Replication Exercise: "A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors"

Theme


PAY FOR EQUAL PAY


It's equality.


WE'RE 50/50, RIGHT?
#UNCHARTEDMOVIE


NOBODY'S BETTER


EQUALITY
NETFLIX


We are truly equal


FREEDOM IS EQUAL HOUSING TOO-

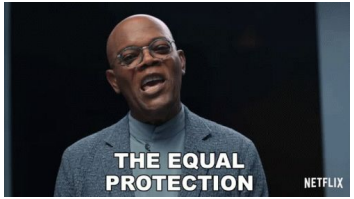

EVERYONE'S EQUAL


EVERYBODY IS TO BE TREATED EXACTLY THE SAME


I GUESS WE'RE EVEN NOW


EQUAL MEANS EQUAL.


WE'RE ALL EQUAL HERE, OKAY?


THE EQUAL PROTECTION
NETFLIX


WE ARE ALL THE SAME

Theme

- Diachronic Analysis: A linguistic analysis approach which considers the development and evolution of a language through history. (i.e. temporal analysis, e.g. ‘equality’ then vs. ‘equality’ now)
- Big corpus vs. Small corpus: How to build good models with small corpus.
- Arbitrary cut points introduce unknown effects into the produced trends
- Language instability that causes semantic relationships to also ebb and flow.
- Spatial noncomparability which precludes direct comparison of cosine similarity across distinct corpora.

Data

- corpus of newspaper articles (n=3105) from The New York Times (NYT), Reuters, and the Associated Press, accessed via the NYT Articles API. All articles from 1855 to 2016 with the word “equality” in the headline were downloaded.
- seven 25-year time slices (eras)
- significant variation of document counts due to both the varying newsworthiness of equality and the general increase in the volume of news articles produced in later slices.

Era	DocumentCount
1855	63
1880	96
1905	432
1930	995
1955	576
1980	221
2005	321

Producing the Gold Standard

- Created a codebook for “equality”
- Hand-coded a random sample of 400 articles
- Fed these 400 coded articles to the ReadMe V1 supervised topic model, and applied to the full corpus, which was splitted into seven 25-year eras
- Estimated the proportion of documents in each topic for each era
- Selected five topics, where the topic could be approximated by a single word, to serve as the “gold standard”

ReadMe V1 (R Package)

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, *54*(1), 229-247.

- Mainly relies on word frequency statistics features, without using word embeddings.
- Define topic categories and manually label a small training set.
- Input the training set and unlabeled test set into the ReadMe algorithm.
- Estimate the topic probabilities for each document in the test set.
- Aggregate the probabilities to estimate overall topic proportions.
- Analyze topic prevalence changes over time by dividing the test set into time periods.

Assumptions and Choices

Assumptions:

- Every text in the corpus has two topics: ‘equality’, and a 2nd topic.
- Proportions of secondary topics track the strengths of the semantic relationships in the corpus between them and Equality.
- Semantic relationships between the words (“gender,” “race,” “african_american,” “german,” and “treaty”) and “equality” mirror semantic relationships between the topics (gender, race, African American, Germany, and international relations) and the equality topic.

Data choices:

- “Equality” in the headline = Equality topic. Secondary topic is assigned by ReadMe, using an initial set of human-coded texts.

- Choose topics (gender, race, African American, Germany, and international relations) due to a significant overlap in representative keywords.

“required preprocessing .., performing some language stabilization and selective stemming.”

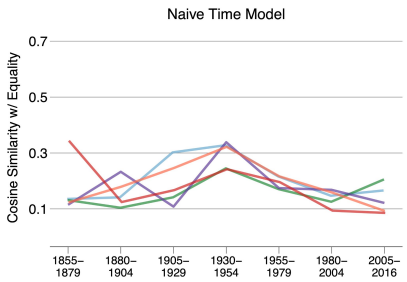
Word2vec Models

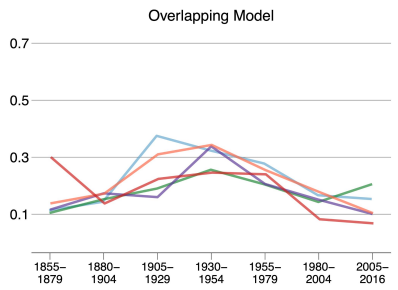
- Naive time series model: simple slicing
- Overlapping time series model: overlapping edges between slices
- Chronologically trained model: use previous slice's vector to build the next's, with the first slice's using the full corpus vector
- Aligned time series method: simple slicing with post-modelling alignment using a selected anchor slice

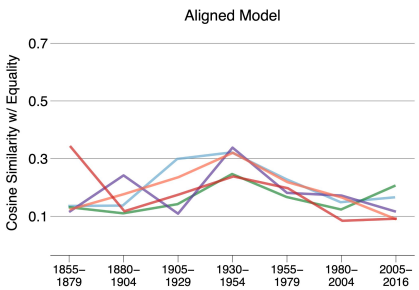
Bootstrap resampling was applied to all 4 models to generate stable cosine similarities and confidence intervals.

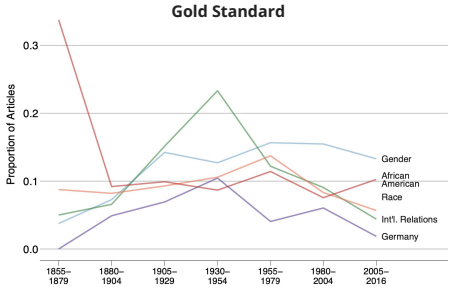
("*cosine similarities stabilize at between n=25 and n=50 bootstrapped samples*")

Similarities and differences

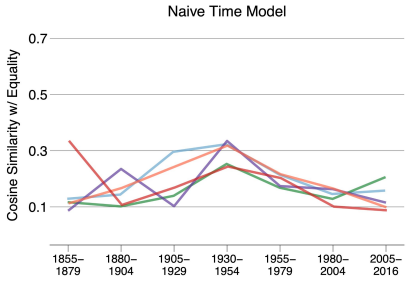


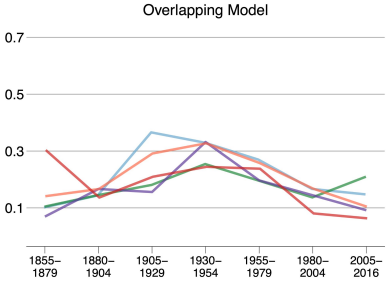


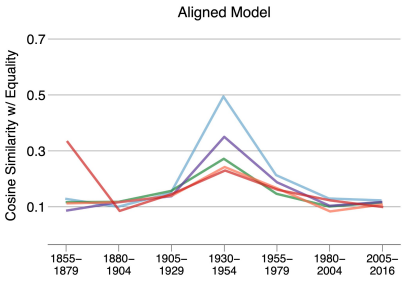


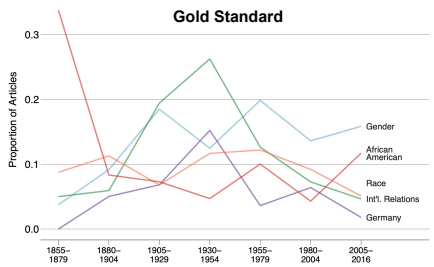


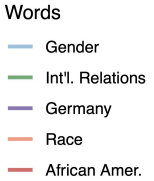
Our results ⬆️ vs. Rodman ⬇️











Similarities and differences (Chrono model)

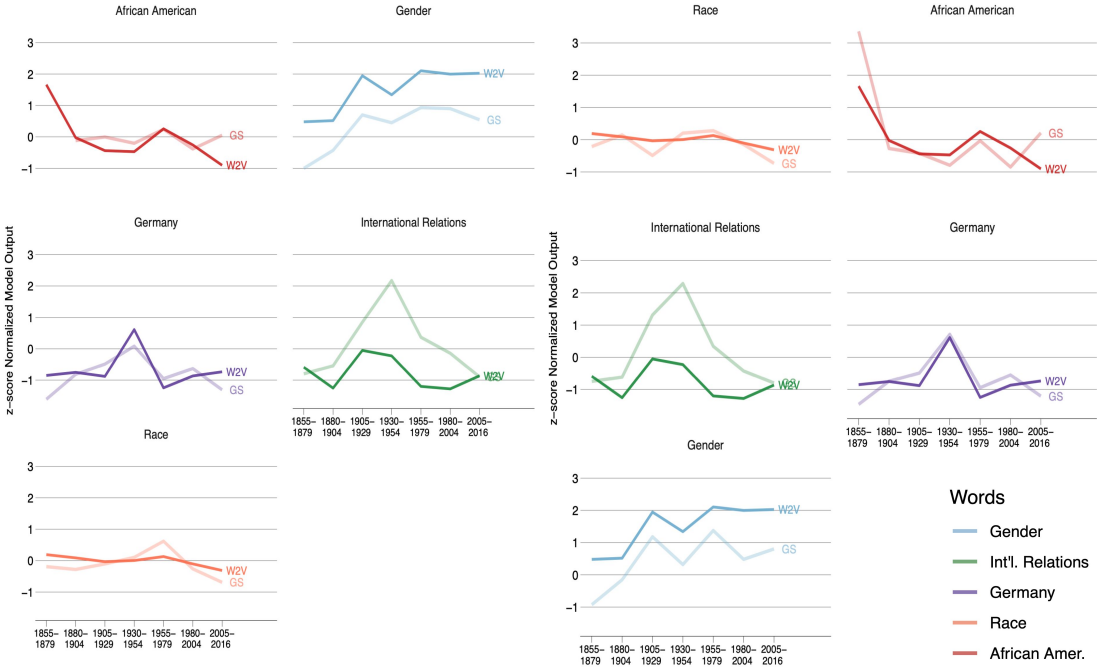


Table 1: Rodman's ANOVA

	Model	Deviance	Squared.Deviance	Correlation
1	naive	26.613	25.589	0.555
2	overlap	25.163	25.602	0.564
3	**chrono**	22.689	21.860	0.612
4	aligned	25.385	25.775	0.515

Table 2: Replication ANOVA

	Model	Deviance	Squared.Deviance	Correlation
1	naive	27.856	27.182	0.532
2	overlap	25.558	25.563	0.520
3	**chrono**	24.179	23.312	0.568
4	aligned	27.508	26.797	0.536

Autopsy

- Compatibility
- Faster processing time with more modern/powerful equipment (~12 hours vs. ~43 hours originally)
- Less stable Gold Standard baseline can largely determine the relative superiority of the Chrono model
- Well-documented model rationales, assumptions, and decisions ⇒ Easier, more comparable replication.

Run time per script (in order of replication):

- `01-gold_standard_model.R` = 5:18:42
- `02-naive_time_word2vec.py` = 11:08:28
- `03-overlap_word2vec.py` = 11:49:48
- `04-chrono_word2vec.py` = 5:25:54
- `05-aligned_word2vec.py` = 9:32:07
- `06-figures.R` = 0:00:11

Extension

- Application of ReadMe V2
  - Jerzak, C. T., King, G., & Strezhnev, A. (2023). An improved method of automated nonparametric content analysis for social science. *Political Analysis*, *31*(1), 42-58.
- Similar examinations of other words/concepts (e.g. “soul”, “rich”, “war”, “power”, “scandal”, “king”, “justice”, “normal”, etc.)
- Gather and replicate this study using a richer, larger, more comprehensive corpus of newspaper articles.