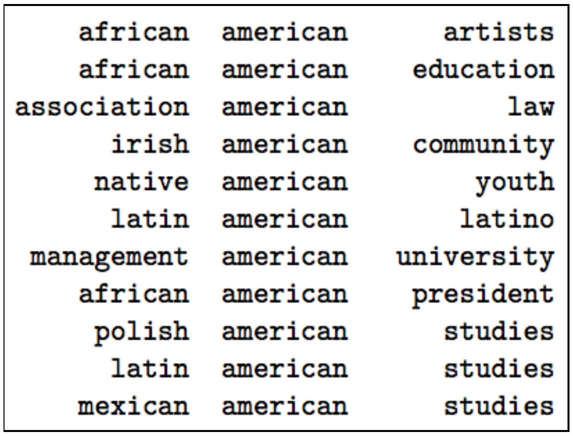
Methods Issues & Limitations: Word order matters.

Topic modeling uses a “bag-of-words” (BoW) approach to statistically analyze word co-occurrences in datasets. BoWs are simplified representations of text that do not contain information about word order, grammar, or (for the most part) punctuation; rather, they only preserve the number of times a given word appears in a document. While BoWs are particularly apt for topic modeling (which measures a word’s frequency to determine what other words it should be clustered with), these representations significantly diverge from natural language usage. For example, by the time the phrase “I like the humanities” passes through the preprocessing workflow WE1S uses to transform documents into BoWs, it will simply read: “{‘humanities’:1}.”

The interpretive payoff that topic modeling offers often outweighs the cost of creating such distance between our data and its original form. But BoWs also occlude potentially significant patterns in language. A key part of WE1S involves studying relationships between the humanities and racial, ethnic, gender, and other social groups such as first generation students. One common way to begin finding these relationships is to search our models for instances where textual data signifies identity, as in “first generation students” or “African American.” But multi-word strings are the very things our preprocessing breaks apart. A phrase like “first generation students” is not typically retained in BoWs; and if preprocessing removes it, it will not show up in our models. *Data preparation, then, may thus have highly consequential ramifications for interpretation. For WE1S’s work with social groups, preprocessing erases the context-dependent markers of identity we would otherwise want to find and study.*



african	american	artists
african	american	education
association	american	law
irish	american	community
native	american	youth
latin	american	latino
management	american	university
african	american	president
polish	american	studies
latin	american	studies
mexican	american	studies

Trigrams in Collection 15. A typical preprocessing workflow would break apart these phrases.

Techniques like named-entity recognition (NER) or phrase detection can partly ameliorate this problem, but multi-word markers of identity tend to be permutational, combinatorial, and fuzzy. Perhaps unsurprisingly, there is no standardized way people talk about identity or identify themselves. While we have experimented with collocation analysis to look for phrases that may go undetected by NER or phrase detection, our work with this method is only in a pilot stage, and it is not without its own complications. For now, it remains unclear to us how researchers might best sensitize their data work to the dynamics of identity.

Resources

WE1S preprocessing workflow:
github.com/whatevery1says/preprocessing

Recommended reading:

* Mimno, David. “Using phrases in Mallet topic models” (2015)

* WE1S Bibliography: Text Analysis

16 July 2020; rev. 12 Nov. 2021 (Su Burtner, Melissa Filbeck, sam goli, Jamal Russell, Tyler Shoemaker)