Extracted Features in the Wild

Projects

[Word Similarity Tool](#), David Mimno

A web-based tool for viewing similar words to a query, for each year from 1800 to 1923.

Search for:	washington							
1923 washington	columbia	ohio	york	chicago	montgomery	west	boston	michigan
1922 washington	dawes	pennsylvania	columbia	york	dent	president	city	washington.
1921 washington	columbia	washington.	national	pennsylvania	president	dent	boston	york
1920 washington	boston	columbia	illinois	york	washington.	ohio	city	pennsylvania
1919 washington	illinois	address	york	vice	national	boston	washington.	tribune
1918 washington	york	illinois	boston	baltimore	american	national	president	pennsylvania
1917 washington	boston	washington.	columbia	union	address	charleston	baltimore	vice
1916 washington	washington.	boston	virginia	columbia	charleston	massachusetts	wilson	president
1915 washington	washington.	charleston	virginia	massachusetts	louis	columbia	baltimore	portsmouth
1914 washington	columbia	denver	philadelphia	maryland	louis	baltimore	virginia	ohio
1913 washington	philadelphia	baltimore	charleston	american	address	pennsylvania	columbia	massachusetts
1912 washington	va.	columbia	charleston	louis	columbian	baltimore	boston	american

For a word that you're investigating, the tool generates a table showing the words that occur in similar contexts as the queried-for word. Instead of entering the word in the search box, you can also enter the word directly into the browser (by adding ?q=[word] after the URL http://mimno.infosci.cornell.edu/wordsim/nearest.html), like so:

http://mimno.infosci.cornell.edu/wordsim/nearest.html?q=caste to have the queried-for word be "caste".

Here's an example of the kinds of useful observations the tool empowers you to make:

With the above query for "caste", that is, http://mimno.infosci.cornell.edu/wordsim/nearest.html?q=caste ,

you can notice a few interesting things: from the generated table for "caste", it appears that "suffering" was a pretty frequent contextual word for "caste" in the early nineteenth century, but then "suffering" seemed to drop out of the context by the late nineteenth century. On the other hand, "degradation" seemed to remain, more or less, part of the context of the word "caste" throughout. You can also notice that the occurrence of "race" in the same context as "caste" becomes more frequent after 1870 or so. Before 1870, the instances of "race" in the same context as "caste" was less pronounced.
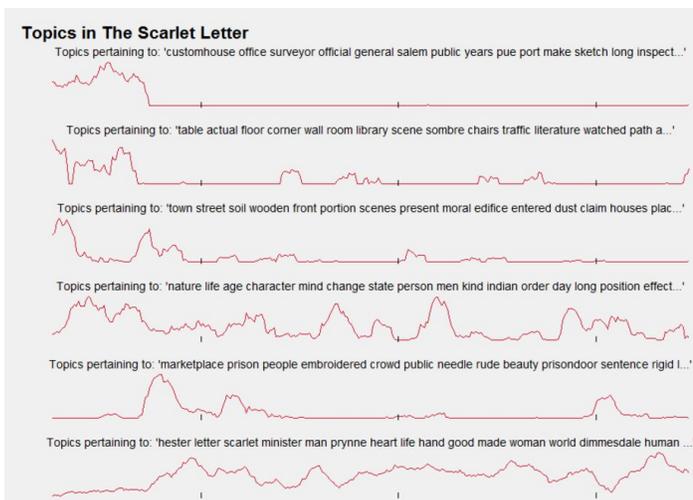
HT+Bookworm

An interactive, faceted, visualization of terms across the HathiTrust collection, built on the EF dataset.

Here is our recent slide deck providing an overview of the HT+Bookworm project.

Within-Book Topic Modeling, Peter Organisciak

An approach for visualizing thematic trends within a book.

A Topic Model of Fiction, Jonathan Goodwin

A topic model of fiction, based on the genre-classified dataset (only 1920-22); it may be extended once extracted features are available after 1922.

Top words

Word	Weight
squire	
farm	
road	
lane	
village	
cottage	

[Berkeley Data Science Module](), Chris Hench and Cody Hennesy

A Jupyter Notebooks-based curriculum for using HTRC Extracted Features in the classroom developed at the University of California, Berkeley.

```
539 lines (538 sloc)   17.2 KB                          Raw  Blame  History

Analyzing volumes for word frequencies

This notebook will demonstrate some of basic functionality of the Hathi Trust FeatureReader object. We will look at a few examples of
easily replicable text analysis techniques — namely word frequency and visualization.

In [ ]:  %%capture
         !pip install nltk

In [ ]:  from htrc_features import FeatureReader
         import os
         import numpy as np
         import pandas as pd
         import matplotlib.pyplot as plt
         plt.style.use('fivethirtyeight')
```

Tools

HTRC Feature Reader

A Python library that scaffolds Pandas use of EF data. With [example]() scripts.

Tutorials and Lessons

Send us your Lessons or Tutorials related to the EF Dataset.

Python code for some simple examples of "literary sleuthing":

- [Estimating the proportion of poetry-to-prose in a volume, based on the proportion of capitalized letters (Coleridge wrote a lot more prose than Keats did!)]()
- [Making use of the incidence of a word's occurrence to draw inferences (*Little Dorrit* by Charles Dickens mentions "prison" a lot more than his *Bleak House* does...)]()
- [Identifying that volume in a workset in which a specified word occurs the most times (Which of the English romantic poets was the greatest "dream"-er among them all?)]()

Blog Posts

Underwood, Ted. June 3, 2014. "A window on the twentieth century may be about to open." *The Stone and the Shell.* Blog. [http://tedunderwood.com/2014/06/03/a-window-on-the-twentieth-century-may-be-about-to-open/]()

Mimno, David. 2014. "Word counting, squared." *David Mimno*. Blog. http://www.mimno.org/articles/wordsim/

Forster, Chris. 2015. "A Walk Through the Metadata: Gender in the HathiTrust Dataset." (Based on genre-classified subsets.) http://cforster.com/2015/09/gender-in-hathitrust-dataset/

Underwood, Ted. 2015. "How Scholars Can Support Digital Libraries." Europeana Research. http://research.europeana.eu/blogpost/text-mining-2-how-scholars-can-support-digital-libraries