Word Trends in the Middlebury Campus: 2011-2016

Group Members: Caroline Cating, Sarah Koenigsberg, Rebecca Conover

Purpose

The purpose of our investigation is to examine the evolution of the overall dialogue at Middlebury College in the past year (ideally up to the past 5 years). This will include looking at the most popular words as well as which words have increase or decreased the most in popularity over time. As such, we can shed light on which topics have waxed and waned in significance to the Middlebury community.

Scientific Questions

1) Which words are the most popular amongst the Middlebury community over the past year?
2) Which words (or groups of words) have increased the most in popularity?
3) Which words (or groups of words) have decreased the most in popularity?

Example: Has awareness of race and racial tensions on campus increased over the past five years? We could answer this question by looking at the change in frequency of pertinent words such as "race", "diversity", "intolerance" and "microagression".

Data Sources

Text data scraped from the **front pages** of The Middlebury Campus in 2016. taken from the archives on The Campus website.

We scraped data from front pages and compiled indivdual text files for each issue.

Then we wrote several functions that opened each txt file, **split** lines into vectors of words and **striped** white space and punctuation. Then we **counted** the occurance of each word in the file and created a data frame that contained words, counts, and dates. Then we **merged** the data frames created from each file to one large frame.

We did not do this very efficiently, so we only took data from the past year and the front pages, giving us 10,000 obeservations. We may decide to go back further (campus has usable archives from as far back as 2011) or to expand the scope to entire issues. This would not be difficult to do, as we created functions that we could be adapted.

Depending on the direction that our analysis takes, we may decide to encorperate data from a text corpus that examine national news trends–data format not ideal here, but will pursue further.

Data Format

For now: 1 table with 11,000 observations.

The observational units are word count on the front page per word for each issue over the past year.

Rows: 11,000 for each observational unit (could change if we expand the scope). Variables: words, counts, year, month, day.

Sample of Data Frame

x	freq	year	month	day
been	2	2016	5	5
been	2	2016	9	29
been	2	2016	11	10
been	3	2016	4	14
been	3	2016	4	28
been	4	2016	2	18
been	4	2016	5	12
been	4	2016	10	27
been	5	2016	3	24
been	6	2016	1	28
been	9	2016	3	3
before	1	2016	9	16
before	1	2016	9	22
before	1	2016	9	29
before	1	2016	11	3
before	2	2016	5	12
before	2	2016	11	16
being	1	2016	1	28
being	1	2016	5	12
being	1	2016	9	16
being	1	2016	9	29
being	1	2016	10	13
being	2	2016	2	18
being	2	2016	2	25