Faces of Russia

Analyzing Changes in Perceptions of Russia in English-Language Literature via Distant Reading (1800-1923)

Outline

1. Research Question(s)
    a. Importance
2. Data Set
    a. Acquisition and Inclusion Criteria
    b. Normalization
3. Tools
4. Presentation
5. Timeline
6. Budget & Sustainability
7. Preview
8. Questions (for the audience and for me)

Research Question(s)

Research Question(s)

- How was Russia perceived in English-language literature in the 19th and early 20th centuries?
- Are there recurring topics or themes found in these texts?
- How do these themes change over time? Can these changes be linked to historical events, both in Russia and in the English-speaking world?

The importance of this topic lies in its potential to allow for scholars to "index" into the corpus of English-language literature on Russia by topic. This could be a tool that fosters deeper analysis of more specific themes.

Data Set

Data Acquisition

- Use Project Gutenberg's collection of texts via Clemens Wolff's gutenberg.py API
- Criterion for including a text in the corpus:
  - Returned by Project Gutenberg's search feature for the keyword "Russia"
  - Manually screened for being originally written in English (no translators listed)
  - Manually screened for being on-topic (avoid "Prussia" unless also mentioning "Russia")
  - Currently, the whole text is included (even if "Russia" is only mentioned once)
- Texts are downloaded as plain text files onto a user's local machine to prepare for normalization
- In addition to content, also collect a text's title, author, and publication date*.

Free eBooks - Project Gutenberg

Book search · Book categories · Browse catalog · Mobile site · Report errors · Terms of use

Some of the Latest eBooks

Welcome

Project Gutenberg is a library of over 60,000 free eBooks. Choose among free epub and Kindle eBooks, download them or read them online. You will find the world's great literature here, with focus on older works for which U.S. copyright has expired. Thousands of volunteers digitized and diligently proofread the eBooks, for enjoyment and education.

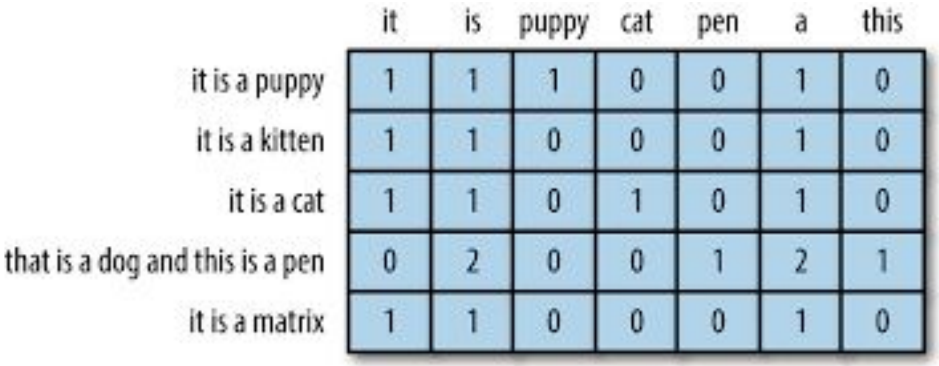
Looking for something to read? Project Gutenberg eBooks are mostly older literary works. Most were published before 1924, with some published in the decades after. Use one of the Search methods on this page, or try using the Bookshelves to browse by genre, age group, and topic.

Project Gutenberg Mobile Site

New website available for testing. Visit https://dev.gutenberg.org (or http://dev.gutenberg.org) to test the site (it may have occasional outages, as improvements are made). There is a new website page that lists some known issues, and part of the motivation for the change. If you visit the new website, please consider providing your input and suggestions via an anonymous online survey afterwards.

Data Normalization and Modeling

- Remove Project Gutenberg’s header and footer material from each text
- Lemmatize each text and remove stop words
    - Lemmatization groups together inflected forms of a word so they can be analysed as a single item (“good”, “better”, “best” → “good”)
    - Stop words are the most common words in a language and therefore unimportant for this analysis (“the”, “a”, “and”, “to”, etc.)
- Vectorize the resulting combined texts into a bag of words to prepare for topic modeling with MALLET
    - After lemmatization and removal of stop words, the current corpus consists of 3,359,718 words



	it	is	puppy	cat	pen	a	this
it is a puppy	1	1	1	0	0	1	0
it is a kitten	1	1	0	0	0	1	0
it is a cat	1	1	0	1	0	1	0
that is a dog and this is a pen	0	2	0	0	1	2	1
it is a matrix	1	1	0	0	0	1	0

Tools

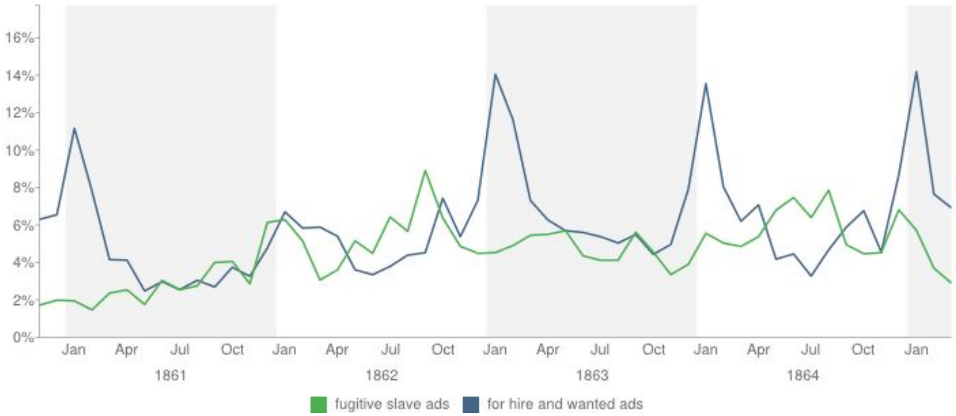
Tools

- Data Collection
  - Project Gutenberg, gutenberg API via Python
- Data Normalization/Cleaning
  - Gensim, NLTK via Python
- Analysis
  - MALLET via Gensim Python wrapper
  - Visualization in Plotly via Python*
- Presentation
  - Website design via Jekyll and Ruby*
  - Website hosting via GitHub pages*
- Preservation
  - Source code and documentation availability via GitHub*

Presentation

Presentation

- Organize and publish the results of the modeling via a website
  - GitHub Pages provides this functionality in a simple, cost-free way
- Explain topic modeling as a tool for distant reading
- Highlight the most interesting topics and how these change over time
  - Take inspiration from *Mining the Dispatch*, created by Richard K. Nelson at the University of Richmond
- Display texts that best exhibit these most interesting topics



Timeline

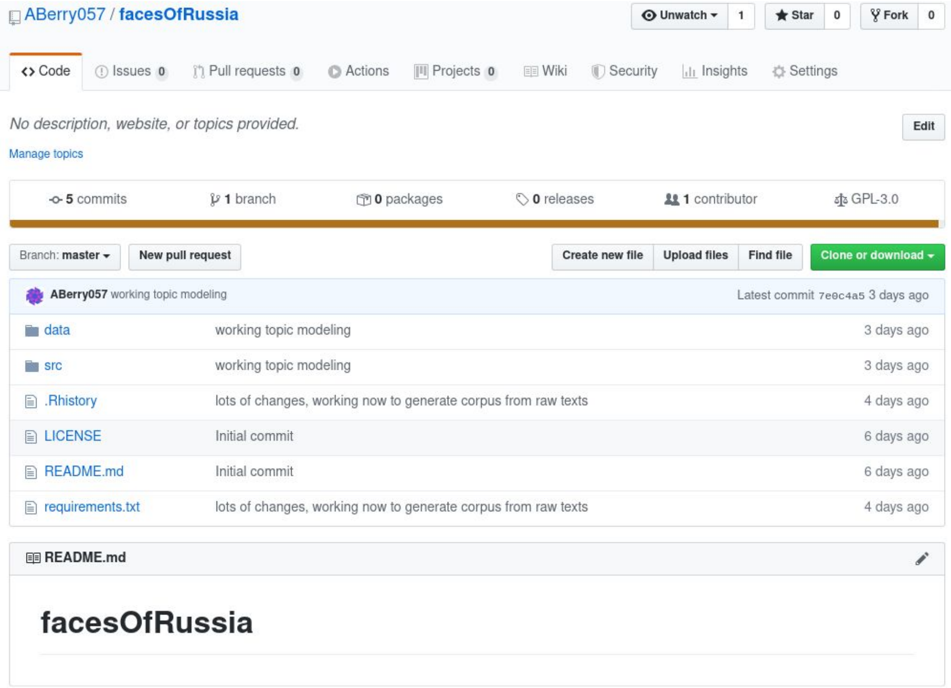
Timeline - Check-In Dates and Deadline

- 4/1 - Present initial ideas and solicit feedback
- 4/8 - Decide on and implement approach to publication dates, author origin
- 4/15 - Finalize corpus selection criterion and gather final collection of texts
- 4/22 - Select number of topics, run analyses for different time periods as needed, visualize results
- 4/29 - Find exemplary texts, build first version of website to present tool/analysis, write commentary
- 5/8 - Final project submission, dot the i's and cross the t's
- Post 5/8 - Apply same methods to other subjects ("United States", "Brown University", "democracy", etc.) and compare results, support others interested in work

Budget & Sustainability

Budget & Sustainability

- Host source code and documentation of process in public GitHub repository
- Safety in L.O.C.K.S.S. strategy (lots of copies keep stuff safe)
  - This only works if the entire process is replicable, and I will therefore need to document all of my work.
- Website containing analysis and commentary will also be hosted by GitHub and will follow the same LOCKSS approach
- Don’t foresee any costs associated with this project, with possible exception of accessing collections behind a paywall (though Brown does a good job of helping avoid that)



ABerry057 / facesOfRussia

No description, website, or topics provided.

5 commits · 1 branch · 0 packages · 0 releases · 1 contributor · GPL-3.0

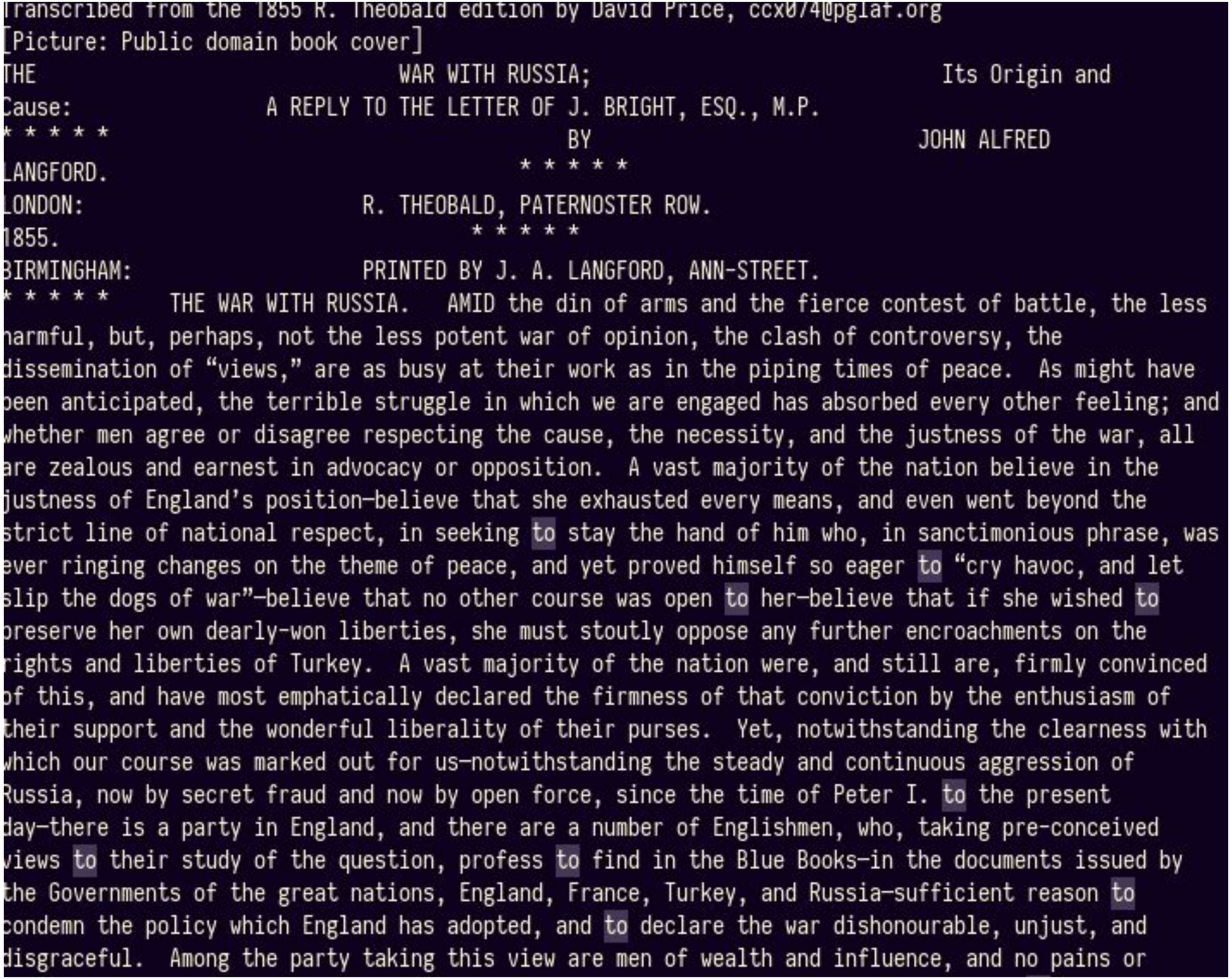
ABerry057 working topic modeling — Latest commit 7e0c4a5 3 days ago

data	working topic modeling	3 days ago
src	working topic modeling	3 days ago
.Rhistory	lots of changes, working now to generate corpus from raw texts	4 days ago
LICENSE	Initial commit	6 days ago
README.md	Initial commit	6 days ago
requirements.txt	lots of changes, working now to generate corpus from raw texts	4 days ago

README.md

facesOfRussia

Preview of Results



```
Transcribed from the 1855 R. Theobald edition by David Price, ccx074@pglaf.org
[Picture: Public domain book cover]
THE                              WAR WITH RUSSIA;                             Its Origin and
Cause:                A REPLY TO THE LETTER OF J. BRIGHT, ESQ., M.P.
* * * * *                                            BY                             JOHN ALFRED
LANGFORD.                                        * * * * *
LONDON:                         R. THEOBALD, PATERNOSTER ROW.
1855.                                        * * * * *
BIRMINGHAM:                     PRINTED BY J. A. LANGFORD, ANN-STREET.
* * * * *     THE WAR WITH RUSSIA.   AMID the din of arms and the fierce contest of battle, the less
harmful, but, perhaps, not the less potent war of opinion, the clash of controversy, the
dissemination of “views,” are as busy at their work as in the piping times of peace.  As might have
been anticipated, the terrible struggle in which we are engaged has absorbed every other feeling; and
whether men agree or disagree respecting the cause, the necessity, and the justness of the war, all
are zealous and earnest in advocacy or opposition.  A vast majority of the nation believe in the
justness of England’s position—believe that she exhausted every means, and even went beyond the
strict line of national respect, in seeking to stay the hand of him who, in sanctimonious phrase, was
ever ringing changes on the theme of peace, and yet proved himself so eager to “cry havoc, and let
slip the dogs of war”—believe that no other course was open to her—believe that if she wished to
preserve her own dearly-won liberties, she must stoutly oppose any further encroachments on the
rights and liberties of Turkey.  A vast majority of the nation were, and still are, firmly convinced
of this, and have most emphatically declared the firmness of that conviction by the enthusiasm of
their support and the wonderful liberality of their purses.  Yet, notwithstanding the clearness with
which our course was marked out for us—notwithstanding the steady and continuous aggression of
Russia, now by secret fraud and now by open force, since the time of Peter I. to the present
day—there is a party in England, and there are a number of Englishmen, who, taking pre-conceived
views to their study of the question, profess to find in the Blue Books—in the documents issued by
the Governments of the great nations, England, France, Turkey, and Russia—sufficient reason to
condemn the policy which England has adopted, and to declare the war dishonourable, unjust, and
disgraceful.  Among the party taking this view are men of wealth and influence, and no pains or
```

Preliminary Topic Modeling Results (10/20)

```
0	2.5	spirit big society hotel detail mentioned altogether entire farther pocket turkish waited prevent grass car crime pity gift luka dull

1	2.5	fro respectful seventeenth altar intently mushroom kronstadt satan convenience employing distrust accumulated vestment brand desperately perspective vigour gallon buyer exceeded

2	2.5	time eye room hour moscow looked peasant peter white stood open fact full kind winter petersburg hope fire table service

3	2.5	gave beauty discovered information price demand formed apartment cast dangerous public art shook risk fled composed average heat advice commerce

4	2.5	soldier air master hold pretty send companion inhabitant mongol skin natural low driven stay degree government twelve perfectly tribe ear

5	2.5	place part word war de officer black month book person dark mountain tree week met wind colonel carriage call fine

6	2.5	music paul killed fighting simple worth style succeeded fire intended aunt star plot ball message direct instruction frederick crowded englishman

7	2.5	give order church mr god son knew story short fellow scene added terrible ground officer post mine sleep live bridge

8	2.5	gloom hebrew varying lowered era augustus vitebsk jaw watchful softened ay circulated brightness alcohol christendom dire tint unrest shaved chicago

9	2.5	made country general city prince ivan find state le high street girl arm letter troop form ten boat strong read
```

Questions (for the audience and for me)

- Is it enough to only source data from Project Gutenberg? If not, what other sources would you suggest?
- What is your opinion on the corpus-inclusion criterion? Would you take a more strict or more lenient approach?
- How would you solve the problem of obtaining dates for the texts? Using the author’s lifetime, an external source, or something else? (A similar dilemma exists for author origin)
- What would you like to see emphasized in the presentation of the results and commentary on the website?