



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.

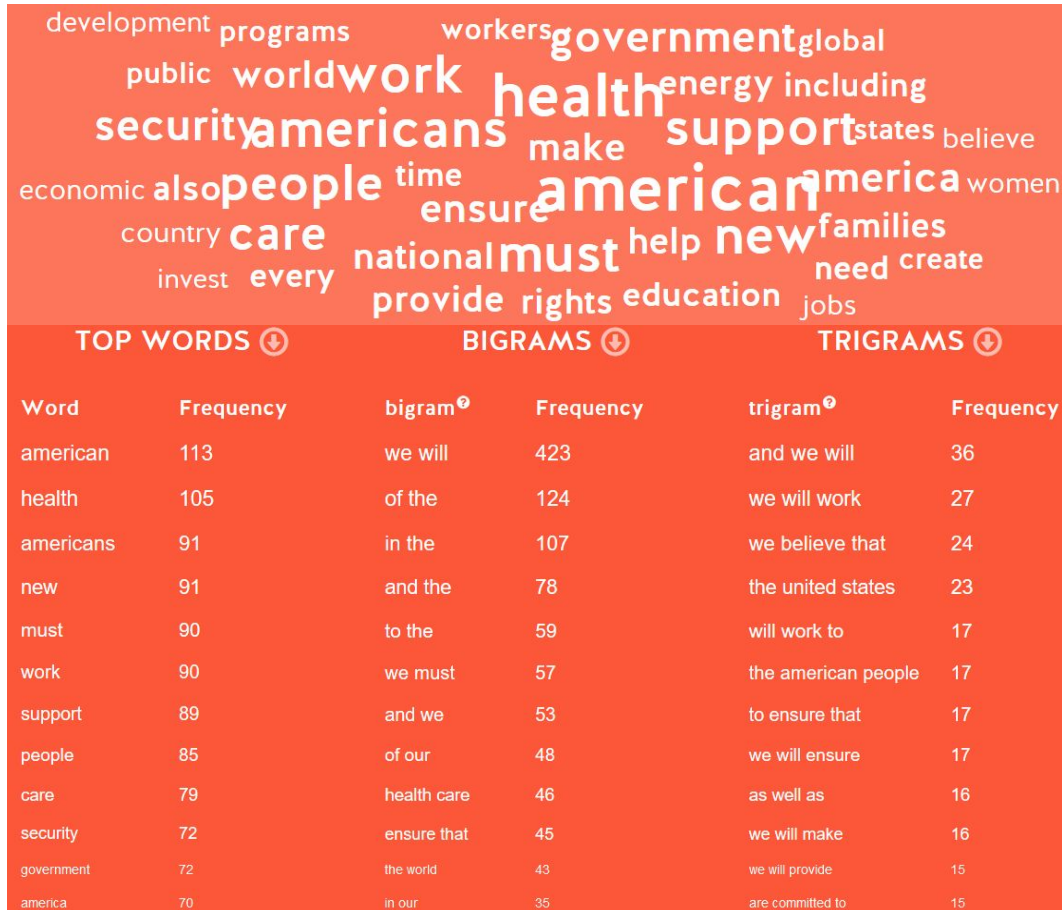
The output can then be downloaded as a .csv to do further analysis



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

A screenshot of the wordCOUNTER web application interface. The interface has an orange background. At the top, there are four buttons: 'use a sample', 'paste text', 'upload a file', and 'paste a link'. The 'upload a file' button is highlighted with a black rectangular box. Below these buttons is a large white text input area. The text 'upload a file' is pasted into this area and is also highlighted with a black rectangular box. Below the input area, there are two checkboxes, both of which are checked: 'ignore case' and 'ignore stopwords'. At the bottom of the interface is a large orange button labeled 'COUNT' in white capital letters. A dashed vertical line with arrows at both ends connects the 'upload a file' button in the top bar to the 'upload a file' text in the input area, indicating the sequence of actions.

WordCounter Results



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.

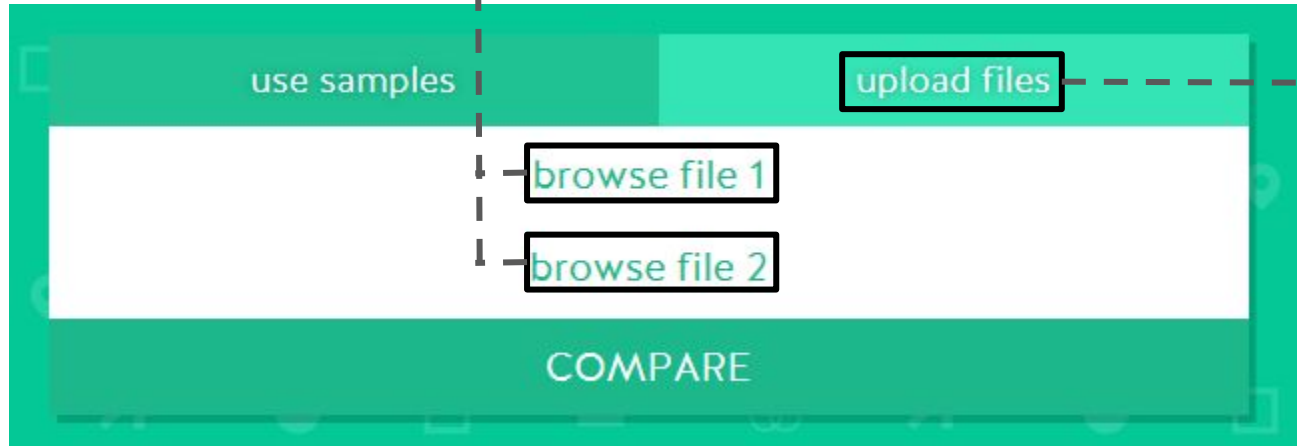
What is SameDiff?

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.



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.'

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:	<input type="text"/>	25,966 words
2016trump.txt:	<input type="text"/>	35,621 words

Normalization[?] ☒ ON ☐ OFF

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

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

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