

Linguistic Self-Expression on Dating Websites



Elizabeth Dellea
Emerging Media, '16 • Ithaca College



Kelly Sadwin
Computer Science, '16 • Ithaca College

Do generalizations about gender and language hold true in online communications?

Past linguistic studies and common beliefs tell us that men and women use language differently. Our research investigates such claims in an online context.

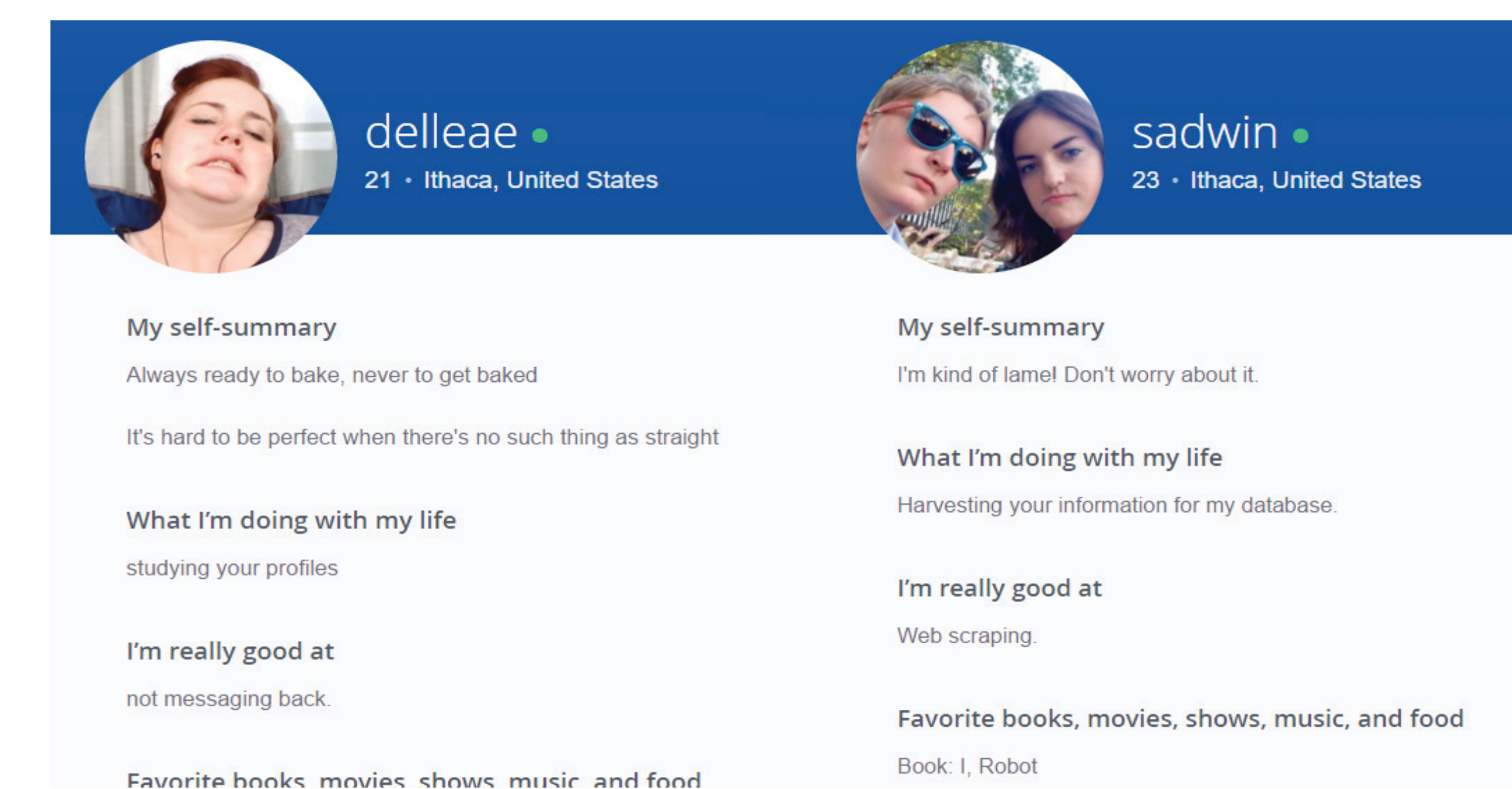
So, where can we find substantive samples of people writing about themselves? Dating websites, of course!

Web Scraping to Gather Data

We created a pair of bots on OkCupid and cast as wide a net as possible for our potential “matches”.

- One male bot, one female bot
- Both bisexual and looking for everyone
- Ignore location for matches from all over the world
- Looking for ages 18-30
- Answered minimum number of survey questions

We designed a web scraper in Python to download over 1500 matches’ profiles and their demographic tags.



Word Clouds

Using ltc tf-idf to increase the importance of words unique to the demographic, we generated the following word clouds. We'll let them speak for themselves.



Features

Using Natural Language Processing techniques provided by the Python modules NLTK and TextBlob, we recorded the following linguistic features for each profile:

Word count: tokenize and count total words.

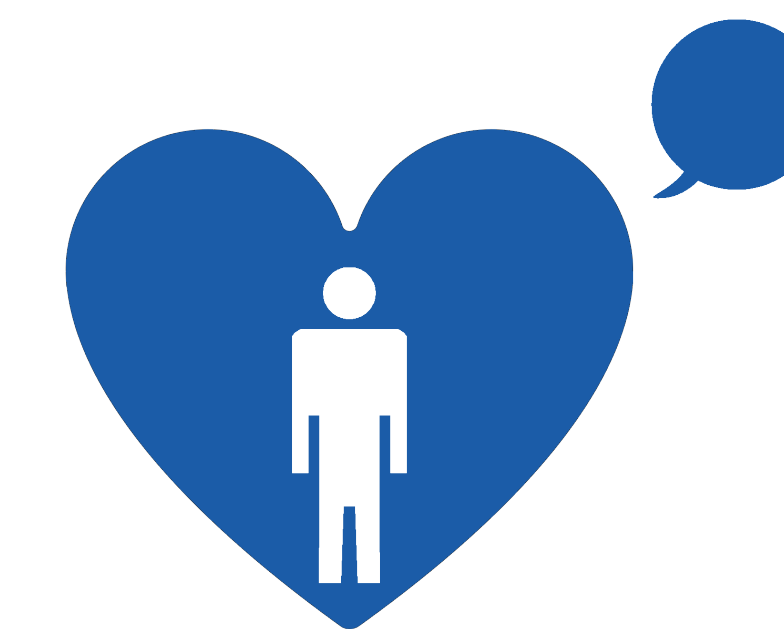
Average word length: tokenize and average character length of words.

Unique word count: turn list of tokens into a set and find its length.

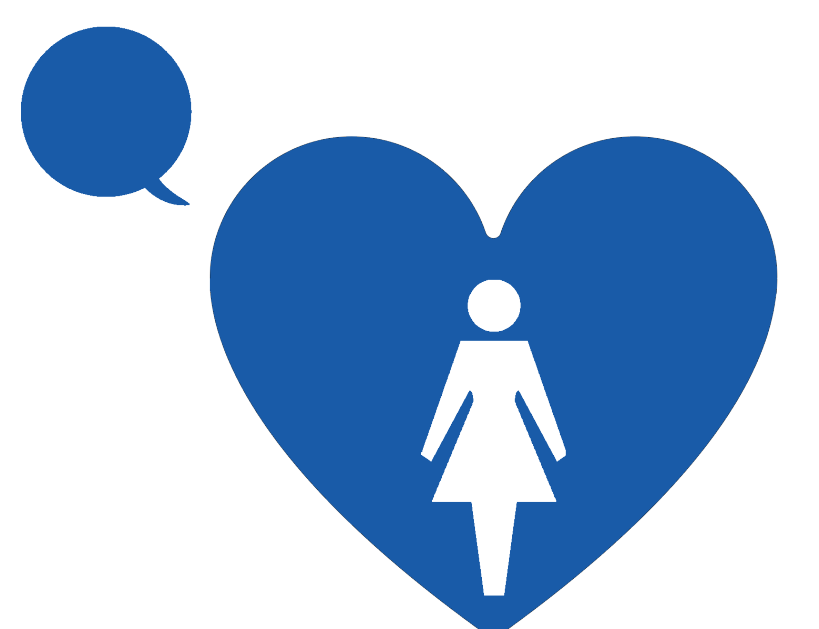
Polarity: use TextBlob's pretrained algorithm to determine a polarity score between -1 (negative) and 1 (positive).

Subjectivity: use TextBlob's pretrained algorithm to determine a subjectivity score between 0 (objective) and 1 (subjective).

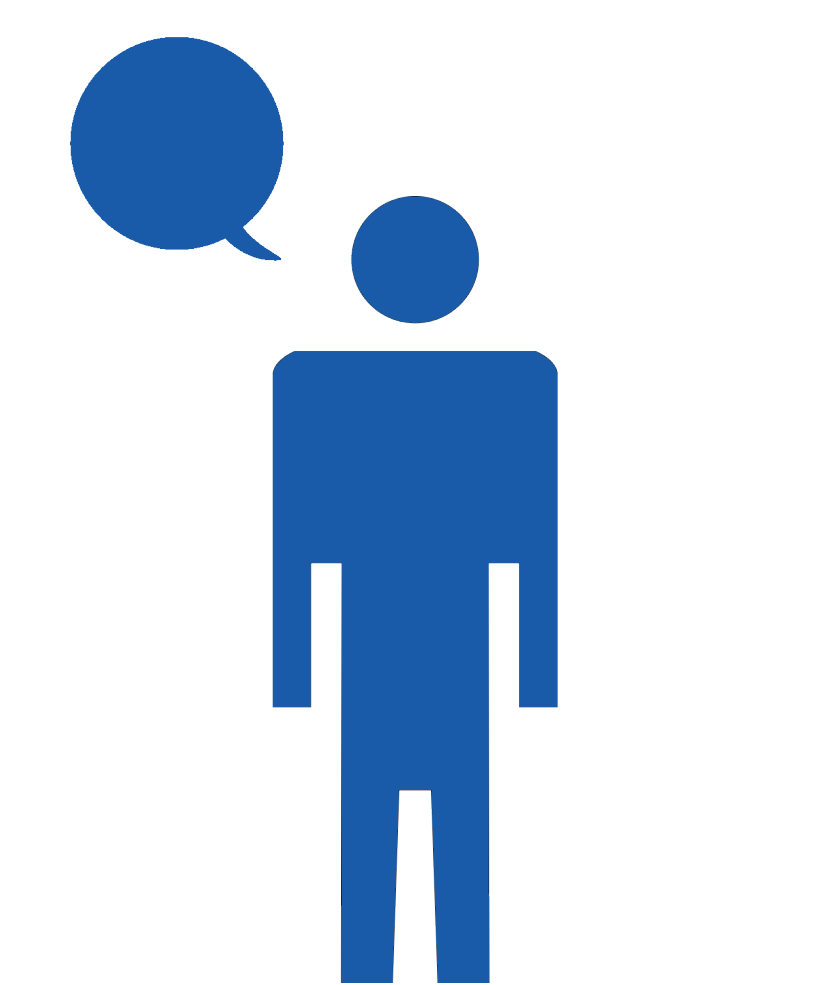
Through statistical tests, we determined a few factoids.



Users seeking men use longer words and more objective language than users seeking women.



Straight users write shorter profiles, with a smaller variety of words than non-straight users, but they use more positive and more subjective language.



Interestingly enough, there were no statistically significant differences found between men and women.

References and Special Thanks

Python modules used: wordcloud, beautifulsoup, textblob, nltk, matplotlib, and numpy

Iconography by Edward Boatman, Matt Brooks, and Hayashi Fumihiro (nounproject.com)

Special thanks to our faculty adviser, John Barr, Ph. D. Thanks to Doug Turnbull, Ph. D., for course material on web scraping, tf-idf, and NLP.

Additional thanks to Laurie Arliss, Ph. D., professor of Communications and Women's & Gender Studies, for her insight.