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

19th Century Digital Humanities

Topic Modeling Part 2: Graphing the Results

MARCH 27, 2015MARCH 30, 2015 ~ ANNIESWAFFORD

Instructions: For this assignment, you will take the data from the Topic Modeling Tool and use Google Fusion Tables to graph the 10 topics you have identified. You will then look for trends in the graphs (e.g. does "violence" rise in the Holmes stories over time? Does "writing" appear more in the early days of the stories?) and start to theorize about them in a 300-word blog post. Refer to

http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/ (http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/) and

http://dsl.richmond.edu/dispatch/Topics (http://dsl.richmond.edu/dispatch/Topics) for examples of how to analyze graphs. Make sure to include enough screenshots so that all your 10 topics appear (you can graph them individually or in pairs if you find interesting relations between them).

Due: ~~3/31 by 10am~~ 4/3 by 8pm (4% of final grade)

Preparing Data:

1. Download the zip file of your Holmes topic modeling from our last class and save it on the desktop.
2. Right-click the zip file, and go to WinZip->"extract to here."
3. Your unzipped folder should contain two folders: output_csv and output_html. If it doesn't, you're missing some vital data, and you'll need to quickly redo your topic modeling. If you have both folders, you're ready to go!
4. Right now, the data is a bit messy: TopicsInDocs.csv in your output_csv folder tells us the topic distribution for each chunk of each short story, but we need average the chunks of each story together to compare entire stories to each other. We also need to add the full title and date of publication for each story. Normally this would be a time-consuming process, but I asked Daniel Lepage to build a small web tool to do this for us, and he kindly agreed.
5. Navigate to the web app at http://holmes-processor.appspot.com/ (http://holmes-processor.appspot.com/) and upload the file called "TopicsInDocs.csv."
6. The web tool will output a new spreadsheet: this spreadsheet has the story abbreviation, title, publication date, and percent of each topic for each story from your original spreadsheets.

7. Now that your data is organized, you're ready to upload it into Google Fusion Tables.

Importing Data with Google Fusion Tables:

1. Make sure you're logged out of your Hawkmail account.
2. Log into your non-Hawkmail Gmail account.
3. Navigate to http://tables.googlelabs.com (http://tables.googlelabs.com) and click "Create a Fusion Table" to start.
4. To upload your spreadsheet, select "From this computer," "choose file," and then browse until you find the spreadsheet on the Desktop. Highlight it, click "open," and then click "Next."
5. It will give you a preview of the spreadsheet and ask if the column names are in row 1; double check that they are, and then click "Next."
6. Give your project a title and a description, and then click "Finish." You've now imported the data!

Refining Data:

1. Now, you need to make sure that the columns are all the correct data type
   1. Click on "edit," "Change Column," and look at all the information. "Story ID" and "Title" should be "text," "Publication Date" should be "Date/Time," and everything else should be "Number."
   2. If you've changed anything, click "Save." If not, click the arrow to the left of the "Save" button to go back to the spreadsheet. Now you're ready to graph!

Graphing Data:

1. Click the red "+" sign and select "Add chart" to create a line graph.
2. Select the second chart option ("Continuous Variable Chart").
   1. **NOTE**: Do not select the "Categorical Chart" line graph further down the left-hand menu. It will not work correctly with our data.
3. You should now see the graph of "Topic 1." Make sure "Publication Date" is the label for the x-axis on the bottom of the graph (and change it if it's not). The bottom of the graph has little scroll bars on either side of it; you can click and drag them to zoom in on different parts of the graph.
4. Minimize the window, and look at the all_topics.html file that you used to choose the 10 topics you identified for today's class. Write down the topic numbers.
5. Go back to Google Fusion Tables's "Continuous Variable Chart" page, and click the button labeled "Choose": this will let you select the topic numbers of your 10 favorite topics.
6. Select them one at a time (uncheck them to make them disappear), and then try comparing them to your other categories. Look for trends.
7. When you're happy with your chart, click "Done" to finish it and click the red "+" to start your next chart.
8. Take screenshots (http://www.take-a-screenshot.org/ (http://www.take-a-screenshot.org/)) of the charts, and make sure that each of the 10 topics appears at least once in your images.

Writing the Blog Post:

1. Now that you have your images, it's time to analyze them.
2. Write a 300-word blog post that points out some trends that you've found across the stories.
3. If there aren't any trends, provide evidence to back up that assertion.
4. Is there any correlation between historical events (such as Women's suffrage, the Second Boer War, Queen Victoria's death) and spikes in your topics? (Check BRANCH

(http://www.branchcollective.org/ (http://www.branchcollective.org/)) and

http://myweb.fsu.edu/cupchurch/Resources/Timeline_19thcBrit.html

(http://myweb.fsu.edu/cupchurch/Resources/Timeline_19thcBrit.html) for ideas.) Do you think there's a connection? Why/why not? What additional research would you need to do to decide?

5. Include your screenshots throughout your post, and make sure to label them with your topic titles.
6. You did it! Proofread the post, submit it, and be proud of what you've accomplished!

POSTED IN <u>ANNOUNCEMENTS</u>
 <u>CLASS ASSIGNMENTS</u>   <u>DATA VISUALIZATION</u>   <u>DISTANT READING</u>   <u>SHERLOCK HOLMES</u>   <u>TOPIC MODELING</u>