הבינתחומי הרצליה
IDC HERZLIYA

Advanced Machine Learning course

Final project report

Utilizing the notion of temporal changes in social stigmas[1]

Alon Mannor[2]

Under the tutorship of Abraham Israeli

[1] All code & data related to this report can be found at: https://github.com/Amannor/EmbeddingDynamicStereotypes/tree/final-proj-temporal-social-stigmas

[2] To whom correspondence may be addressed: Alon.mannor@gmail.com

September 2020

Table of contents

Introduction 3

Related Work 4

Historical and current context 4

100 years of word embedding 5

Extension project 6

Method 6

Results 7

Graphs 8

Next steps 11

Epilogue 12

Introduction

Word embeddings are a very powerful and useful tool for text analysis. It's a family of algorithms all centered around mapping words to real number vectors of finite dimension on the basis of co-occurrence of different words in very large text corpora. Various algebraic operations can be performed on the resulting vectors (e.g. addition and subtraction) – s.t. the end result of said operations represent a semantical relation between the word vectors. Thus, for example, the vector for "king" minus "man" plus "woman" will result in a vector almost identical to that of the word "queen".

This method introduced many a new possibilities thanks to its novel and efficient approach. This is true, among other things, in the realms of Natural Language Processing.

More specifically, word (and later whole phrase) embeddings, when used as the underlying input representation, have been shown to boost the performance in NLP tasks such as syntactic parsing[3] and sentiment analysis[4].

Pertaining to this specific project, the method chosen was embedding a set of words simply by taking the average vector of the words that make up the set (see below).

[3] Socher, Richard; Bauer, John; Manning, Christopher; Ng, Andrew (2013). *Parsing with compositional vector grammars* (PDF). Proc. ACL Conf.

[4] Socher, Richard; Perelygin, Alex; Wu, Jean; Chuang, Jason; Manning, Chris; Ng, Andrew; Potts, Chris (2013). *Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank* (PDF). EMNLP.

Related work

Word embedding was developed as a framework to represent words as a part of the artificial intelligence and natural language processing pipeline[5].

Historical and current context

The relation between language and the subjective perception of reality is a long studied one. Numerous works have shown that the way we speak and the language we construct influences our very basic perception of what we view and experience as real[6].
With that in mind, it's no wonder that analyzing trends in distances between word vectors (and perhaps vectors for whole sentence, words set etc.) has gained popularity in social sciences in recent years, as it allows for a relatively straightforward and easy-to-use tool to show semantic relations between large (often historical) text corpora. These in turn can be applied to social trends pertaining to the sets of words chosen (e.g. evaluating just "how much more" the society associates women with certain professions in the wake of the suffragette movement).

There have been quite a few articles dealing with "built-in" biases in text corpora that later reflect a bias in the resulting words vector space. These can be grouped into three groups:

Identifying and validating the problem

Some have shown that such bias exists in the resulting word vectors and that this correlates with real-world data [16][7]. This was extended for sentence vectors as well[8]. For example, the Word Embedding Association Test shows that GloVe and word2vec word embeddings exhibit human-like implicit biases based on gender, race, and other social constructs[9]. This phenomenon has been shown to occur in many languages[10]. Interestingly, this has also been applied (or perhaps "extended") to the study of social classes (e.g. in the wake of economic transformations[11]).

Tackling the problem by suggesting tools

Others have suggested a way to detect automatically such biases and called for the community to incorporate detection mechanisms into their work[12]. Another method has been demonstrated to keep the bias in some dimensions of the vectors while other dimensions are debiased[13].

Criticism of the tools offered so far

These criticisms range from the "general" \ "big picture" to the specific issue of biases in word vectors. The former can be exemplified by detecting potential problems and offering solutions to the word

[5] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26*, eds Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (Curran Associates, Inc, Lake Tahoe, NV), pp 3111–3119.

[6] See for example the Sapir–Whorf hypothesis: Wikipedia contributors. "Linguistic relativity." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 14 Sep. 2020. Web. 27 Sep. 2020.

[7] Rudinger, R., Naradowsky, J., Leonard, B., & Van Durme, B. (2018). Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301.*

[8] May, C., Wang, A., Bordia, S., Bowman, S. R., & Rudinger, R. (2019). On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903*.10561.

[9] Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183-186.

[10] Lewis, M., & Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature human behaviour*, 1-8.

[11] Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review, 84*(5), 905-949.

[12] Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist—it's time to make it fair.

[13] Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K. W. (2018). Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496.*

embeddings evaluation methods as a whole[14]. The latter have, for example, criticized current debiasing methods and claimed that much more work is needed before a "good enough" level is reached[15].

100 years of word embedding

The work in this report builds on the words-set-to-vector approach described in the article "Word embeddings quantify 100 years of gender and ethnic stereotypes"[16]. In the article, the authors used a variety of methods to measure the "semantic proximity" between two sets of words. Most notably taking the average distance between all possible pairs in the Cartesian product of the two sets. This was accomplished in several ways – the two major ones being averaging the Euclidean distance and the other by averaging the cosine similarity.

The rationale behind this method was to provide a more accurate way to show social trends pertaining to biases. As the title of the article suggests, the biases checked were gender (men\women), ethnic (Hispanic\Asian) and religious (Christian\Muslim). Every group was represented by a set of words and there were different "thematic" sets of words – from which the distance was calculated (e.g. professions, words related to terrorism, external appearance etc.).

[14] Bakarov, A. (2018). A survey of word embeddings evaluation methods*. arXiv preprint arXiv:1801*.09536.

15 Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862.*

[16] Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences, 115*(16), E3635-E3644. Hereinafter "the article".

My extension project

Method - Utilizing the notion of temporal changes in social stigmas

Social stigmas have always fascinated me. What causes a group of people to “decide” collectively (and perhaps subconsciously) to label a person or another group of people? Does the answer lie in the fields of psychology\sociology, politics or evolutionary sciences? Moreover, do they change over time, and if so - how?

These questions came to my mind as I read the article and tried to think about a subgroup of people that suffer from collective stigma.

Medical illnesses, specifically mental-related illnesses, have been a source for much interest and mystique for human societies since the dawn of civilization. These, in turn, served as a hotbed of social stigmas. In modern times, these stigmas are discussed in a broader sense. One such important work is Madness and Civilization (by Michel Foucault[17]) – it traces the development of the different (socio-political) aspects associated with the word “Madness”.

With this state of mind, I set out to try to see what, if any, temporal trends exist. I used three sets of words:

i) Negative stigmas and connotations, usually directed towards people with medical mental conditions[18]
ii) Words pertaining to medical mental conditions[19]
iii) A “control group” of words pertaining to medical physical conditions[20]

The trends I chose to focus on were temporal ones. As such, out of the six data sets that were used for the article, I focused on the three temporal ones*:

- NYT[21] - Embeddings from the *New York Times* Annotated Corpus for every year between 1987 and 2004. Each vectors set represents a text from a 3-year period (1987-1990, 1988-1991, etc.).

- SVD[22] - Vectors trained on a combined corpus of genre balanced Google Books and the COHA[23]. For each decade, a separate embedding is trained from the corpus data corresponding to that decade.

[17] Wikipedia contributors. "Madness and Civilization." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 25 Sep. 2020. Web. 26 Sep. 2020.
[18] fear, loathing, afraid, terrified, scared, uneasy, deter, anxious, alarmed, spooked, horrified, nervous, despise, disgust
[19] depression, anxiety, agony, helpless, hopeless, melancholy, desperation, blue, unhappiness, sorrow, sadness, stressed, upset, adrift, stuck, worthless, lost, blackness, alone
[20] injury, injured, ailing, bad, down, ill, indisposed, peaked, peaky, sickened, unwell, symptomatic, cruddy, sickish, nauseated, nauseous, qualmish, queasy, queazy, squeamish, airsick, carsick, seasick, dizzy, shaky, woozy, achy, feverish, diseased, disordered, decrepit, feeble, fragile, frail, infirm, invalid, sickly, weak, weakly, afflicted, debilitated, disabled, halt, incapacitated, lame
[21] Raw vectors: http://stanford.edu/~nkgarg/NYTembeddings/. direct link for normalized, ready to use: https://drive.google.com/file/d/1JNy19NfBwNj5JWj71UipA-zrFVIsYp1p/view?usp=sharing
[22] Raw vectors: http://snap.stanford.edu/historical_embeddings/coha-word.zip direct link for normalized, ready to use: https://drive.google.com/file/d/1YajeJU2tQOG6GEgX_HXe8uOBveFxDjcc/view?usp=sharing
[23] Davies M (2010) The 400 million word corpus of historical American English (1810–2009). *Selected Papers from the Sixteenth International Conference on English Historical Linguistics (ICEHL 16), Pe´ cs, 23–27 August 2010*, eds Hegedu˝ s I, Fodor A (John Benjamins Publishing, Amsterdam), Vol 325.

This source is calculated using SVD. In the original article, they only looked at the data from the 1900s and so did I.

- SGNS[24] – same as SVD, the only difference being this is calculated skip-gram with negative sampling (SGNS)(also known as word2vec). In the original article, they only looked at the data from the 1900s and I added the 1800s to this set (available in the original data source).

I calculated the distance between set (i) and each one the sets (ii) and (iii) for each one of these temporal vector sets. The distance between two word sets was calculated as the average of all the ordered pairs that make up the Cartesian product of the two sets. For each set distance, two kinds of distance were calculated – the Euclidean and the cosine similarity.

Results

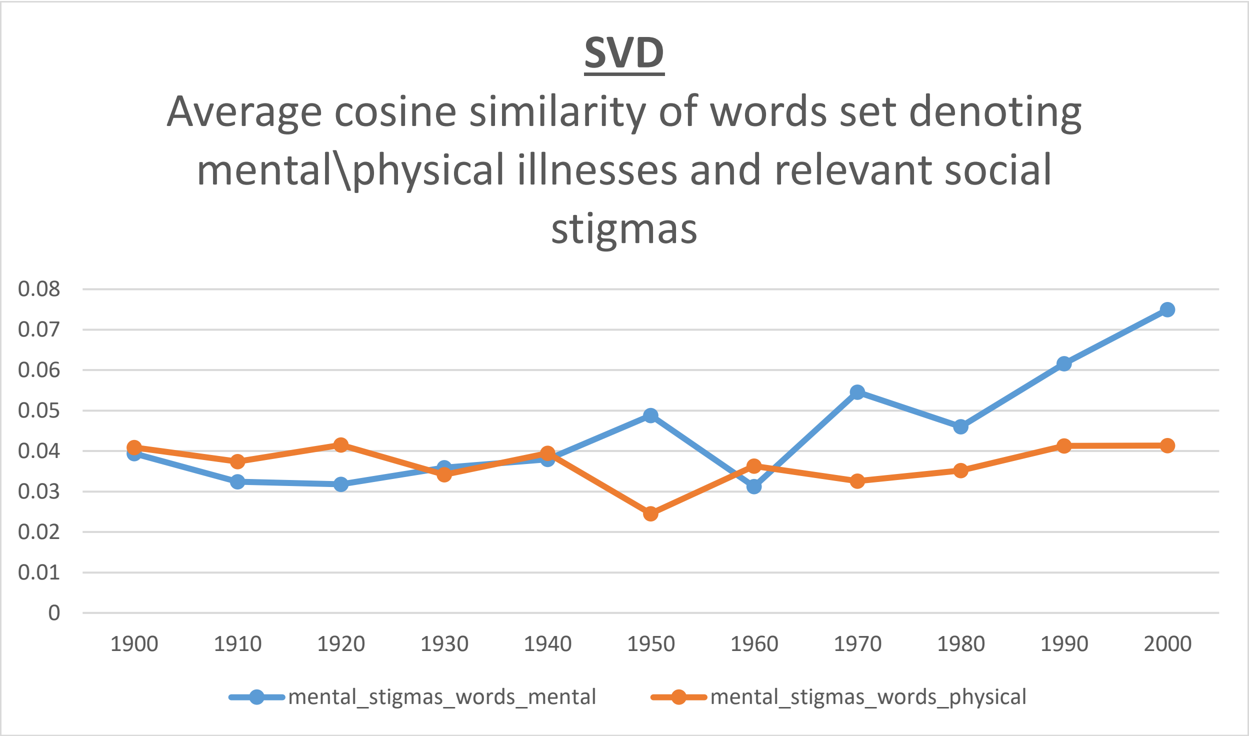
The results are shown in the three pairs of graphs below, and can be grouped to two different result "equivalence classes":

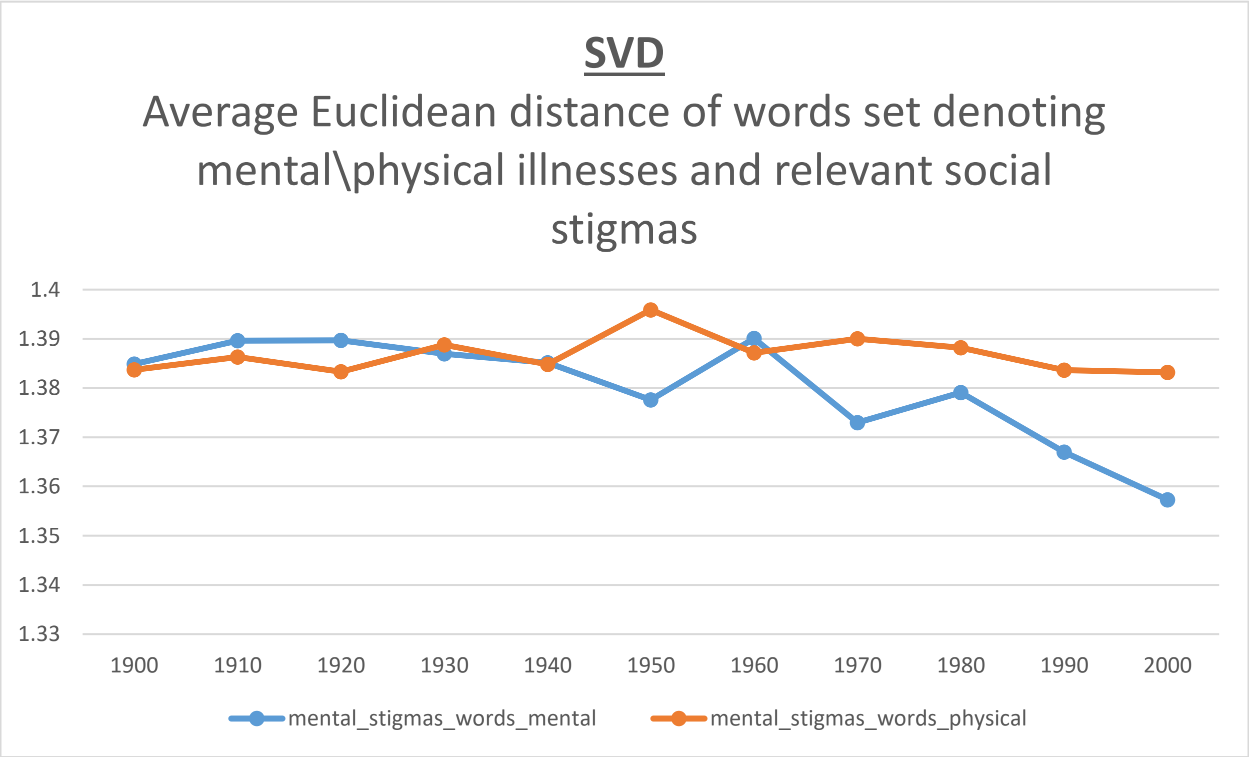
- The results for the SVD and SGNS vector sets showed considerable overlap between the distance of each of the two groups ((ii) and (iii)) to the stigmas. Accordingly, they were also semantically similar cosine-similarity wise. The SVD graph shows a trend throughout the last three decades of the 1900s of the stigmas and the words associated with mental conditions becoming slightly closer. This trend might because the words chosen associated with mental conditions are words that are more prevalent in modern times. However, this conclusion is not supported in the SGNS results and hence should be researched further (see Next Steps section).
  Overall for both SGNS and SVD, the trends suggest that further research is still needed, especially in the words chosen (see Next Steps section).
  Note that pair the aforementioned description, the SGNS graph contains data for one more century (the 1800s). I added the time span despite it not being taken in the original article since I wanted to eliminate the possibility of any "surprising" results "hiding" in the data from that century. The graphs show that indeed that data trends correspond to those from the 1900s.
  For a possible explanation of the "jump" between 1810 and 1820 in SVD please see "Wording" in the Next steps section.

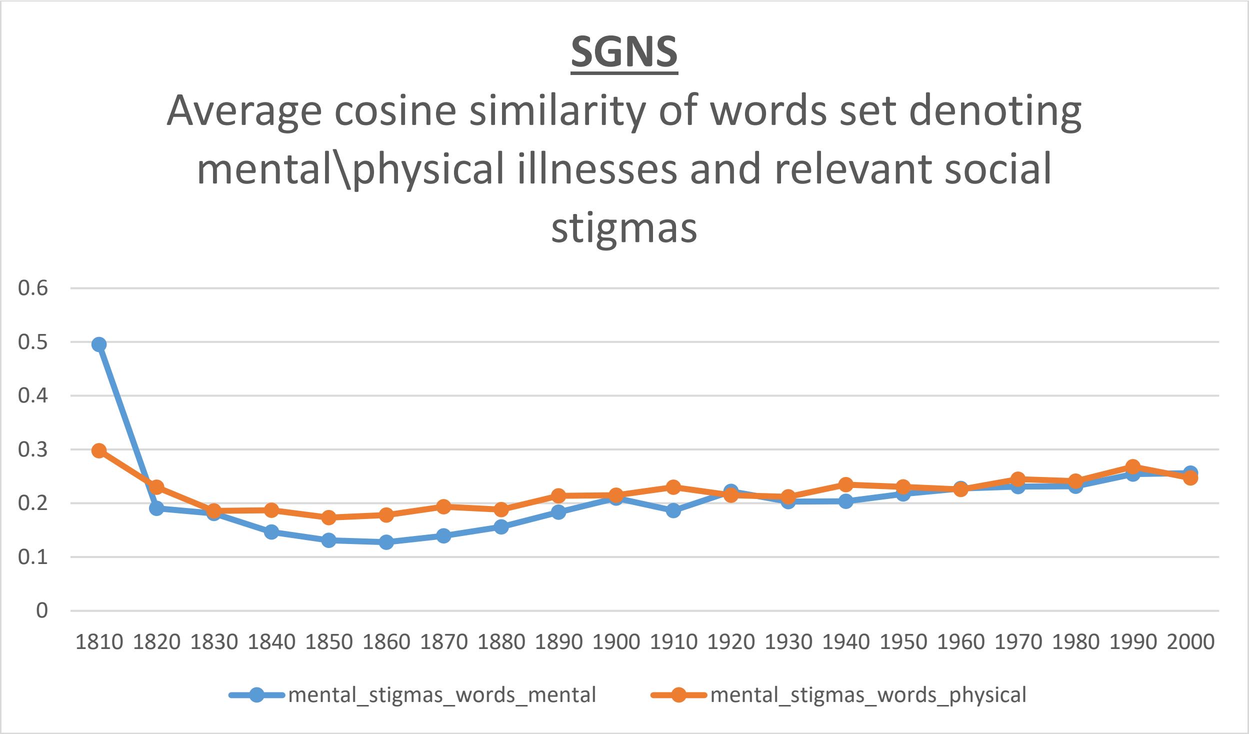
- The results for the NYT vector set were much more statistically obvious as they exhibit a rather constant gap between the relevant distances as well as in the cosine similarity. The set of words associated with various mental conditions was closer (than their physical equivalent) to the set of words describing negative social stigmas. Same trend was shown in the cosine similarity graph – the "mental conditions" words set is more similar to the negative social stigmas words set than the "physical conditions" words set. This reaffirms our intuition as to the overall negative stigmas that mental conditions bring upon the people suffering from it.

[24] Raw vectors: http://snap.stanford.edu/historical_embeddings/coha-word_sgns.zip direct link for normalized, ready to use: https://drive.google.com/file/d/1HdSxw_un9en7G14Kkm38d2ko0uSrSL0A/view?usp=sharing

Graphs

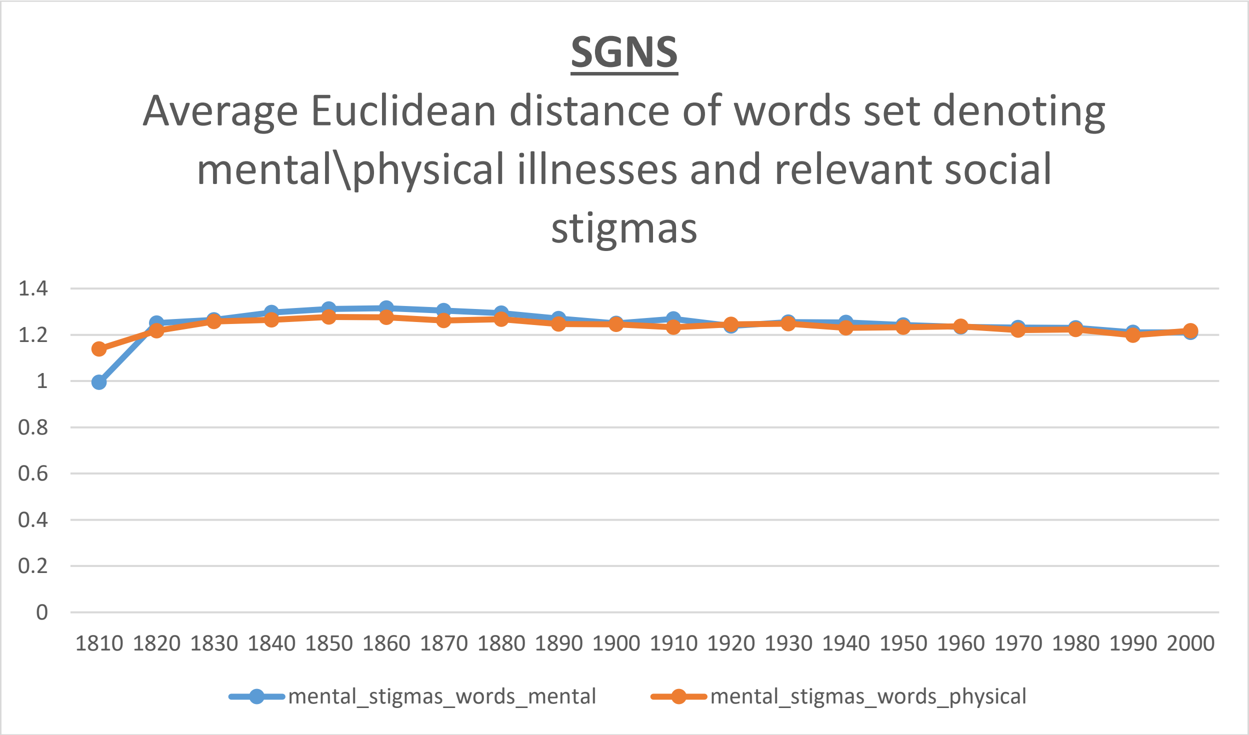

SVD
Average cosine similarity of words set denoting mental\physical illnesses and relevant social stigmas
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
mental_stigmas_words_mental
mental_stigmas_words_physical


SVD
Average Euclidean distance of words set denoting mental\physical illnesses and relevant social stigmas
1.4
1.39
1.38
1.37
1.36
1.35
1.34
1.33
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
mental_stigmas_words_mental
mental_stigmas_words_physical



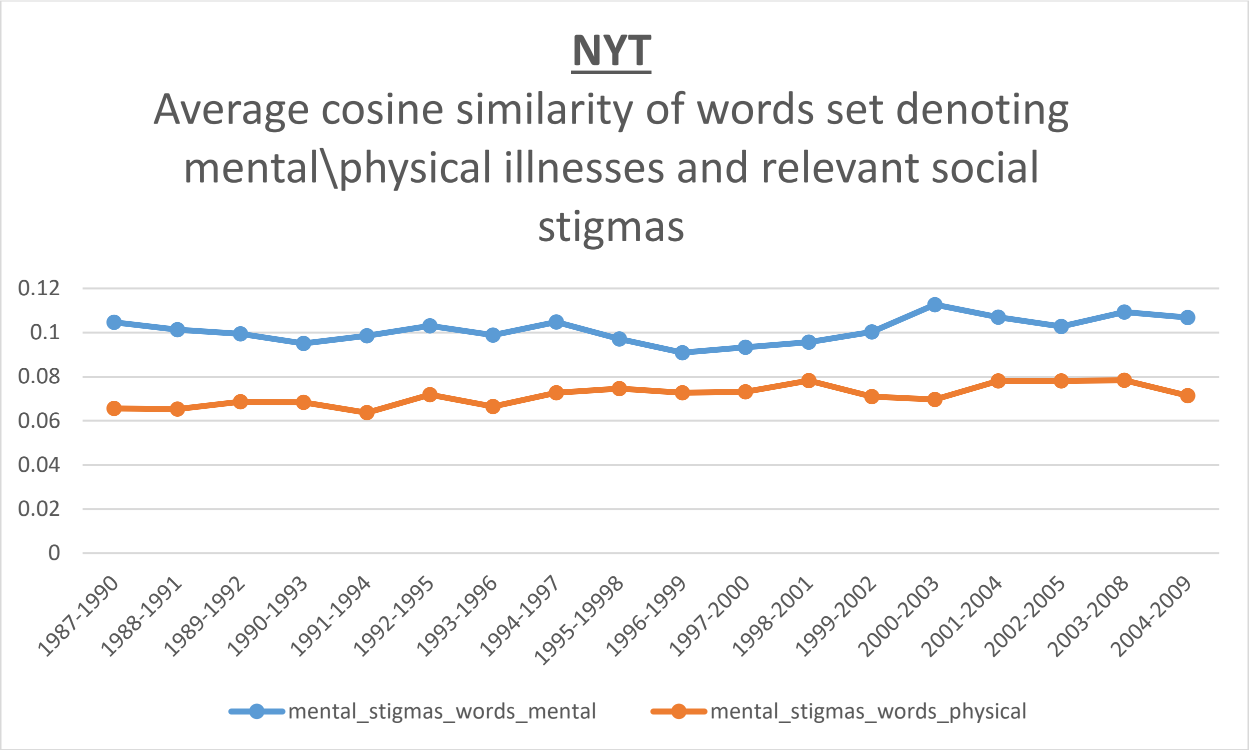
SGNS

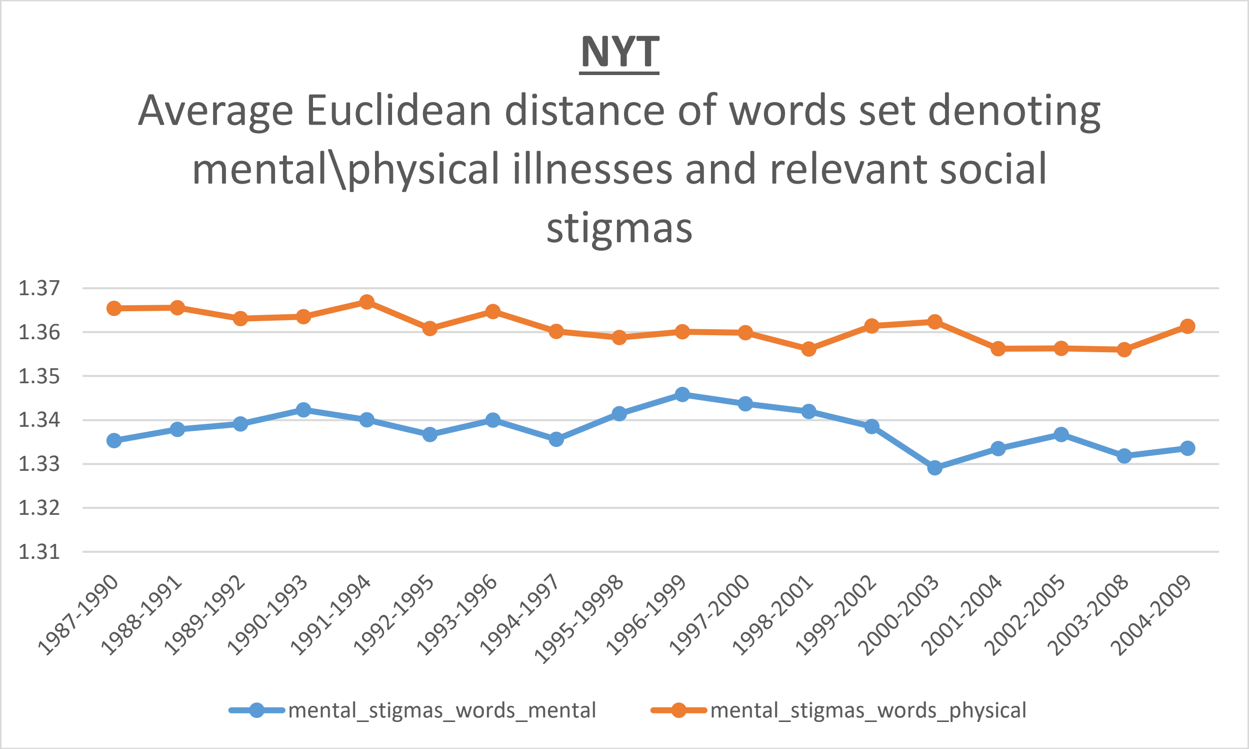
Average cosine similarity of words set denoting mental\physical illnesses and relevant social stigmas



SGNS

Average Euclidean distance of words set denoting mental\physical illnesses and relevant social stigmas


NYT
Average cosine similarity of words set denoting mental\physical illnesses and relevant social stigmas
0.12
0.1
0.08
0.06
0.04
0.02
0
1987-1990
1988-1991
1989-1992
1990-1993
1991-1994
1992-1995
1993-1996
1994-1997
1995-19998
1996-1999
1997-2000
1998-2001
1999-2002
2000-2003
2001-2004
2002-2005
2003-2008
2004-2009
mental_stigmas_words_mental
mental_stigmas_words_physical


NYT
Average Euclidean distance of words set denoting mental\physical illnesses and relevant social stigmas
1.37
1.36
1.35
1.34
1.33
1.32
1.31
1987-1990
1988-1991
1989-1992
1990-1993
1991-1994
1992-1995
1993-1996
1994-1997
1995-19998
1996-1999
1997-2000
1998-2001
1999-2002
2000-2003
2001-2004
2002-2005
2003-2008
2004-2009
mental_stigmas_words_mental
mental_stigmas_words_physical

Next steps

The work presented in this report should be seen as a torch that lights a general direction of research, the exact bounds, breadth and even goals of which are currently broad and it seems they will remain a work-in-progress for the foreseeable future.

Interesting possible future steps:

- Wording: The words to represent the illnesses as well as the stigmas are from modern common language. When analyzing words from decades \ centuries ago this point needs to be thought of very thoroughly, possibly with the help of experts. Perhaps even considering having different sets of words for different time spans. In case of different word sets for different times – a statistical approach should be taken to make sure that we're comparing "apples to apples and not oranges". E.g. if we have 2 sets for time periods A and B, then the overall temporal distance between sets A & B should be taken into consideration when calculating the distances between the stigmas and the illnesses words.
- Temporal vs. non-temporal: The data sets supplied by the original article I based my work on were composed of half temporal and half non-temporal. An interesting approach would be to try to think how we do measure social stigmas in non-temporal datasets. See next bullet point.
- A different approach to identify stigmas: Another method might be measuring what the most similar words to a set of words (e.g. associated with mental illnesses etc.) and seeing what the closest words between those are. A further direction can be taken that out of the closest words in the vocabulary – only specific words be taken (e.g words describing professions, feelings etc.).
- Consulting and collaborating with professionals from other fields:
    - Sociologists: People who research social trends (and specifically the underlying power struggles around social norms \ prejudice \ stigmas).
    - The medical community: The DSM[25] (and more recently the ICD[26]) are highly regard by the relevant professional community as the textbooks on the matters of mental medical conditions. The definitions and data there should be used to better understand and "fine-tune" what it is we're talking about when we talk about mental conditions.

[25] The ***Diagnostic and Statistical Manual of Mental Disorders*** **(DSM)** is the main publication diagnostic tool published by the American Psychiatric Association (APA). It lists all the known mental medical conditions

[26] The **International Classification of Diseases** (**ICD**) is a globally used diagnostic tool for epidemiology, health management and clinical purposes. The ICD is maintained by the World Health Organization

Epilogue

This is my first such project on this scale. I had a fun & interesting time working on it, although I must admit that wrapping my head around the original article, struggling to setup the environment, run the code & recreate their findings was somewhat daunting & challenging.

Nonetheless, I was successful and enjoyed it very much. I would like to thank the course staff Dr. Shai fine & especially Abraham Israeli for their patience and devotion. The course was jam-packed with information and I had a fun time trying to chase the train of knowledge.

This subject of utilizing word vectors to characterize common language and use that to show social trends, inclinations etc. is fascinating me and I will be happy to incorporate it into my future research.

A.M.