Topics on digital history blogs

Max Kemman – maxkemman@gmail.com

Data set part of the analysis for my PhD thesis *Trading Zones of Digital History*, defended on 26 April 2019 at the University of Luxembourg, supervised by Prof. Andreas Fickers.

Please cite this data set as:
Kemman, M. (2019) Topics on digital history blogs [dataset]. Figshare.
https://doi.org/10.6084/m9.figshare.7813571

Files in this data set

This dataset contains

0. data_description.pdf – PDF describing the dataset
1. blogs.ods – Spreadsheet listing blogs scraped for analysis including statistics
2. topic-words.csv – spreadsheet listing 60 extracted topics and the top 50 terms associated with each topic
3. topics-metadata.csv – spreadsheet listing per document the metadata and topic weights for 60 topics
4. docs-in-topics.csv (see below under TMT output)
5. topics-in-docs.csv (see below under TMT output)

This data set covers the output of an LDA topic modelling exercise analysing the blogs of 73 historians in the period 2008-2017. I ran two analyses, with 5 topics for a general overview, and with 60 topics for a granular perspective. Unfortunately, the data with the 5 topics became corrupted after analysis and could not be restored. This data set therefore only includes the output with 60 topics, which was the most interesting for analysis.

Due to copyrighted material, the text files analysed are not included in the data set. In the spreadsheets I have therefore removed columns listing file names.

Selection of blogs

I collected blogs through a combination of manually reviewing the list of Digital Humanities Now,[1] the subject category History & Archaeology on Hypotheses,[2] searching on Google for "digital history blogs", and adding any blogs of interest that I had already been following myself.

[1] Digital Humanities Now is an aggregator of digital humanities blogs, which frequently highlights blog posts of interest. I consulted this list of approximately 550 blogs on 5 November 2018. For an archived version of the page see https://web.archive.org/web/20181105181437/http://digitalhumanitiesnow.org/subscribed-feeds/

[2] Hypotheses is a platform originating from France where scholars from the humanities and social science can host their online blogs. The subject category History & Archaeology contained approximately 1100 blogs and was consulted in July 2018 via https://www.openedition.org/catalogue-notebooks?limit=30&discipline%5B%5D=History+%26+Archaeology

In my review of blogs, I selected blogs that were

1) from individuals, who
2) self-identified as historians, either in an academic position as a historian or having a PhD in history,
3) wrote only in English,
4) were related to digital history, and
4) had written at least twenty blog posts in the period of 2008 to 2017.

Per author I scraped all the blog posts from their personal blogs, storing per blog post the title, content, and date of publication. I used the Google Chrome extension Web Scraper https://www.webscraper.io/ in combination with the uBlock Origin plugin https://github.com/gorhill/uBlock. The collection did not include comments, images, or URLs, focusing solely on the discursive practices of blog authors.
While the earliest blog posts appeared in 2004 already, and the most recent were from 2018, I chose to cover the period 2008 to 2017.
The resulting corpus consisted of the works of 73 authors who jointly wrote 10,918 blog posts, containing over 5.8 million words and 128 thousand unique word forms. Words were counted based on blog post titles and contents with Voyant Server 2.4.0-M7.

Preprocessing of data

In data cleaning, one step of importance was removing automatically inserted content such as “suggested posts” and “social sharing”. Many blog platforms, especially WordPress, include such content as a way to engage readers, by automatically considering the content of a blog post and suggesting similar posts. However, since this is inserted at the time of reading, a post can contain suggestions from other posts that were published earlier or later. As I am interested in the chronological development of discourse, such anachronistic content would negatively affect the analysis.

I did no stemming or removal of words.

All blog posts were stored as separate .TXT files. Metadata of all blog posts (including author, year, month, title, URL) per post was stored in a .CSV file.
Analysis
Analysis was done sequentially using Topic Modeling Tool (TMT, https://senderle.github.io/topic-modeling-tool/) and spreadsheet software.

Topic Modeling Tool (TMT)

With TMT, I associated the metadata file with the text files. I then ran TMT over all text files. I used the following settings:

1500 iterations
8 threads
50 words per topic
α=5
β=0.01
seed=28.

For number of topics *k* I conducted trials where *k* was 5, 40, 50, 60, 70, 80, 90, 100 and 110. Through close reading the different topics, I found *k*-values of 5 and 60 to be of main interest for discussion. The results for these two runs are included in this dataset.
By setting a seed the results between different trials should be repeatable and thereby comparable. These settings were decided following the documentation on https://web.archive.org/web/20181012092326/https://senderle.github.io/topic-modeling-tool/documentation/2018/09/27/optional-settings.html

TMT output

TMT generates the following output per run:

1. **topics-metadata.csv** – showing the relation between each blog post and the topics assigned to it.
2. **topic-words.csv** – showing the relation between each topic and the top 50 words associated with it.
3. **docs-in-topics.csv** – showing for each topic 500 documents ranked by strongest association.
4. **Topics-in-docs.csv** – showing for each document the topics ranked by strongest to weakest association.

In my analysis I have only used the first two files.

Charts

From the documents-topics.csv file, I created pivot tables to summarise topic weights per author or per year. For a more elaborate description and how to create the pivot tables from topics-metadata.csv see the TMT documentation
https://web.archive.org/web/20190131194600/https://senderle.github.io/topic-modeling-tool/documentation/2017/01/06/quickstart.html
The results of these pivot tables were visualised using the chart tools available in Microsoft Excel. I do not include the individual pivot tables in this dataset. Pivot tables used the following settings:

- Collective trends per year:
    - Data: select all rows
    - Pivot row: year
    - Pivot columns: **average** per topic
- Individual trends per year
    - Data: select rows related to 1 author
    - Pivot row: year
    - Pivot columns: **sum** per topics

Topics deemed of interest

In my analysis, I manually interpreted topics and grouped topics together, leading to the following groups. The number correspond to the topic IDs in the TMT output CSV files.

- Topics related to historical material
  - 16 Sources
  - 47 Data
- Topics related to historical subjects
  - 9 Industries
  - 10 Justice
  - 30 London-Court
  - 32 American-Settlers
  - 34 American-States
  - 35 Rome
  - 44 Science
  - 45 American-Republic
  - 48 Violence
  - 50 Finance
  - 51 Religion
  - 53 Slavery
  - 55 Racism
- Topics related to methodology
  - 15 Web-History
  - 24 Hermeneutics
  - 26 Ngrams
  - 29 Maps
  - 37 Topic-Modelling
  - 57 Networks
  - 58 Oral-History
- Topics related to digital humanities
  - 1 DH-conferences
  - 7 Social-Media
  - 38 DH-practices
  - 56 Web-Projects
- Topics associated to academia
  - 6 Teaching
  - 8 Publishing
  - 14 Faculty
  - 18 Online-Education
  - 42 Conferences
  - 43 Funding