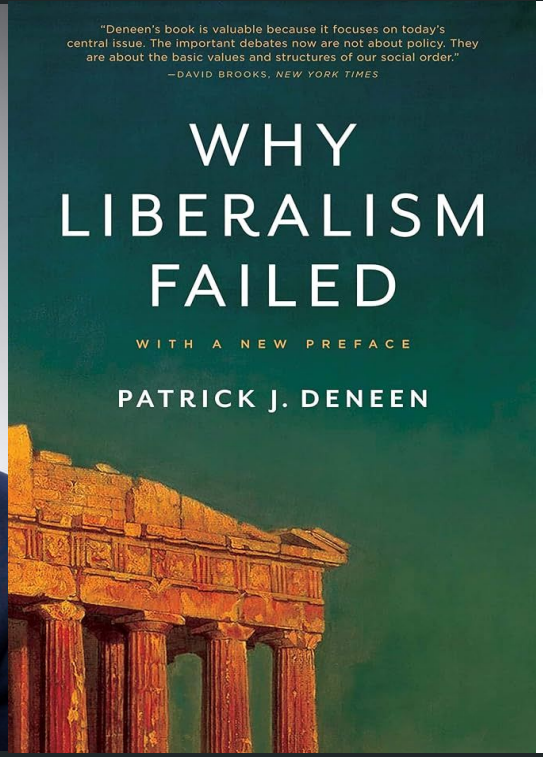
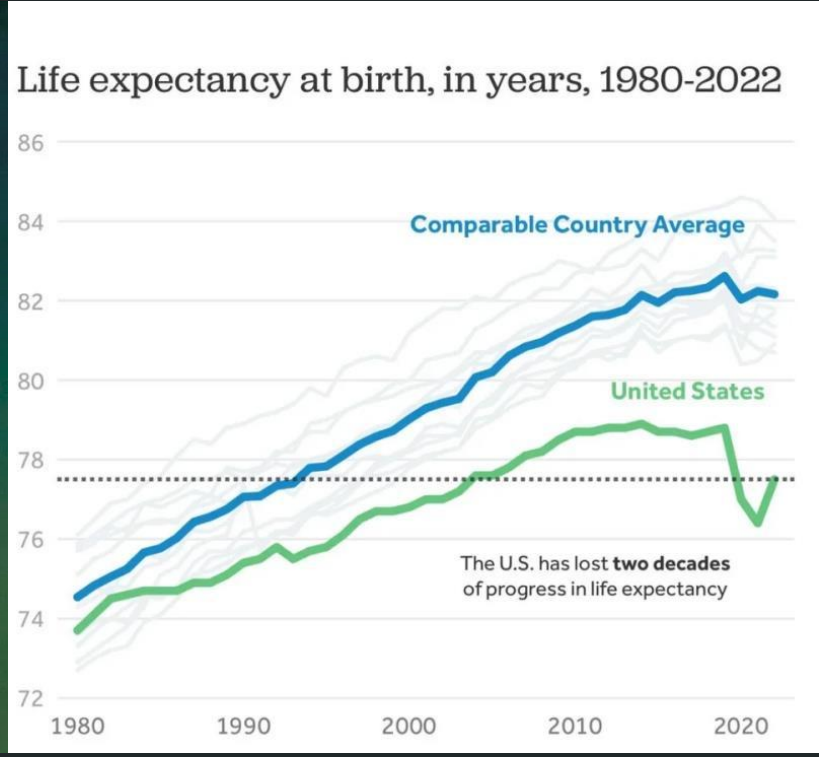
Final Project: Purpose & Meaning

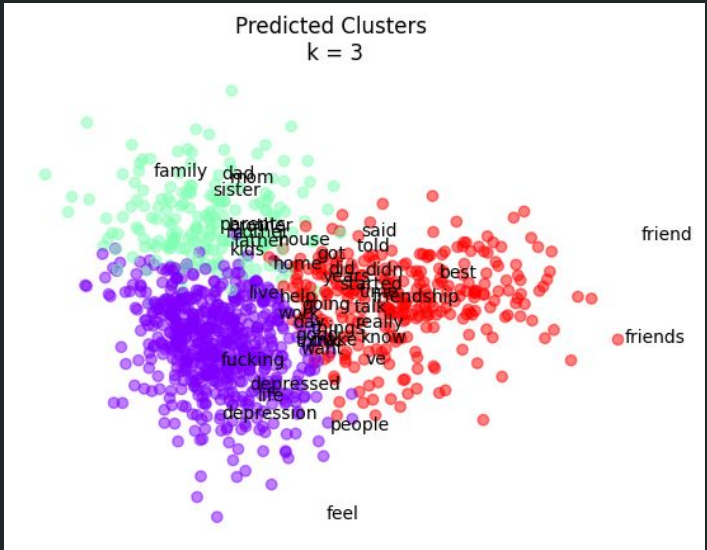
Dan Gilles

Inspiration



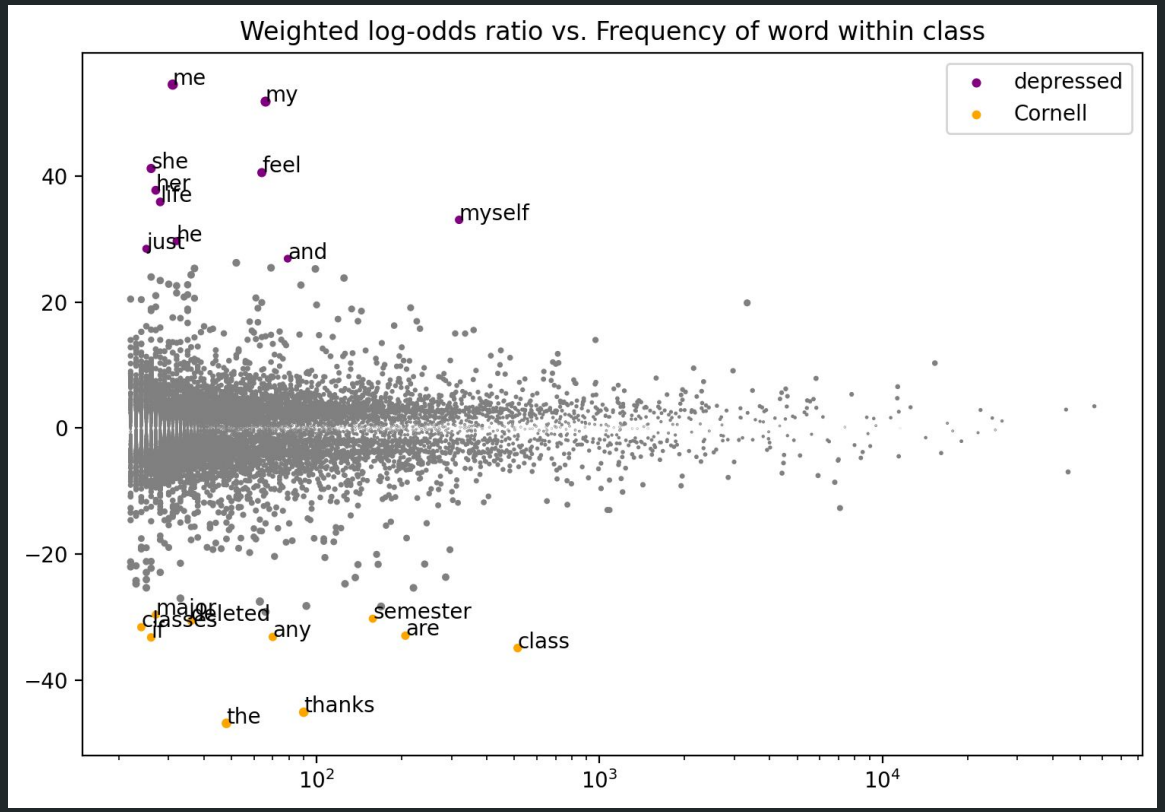


K-Means analysis




```
Top terms per cluster: (3 clusters)
['Cluster 0:', ' feel', ' life', ' want', ' depression', ' people', ' ve', ' know', '
think', ' really', ' time', ' depressed', ' things', ' help', ' day', ' fucking', '
make', ' work', ' good', ' going', ' years', ' happy', ' better', ' feeling', ' hate',
' love', ' wish', ' job', ' talk', ' makes', ' need', ' try', ' bad', ' right', '
sad', ' thoughts', ' having', ' anymore', ' shit', ' friends', ' year', ' today', '
tired', ' way', ' fuck', ' doing', ' school', ' long', ' getting', ' say', ' does',
'Cluster 1:', ' friend', ' friends', ' best', ' know', ' ve', ' feel', ' time', '
really', ' people', ' want', ' said', ' didn', ' told', ' friendship', ' talk', '
make', ' years', ' did', ' things', ' started', ' school', ' got', ' think', ' group',
' person', ' new', ' life', ' say', ' going', ' girl', ' guy', ' way', ' doesn', '
help', ' day', ' tell', ' hang', ' bad', ' talking', ' boyfriend', ' good', ' year', '
close', ' recently', ' months', ' sure', ' asked', ' thing', ' advice', ' need',
'Cluster 2:', ' family', ' dad', ' mom', ' sister', ' parents', ' mother', ' brother',
' father', ' house', ' time', ' want', ' home', ' know', ' years', ' kids', ' got', '
told', ' live', ' really', ' said', ' husband', ' help', ' feel', ' old', ' things', '
didn', ' year', ' wife', ' ve', ' say', ' cousin', ' doesn', ' away', ' money', '
going', ' think', ' went', ' daughter', ' does', ' child', ' work', ' day', ' ago', '
school', ' car', ' mum', ' right', ' died', ' asked', ' care']
```

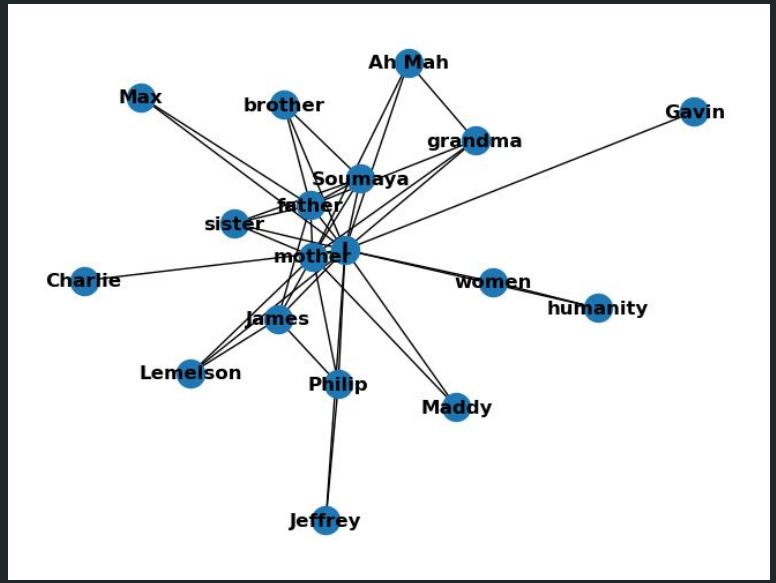
Word frequency differences

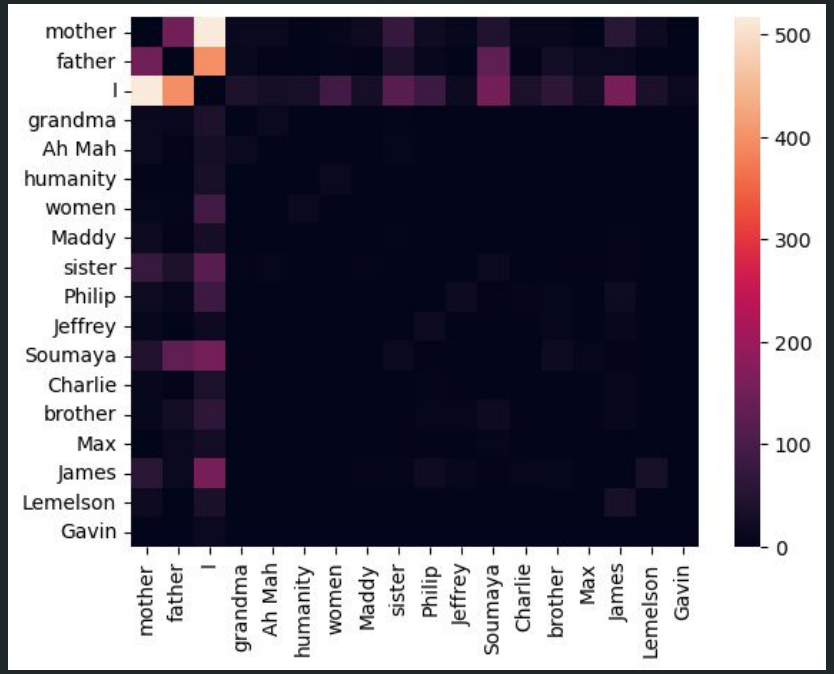


ngram	z-score	class
the	-46.813787	Cornell
thanks	-45.052157	Cornell
class	-34.871888	Cornell
if	-33.188343	Cornell
any	-33.114831	Cornell
...	...	...
her	37.721990	depressed
feel	40.518676	depressed
she	41.175328	depressed
my	51.782092	depressed
me	54.478234	depressed

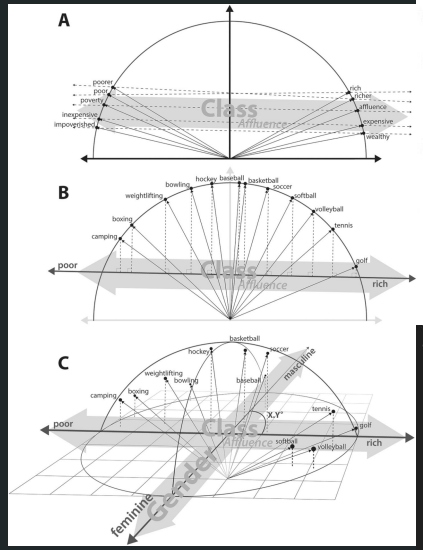
15000 rows × 2 columns

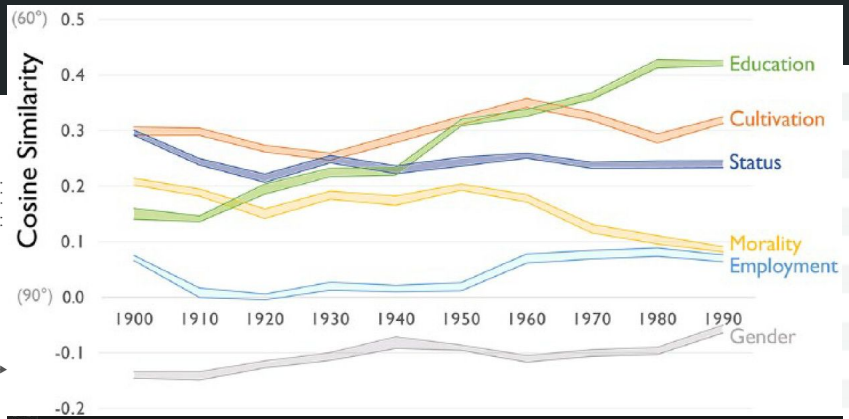
Network analysis



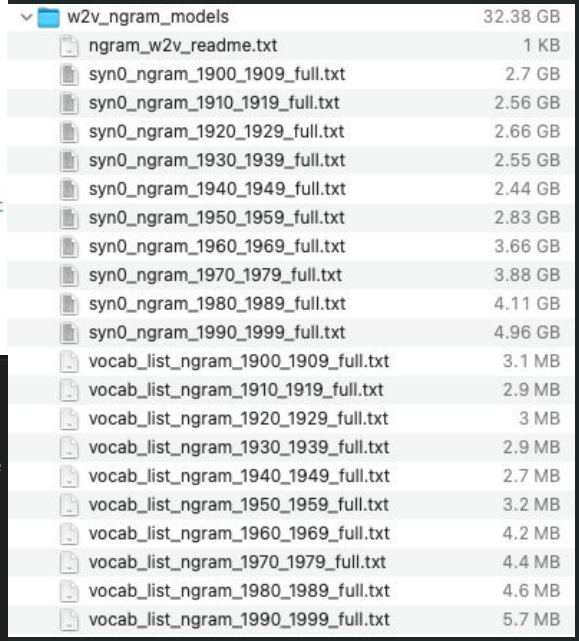


Word embeddings

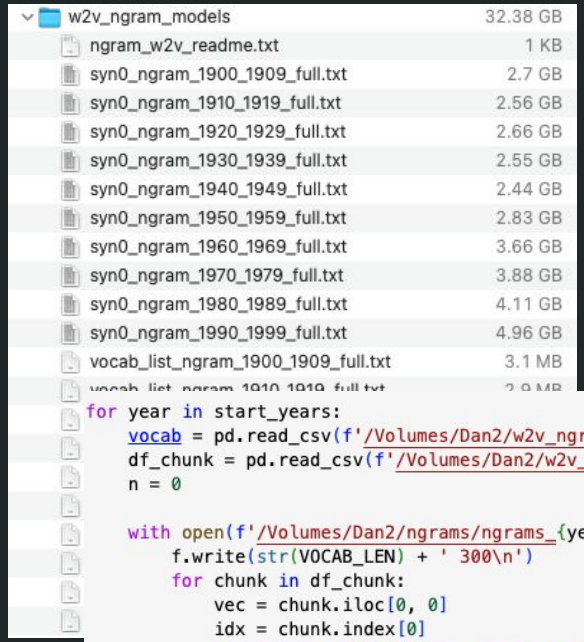





```
...  '-2.4909955263113781738e-01 1.5136285126209259O3e-01 1.6276098787784576 42e-02 -2.8991273045539855 96e-01 1.9946383312344551 09e-02
     1.0475208610296249 39e-01 -9.9358685314655306 96e-02 -2.2540719807147979 74e-01 3.0882361531257629 39e-01 3.8205713033676174 6e-02
     6.7393556237220764 16e-02 -3.0086010694503784 18e-01 -7.8139871358871459 96e-02 3.2221338152885437 01e-01 1.3174772262573242 19e-01
     1.6053904059299087 524e-01 2.6411458849906921 39e-01 -2.0567528903484344 8e-02 1.0612437129020690 92e-01 1.4588931202888488 77e-01
     3.0556660145521163 94e-02 -3.5023525357246398 93e-02 5.0224523991346359 25e-02 4.8054236918687820 43e-02 4.4832348823547363 28e-02
     1.5642687678337097 17e-01 -2.4818217754364013 67e-01 -9.4318687915802001 95e-02 1.2736117467284202 58e-02 2.6948615908622741 70e-01
     -8.3998404443264007 57e-02 -2.5799274444585007 812e-01 -5.3006120026111602 78e-02 -1.7246803641319274 90e-01 -2.3312166333198547 36e-02
     -8.2609206438064575 20e-02 -7.9707957804203033 45e-02 -3.9628051221370697 02e-02 -1.5646077692508697 51e-01 2.6458808779716491 70e-01
     -2.5030508637428283 69e-01 -2.2567684948444366 46e-01 7.9212367534637451 17e-02 1.2948718853294849 40e-02 -2.5091329216957092 29e-01
     -3.7487979978322982 79e-02 8.9779190719127655 03e-02 -2.4092243611812591 55e-01 -6.0012038797140121 46e-02 -2.0811735093595974 1e-01
     -2.2428885102272033 69e-01 -4.3676853179931640 62e-01 4.2190477252006537 6e-02 1.4810806512832641 60e-01 -8.9434310793876647 95e-02
     -1.9157210364937782 29e-02 -6.9691382348537445 07e-02 -2.1363325417041778 56e-01 6.7273736000061035 16e-02 6.3982144929468631 74e-03
     2.2337548434737434 48e-01 3.8089296221733093 26e-01 3.1136328354477882 39e-02 -1.1149107664823532 10e-01 3.5562223196029663 09e-01
     1.6730532050132751 46e-02 -2.2529672086238861 08e-01 -1.0983175039291381 84e-01 -4.0040757507085800 17e-02 -2.4377329647541046 14e-01
     3.6149360239505767 82e-02 2.8548163175582885 74e-01 -1.7289797961711883 54e-01 -1.5224343538284301 76e-01 -1.8060381710529327 39e-01
     1.3446143828332424 16e-02 -7.8940816223621368 41e-02 8.3853602409362792 97e-02 -1.5266552567481994 63e-01 6.3879333436489105 22e-02
     2.2094126045703888 794e-01 -2.3508111480623483 66e-03 2.7227789163589477 54e-01 6.3216082751750946 04e-02 -1.0829608142375946 04e-01
     3.6902809143066406 25e-01 7.1154028177261352 54e-02 -2.9226597398519515 99e-03 -2.5149264931678771 97e-01 1.4880640322385788 0e-02
     1.1370959877967834 47e-01 -3.5148447751998901 37e-01 3.6239951848983765 6e-01 1.3820353150367736 82e-01 5.9656158089637756 35e-02
     -1.0127353668121890 62e-01 1.3107146322727203 37e-01 -1.7281359061598777 7e-02 4.3267101049423217 77e-01 3.4095700830221176 15e-02
     -1.1618678458034992 22e-02 1.4928348362442583 30e-01 4.8456311225891113 28e-02 1.3688321411609649 66e-01 6.2696345150470733 64e-02
     7.0863433182239533 247e-02 6.2775090336799621 58e-02 -1.7507240176200866 70e-01 1.0225379467010490 85e-01 -3.5697105526924133 30e-01
     1.3890580739825963 97e-03 -2.1330550312995910 64e-01 -2.7272798120975494 38e-02 -1.6921156644821166 99e-01 -1.8563507497310638 43e-01
     -7.8381203114986413 68e-02 -5.2917748689651489 26e-02 1.1723355948925018 31e-01 1.8796463310718536 38e-01 7.8844815492630004 88e-02
     2.2312869131565093 99e-01 1.2926700711250305 18e-01 -1.8113265931606292 72e-01 -1.2808436155319213 87e-01 -2.8077873576083374 0e-01
     -1.9975920021533966 06e-01 9.2348530888557434 08e-02 -4.5345988124608099 53e-02 1.8689785152673721 31e-02 -2.6216223835945129 39e-01
```



Word embeddings



	0
0	the
1	of
2	,
3	to
4	.
...	...
334024	urticastrum

```
for year in start_years:
    vocab = pd.read_csv(f'/Volumes/Dan2/w2v_ngram_models/vocab_list_ngram_{year}_{year+9}_full.txt
    df_chunk = pd.read_csv(f'/Volumes/Dan2/w2v_ngram_models/syn0_ngram_{year}_{year+9}_full.txt',
    n = 0

    with open(f'/Volumes/Dan2/ngrams/ngrams_{year}_{year+9}_{VOCAB_LEN}.txt', 'w') as f:
        f.write(str(VOCAB_LEN) + ' 300\n')
        for chunk in df_chunk:
            vec = chunk.iloc[0, 0]
            idx = chunk.index[0]
            # vectors.loc[n] = [idx, vocab.iloc[idx][0], vec]
            f.write(f'{vocab.iloc[idx][0]} {vec}\n')
            if n % 1000 == 0:
                print(year, n)
            n += 1
            if n == VOCAB_LEN:
                break
```

```
for year in start_years:
    print(year, vars()[f'vecs_{year}'].
    most_similar(positive=['woman',
    'king'], negative=['man'], topn=1))
```

✓ 0.1s Python

```
1900 [('inside', 0.44141054153442383)]
1910 [('queen', 0.7039340138435364)]
1920 [('mouth', 0.47583141922950745)]
1930 [('hydrogen', 0.41790416836738586)]
1940 [('employed', 0.4347355365753174)]
1950 [('soon', 0.5123521685600281)]
1960 [('held', 0.44984015822410583)]
1970 [('growth', 0.5318295359611511)]
1980 [('wife', 0.42465054988861084)]
1990 [('pounds', 0.4013866186141968)]
```

```
for year in start_years:
    print(year, vars()[f'vecs_{year}'].
    most_similar(positive=['woman',
    'king'], negative=['man'], topn=1))
```

✓ 0.0s Python

```
1900 [('queen', 0.7076151371002197)]
1910 [('queen', 0.7039340138435364)]
1920 [('queen', 0.7088885307312012)]
1930 [('queen', 0.7042350769042969)]
1940 [('queen', 0.6755873560905457)]
1950 [('queen', 0.7101660370826721)]
1960 [('queen', 0.7316992878913879)]
1970 [('queen', 0.6998559236526489)]
1980 [('queen', 0.6730139255523682)]
1990 [('queen', 0.6461030840873718)]
```

Antonym pairs

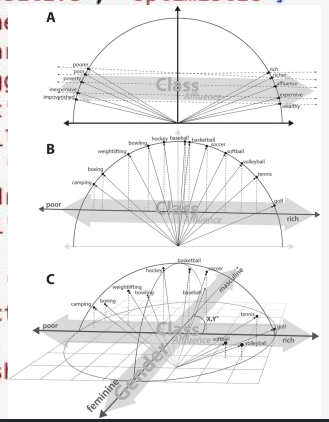
```
axes = ['purpose', 'career', 'happiness', 'satisfaction', 'anxiety', 'positivity', 'success', 'power', 'age', 'self', 'alone', 'wealth']
purpose_words_pos = ['purpose', 'significance', 'significant', 'meaning']#, 'worthwhile', 'meaningful']
purpose_words_neg = ['purposeless', 'insignificance', 'insignificant', 'meaningless']#, 'worthless', 'meaningless']
career_words_pos = ['employed', 'skilled', 'occupied', 'professional']
career_words_neg = ['unemployed', 'unskilled', 'unoccupied', 'idle']
happiness_words_pos = ['happy', 'joyful', 'glad']
happiness_words_neg = ['sad', 'sorrowful', 'grieving']
satisfaction_words_pos = ['satisfied', 'content', 'satisfaction']
satisfaction_words_neg = ['dissatisfied', 'discontent', 'dissatisfaction']
anxiety_words_pos = ['calm', 'unafraid', 'fearless']
anxiety_words_neg = ['anxious', 'afraid', 'fearful']
positivity_words_pos = ['hopeful', 'positive', 'optimistic']
positivity_words_neg = ['hopeless', 'negative', 'pessimistic']
success_words_pos = ['successful', 'thriving']
success_words_neg = ['failing', 'struggling']
power_words_pos = ['powerful', 'robust', 'strong']
power_words_neg = ['powerless', 'fragile', 'weak']
age_words_pos = ['young', 'youthful', 'energetic']
age_words_neg = ['old', 'decrepit', 'drained']
self_words_pos = ['self', 'myself', 'i', 'me', 'mine']
self_words_neg = ['family', 'friends', 'us', 'we', 'our']
alone_words_pos = ['alone', 'lonely', 'isolated']
alone_words_neg = ['together', 'connected', 'united']
wealth_words_pos = ['wealthy', 'rich', 'affluent']
wealth_words_neg = ['poor', 'impoverished', 'destitute']
```

Antonym pairs

```
axes = ['purpose', 'career', 'happiness', 'satisfaction', 'anxiety', 'positivity', 'success', 'power', 'age', 'self', 'alone', 'wealth']
purpose_words_pos = ['purpose', 'significance', 'significant', 'meaning']#, 'worthwhile', 'meaningful']
purpose_words_neg = ['purposeless', 'insignificance', 'insignificant', 'meaningless']#, 'worthless', 'meaningless']
career_words_pos = ['employed', 'skilled', 'occupied', 'professional']
career_words_neg = ['unemployed', 'unskilled', 'unoccupied', 'idle']
happiness_words_pos = ['happy', 'joyful', 'glad']
happiness_words_neg = ['sad', 'sorrowful', 'grieving']
satisfaction_words_pos = ['satisfied', 'content', 'satisfaction']
satisfaction_words_neg = ['dissatisfied', 'discontent', 'dissatisfaction']
anxiety_words_pos = ['calm', 'unafraid', 'fearless']
anxiety_words_neg = ['anxious', 'afraid', 'fearful']
positivity_words_pos = ['hopeful', 'positive', 'optimistic']
positivity_words_neg = ['hopeless', 'negative', 'pessimistic']
success_words_pos = ['successful', 'thriving']
success_words_neg = ['failing', 'struggling']
power_words_pos = ['powerful', 'robust', 'strong']
power_words_neg = ['powerless', 'fragile', 'weak']
age_words_pos = ['young', 'youthful', 'energetic']
age_words_neg = ['old', 'decrepit', 'drained']
self_words_pos = ['self', 'myself', 'i', 'me', 'mine']
self_words_neg = ['family', 'friends', 'us', 'we', 'our']
alone_words_pos = ['alone', 'lonely', 'isolated']
alone_words_neg = ['together', 'connected', 'united']
wealth_words_pos = ['wealthy', 'rich', 'affluent']
wealth_words_neg = ['poor', 'impoverished', 'destitute']
```

Axis vectors

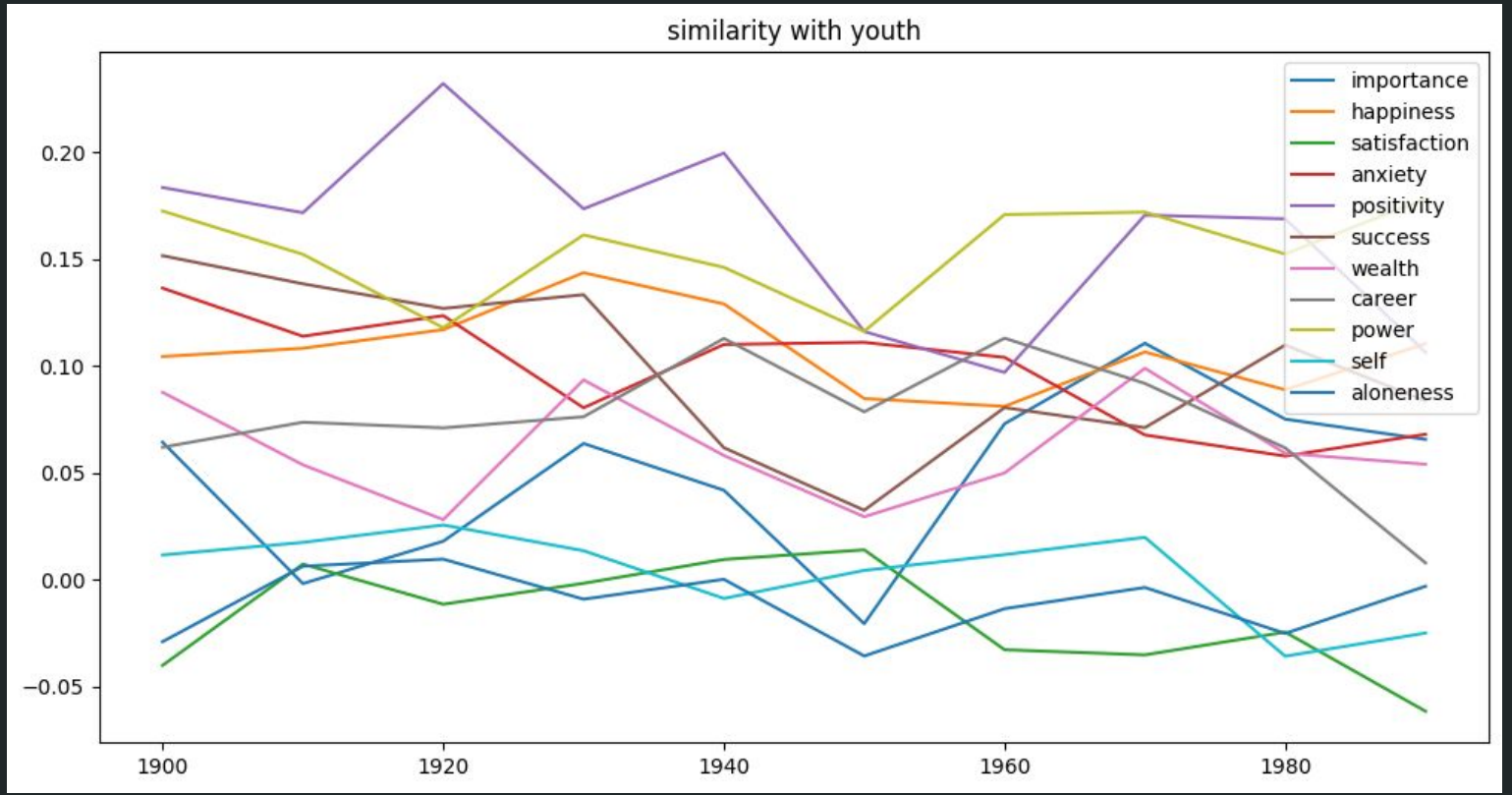
```
axes = ['purpose', 'career', 'happiness', 'satisfaction', 'anxie
purpose_words_pos = ['purpose', 'significance', 'significant', '
purpose_words_neg = ['purposeless', 'insignificance', 'insignifi
career_words_pos = ['employed', 'skilled', 'occupied', 'professi
career_words_neg = ['unemployed', 'unskilled', 'unoccupied', 'ic
happiness_words_pos = ['happy', 'joyful', 'glad']
happiness_words_neg = ['sad', 'sorrowful', 'grieving']
satisfaction_words_pos = ['satisfied', 'content', 'satisfaction'
satisfaction_words_neg = ['dissatisfied', 'discontent', 'dissati
anxiety_words_pos = ['calm', 'unafraid', 'fearless']
anxiety_words_neg = ['anxious', 'afraid', 'fearful']
positivity_words_pos = ['hopeful', 'positive', 'optimistic']
positivity_words_neg = ['hopeless', 'ne
success_words_pos = ['successful', 'thr
success_words_neg = ['failing', 'strugg
power_words_pos = ['powerful', 'robust
power_words_neg = ['powerless', 'fragi
age_words_pos = ['young', 'youthful',
age_words_neg = ['old', 'decrepit', 'd
self_words_pos = ['self', 'myself', 'i
self_words_neg = ['family', 'friends',
alone_words_pos = ['alone', 'lonely',
alone_words_neg = ['together', 'connec
wealth_words_pos = ['wealthy', 'rich',
wealth_words_neg = ['poor', 'impoveris
```



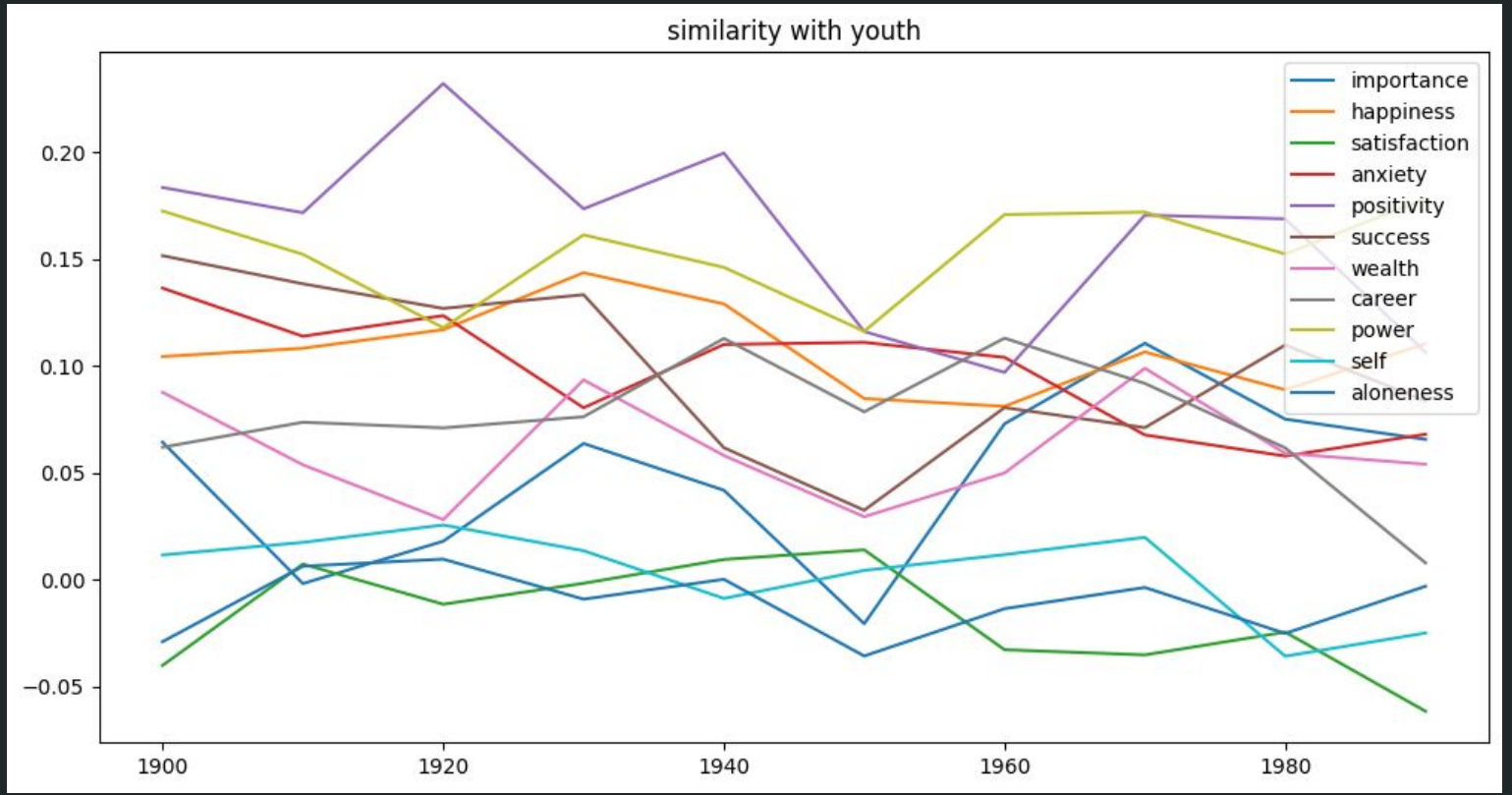
	vecs	purpose_axis	career_axis	happiness_axis	satisfaction_axis	a
1900	[[-0.15527616, -0.044475354, 0.24905019, 0.069...	[0.27155071, -0.05503708, -0.11242683, 0.47478...	[-0.046234615, -0.25147384, 0.20027569, -0.060...	[0.26558653, -0.0071866014, -0.042752195, -0.1...	[-0.15267624, -0.29770717, -0.28999034, -0.073...	]
1910	[[0.13818951, 0.08336567, 0.43356463, 0.008813...	[-0.09118675, 0.10703693, 0.21998593, 0.029964...	[0.18839166, -0.13435283, -0.15498304, 0.25415...	[-0.04038347, -0.29692504, -0.12958808, 0.2171...	[-0.5792359, -0.42665628, -0.16115755, -0.3567...	
1920	[[0.08622483, 0.059710536, 0.03332708, 0.30869...	[-0.5595601, 0.016212419, 0.11945161, -0.00287...	[0.051258057, 0.0387393, 0.17393678, 0.1084501...	[-0.0063034794, 0.15272544, -0.36101255, -0.08...	[-0.15660228, 0.040637117, 0.13808829, 0.17992...	
1930	[[0.18046615, 0.025464488, 0.057560947, -0.195...	[-0.23274247, 0.0610854, -0.4943271, -0.062550...	[0.10255523, -0.011350911, 0.009990338, 0.1768...	[0.06374652, 0.051403016, -0.2491311, 0.150112...	[0.18366778, 0.115343146, 0.2480477, -0.079701...	
1940	[[-0.24909955, 0.15136285, 0.016276099, -0.289...	[			[-0.087414175, -0.0725108, 0.29694104, 0.75403...	
1950	[[0.04905158, 0.21111171, 0.059865676, -0.3999...	[			[0.5251739, -0.11746201, -0.047229905, -0.6627...	
1960	[[-0.10262189, -0.20712672, -0.25004148, -0.36...	[			[-0.03319764, -0.19160075, 0.27791998, -0.1405...	
		0.35192...	-0.25...	-0.5432...		

$$\frac{\sum_{p}^{|P|} \overrightarrow{p_1} - \overrightarrow{p_2}}{|P|}$$

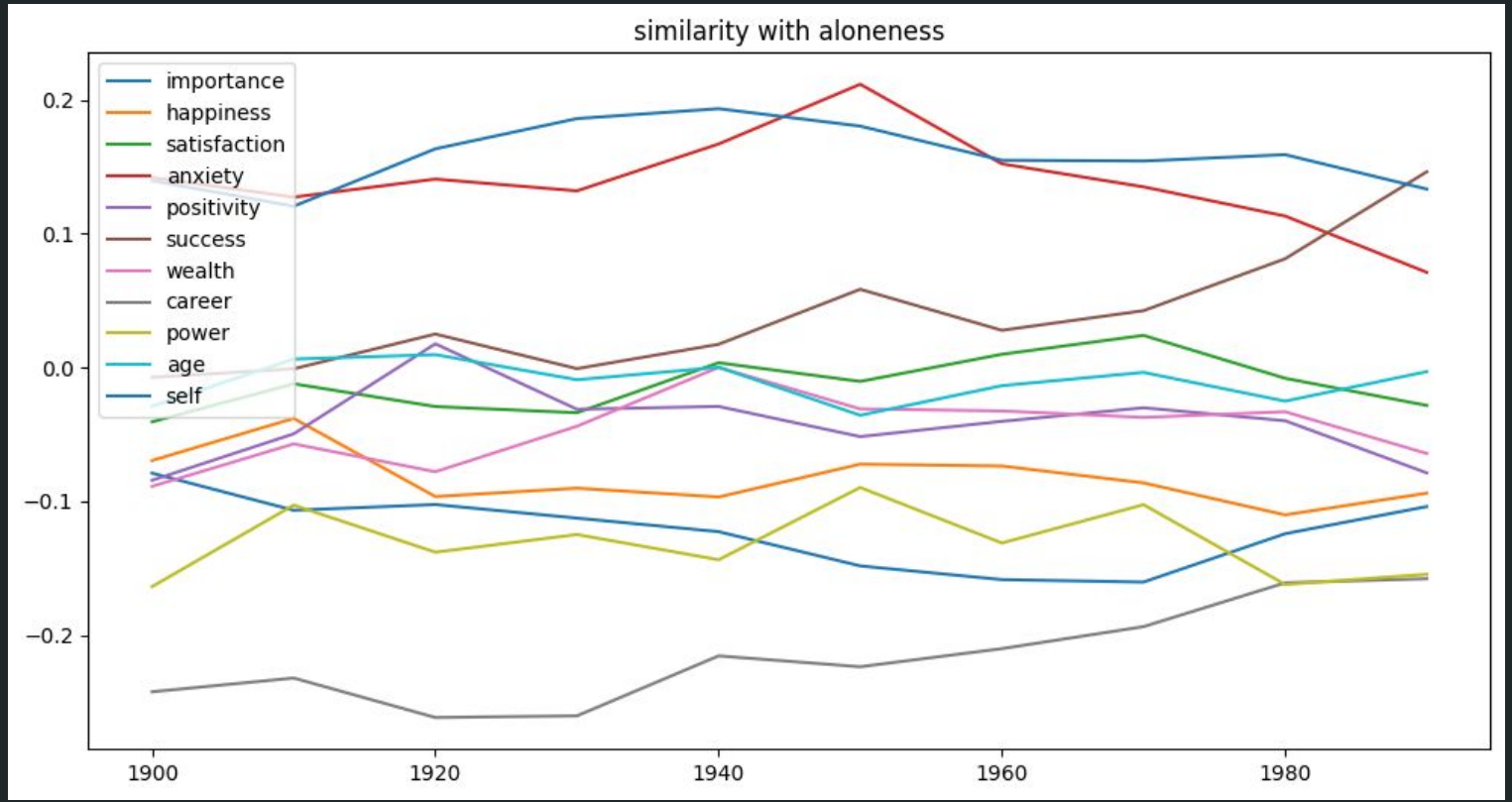
Axis similarity: youth



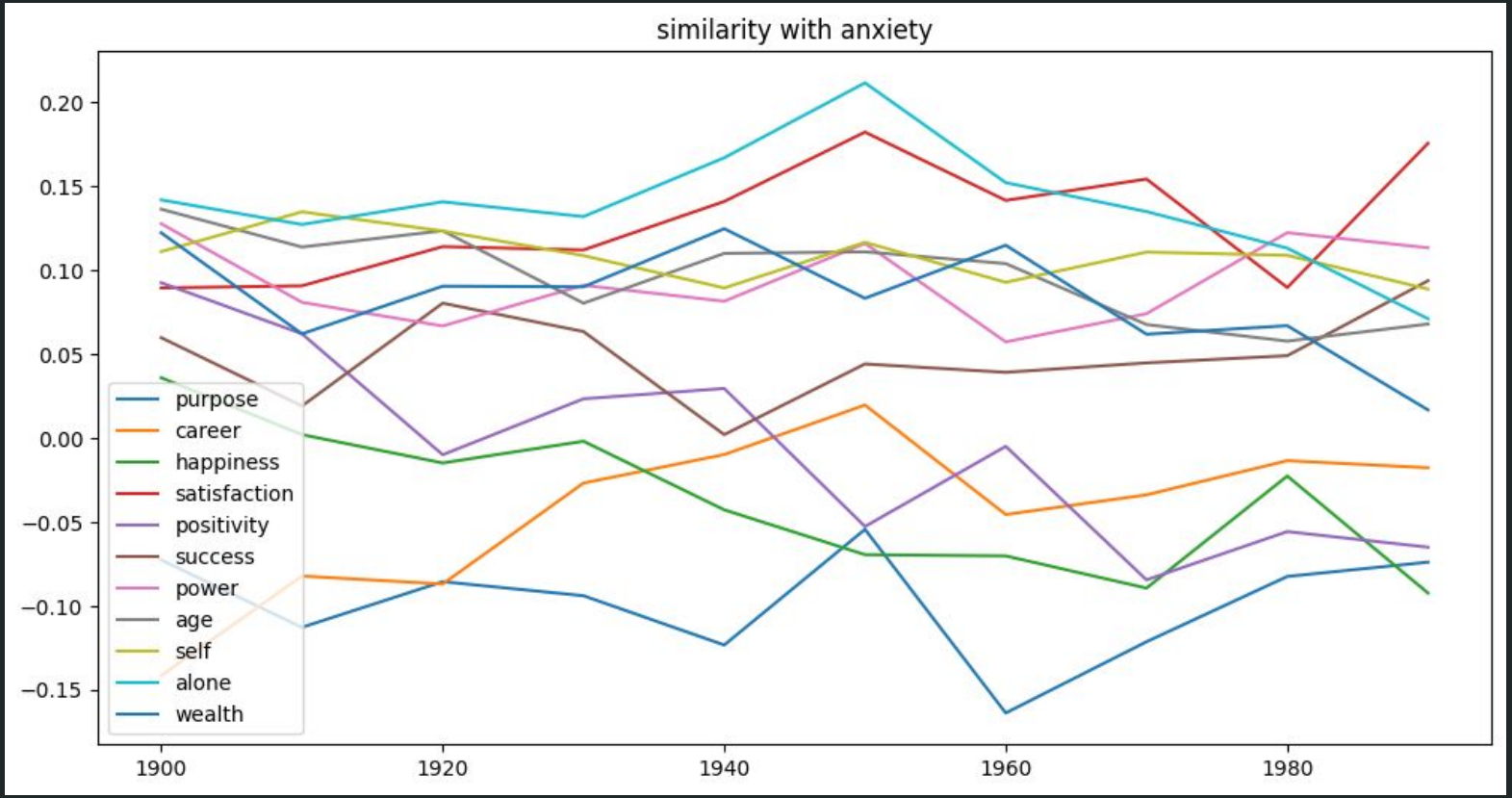
Axis similarity: youth



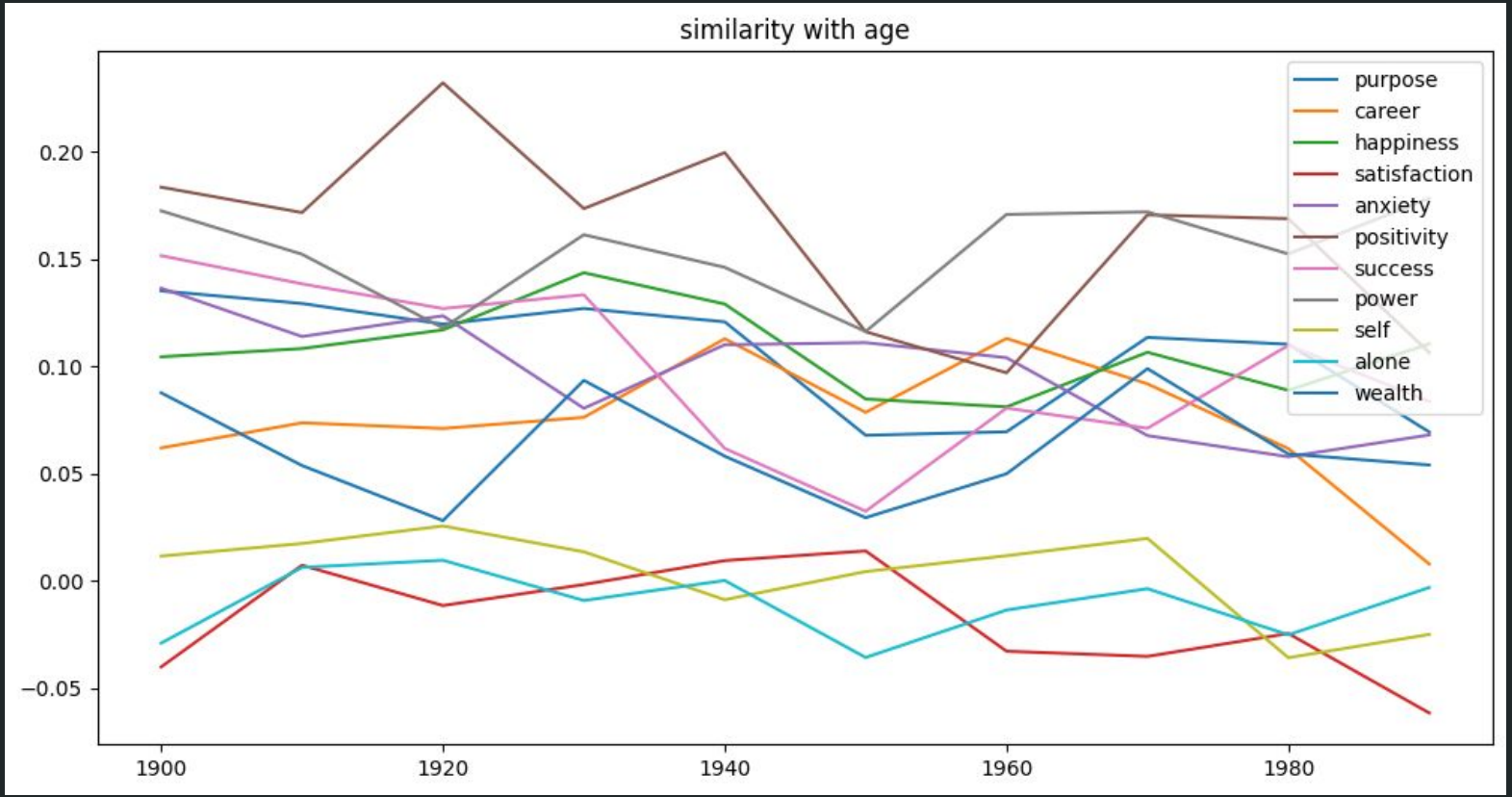
Axis similarity: aloneness



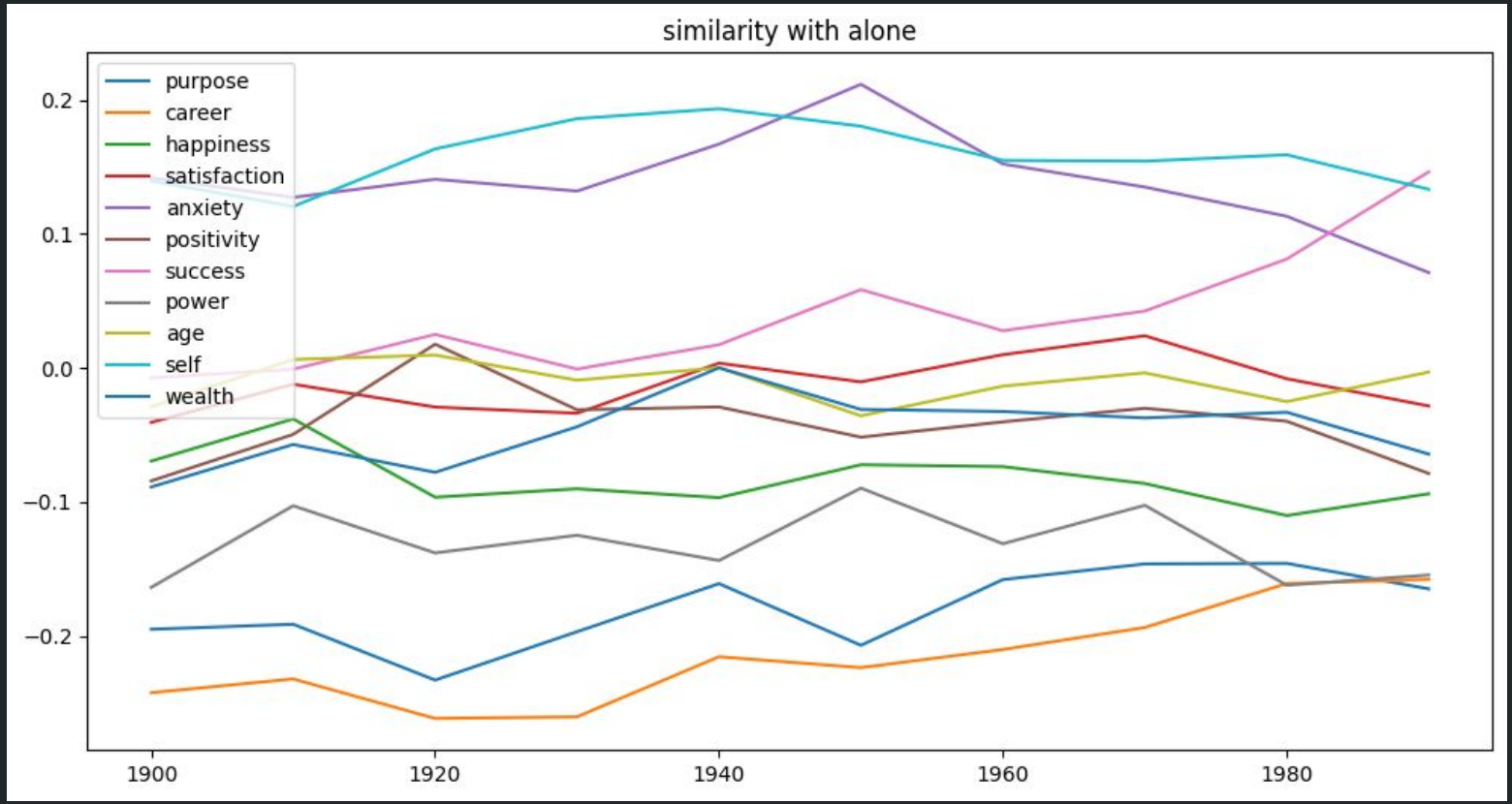
Axis similarity: anxiety



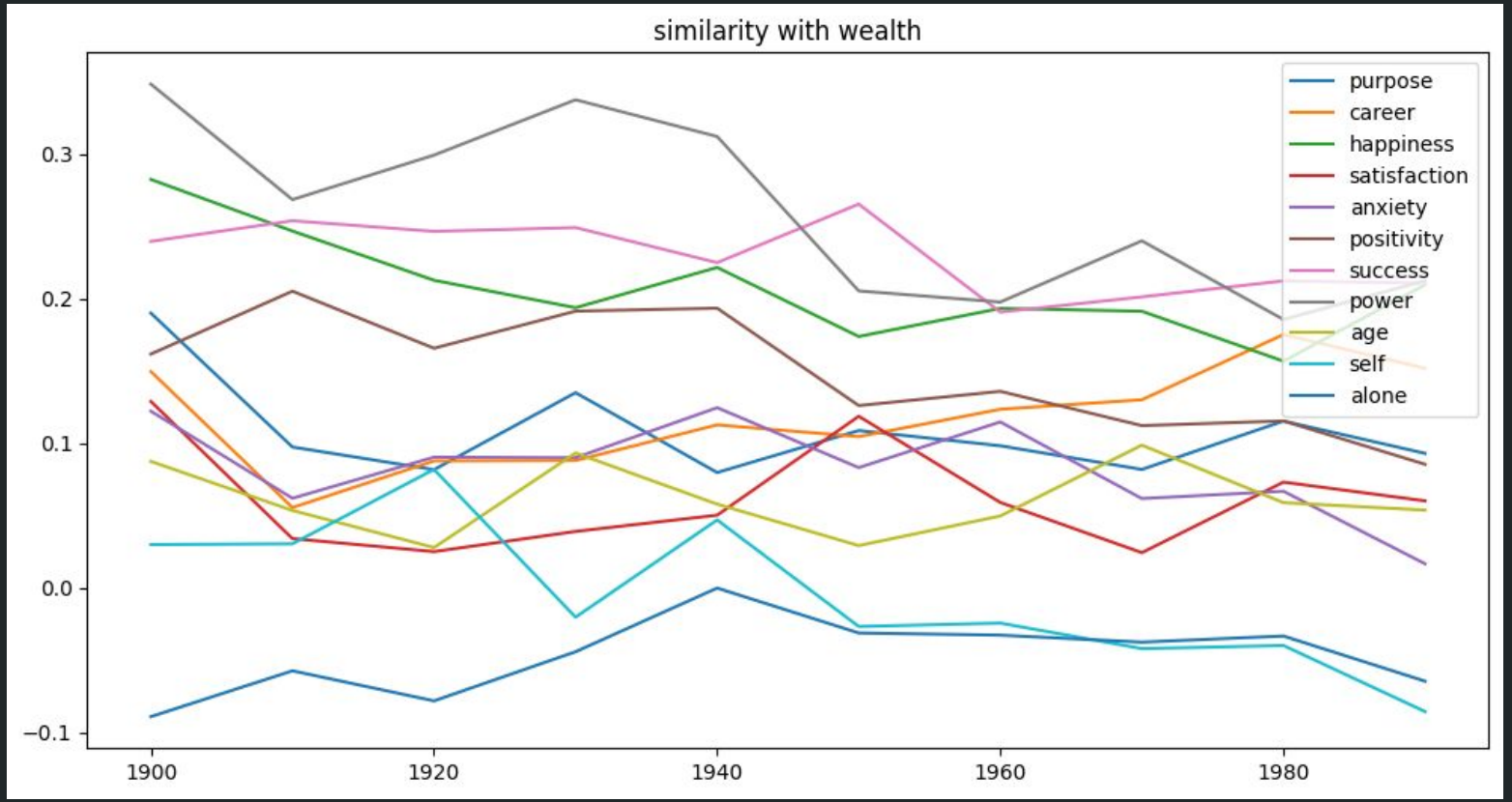
Axis similarity: age



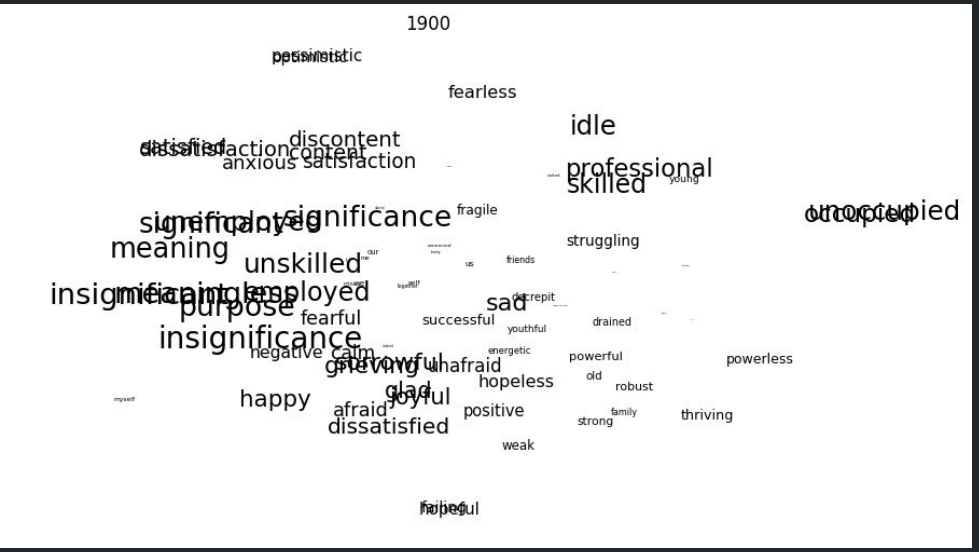
Axis similarity: aloneness



Axis similarity: wealth



Linguistic change



Next steps

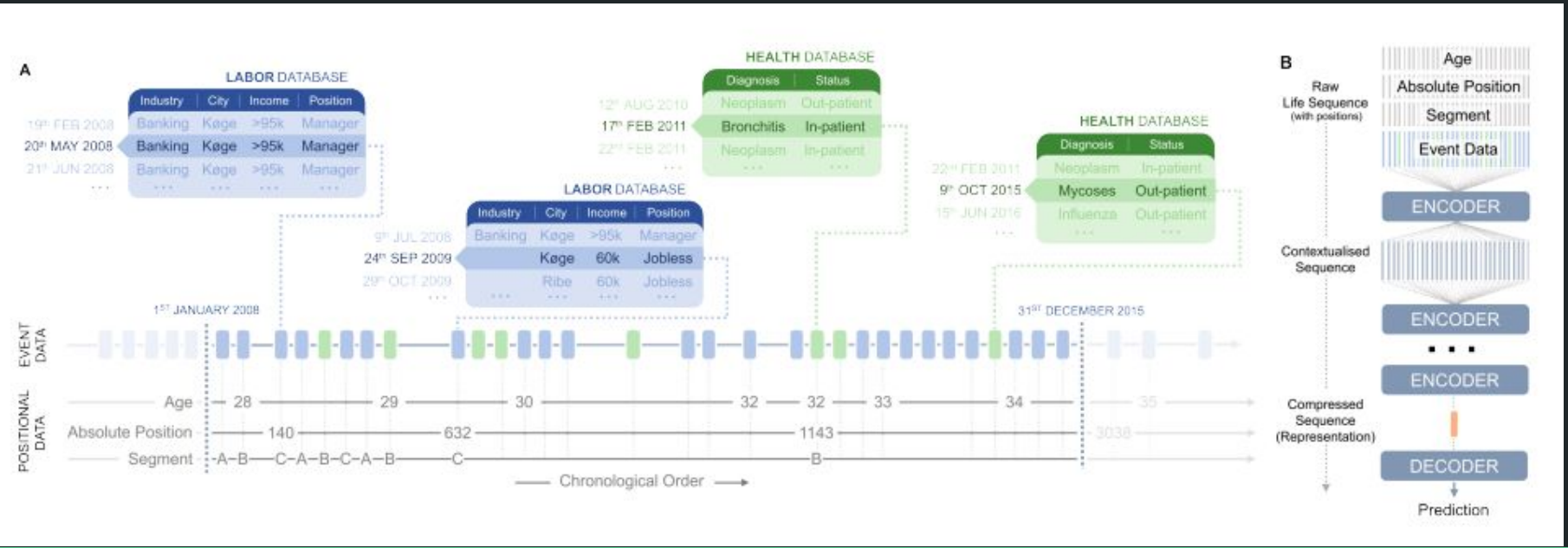
-Historical thesaurus

-Word projections

-Doc2Vec

-Discourse atoms

-Life2Vec



Next steps

-Historical thesaurus

-Word projections

-Doc2Vec

-Discourse atoms

-Life2Vec