Northeastern University
*NULab for Texts, Maps, and Networks*

# DataBasic Text Analysis: WordCounter and SameDiff

Garrett Morrow

# DataBasic

DataBasic website: https://databasic.io/en/

Data Basic is an easy-to-use package of data analysis tools for beginners.

Four tools:

1. WordCounter
2. WTFcsv
3. SameDiff
4. ConnectTheDots

This presentation will focus on the text analysis tools: WordCounter and SameDiff

# Terminology

**Corpus/corpora**: a text or collection of multiple texts that is used for analysis. For example, one could create a corpus of all of Barack Obama's speeches to trace language over the course of his presidency.

**N-gram**: a continuous sequence of n items in a text, corpus, etc. For example, if looking at Obama's speeches, a bigram (2 items in a sequence) could be 'United States,' while a trigram (3 words) could be 'yes we can.'

**Stopwords**: commonly used words commonly omitted in text analysis. Stopwords typically do not add meaning to a sentence, but can supply context. Examples: the, but, this, that, an.

# What is 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

word COUNTER

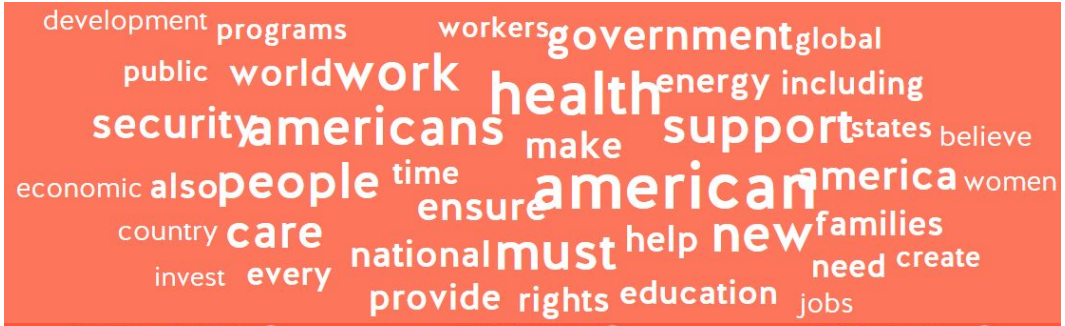
**1.** Select ‘upload a file’ from the top bar, then again on the middle bar.

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

Note: Ignore case and stopwords are enabled by default

**3.**Click on count

# 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.

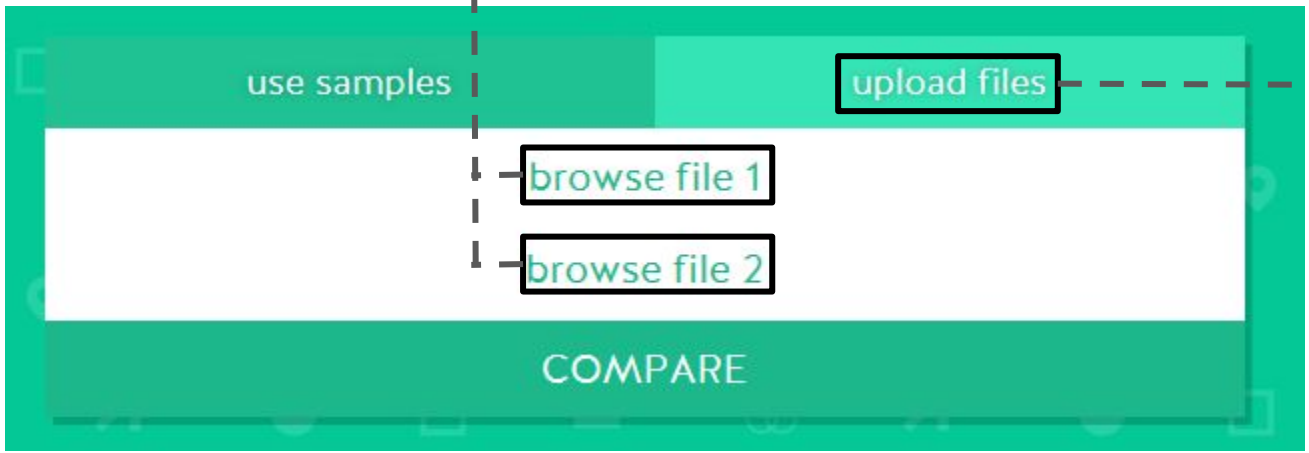
# What is SameDiff?

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.

samediff

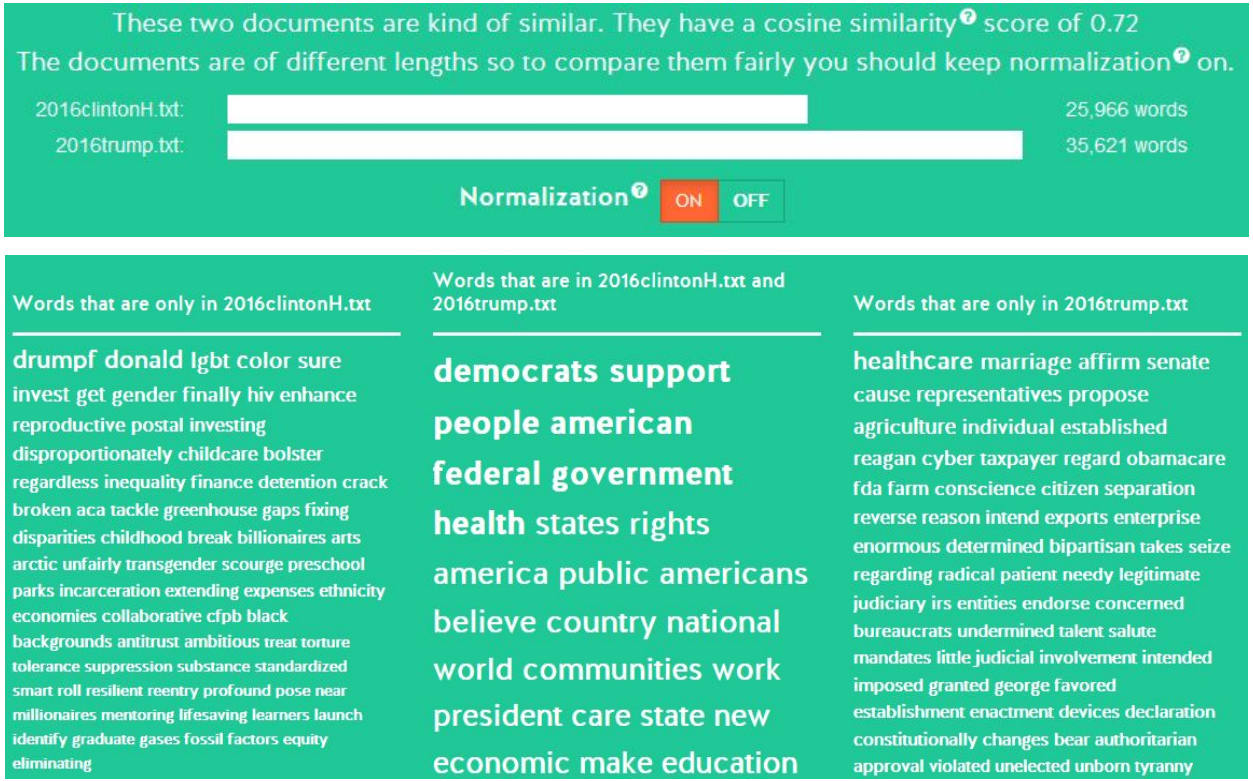
**2.** Navigate to the texts individually
**3.** Click Compare




**1.** Select upload files

# SameDiff Results

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




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.

# Questions & Contact Information

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