社会学におけるテキスト分析とNDL全文データの可能性

#NDL全文使ってみた　～「次世代デジタルライブラリー」＆「NDL Ngram Viewer」

東京大学

瀧川裕貴

全文テキストに関わる過去の自分自身の研究

1. 「国会会議録」全文テキストを用いた道徳社会学の研究
2. 『社会学評論』全文テキストを用いた戦後社会学史の研究

「国会会議録」全文テキストを用いた道徳社会学の研究

- データ：1947-2016年の衆参両院，本会議と各委員会会議録のほぼすべて．発言数8,328,057
- 目的：日本の政治家による言論の感情および道徳原理の用い方の立場（与党・野党）による違いや変化を探る

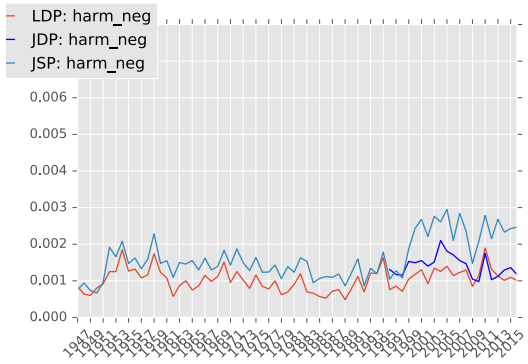
「国会会議録」全文テキストを用いた道徳社会学の研究

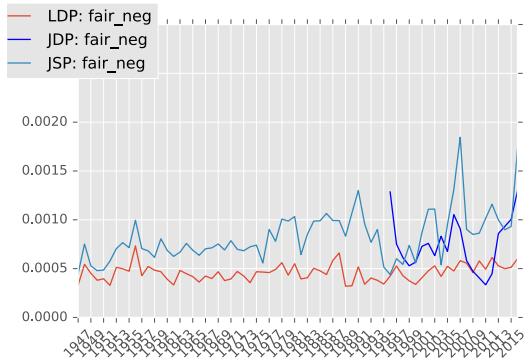
- 方法：辞書を用いた頻度分析
  - 日本語評価極性辞書（東山ほか2008）
  - J.Haidtの道徳基盤辞書の独自翻訳版

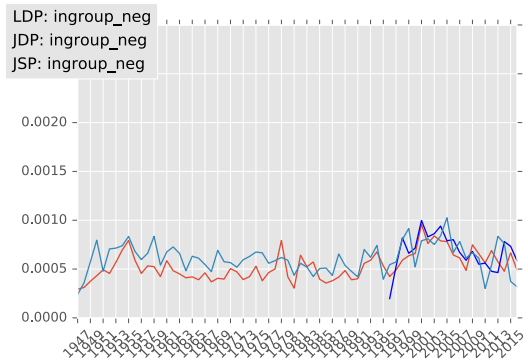
- The moral dictionary and its translated version

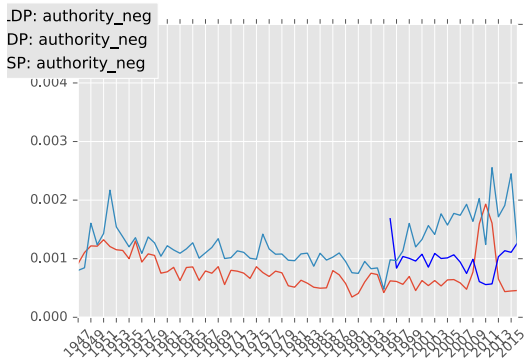
安心* 01	safe*
平和* 01	peace*
思いやり* 01	compassion*
共感* 01	empathy*
同情* 01	sympathy*
ケア01	care
思いやり01	caring
保護* 01	protect*
保護01	shield
避難所01	shelter
友好01	amity
安全* 01	secure*
利益* 01	benefit*
防衛 * 01	defen*
守る* 01	guard*
保存する01 07 0	preserve

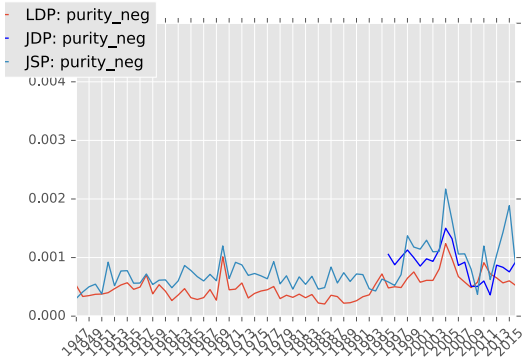
「国会会議録」全文テキストを用いた道徳社会学の研究











- 日本の主要政党による道徳原理の用い方とその変化

『社会学評論』全文テキストを用いた戦後社会学史の研究

- データ：『社会学評論』に掲載された学術論文．1号（1951年）から65号（21015年）までの全論文（書評等を除く）1646本
- 目的：戦後日本社会学が扱った研究テーマや研究領域・研究関心の変遷を明らかにする

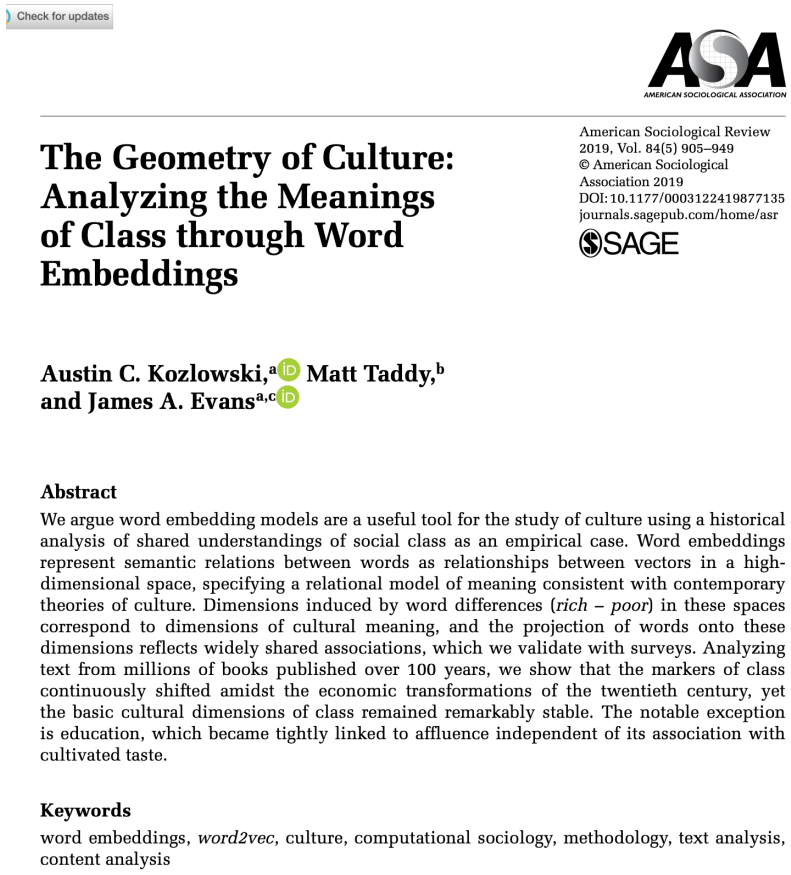
『社会学評論』全文テキストを用いた戦後社会学史の研究

- 方法：トピックモデルという機械学習の方法を用いて，全文データから自動的にトピックを抽出


01:数理社会学
02:日本社会学
03:家
04:ジャンク1
05:情報とシステム
31:社会学教育
32:人口移動
33:ジャンク3
34:社会意識
35:ルーマンと社会理論
06:医療
07:日本論
08:組織
09:自己論
10:子ども
36:メディア論
37:ジャンク4
38:社会調査論
39:同族
40:現象学的社会学
11:職業調査
12:部落・村
13:行為・規範・意味
14:家・相続
15:計量社会学
41:教育社会学
42:マルクス主義
43:労働社会学
44:産業社会・脱産業社
45:集団
16:研究動向
17:都市と地域
18:構築主義
19:社会階層・階級
20:コミュニケーション
46:女性・性別役割
47:農業
48:外国人・移民
49:デュルケム社会学
50:住宅と都市
21:文化・芸術
22:ケア
23:機能主義
24:村（歴史）
25:視覚・聴覚メディア
51:社会運動
52:学校
53:記憶・環境運動
54:家族（世帯）
55:研究・報告
26:宗教社会学
27:明治・近代
28:ジャンク2
29:支配と権力
30:経済社会学
56:語り・インタビュー
57:住民運動
58:死
59:国家・市民権
60:野宿・寄せ場
Proportion
Proportion
0.0
0.1
0.2
0.3

社会学における全文テキスト研究

- データ：The Google Ngram corpus（全文ではなく，5-gram[5つの連続する単語の集まり]を使用）
- 目的：社会階級の文化的意味の変遷を探る



Check for updates

ASA
AMERICAN SOCIOLOGICAL ASSOCIATION

American Sociological Review
2019, Vol. 84(5) 905–949
© American Sociological Association 2019
DOI: 10.1177/0003122419877135
journals.sagepub.com/home/asr
SAGE

The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings

Austin C. Kozlowski,[a] Matt Taddy,[b] and James A. Evans[a,c]

Abstract

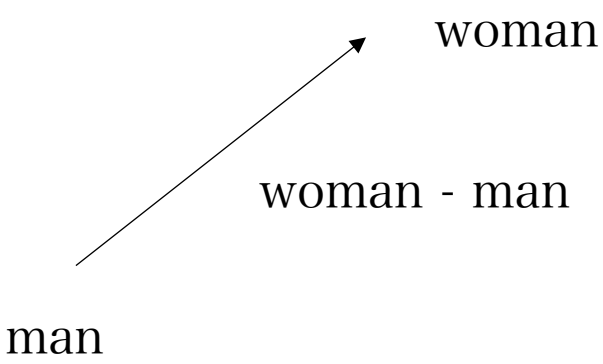
We argue word embedding models are a useful tool for the study of culture using a historical analysis of shared understandings of social class as an empirical case. Word embeddings represent semantic relations between words as relationships between vectors in a high-dimensional space, specifying a relational model of meaning consistent with contemporary theories of culture. Dimensions induced by word differences (*rich – poor*) in these spaces correspond to dimensions of cultural meaning, and the projection of words onto these dimensions reflects widely shared associations, which we validate with surveys. Analyzing text from millions of books published over 100 years, we show that the markers of class continuously shifted amidst the economic transformations of the twentieth century, yet the basic cultural dimensions of class remained remarkably stable. The notable exception is education, which became tightly linked to affluence independent of its association with cultivated taste.

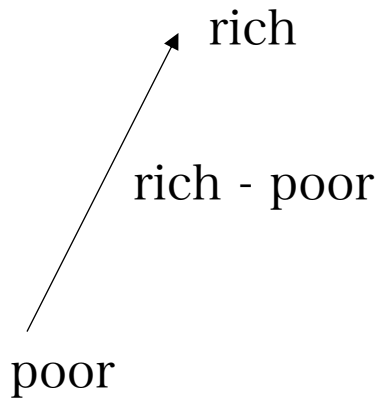
Keywords

word embeddings, *word2vec*, culture, computational sociology, methodology, text analysis, content analysis

社会学における全文テキスト研究

- 方法：単語埋め込みモデルの応用による文化的意味次元の抽出
  1. 単語の持つ超高次元情報を低次元ベクトルに効率的に縮約（分散意味表現)
  2. さらに対義語の単語ベクトルを組み合わせて特定の文化的意味次元を表す解釈可能なベクトルを構成



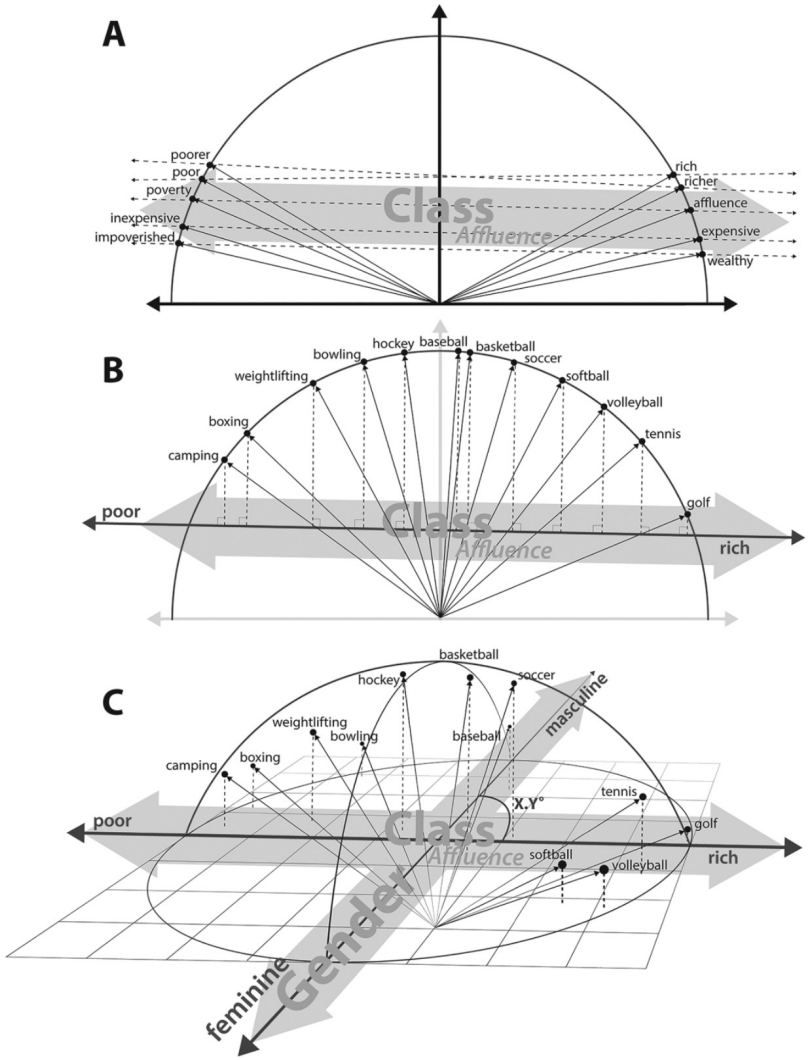


社会学における全文テキスト研究

A: 経済的豊かさの軸の構成

B: 経済的豊かさという意味次元上でのスポーツの位置

C: 経済的豊かさおよびジェンダーという意味次元上でのスポーツの位置



NDL全文テキストを用いた研究計画

- 単語埋め込みモデルを用いた社会的実践・思想・概念の文化的意味づけの構造と歴史的変遷の分析