Content Analysis of Craigslist's Missed Connections

Stephanie Tong

stong1108@gmail.com

github.com/stong1108/CL_missedconn

missingpeople.solutions

What is Craigslist's Missed Connections?

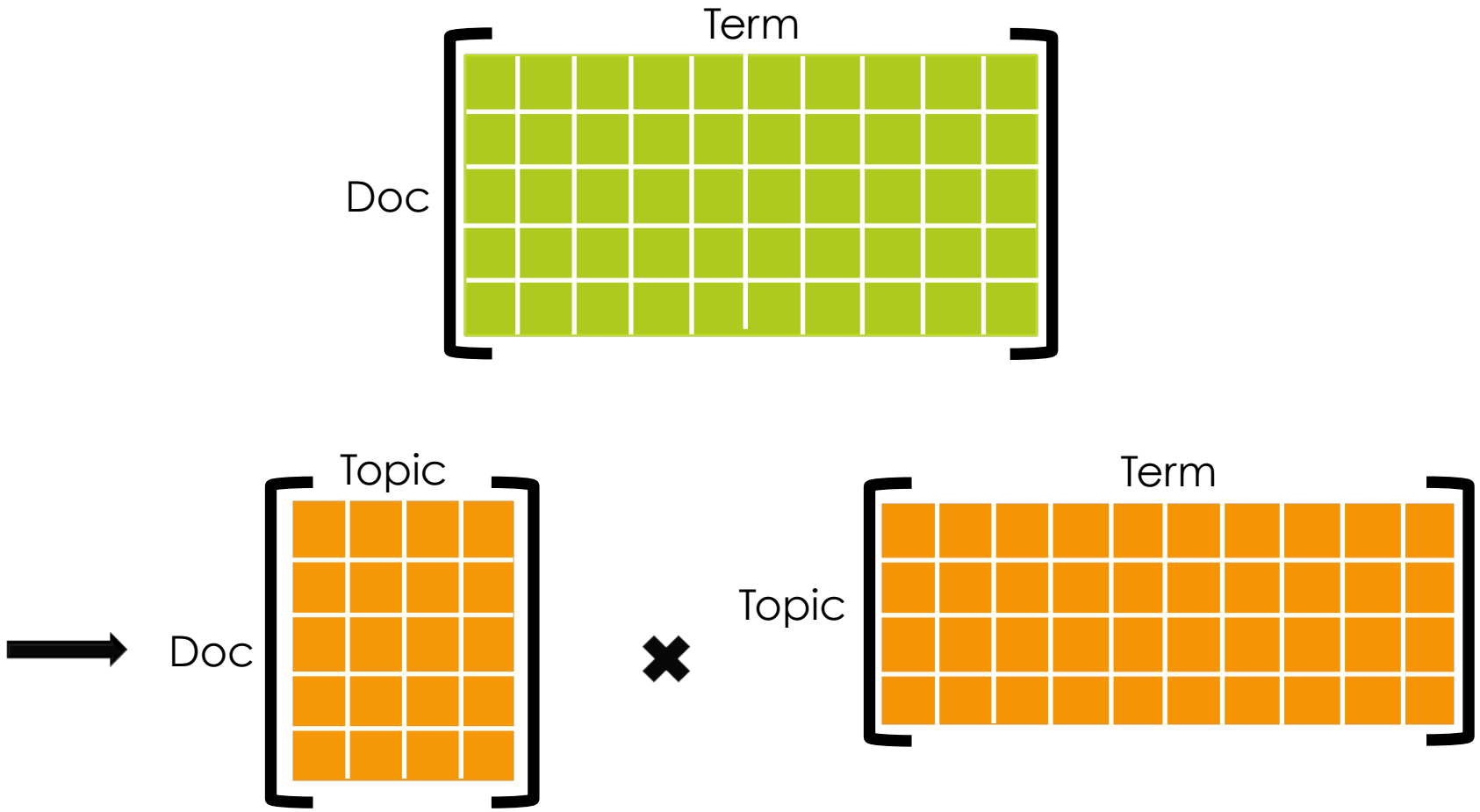
- Section of craigslist.org where people post about strangers they noticed, generally to reconnect
- No account required → safety in anonymity, unlike social media
- Varied collection of posts
  - Funny stories (real or made-up)
  - Creative outlet (poems)
  - Re-living past experiences
  - Carnal desires

Data Pipeline

- Build web crawlers to scrape and store ~23,000 Missed Connections
  - Python (requests, BeautifulSoup), PostgreSQL
  - Separate web crawlers for scraping by city and scraping Best-of-Craigslist
- Perform topic modeling to explore common themes
  - NMF, LDA, LSA (word2vec + SVD)
- Create "search" and map visualizations for web app

Topic Modeling

Matrix Factorization



Matrix Factorization

NMF

Non-negative Matrix Factorization

- Each doc is a weighted combination of topics
- TF or TF-IDF

LDA

Latent Dirichlet Allocation

- Each doc contains a topic distribution
- Each topic contains a word distribution
- TF only

LSA

Latent Semantic Analysis

- Compare word meanings
- Low-rank approximation by SVD
- TF of word vectors

Matrix Factorization

NMF

Non-negative Matrix Factorization

- Each doc is a weighted combination of topics
- TF or TF-IDF

LDA

Latent Dirichlet Allocation

- Each doc contains a topic distribution
- Each topic contains a word distribution
- TF only

Not good for docs with similar wording

LSA

Latent Semantic Analysis

- Compare word meanings
- Low-rank approximation by SVD
- TF of word vectors

Matrix Factorization

NMF

Non-negative Matrix Factorization

- Each doc is a weighted combination of topics
- TF or TF-IDF

LDA

Latent Dirichlet Allocation

- Each doc contains a topic distribution
- Each topic contains a word distribution
- TF only

Not good for docs with similar wording

LSA

Latent Semantic Analysis

- Compare word meanings
- Low-rank approximation by SVD
- TF of word vectors

"fluffy, pungent"
"potato, sorghum"

Matrix Factorization

NMF

Non-negative Matrix Factorization

- Each doc is a weighted combination of topics
- TF or TF-IDF

LDA

Latent Dirichlet Allocation

- Each doc contains a topic distribution
- Each topic contains a word distribution
- TF only

Not good for docs with similar wording

LSA

Latent Semantic Analysis

- Compare word meanings
- Low-rank approximation by SVD
- TF of word vectors

"fluffy, pungent"
"potato, sorghum"

Matrix Factorization

NMF

Non-negative Matrix Factorization

- Each doc is a weighted combination of topics
- TF or TF-IDF

Better results by category (ex: m4m)

LDA

Latent Dirichlet Allocation

- Each doc contains a topic distribution
- Each topic contains a word distribution
- TF only

Not good for docs with similar wording

LSA

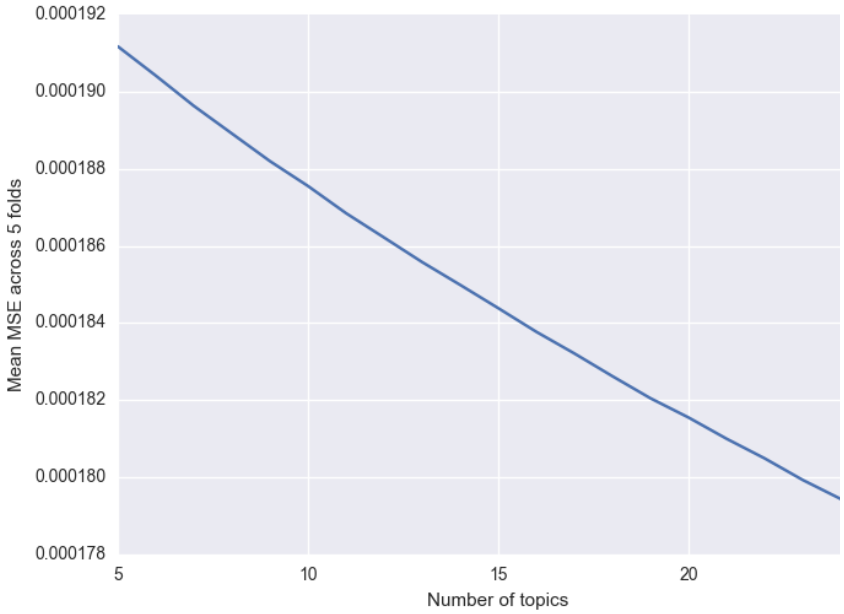
Latent Semantic Analysis

- Compare word meanings
- Low-rank approximation by SVD
- TF of word vectors

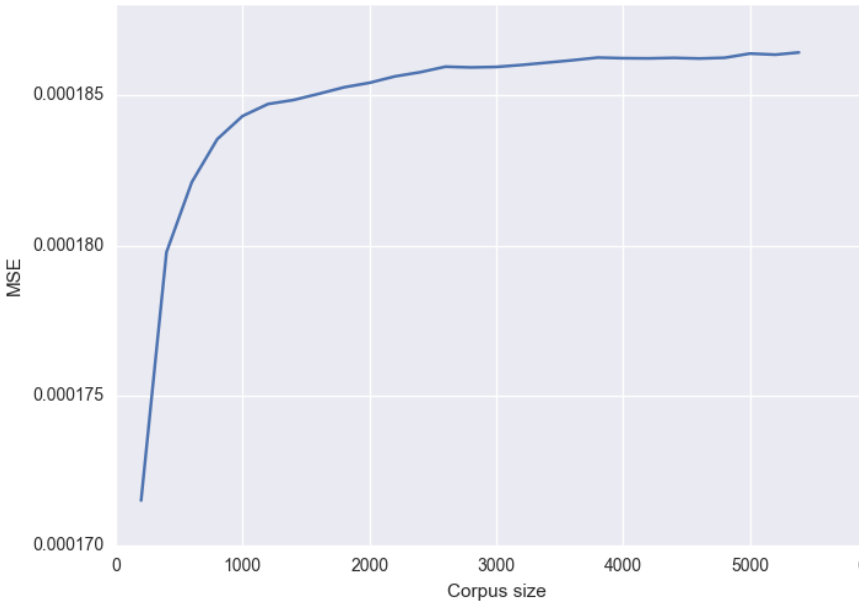
"fluffy, pungent"
"potato, sorghum"

Choosing number of topics & corpus size (m4m)

Error as number of topics increases



Error as corpus size increases



Topics

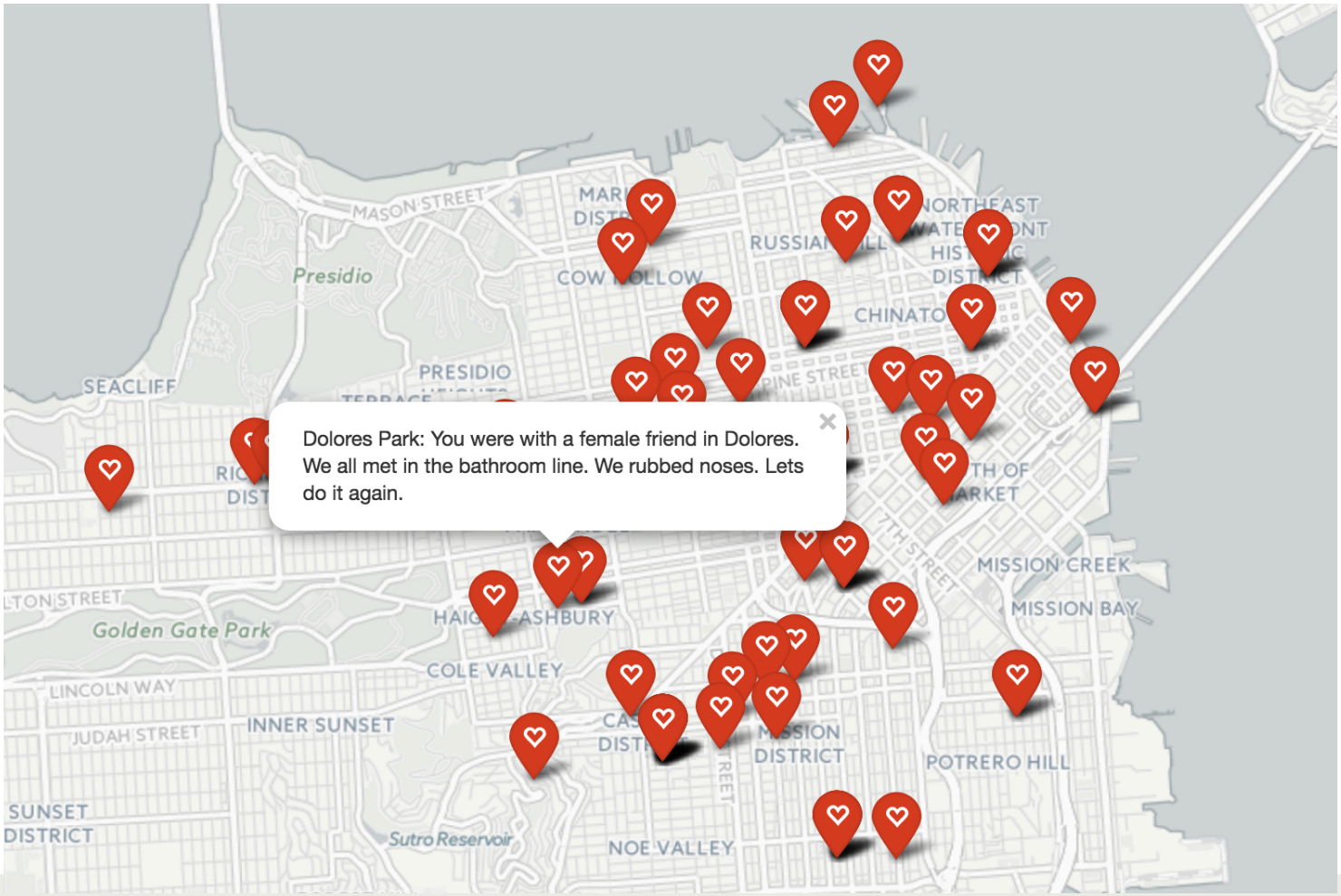
- eye contact
- **Situations**: driving, walking dog, public transit
- **m4m, w4w**: extremely NSFW topics
- **w4m, w4w**: pain, hurt, birthday & happy birthday
- **m4m**: gym, shower, workout

Best of Missed Connections

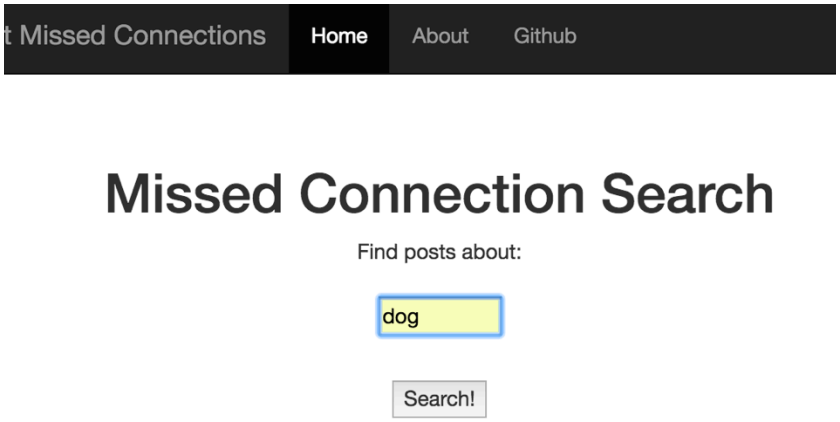
- bathroom, toilet, stall, flush, seat
- dog, poop, <NSFW>, bird
- beer, drink, drunk, bar, girl
- <lots of NSFW>
- <NSFW>, <NSFW>, shouted, shoved, stupid
- car, hit, driver, hit car

Missed Connections Explorer (web app)

Interactive Map



Search


t Missed Connections
Home
About
Github
Missed Connection Search
Find posts about:
dog
Search!
dog photo
please send me photos of your dog. i simply would just like to view photos of dogs at the moment.
a photo of your dog eating would be even better. thank you, s

Future Work

- Spelling corrector…
- Category imputer/authorship attribution
- Build larger corpus
  - word2vec, doc2vec, lda2vec
- Trending topics
  - Pokémon Go

Thank you!

Stephanie Tong

stong1108@gmail.com

github.com/stong1108/CL_missedconn

missingpeople.solutions