Digital Tools for the 21st Century: Sherlock Holmes's London

19th Century Digital Humanities

Topic Modeling Assignment

MARCH 23, 2015MARCH 25, 2015 ~ ANNIESWAFFORD

Preparing Data:

1. Click this link (https://drive.google.com/uc?id=0Bzdac2vdsHyMd0JNV1I4VUpjYnM) to download a zip file of all Sherlock Holmes short stories (text from https://sherlock-holm.es (https://sherlock-holm.es/)). (**NOTE**: Each of the 56 short stories has been broken into smaller40-60 smaller text files to improve the results of topic modeling, resulting in a total of 2845 files; each story can be identified by its abbreviation (http://www.bestofsherlock.com/ref/rfab.htm).)
2. Go to the Downloads folder and unzip the data.
a. Right-click the zip file, and go to WinZip->"extract to here."

Preparing the Topic Modeling Tool

1. Go to https://topic-modeling-tool.googlecode.com/files/TopicModelingTool.jar (https://topic-modeling-tool.googlecode.com/files/TopicModelingTool.jar) to download the Topic Modeling Tool (a graphical user interface for Mallet).
2. Open it.

Topic Modeling Sherlock Holmes Stories:

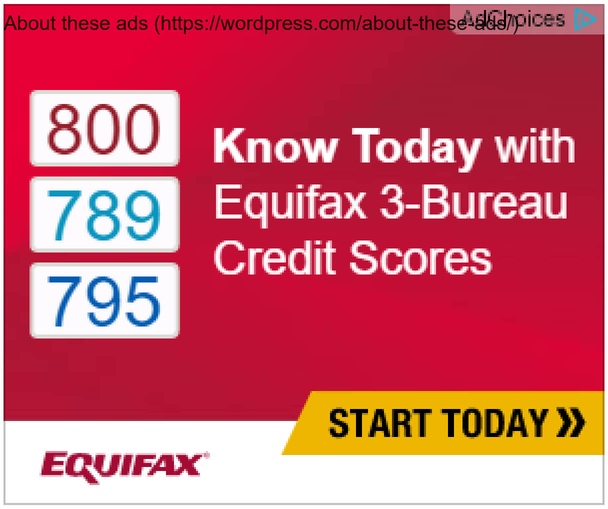
1. Click on the button labeled "Select Input File or Dir" and choose the folder with the Holmes stories from your Downloads folder and select "Open."
2. Click on the button labeled "Select Output Dir."
a. Click the icon that looks like a folder in the upper right part of the window to create a new folder.
b. Click on the new folder title to rename it.
c. Click "Open."
3. Under "Number of Topics," type "**50**," under "Number of Iterations," type "**1000,**" and under "No. of topic words printing," type "**20**."
4. Click on "Learn Topics."
5. Once it finishes running, go to your downloads folder and click on the folder you created.
6. Click on output_html, then all_topics to see your results.
7. Click on each topic to see how it's used.
8. **Repeat steps 3-7** with **different numbers of topics and iterations** to explore how this affects your results. **Give the output folder a new name each time.** Also experiment with checking and unchecking "Remove Stopwords" to see what happens.

Displaying Topic Models:

1. When you find the topics you like, save the data for later by turning it into a zip file (Right-click the file or folder, point to "Send to," and then click "Compressed (zipped) folder.") **WE WILL USE THIS DATA NEXT TIME**.

2. Choose at least 10 of your favorite topics Mallet produced.

3. Post the **10 topics and their words** to the blog, and make sure to start your post by listing the settings you used to generate your topics (**number of iterations, number of topic words printed, and number of total topics**.)
4. Your blog post with the 10 topics and words is due on **March 27th at 10am (NOTE: you do not need to write a 250-word blog post for Friday.)**



POSTED IN ANNOUNCEMENTS
ASSIGNMENTS DIGITAL HUMANITIES DISTANT READING TOPIC MODELING