Northeastern University

NULab for Texts, Maps, and Networks

Computational Text Analysis: Web-browser Tools

May 29, 2019

Cara Marta Messina and Garrett Morrow

For ENGL 3370

Prof. Rachel Lewis

Key Terminology

Corpus/corpora: a collection of texts analyzed for research.

nGram: a contiguous sequence of “n” items in a sample of texts.

Stopwords: words that typically hold no meaning to a sentence and are commonly omitted in text analysis (the, but, this, that...etc)

Before You Start...

Ask yourself:

1. What question(s) am I asking of my corpus?
2. What am I trying to measure and why?
3. Are there any factors that might confuse my results and conclusions?
4. What visualizations might be useful for my corpus and for the variables I have chosen?

What is Voyant?

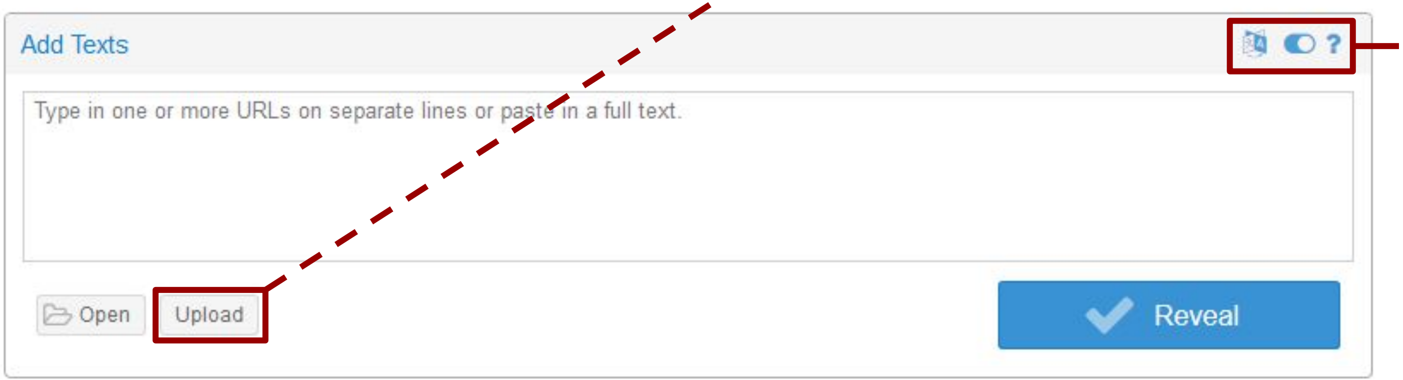
https://voyant-tools.org/

Voyant can read .pdf, .doc, but **.txt is recommended.**

For Voyant documentation/guide see: https://voyant-tools.org/docs/#!/guide/about

VOYANT

see through your text

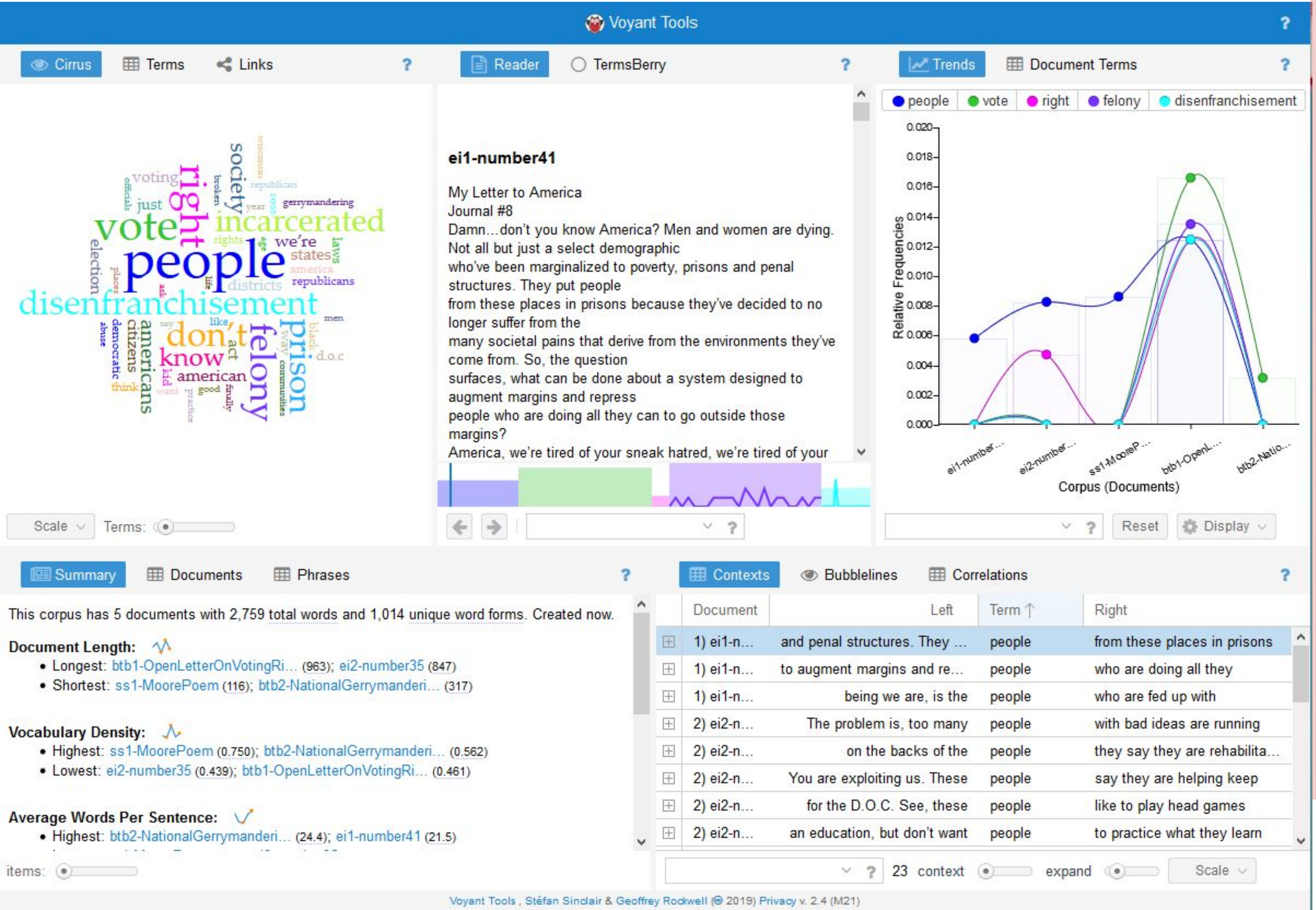


Type in one or more URLs on separate lines or paste in a full text.

Click on upload and navigate to the folder with the text documents you wish to analyze.

Alternatively, insert URLs or full text into textbox.

Click here for help and advanced options



ei1-number41

My Letter to America
Journal #8
Damn…don't you know America? Men and women are dying.
Not all but just a select demographic
who've been marginalized to poverty, prisons and penal
structures. They put people
from these places in prisons because they've decided to no
longer suffer from the
many societal pains that derive from the environments they've
come from. So, the question
surfaces, what can be done about a system designed to
augment margins and repress
people who are doing all they can to go outside those
margins?
America, we're tired of your sneak hatred, we're tired of your

This corpus has 5 documents with 2,759 total words and 1,014 unique word forms. Created now.

Document Length:
- Longest: btb1-OpenLetterOnVotingRi… (963); ei2-number35 (847)
- Shortest: ss1-MoorePoem (116); btb2-NationalGerrymanderi… (317)

Vocabulary Density:
- Highest: ss1-MoorePoem (0.750); btb2-NationalGerrymanderi… (0.562)
- Lowest: ei2-number35 (0.439); btb1-OpenLetterOnVotingRi… (0.461)

Average Words Per Sentence:
- Highest: btb2-NationalGerrymanderi… (24.4); ei1-number41 (21.5)

Document	Left	Term ↑	Right
1) ei1-n…	and penal structures. They …	people	from these places in prisons
1) ei1-n…	to augment margins and re…	people	who are doing all they
1) ei1-n…	being we are, is the	people	who are fed up with
2) ei2-n…	The problem is, too many	people	with bad ideas are running
2) ei2-n…	on the backs of the	people	they say they are rehabilita…
2) ei2-n…	You are exploiting us. These	people	say they are helping keep
2) ei2-n…	for the D.O.C. See, these	people	like to play head games
2) ei2-n…	an education, but don't want	people	to practice what they learn

Voyant Tools, Stéfan Sinclair & Geoffrey Rockwell (© 2019) Privacy v. 2.4 (M21)

Results:

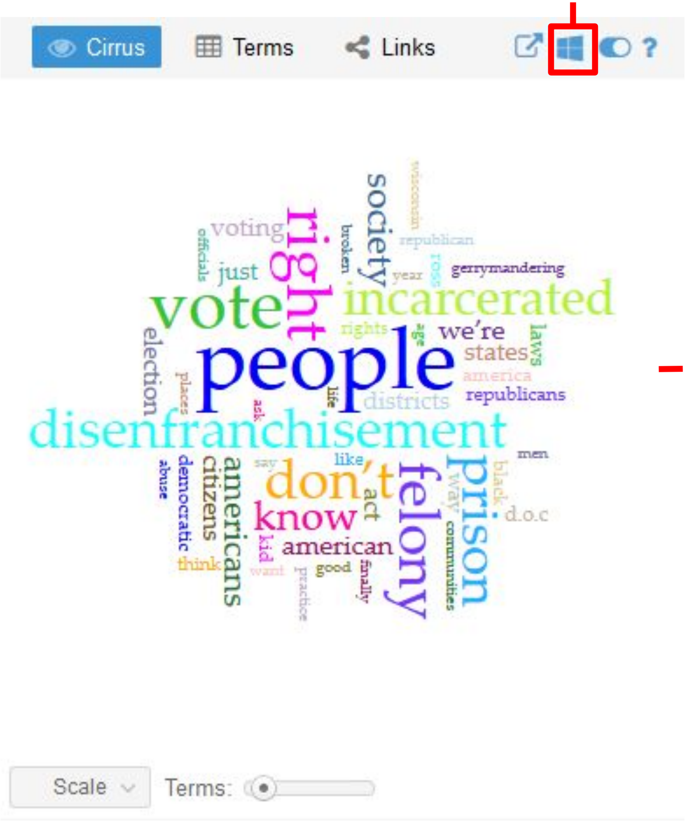
From a corpus of incarceration documents, you can see the default results page with multiple panes:

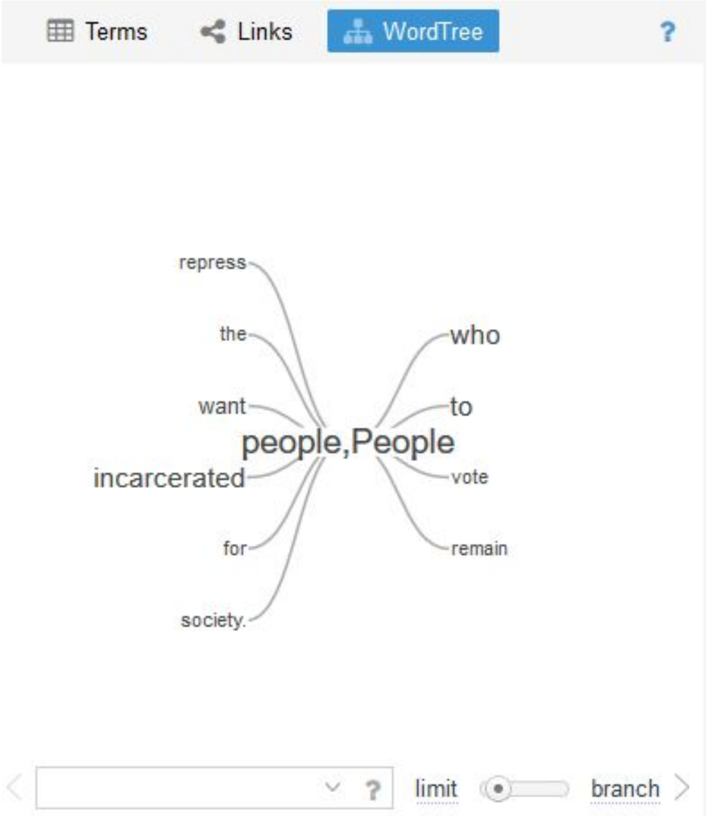
- A word cloud
- Reader section
- Trends
- Document Summary
- Word Context

These boxes can all be changed!

Changing Results Displays in Voyant

Select the panes button and select a new option from the dropdown menu





For our new pane option, we have chosen the WordTree visualization from the 'visualization tools' dropdown sub-menu.

Useful Voyant Visualizations and Analysis

1. **Wordcloud**: A visualization of word frequency where the most frequent words appear the largest and less frequent words are progressively smaller by frequency. (Visualization Tools → Cirrus)
2. **Summary**: Document summary that gives word counts, density, words per sentence, word frequency, and distinct words. (Corpus tools → Summary)
3. **Correlations**: Words that appear in similar contexts and appear in sync, correlated or inversely correlated. Gives correlation and significant factors. (Corpus Tools → Correlations)
4. **Phrases**: The frequency of word sequences that appear together (also known as ‘nGrams’). (Corpus Tools → Phrases)
5. **Contexts:** Displays the sequences of words that appear around a specific word. (Grid Tools → Contexts)

Alternative Web Tools

WordCounter:

WordCounter website: https://databasic.io/en/wordcounter/

WordCounter analyzes a text to count individual words and n-grams.

The output can then be downloaded as a .csv to do further analysis

SameDiff:

SameDiff website: https://databasic.io/en/samediff/

SameDiff compares one corpus or text to another corpus or text, and tells the user how similar or different the texts or corpora are using a cosine similarity.

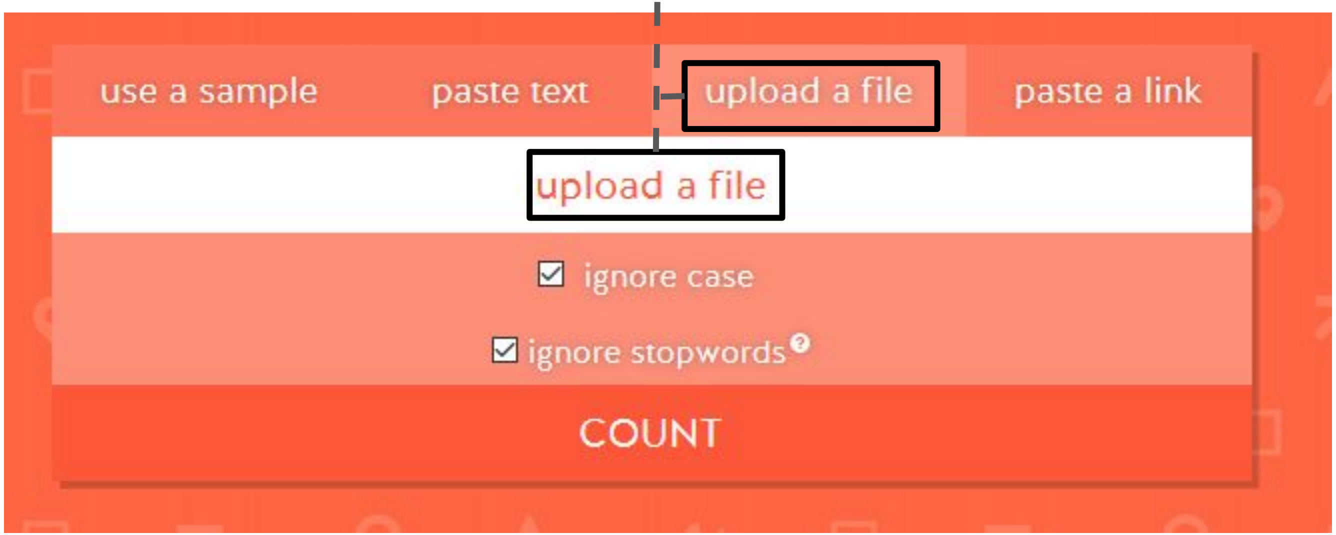
word COUNTER

1. Select ‘upload a file’ from the top bar, then again on the middle bar.

2.Then navigate to the text you would like to analyze.

Note: Ignore case and stopwords are enabled by default

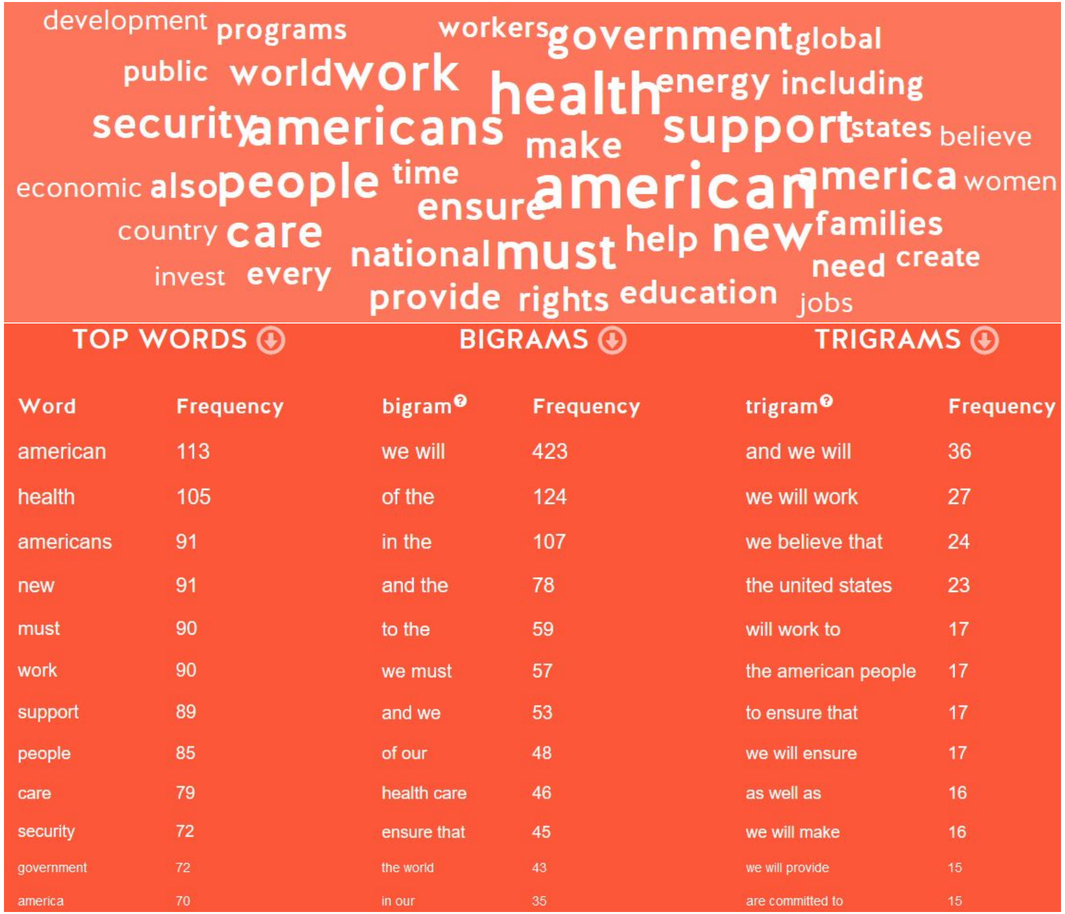
3.Click on count



☑ ignore case

☑ ignore stopwords

WordCounter Results



TOP WORDS BIGRAMS TRIGRAMS

Word	Frequency	bigram	Frequency	trigram	Frequency
american	113	we will	423	and we will	36
health	105	of the	124	we will work	27
americans	91	in the	107	we believe that	24
new	91	and the	78	the united states	23
must	90	to the	59	will work to	17
work	90	we must	57	the american people	17
support	89	and we	53	to ensure that	17
people	85	of our	48	we will ensure	17
care	79	health care	46	as well as	16
security	72	ensure that	45	we will make	16
government	72	the world	43	we will provide	15
america	70	in our	35	are committed to	15

For this example we have used Barack Obama’s Democratic party official Party Platform document from the ‘American Presidency Project.’

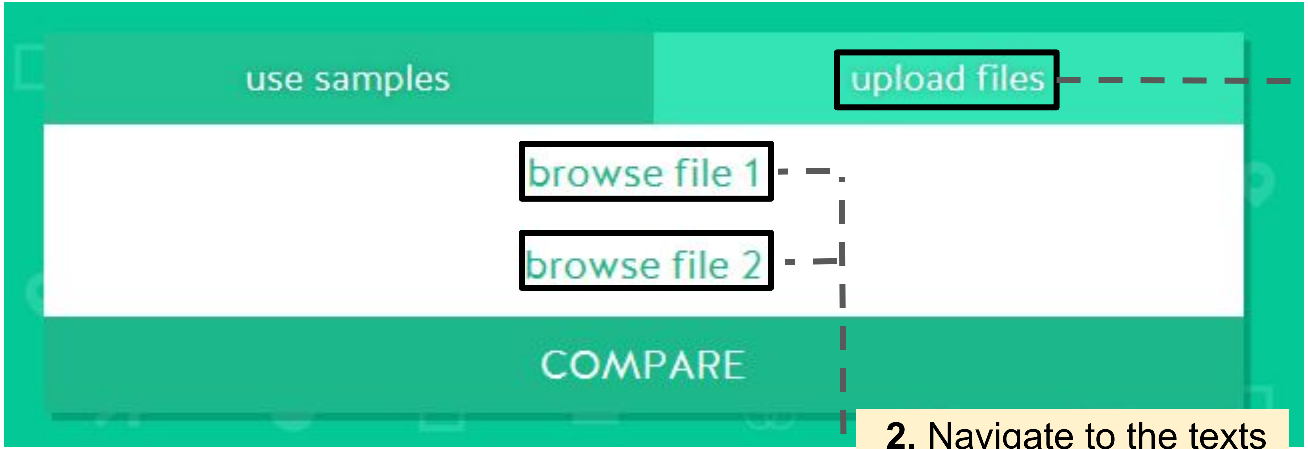
WordCounter outputs a word cloud...

...and a list of top words, bigrams, and trigrams.

WordCounter can also output results in a .csv format accessed by scrolling down the results page.

samediff

Samediff compares two .txt files (two texts) to show the unique and same words used in both texts.

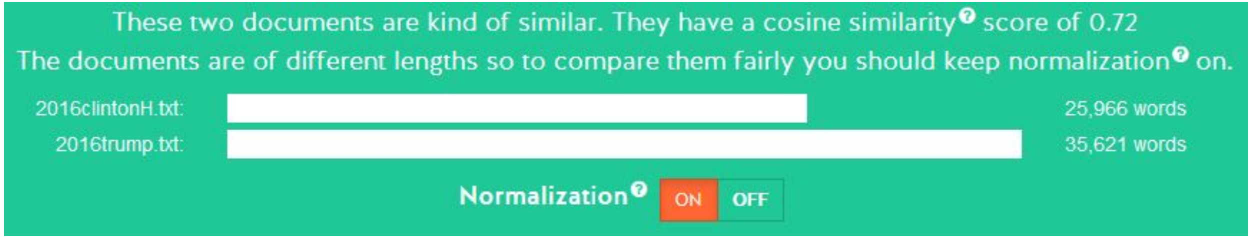


1. Select upload files

2. Navigate to the texts individually
3. Click Compare

SameDiff Results

For this example, we have compared Hillary Clinton and Donald Trump’s 2016 official party platforms from the ‘American Presidency Project.’



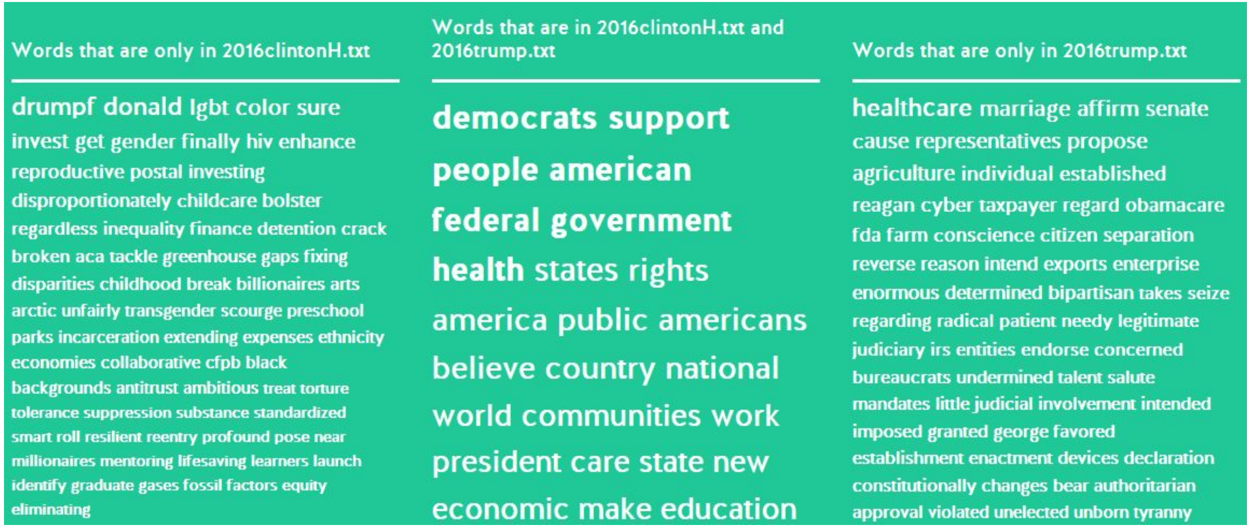
These two documents are kind of similar. They have a cosine similarity score of 0.72

The documents are of different lengths so to compare them fairly you should keep normalization on.

2016clintonH.txt: 25,966 words

2016trump.txt: 35,621 words

Normalization [✓ ON] OFF



Words that are only in 2016clintonH.txt

drumpf donald lgbt color sure invest get gender finally hiv enhance reproductive postal investing disproportionately childcare bolster regardless inequality finance detention crack broken aca tackle greenhouse gaps fixing disparities childhood break billionaires arts arctic unfairly transgender scourge preschool parks incarceration extending expenses ethnicity economies collaborative cfpb black backgrounds antitrust ambitious treat torture tolerance suppression substance standardized smart roll resilient reentry profound pose near millionaires mentoring lifesaving learners launch identify graduate gases fossil factors equity eliminating

Words that are in 2016clintonH.txt and 2016trump.txt

democrats support people american federal government health states rights america public americans believe country national world communities work president care state new economic make education

Words that are only in 2016trump.txt

healthcare marriage affirm senate cause representatives propose agriculture individual established reagan cyber taxpayer regard obamacare fda farm conscience citizen separation reverse reason intend exports enterprise enormous determined bipartisan takes seize regarding radical patient needy legitimate judiciary irs entities endorse concerned bureaucrats undermined talent salute mandates little judicial involvement intended imposed granted george favored establishment enactment devices declaration constitutionally changes bear authoritarian approval violated unelected unborn tyranny

SameDiff outputs an overall similarity score (0.72 in this example) and total word counts.

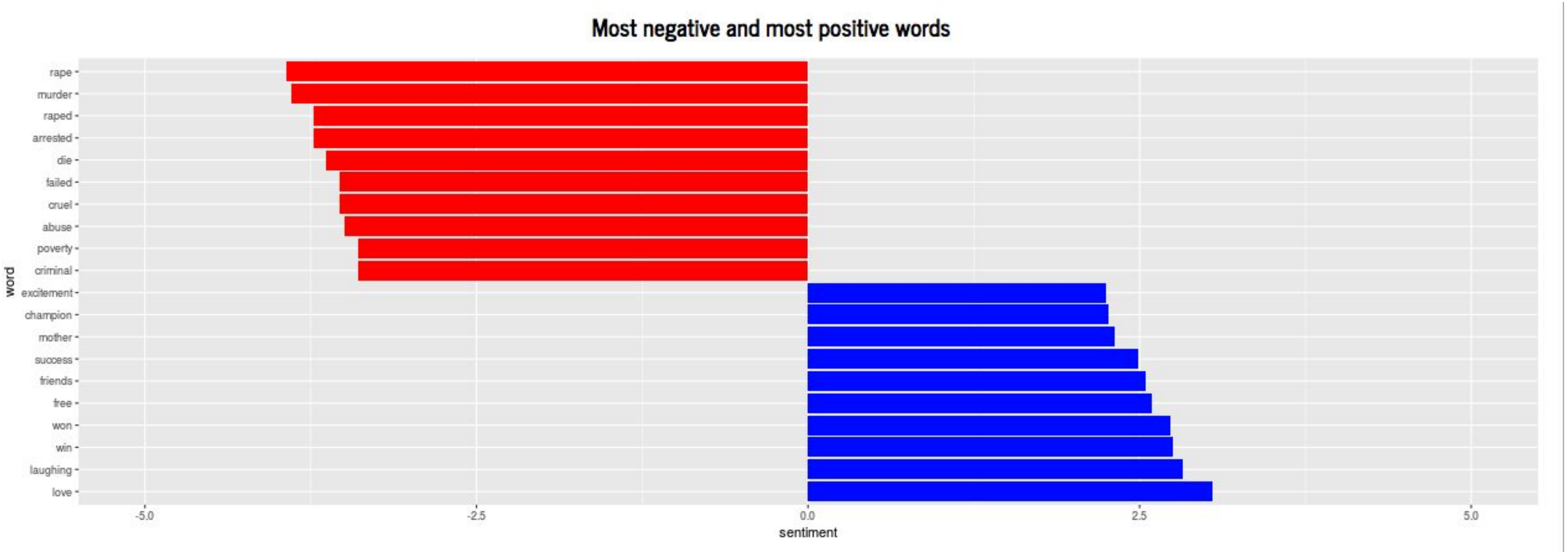
Note that normalization is enabled to account for different total word counts.

SameDiff also outputs the specific words that are similar and the words that differentiate the two texts.

A .csv is also available at the bottom of the results page.

Storybench Drag-and-Drop Textual Analysis

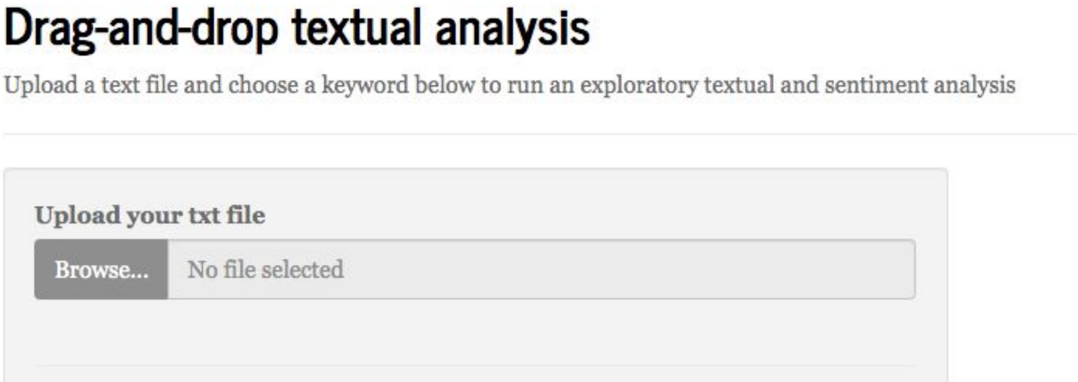
StoryBench is an app that uses **sentiment analysis** to analyze a .txt file. This analysis is based on a pre-determined dictionary that measures if a word is “positive” or “negative.”



How to Use Storybench

Storybench only reads ***one*** .txt file.

- Copy and paste all your .txt files into **one .txt file** titled “all.txt” (or another file name of your choice)
- Click “Upload your txt file” (or drag and drop)
- Explore the results!



Drag-and-drop textual analysis

Upload a text file and choose a keyword below to run an exploratory textual and sentiment analysis

Upload your txt file

Your turn!

Explore the corpus you prepared by using:

1. Voyant: https://voyant-tools.org/
2. SameDiff: https://databasic.io/en/samediff/
3. WordCounter: https://databasic.io/en/wordcounter/
4. Storybench: https://storybench.shinyapps.io/textanalysis/

Questions & Contact Information

Cara Marta Messina
Digital Teaching Integration Assistant Director
PhD Candidate, English
messina.c@husky.neu.edu

Garrett Morrow
Digital Teaching Integration Research Fellow
PhD Student, Political Science
morrow.g@husky.neu.edu