UCSD DSC96

Web Scraping

Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes

Nikhil Garg
Stanford University
nkgarg@stanford.edu
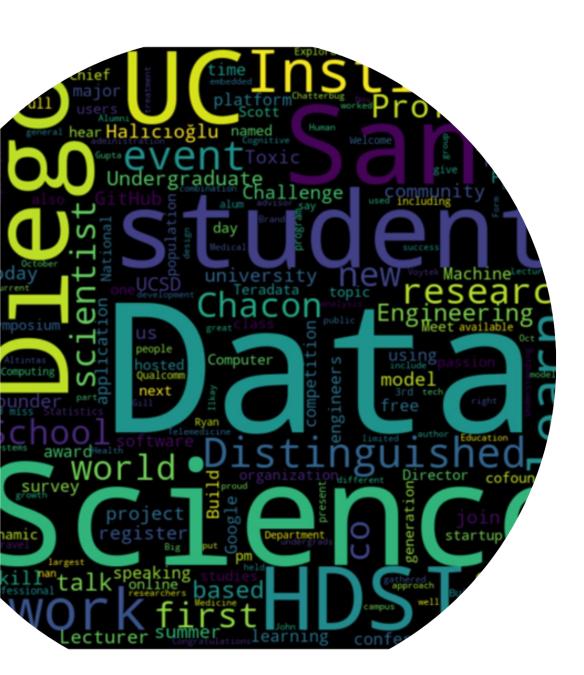
Londa Schiebinger
Stanford University
schieb@stanford.edu

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

James Zou
Stanford University
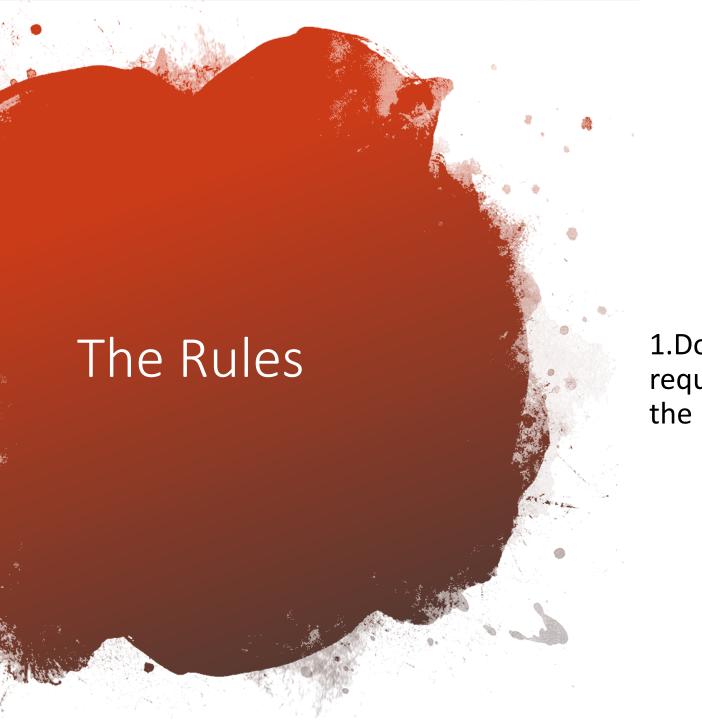jamesz@stanford.edu

November 23, 2017

Abstract

Word embeddings use vectors to represent words such that the geometry between vectors captures semantic relationship between the words. In this paper, we develop a framework to demonstrate how the temporal dynamics of the embedding can be leveraged to quantify changes in stereotypes and attitudes toward women and ethnic minorities in the 20th and 21st centuries in the United States. We integrate word embeddings trained on 100 years of text data with the U.S. Census to show that changes in the embedding track closely with demographic and occupation shifts over time. The embedding captures global social shifts – e.g., the women's movement in the 1960s and Asian immigration into the U.S – and also illuminates how specific adjectives and occupations became more closely associated with certain populations over time. Our framework for temporal analysis of word embedding opens up a powerful new intersection between machine learning and quantitative social science.

UCSD DSC96

Web Scraping

The Rules

1.Don't break anything. Many rapid requests to smaller sites can overload the host server.

The Rules

2.Use a published API if possible - it is more robust and usually much easier!

The Rules

3.Respect the policy published at robots.txt

The Rules

4.Don't spoof your UserAgent (or try to trick the server into thinking you are a person)

The Rules

5.Read the Terms of Service for the site and follow it.

NLP Example

- http://lotrproject.com/statistics/books/wordscount