The Language of Love: Analyzing OKCupid Profiles Using NLP

Goodwin Chen
November 16, 2018

What is OKCupid?

- Online dating site founded in 2004
- Initially web-browser based
- Users have essay questions to guide profile construction
- Users choose who to message by browsing profiles

Motivation

- Gain insight into human psychology
- Confirm or debunk stereotypes
- Just curious!

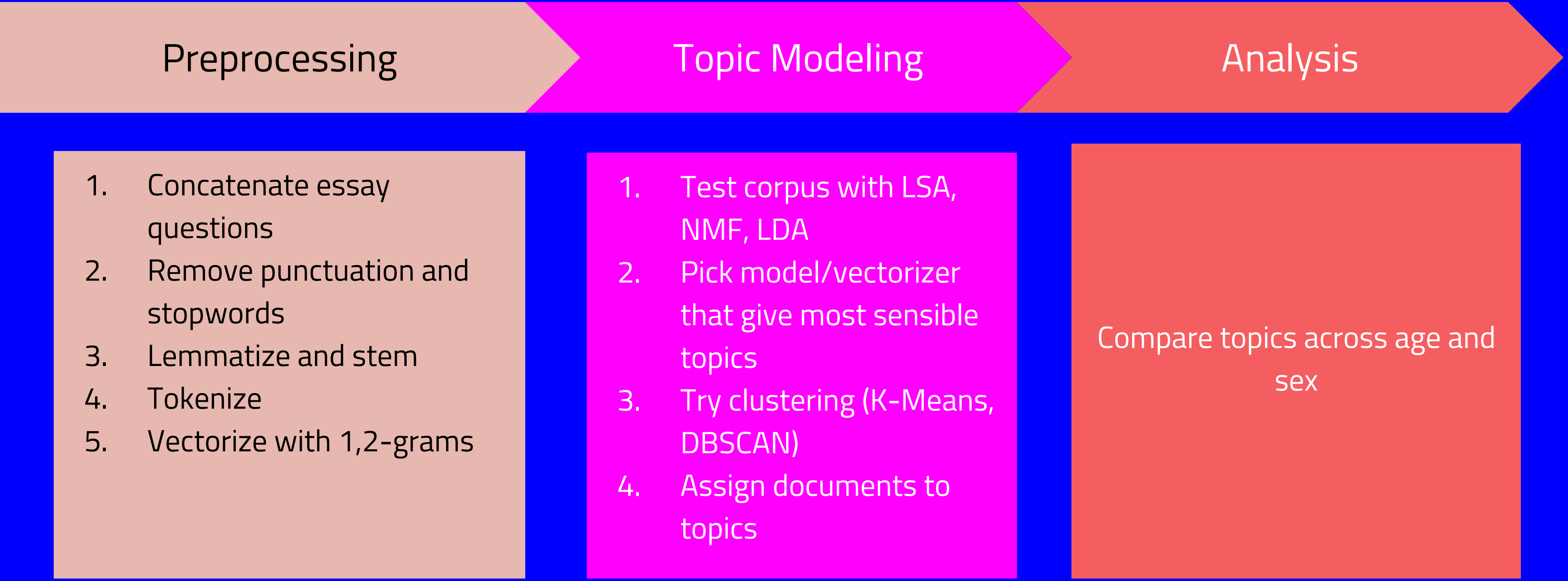
Dataset

- Credit to Albert Kim for obtaining and publishing the dataset: https://github.com/rudeboybert/JSE_OkCupid
- Permission granted by OKCupid to use
- Scraped on 6/30/2012
- 59,946 user profiles
- Has been online within 1 year
- Has at least 1 profile picture
- 25 mile radius of SF
- 10 essay questions with free-form text fields
- Sex, age also included

Essay Questions

1. My self summary
2. What I'm doing with my life
3. I'm really good at
4. The first thing people usually notice about me
5. Favorite books, movies, show, music, and food
6. The six things I could never do without
7. I spend a lot of time thinking about
8. On a typical Friday night I am
9. The most private thing I am willing to admit
10. You should message me if...

Workflow



Preprocessing

1. Concatenate essay questions
2. Remove punctuation and stopwords
3. Lemmatize and stem
4. Tokenize
5. Vectorize with 1,2-grams

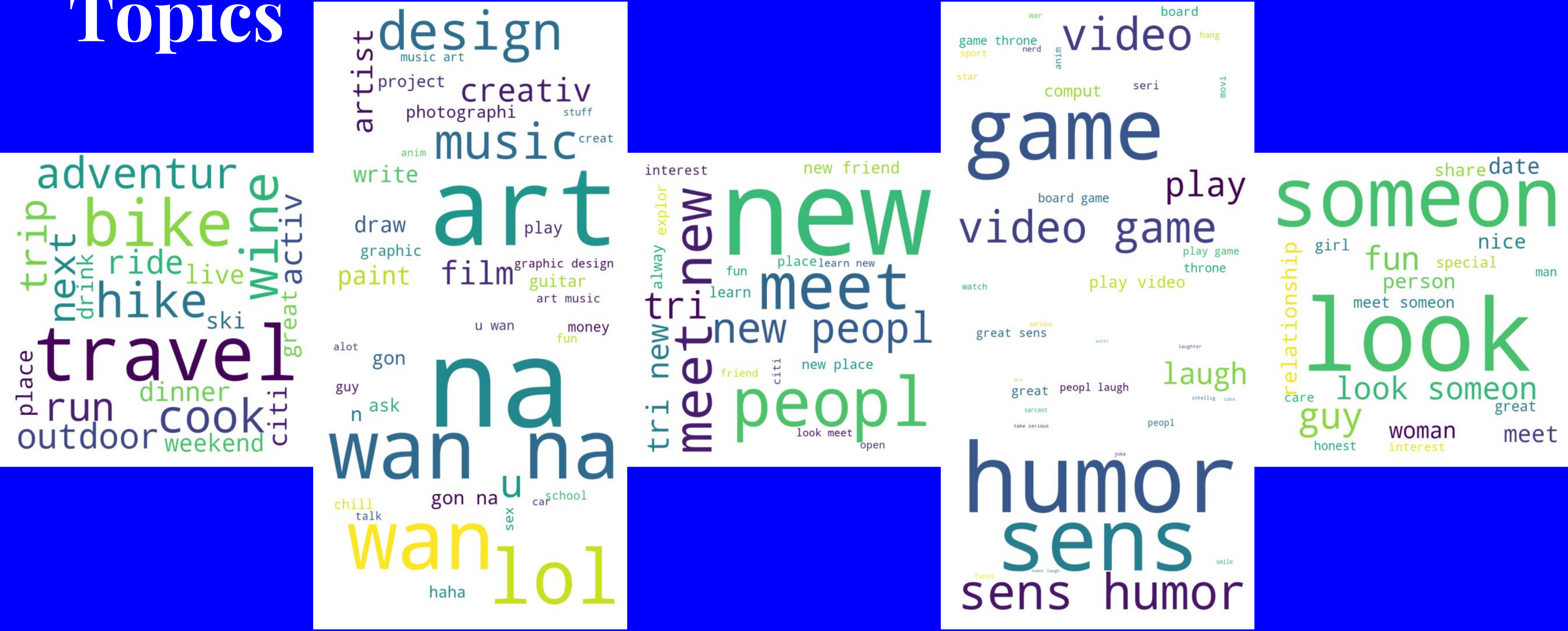
Topic Modeling

1. Test corpus with LSA, NMF, LDA
2. Pick model/vectorizer that give most sensible topics
3. Try clustering (K-Means, DBSCAN)
4. Assign documents to topics

Analysis

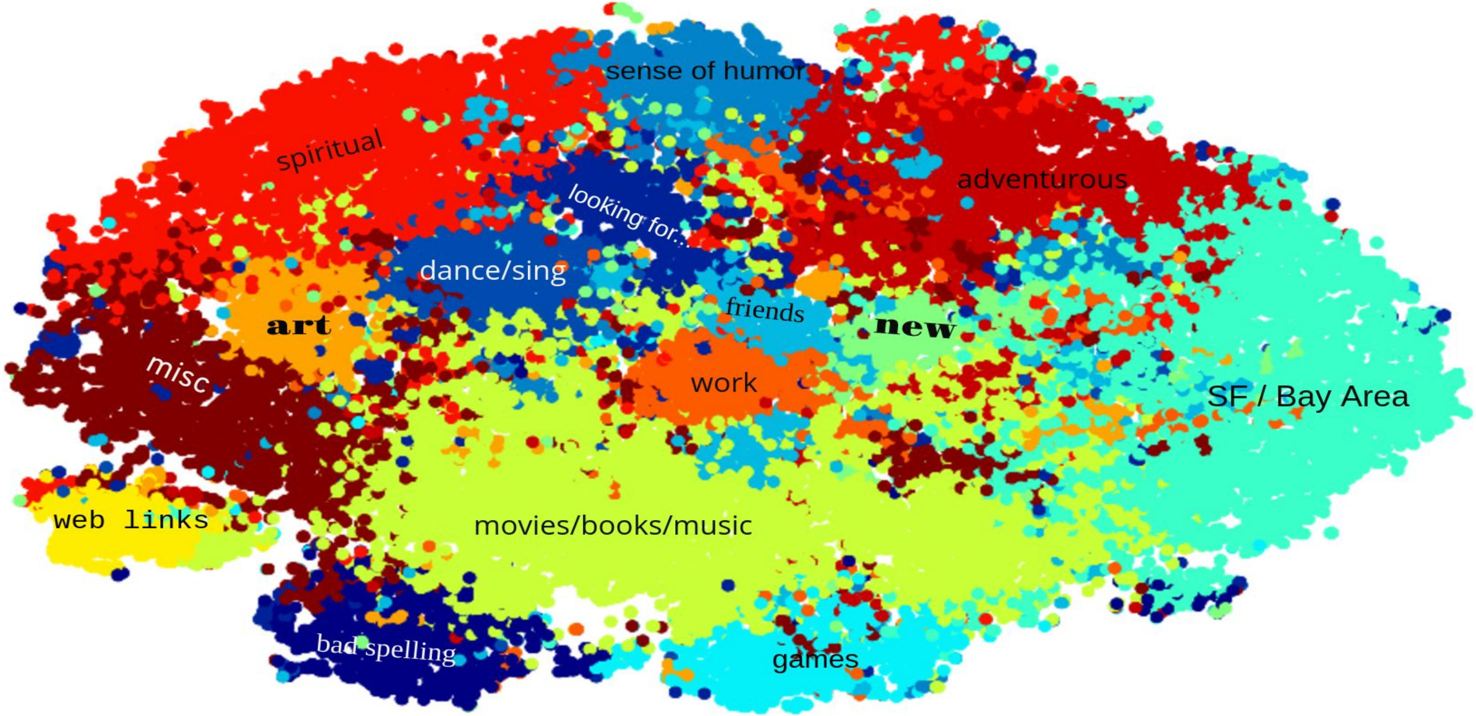
Compare topics across age and sex

Topics



NMF with TF-IDF with 15 topics

t-SNE Plot


sense of humor
spiritual
adventurous
looking for...
dance/sing
friends
art
new
misc
work
SF / Bay Area
web links
movies/books/music
bad spelling
games

Topic Differences by Gender

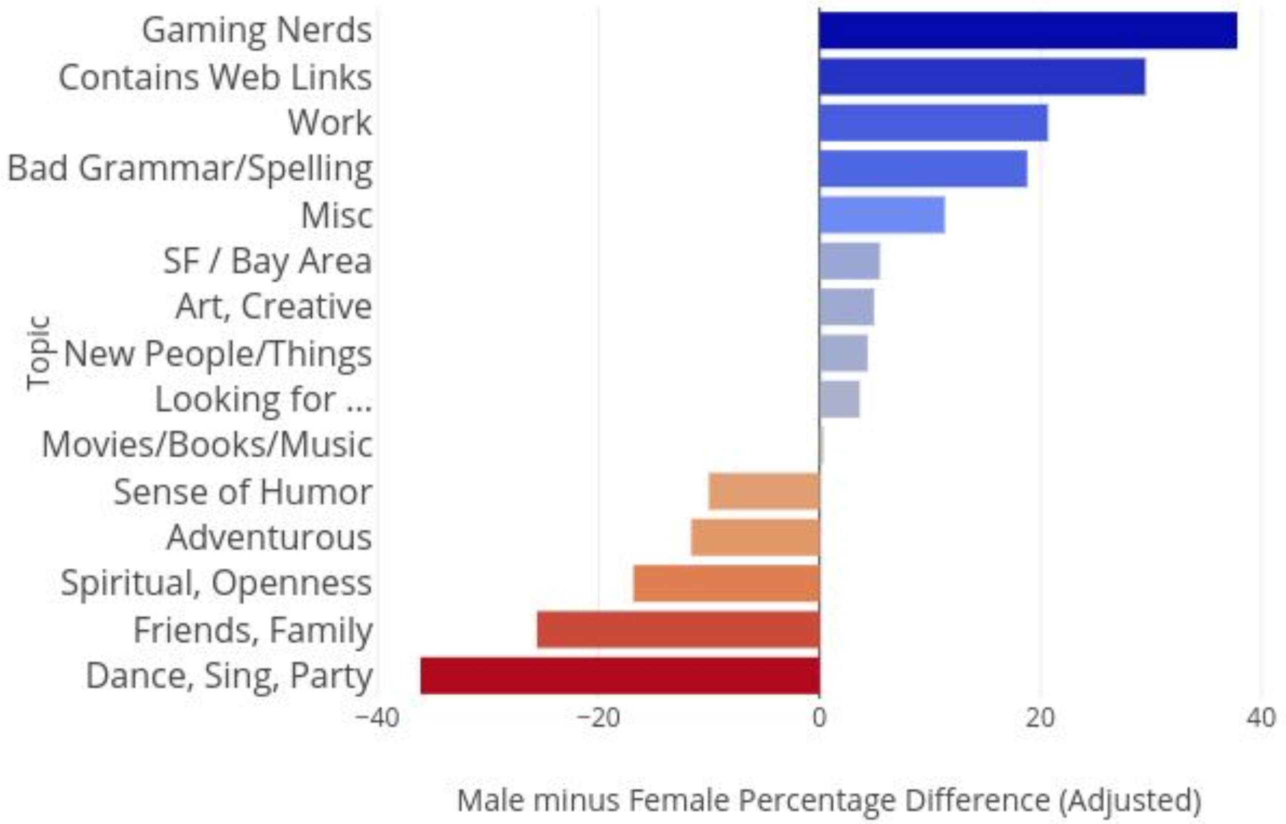
Guys

- Video Games
- Web Links
- Work
- Bad Grammar/Spelling

Girls

- Dance + Sing
- Friends + Family
- Spiritual / Openness
- Adventure / Travel

Topic Preferences by Gender



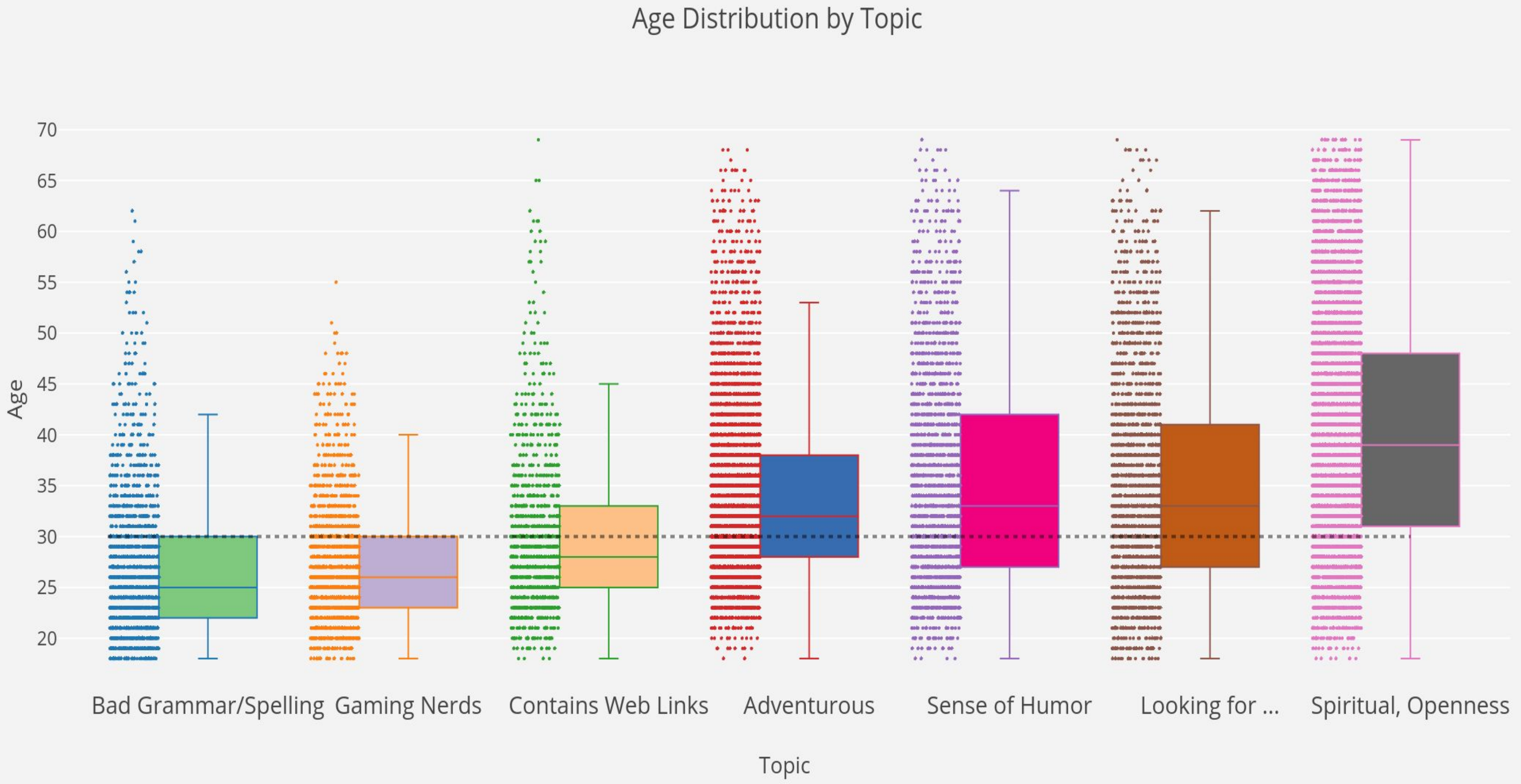
Topic Differences by Age

Younger

- Bad Grammar/Spelling
- Video Games
- Web Links

Older

- Spiritual / Openness
- Looking For ...
- Sense of Humor
- Adventure / Travel


Age Distribution by Topic
Age
70
65
60
55
50
45
40
35
30
25
20
Bad Grammar/Spelling
Gaming Nerds
Contains Web Links
Adventurous
Sense of Humor
Looking for ...
Spiritual, Openness
Topic

General Thoughts

- Close split among multiple topics in most documents
- Certain essay questions guided topics
- OKCupid should (and probably does) combine NLP with internal data to power their matching algorithm
- NLP probably less useful for modern online dating in the app era (fewer words!)

Questions?