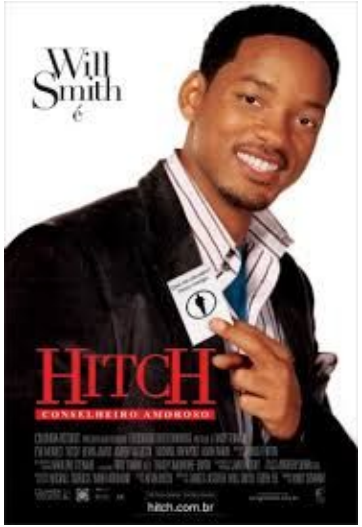


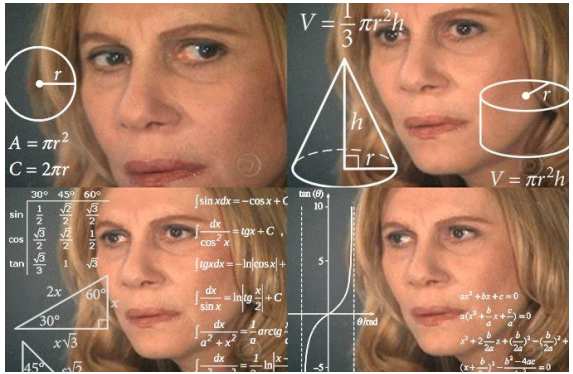
Memorability for Men in Online Dating

Type Right = Swiped Right

Where We Left Off

better define the aims of the project.

gower's.



split the analysis

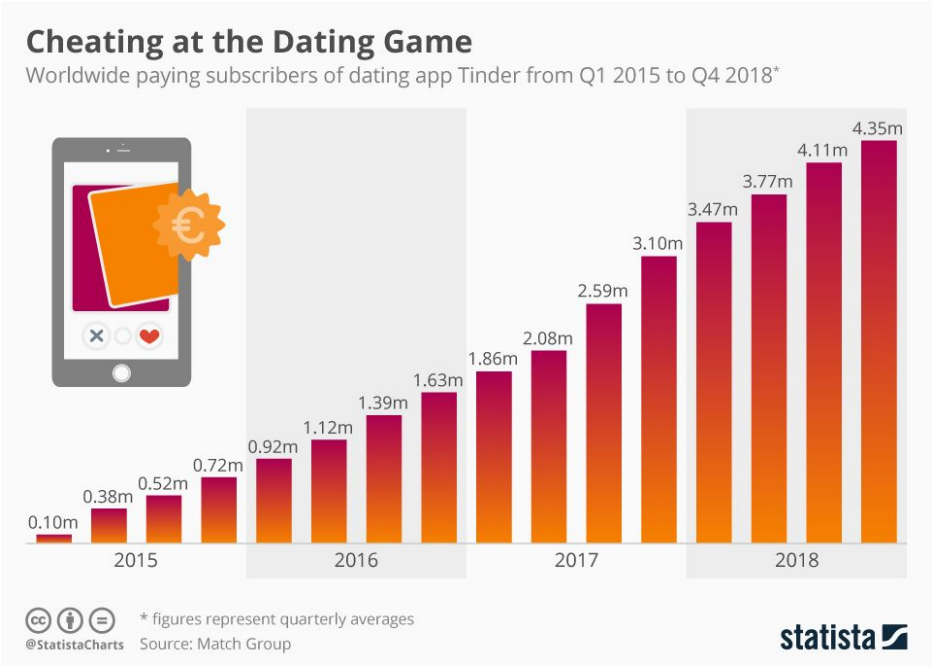
Diagnose the clusterabillity

initial variable selection

It's Raining Men

Percentage Male Users:

52.4%

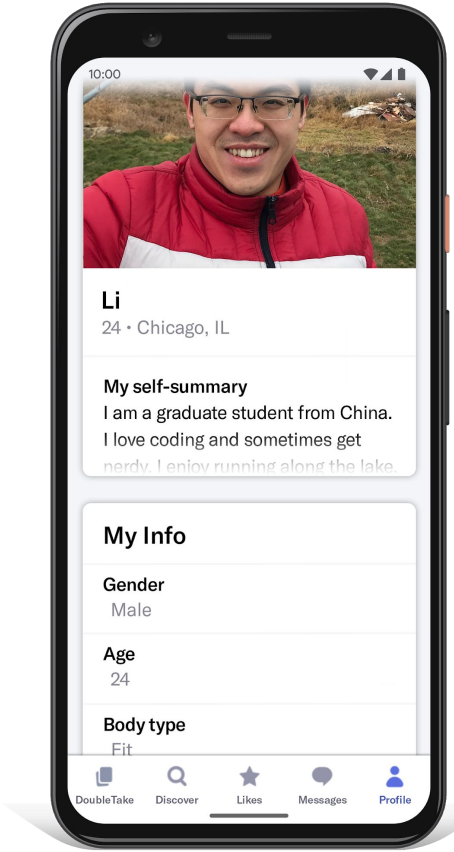


Relevant Research

- Emory (2017)

What's in a profile?

- Text data (short essays)
- Demographic data
- Photos
- Chat Windows
- Matches
- All data is user-submitted
- We focus on approximately 18,000 profiles of men in the San Francisco area from 2012



Relevant Research

- Emory (2017)
- Fiore et al (2010)

The 2 Key Questions:

What are other men like me writing in their profiles?

- Finding the ‘closest’ competitors- demographics or new metrics
- Most Common Topics
- Most Common Words

What can I do to make my essay be memorable and stand out?

- Diverge from your Demographic or Cluster’s Topic Preferences/ Proportions
- Use unusual words
- Humour
- Focused Topics

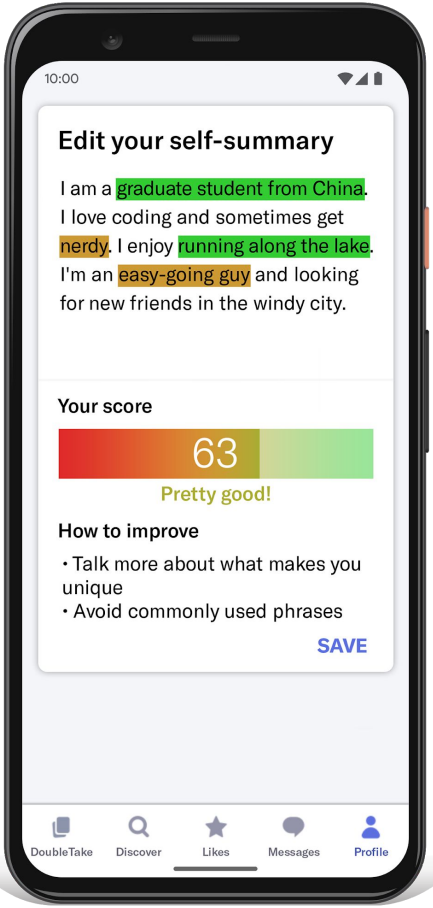
Official Tip for Beginners by OkCupid:

Be authentic. Don't be afraid to share things about yourself that are quirky, slightly embarrassing, or totally unique to you: they'll make great conversation starters!

Source: https://help.okcupid.com/article/15-your-profile

Future Applications

- Predict “memorability score” (rescaled to 0-100 from the number of ‘likes’)
- Cosine Similarity on Vector Space Models
- Topics Network Relationship and Proportions to recommend content for users



Demographic Variables

Categorical Variable Distributions

References:

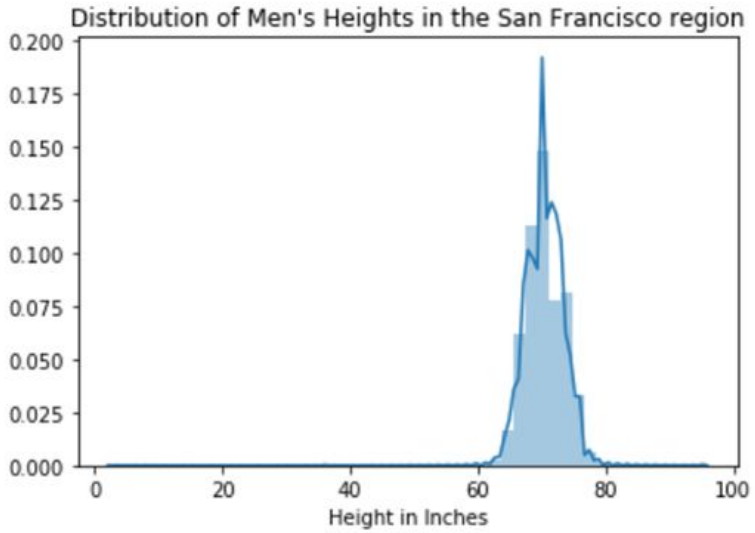
Education: (Torkey, 2018), (Fiore et al, 2015), (Stevens & Schaefer, 1990)

Height: Shepperd & Strattman, 1989

Race: Lindqvist and Lin, 2010

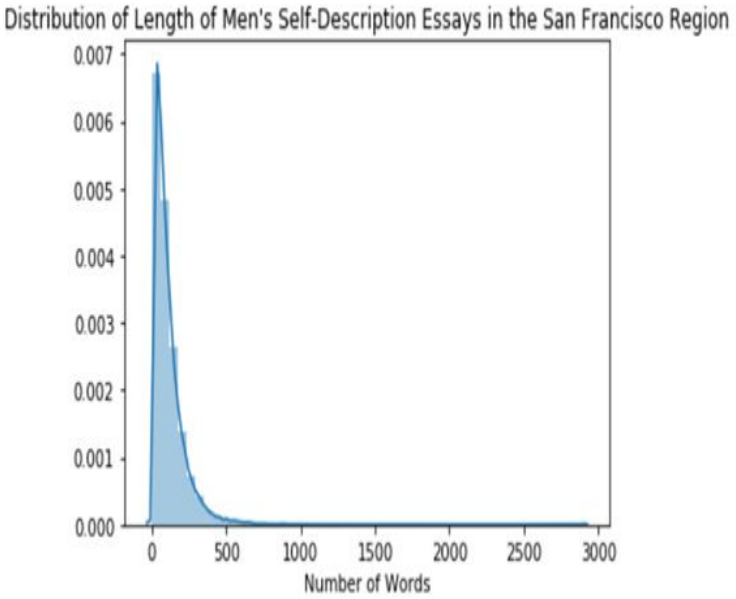
Fitness Level: Burke, 2019

Demographics (Continued)



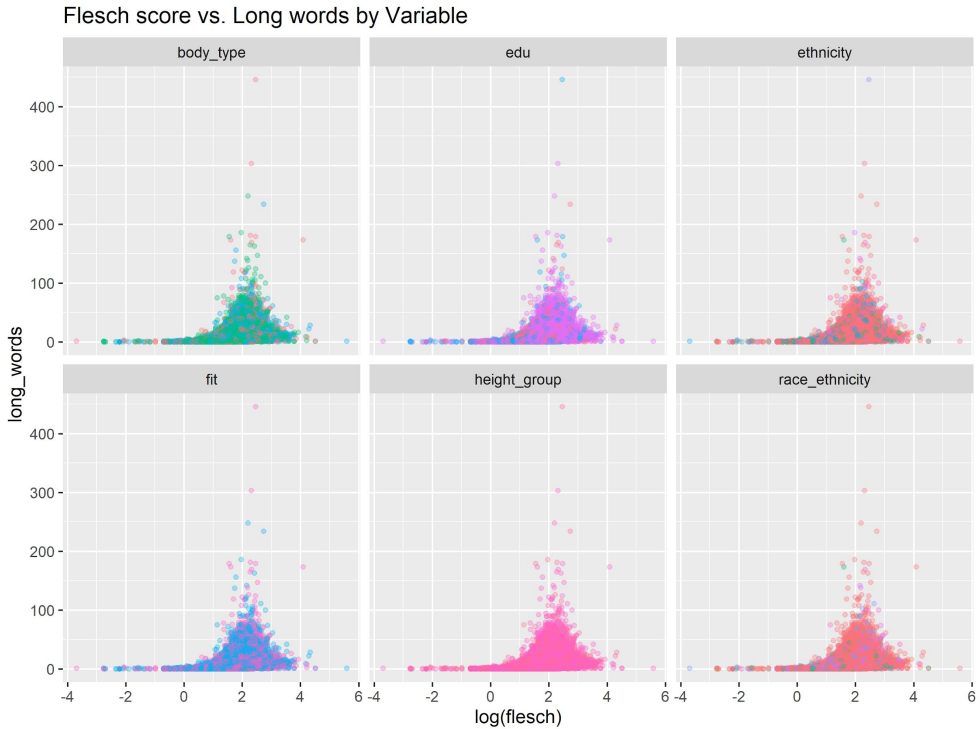
Stat	Height (in inches)
Mean	70.5
Std	3
Min	3
25%	69
50%	70
75%	72
Max	95

Text-Level Variables



User Type	Profile Length (in words)	Flesch Kincaid Reading Level	Proportion of Long Words
General Male User	80	6.73	0.20
Users with more than a High School Diploma	85	7.1	0.21
Users with a High School Diploma or less	66	5.63	0.18

Interaction of Demographics and Text



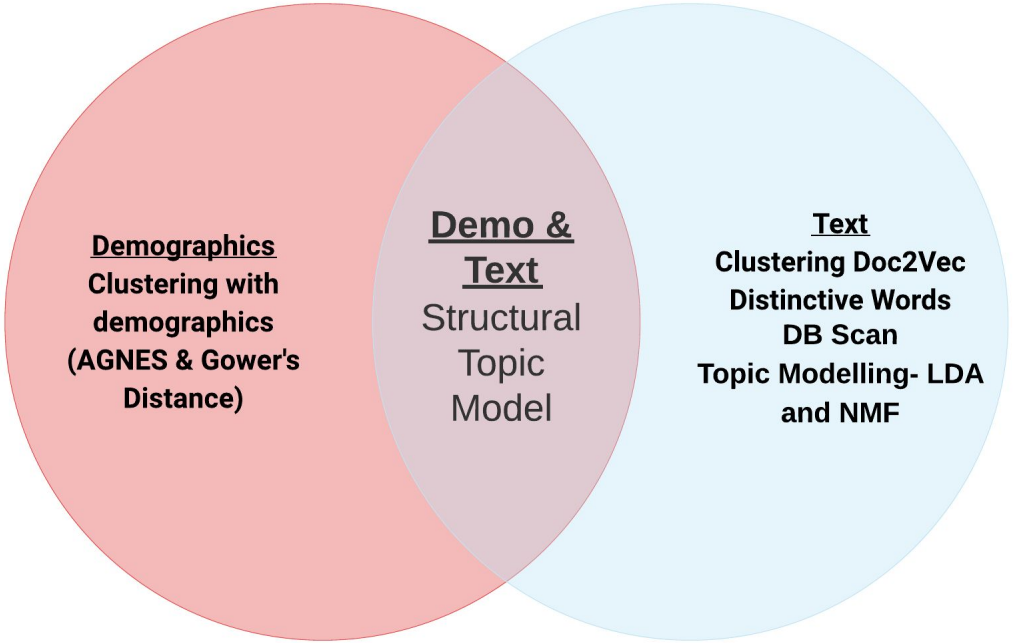
- No clear relationship between categories of demographic variables and basic quantitative measures of profile text

Core Hypotheses and Assumptions

- Demographics CAUSE text patterns
- Norm Exists | Demographics
- Deviations from norm -> More memorable -> More matches

Reference: Parmentier (1994)

Exploring Unsupervised ML Methods



Aspects of Variables

Demographics

- Education
- Race-ethnicity
- Height
- Fitness Level

Text

- Difficulty Level- Sentence Complexity
- Length
- Proportion of Difficult Words (multi-syllable)
- Topics and Topic Distribution
- Humour
- Low Frequency Words Used

Common Vs Uncommon Words

Affecting Ramsay crunches affiliated

Raspberry plantwhisperer FridaKahlo

Ramennoodle conjugate

Oooh Frida Kahlo!? That sounds exciting. I'll check him out

The string 'Frida Kahlo' is an outlier. Does not affect the model. Delete. Stay single forever hahaha

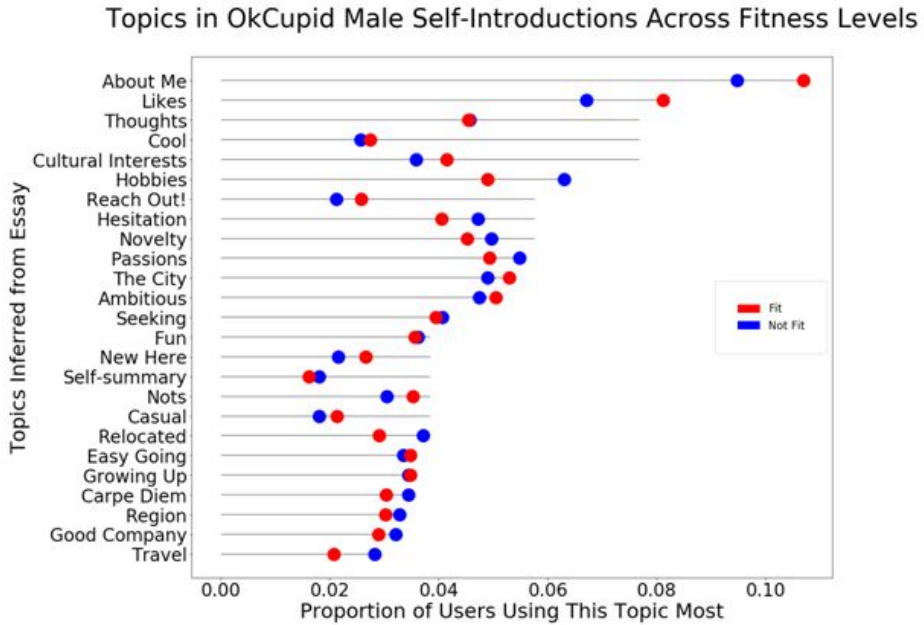
Choice of Number of Topics- I

The Old Model

Shishido et al (2016) used *Negative Matrix Factorization*

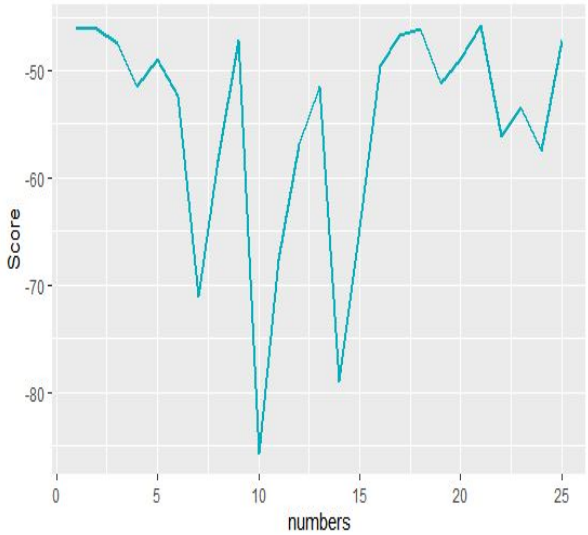
Pros- Clearer Topic Separation for short texts

Cons- Not amenable to Structural Topic Modelling



Choice of Number of Topics II

LDA Model



Choose the topics based on semantic coherence scores

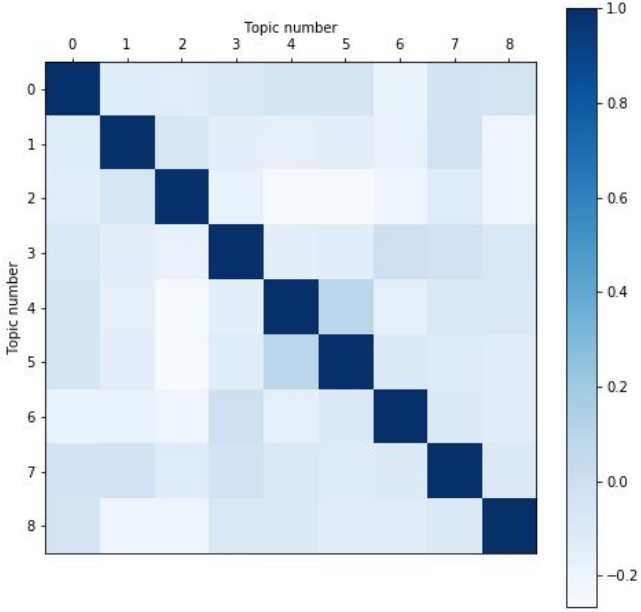
Perplexity - model's "surprise" at the data

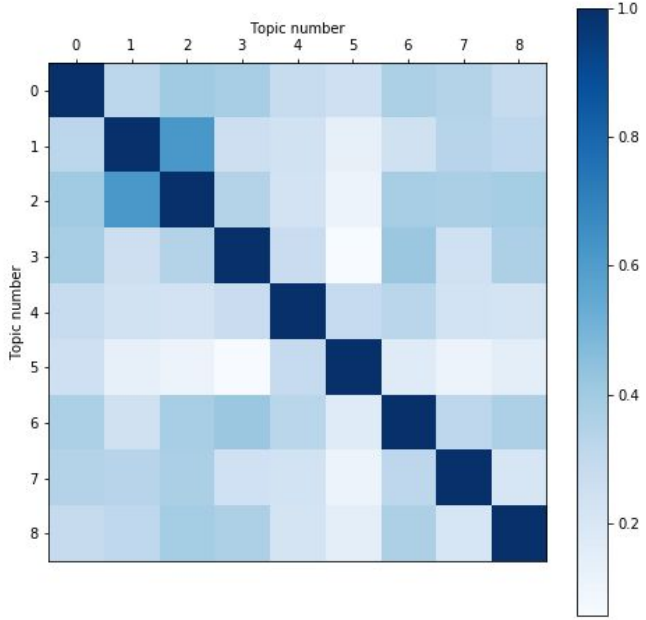
With 9 topics: 3261

With 25 topics: 3265

(Smaller values are better)

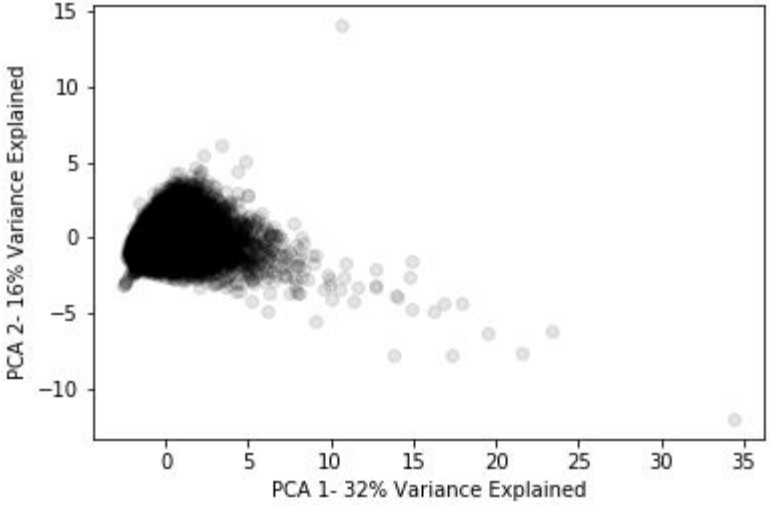
Topic Co-Occurrence vs Correlation

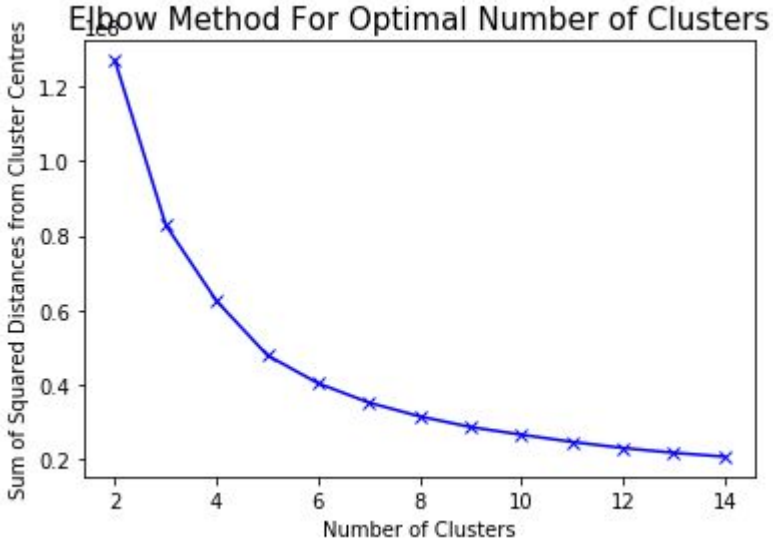




Doc2Vec

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$





How this Helps

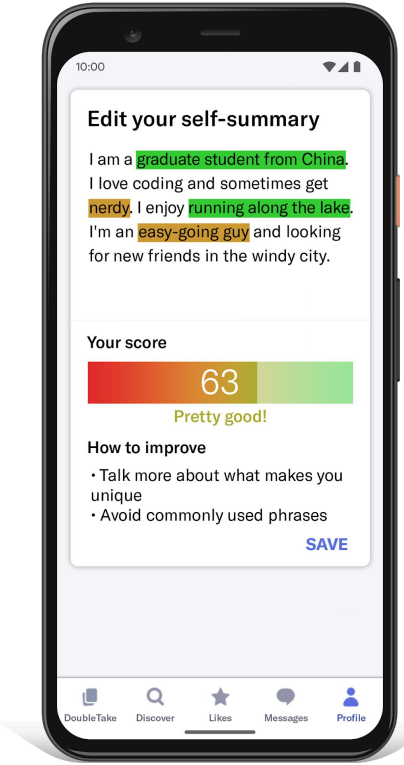
App Recommendation:
Using Doc2Vec

“ Li, you sound similar to 48% users with Ethnicity ‘Asian’ and Education ‘Masters Degree’”

Using Topic Proportions

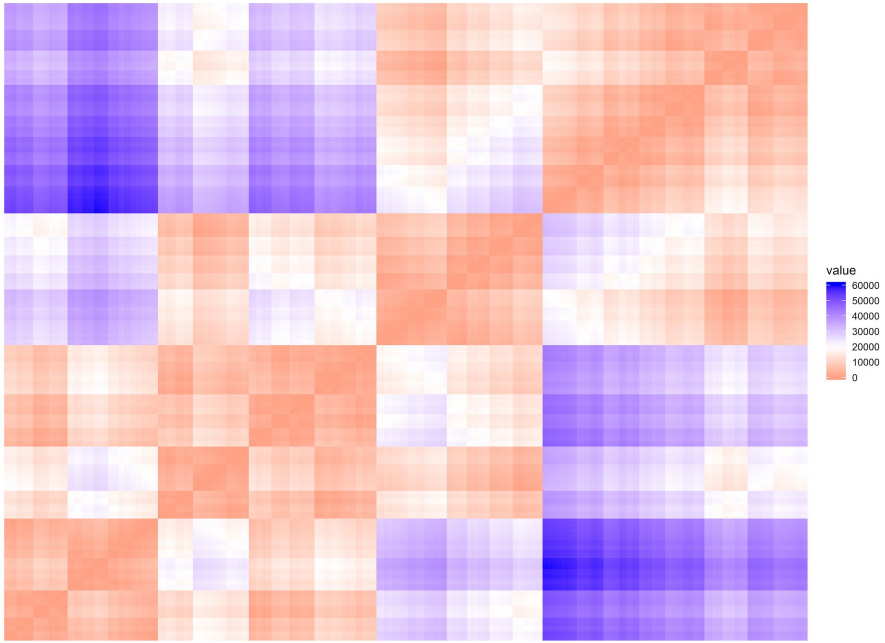
Try talking less about Your City Background and more About Your Hobbies.

Using Unusual Words
Li, you talked about “running”. How about use ‘endurance”- that’s a rare word!

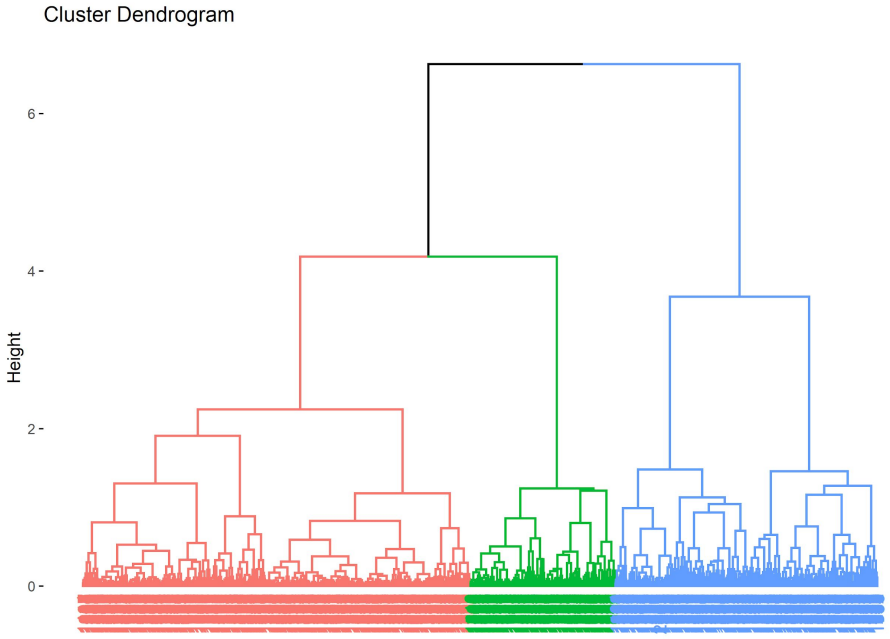


Agglomerative Nesting of Demographic Data

- Demographic data from a random subset of 2,000 profiles
- Hopkins Statistic of 0.770
- Using a Gower's dissimilarity matrix and Ward.D2
- Three clusters selected using Nbclust

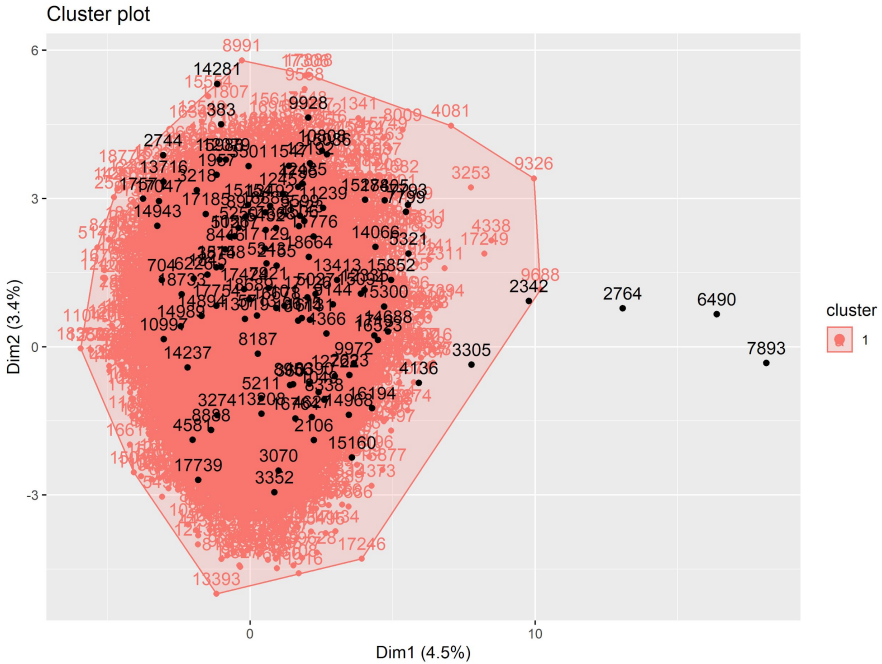


Agglomerative Nesting of Demographic Data



Identification of Doc2Vec Outliers with DBSCAN

- Uses 50 vectors from Doc2Vec, amount of long words, and Flesch scores generated from profiles
- Hopkins Statistic of 0.808
- Epsilon neighborhood size of 9 selected from a 5-NN distribution plot



Identification of Doc2Vec Outliers with DBSCAN

Typical Profiles

1. *“i’ve lived in san francisco for almost 5 years after moving herefrom the uk. i’m laid back and don’t sweat the small stuff…”*
2. *“i’m a playful red-haired guy who’s both shy and outgoing. who likesto play around and be serious...”*

Outlier Profiles

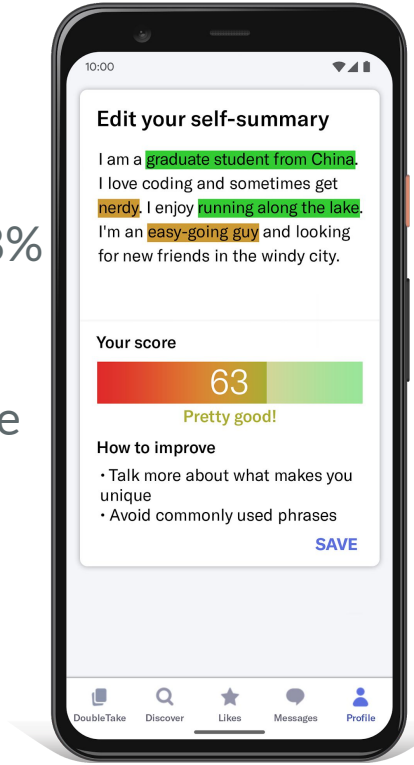
1. *“i’m a super-cerebral dude. i’m prone to pontificate my thoughtsaloud to my empty apartment…”*
2. *“employed full time with benefits. don’t pay child support. nokids.never married. no alimony. no pending law suits. not crazy…”*

How this Helps

App Recommendation:
Using DBScan

“Li, you talked about your interests in a way that matches most 78% users in your demographic.

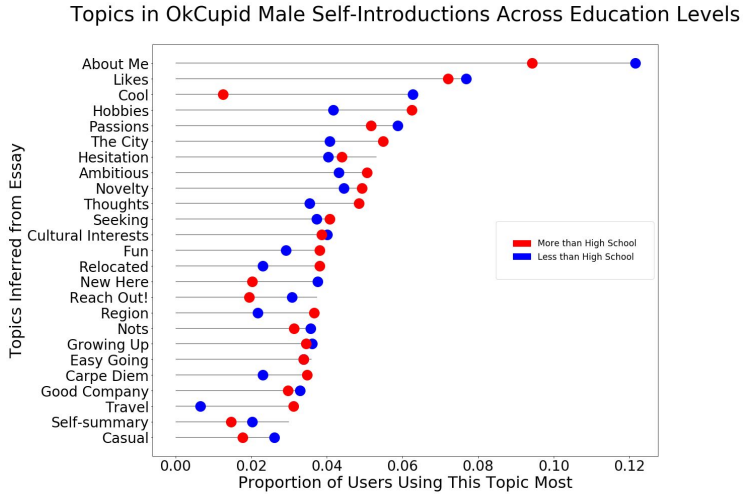
How about you use your ‘running’ and ‘China’ keywords to become more funny?”

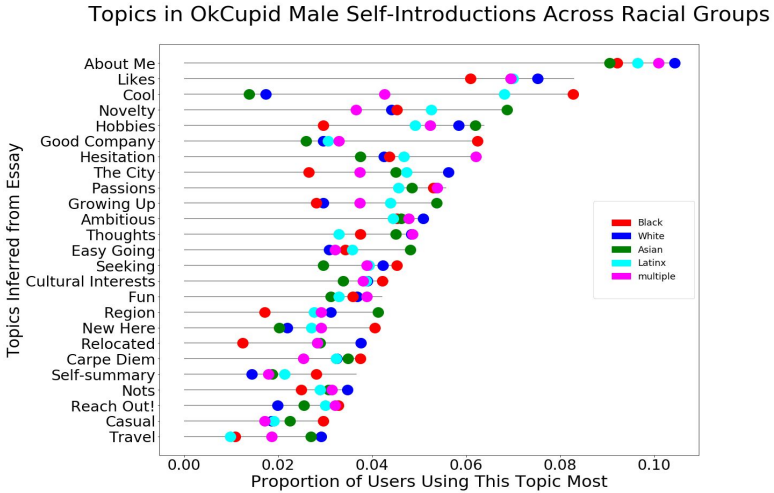


Linking Topic Model and Demographic Variables

Approach 1. Topics comparisons for different factors

Issue: Hard to test if variations are by chance or true underlying differences in subpopulations

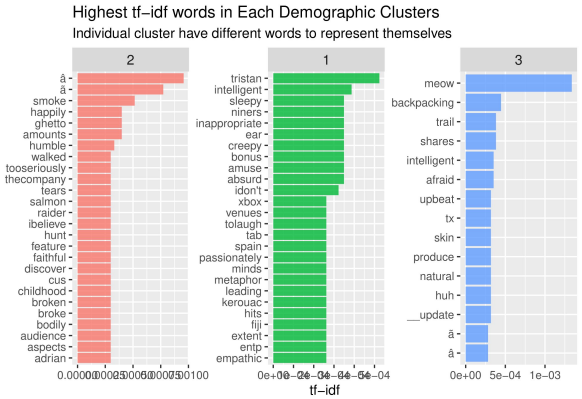


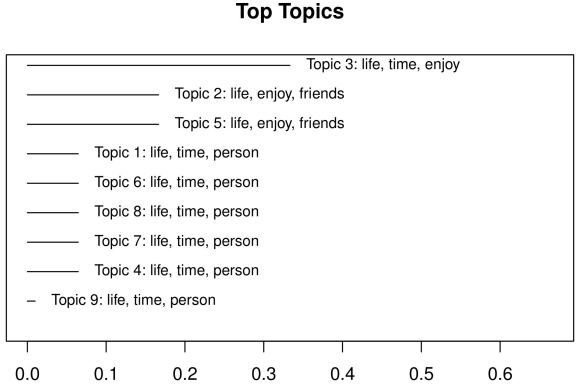


Linking Topic Model and Demographic Variables

Approach 2: Topic Modeling for Clusters

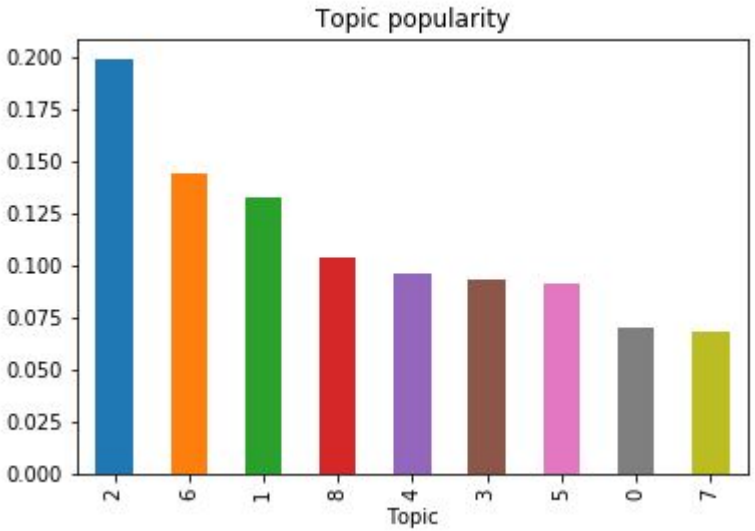
Issue: a. No clear differences; b. Cluster outcomes vary depending on the variables and methods; Clusters are not quite heterogeneous





Assumption: Common Rankings

Across All Demographics:



Top 7 By Race

	0	1	2	3	4	5	6
multiple	topic_2	topic_6	topic_1	topic_8	topic_3	topic_5	topic_4
White	topic_2	topic_6	topic_1	topic_4	topic_3	topic_5	topic_8
Asian	topic_2	topic_8	topic_6	topic_1	topic_5	topic_4	topic_3
Black	topic_2	topic_6	topic_0	topic_8	topic_3	topic_1	topic_4
Latinx	topic_2	topic_6	topic_1	topic_8	topic_0	topic_4	topic_3
other	topic_2	topic_6	topic_8	topic_1	topic_3	topic_4	topic_0

Top 7 By Education

	0	1	2	3	4	5	6
High School or less	topic_2	topic_1	topic_6	topic_0	topic_8	topic_4	topic_3
More than High School	topic_2	topic_6	topic_1	topic_8	topic_5	topic_4	topic_3

Structural Topic Model

Purpose: Estimating metadata and topic relationships

Difference: LDA assumes topics are latent;

STM uncover the "true" structure of the document configurations

Dependent variable: proportion of each document about a topic (“prevalence”)

Independent variables (Metadata/Covariates):

1. User Demographics (fit + edu + height_group + race_ethnicity)
2. DBScan Cluster

Regression: Predict how prevalence shift as a function of the covariates

Significant Covariates

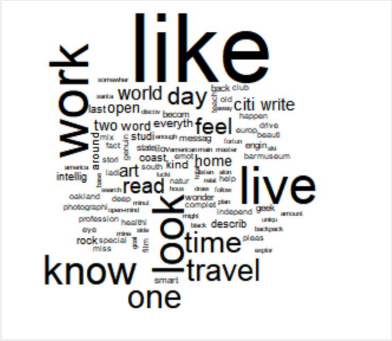
(0: '***' 0.001: '**' 0.01: '*' 0.05 : '.')

Topic / X	fit_not_fit	fit_unknown	edu_More than High School	edu_unknown	height_groupshort	race_ethnicityBlack	race_ethnicityLatinx	race_ethnicitymultiple	race_ethnicityother	race_ethnicityWhite	dbscan_cluster
1			*					***	**	**	***
2	***		***	**			**	*		**	
3	***	*	***	***				**			
4	.		***	*				*		***	**
5	.		*			**					
6			***								
7		.	.		.	.			*		**
8	***		**	***		***	***	***		***	*
9	***		***			***					**

Significant Cov...

...05 : ".")

Topic / X	fit_not_fit	fit_unknown								race_ethnicityWhite	dbscan_cluster
1										**	***
2	***									**	
3	***	*									
4	.		***					*		***	**
5	.		*			**					
6			***								
7		.	.		.	.			*		**
8	***		**	***		***	***	***		***	*
9	***		***			***					**



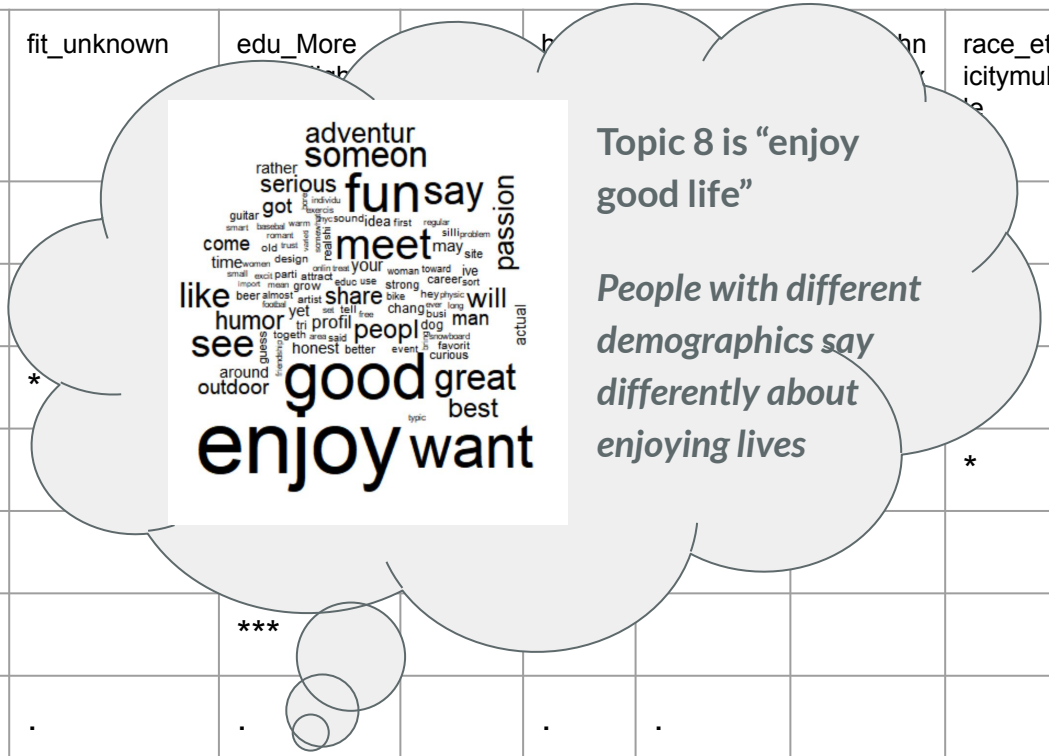
Topic 6 might be “hard working”

Only more_than_high_school factor changes the proportion of Topic 6

Significant Covariates

(0: '***' 0.001: '**' 0.01: '*' 0.05 : '.')

Topic / X	fit_not_fit	fit_unknown	edu_More					race_ethnicitymultiple	race_ethnicityother	race_ethnicityWhite	dbscan_cluster
1									**	**	***
2	***									**	
3	***	*									
4	.							*		***	**
5	.										
6			***								
7		.	.		.	.			*		**
8	***		**	***		***	***	***		***	*
9	***		***			***					**



Closing Notes

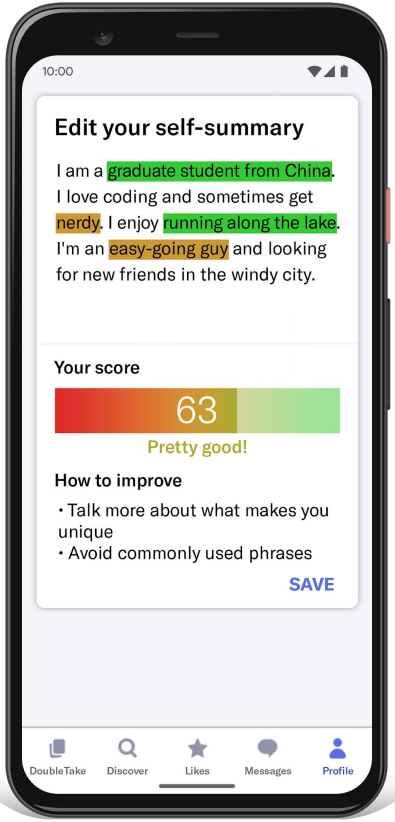
By Final Project:

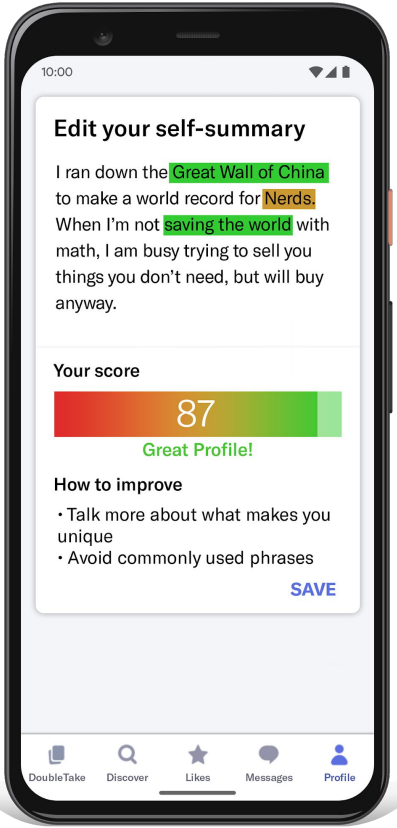
- Validation of STM results
- Testing for Differences in Number of Topics at Level of Individual Profiles
- Test model using AGNES clusters and varying topic number accordingly

For Future Research:

- Factor Modelling
- Association Rules Mining

One More Thing...





Appendix Slides

Type Right =
Swiped Right

References

Burke, R. Social News Daily. ‘Snapchat’s Gender Switch Filter Shows Men What Online Dating Is Like For Women’, Retrieved from: https://socialnewsdaily.com/86938/snapchats-gender-switch-filter-shows-men-what-online-dating-is-like-for-women/

Fiore, A. T., Taylor, L. S., Mendelsohn, G. A., & Hearst, M. (2008, April). Assessing attractiveness in online dating profiles. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 797-806). ACM.

Fiore, A. T., Taylor, L. S., Zhong, X., Mendelsohn, G. A., & Cheshire, C. (2010, January). Who's right and who writes: People, profiles, contacts, and replies in online dating. In *2010 43rd Hawaii International Conference on System Sciences* (pp. 1-10). IEEE.

Emory, Leah(2017), Bustle. ‘How Many People Who Meet On Dating Apps Get Married? Swiping Isn't Just For Hookups’, Retrieved from - https://www.bustle.com/p/how-many-people-who-meet-on-dating-apps-get-married-swiping-isnt-just-for-hookups-44359

Stevens, G., Owens, D., & Schaefer, E. C. (1990). Education and attractiveness in marriage choices. *Social Psychology Quarterly*, 62-70.

Shepperd, J. A., & Strathman, A. J. (1989). Attractiveness and height: The role of stature in dating preference, frequency of dating, and perceptions of attractiveness. *Personality and Social Psychology Bulletin*, *15*(4), 617-627.

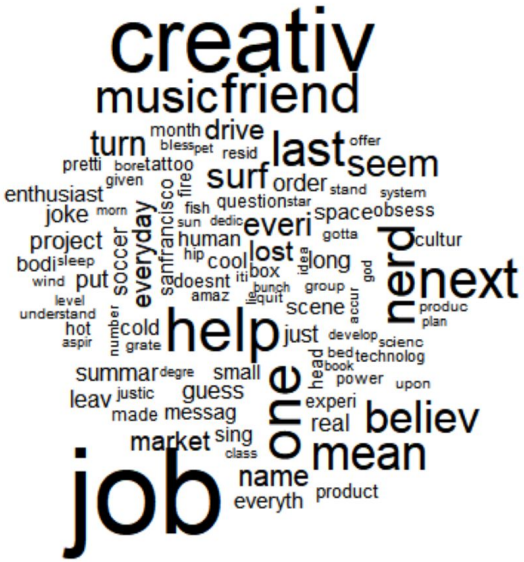
Shishido, Juan, Jaya Narasimhan, and Matar Haller. "Tell Me Something I Don’t Know: Analyzing OkCupid Profiles." (2016).

Torpey, Elka, 2018 , ‘Measuring the value of education’, Bureau of Labour Statistics. Retrieved from:

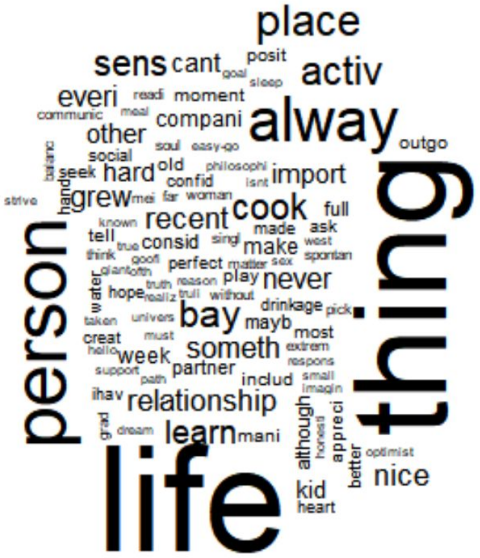
https://www.bls.gov/careeroutlook/2018/data-on-display/education-pays.htm

Word Cloud

Topic 1



Topic 2



Topic 3

STM Effect Coefficients

```
Topic 1:

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.106315   0.006124  17.360  < 2e-16 ***
fitnot_fit                0.002279   0.002008   1.135 0.256366
fitunknown                0.021201   0.020670   1.026 0.305050
eduMore than High School  0.006156   0.002618   2.352 0.018689 *
eduunknown                0.004350   0.003772   1.153 0.248830
height_groupshort        -0.001230   0.002164  -0.568 0.569856
race_ethnicityBlack       0.008091   0.006202   1.305 0.192028
race_ethnicityLatinx      0.002420   0.004088   0.592 0.553919
race_ethnicitymultiple    0.016606   0.004385   3.787 0.000153 ***
race_ethnicityother       0.014693   0.005358   2.742 0.006110 **
race_ethnicityWhite       0.009028   0.003032   2.978 0.002906 **
dbscan_cluster           -0.047060   0.005259  -8.948  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Common Mistakes

Information Overload:

i am very skeptical of the premise of online dating i am here to shake things up i am a kind of a somebody until you get me outdoing something then i get into it and always have a good-time love to cook and bake i play the bass guitar and have been teaching myself to play bluegrass guitar though i'm still pretty terrible i love the ocean but i'm a terrible swimmer i fish i read a lot and love to sit on the beach ideally all three at once i make my own beer and am pretty sure you'll like it if you like beer i almost never drink hard alcohol i coach high school track and field jumping and hurdling in particular at the high-school i went to and love iti'm a self defined nerd who likes sci-fi and can probablytroubleshoot nearly any problem you have with your mac i'm into the environment and have a degree in conservation and resource studies i love to hike and sit in beautiful high up places i am an only child but please don't ask me if i liked it i really can't tell you what having siblings would be like so i don'tknow i was born and raised in the north bay and hope that i never have to live anywhere where it snows even for part of the year am a storehouse of useless random trivia i'm pretty sureit's taking away from important information but i can't help it i love learningi'm looking to make some new friends in the area and also to meet a nice easygoing girl who will accept my quirks hit me up if you want to do something outdoors or just want to get a coffee or a beer and chat

Not Much To Work With:

“everything simple” “kicking it”

Not Targeted:

“u like to dance or eat”

Relevant Research: Fiore et al (2010)