proportions of this topic equal to or greater than 21.5% for each time period (a month in [2]). You can also plot the averaged proportion of this topic over all documents in each time period.

The following graph is of the latter variety. It shows the number of pieces from the paper where the proportion from the fugitive slave ad topic is equal to or greater than 21.5%:



III How to choose the number of topics

You can apply your domain knowledge to manual tuning by changing the number of topics *K*, examining the result, and choose the best Change N, check result, choose the best *K* that gives the most meaningful result.

You can also use some mathematical measures to guide the tuning of *K.* For example, the "ll/token" in Mallet measures the log-likelihood of the data for different *K*, and the best *K* corresponds to the highest log-likelihood. See Figure 3 in [3] for an example.

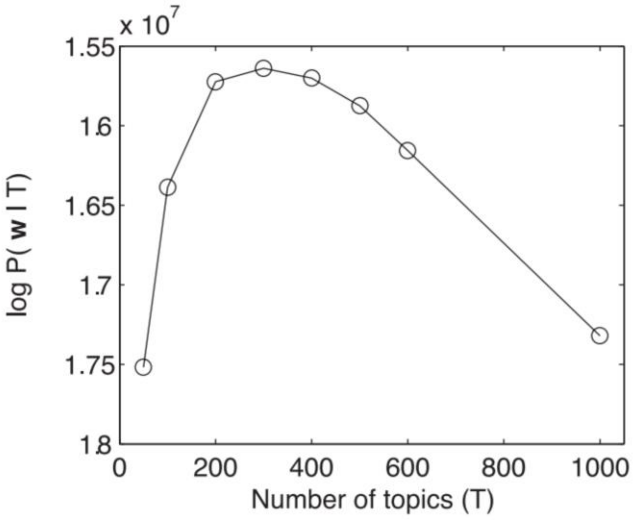
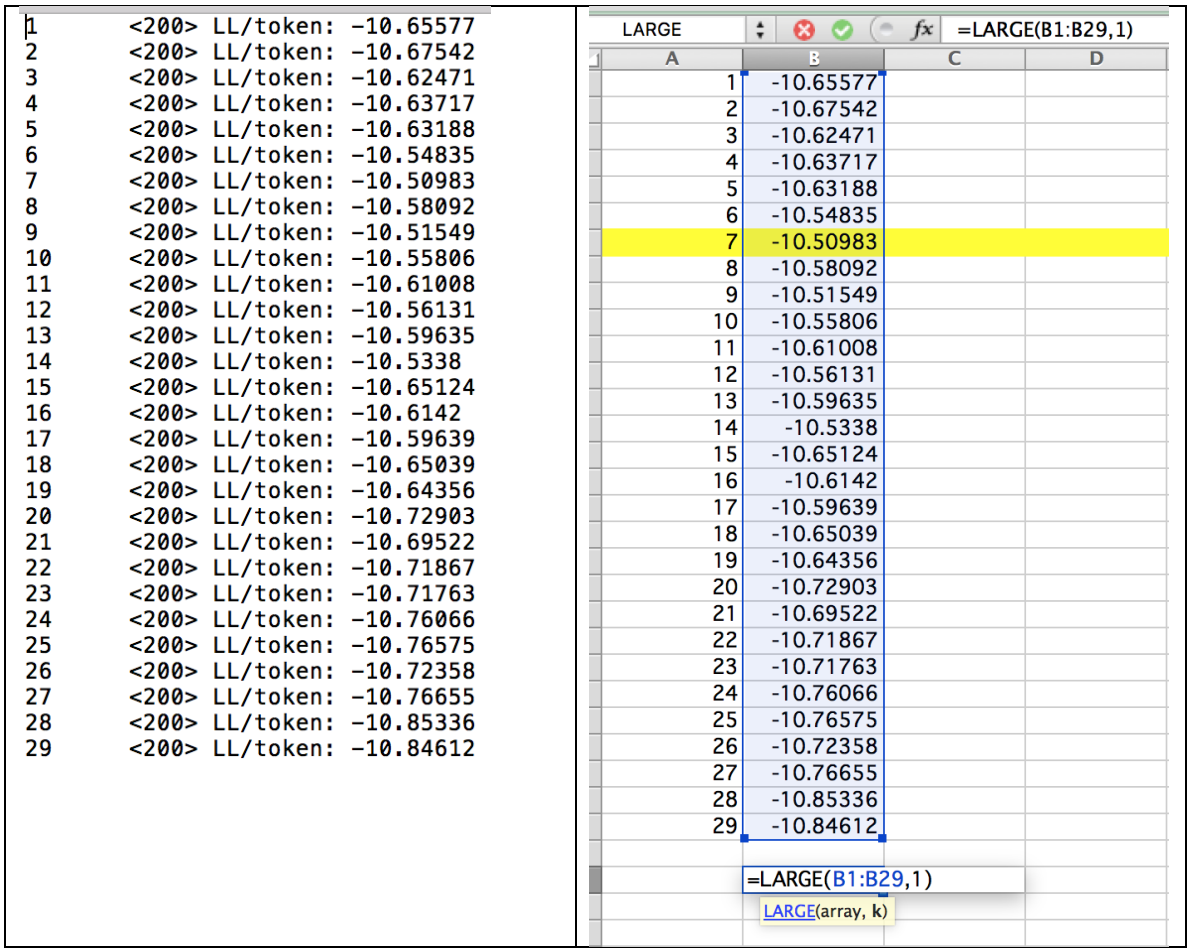


Fig. 3. Model selection results, showing the log-likelihood of the data for different settings of the number of topics, *T*. The estimated standard errors for each point were smaller than the plot symbols.

In Mallet, you can set *K* in a range, e.g. (1, 30), build a topic model for each *K*, and then find the "ll/token" value after each run. Collect all the values into one data file and then find the K that corresponds to the highest "ll/token" value. In the following example, the best *K* is 7.



There are also third-party tools such as the Topic-Stability Tool created by Derek Greene (https://github.com/derekgreene/topic-stability).

IV Interpreting Topic Models

Interpreting the topic model can be challenging. You can read [4] to gain some ideas on how to interpret the topic models that you generated. See more topic modeling use cases in [5][6][7].

References:

[1] Shawn Graham, Scott Weingart, and Ian Milligan. Getting Started with Topic Modeling and MALLET. http://programminghistorian.org/lessons/topic-modeling-and-mallet.html (last access 02/25/2018)

[2] Robert K. Nelson. Mining the *Dispatch*. http://dsl.richmond.edu/dispatch/pages/intro (last access 02/25/2018)

[3] Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1), 5228-5235.

[4] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (pp. 288-296).

[5] Zou, H., Chen, H. M., & Dey, S. (2015, March). Understanding library user engagement strategies through large-scale Twitter analysis. In *Big Data Computing Service and Applications (BigDataService), 2015 IEEE First International Conference on* (pp. 361-370). IEEE.

[6] Campbell, J. C., Hindle, A., & Stroulia, E. (2016). Latent Dirichlet allocation: extracting topics from software engineering data. In *The art and science of analyzing software data* (pp. 139-159).

[7] Chen, Y., Yu, B., Zhang, X., & Yu, Y. (2016, April). Topic modeling for evaluating students' reflective writing: a case study of pre-service teachers' journals. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 1-5). ACM.