



Detecting the Absurd

A subreddit case study.



First World Problems

- Trivial problems from developed nations
- 261k members



Fifth World Problems

- Surreal Problems
- 151k members



How do we tell if a post is surreal or not?

- Collect posts from both subreddits
- Use a model to predict what subreddit a post comes from



Data

- Collection
- Cleaning
- Visualization



Collection

- Pushift Api
- ~ 10k posts with minimum score of 1000 per subreddit



Cleaning

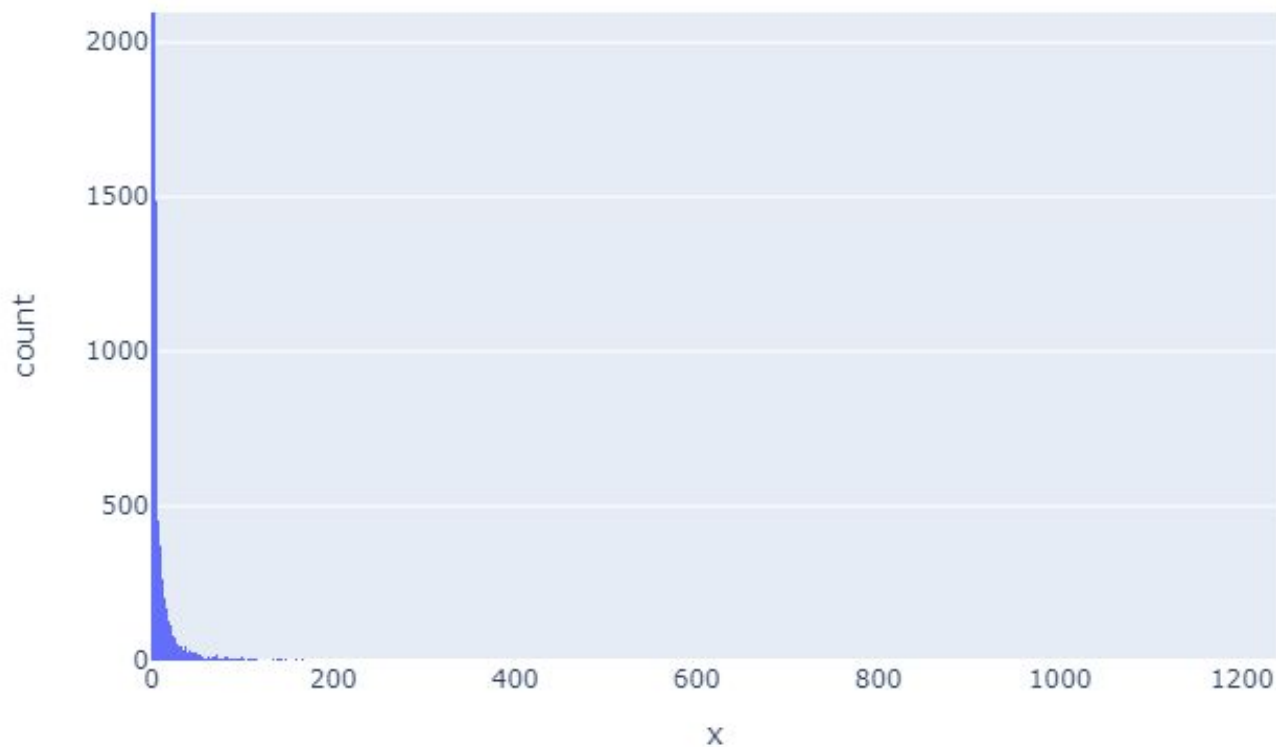
- Dropping posts with invalid selftext
- Lemmatizing words

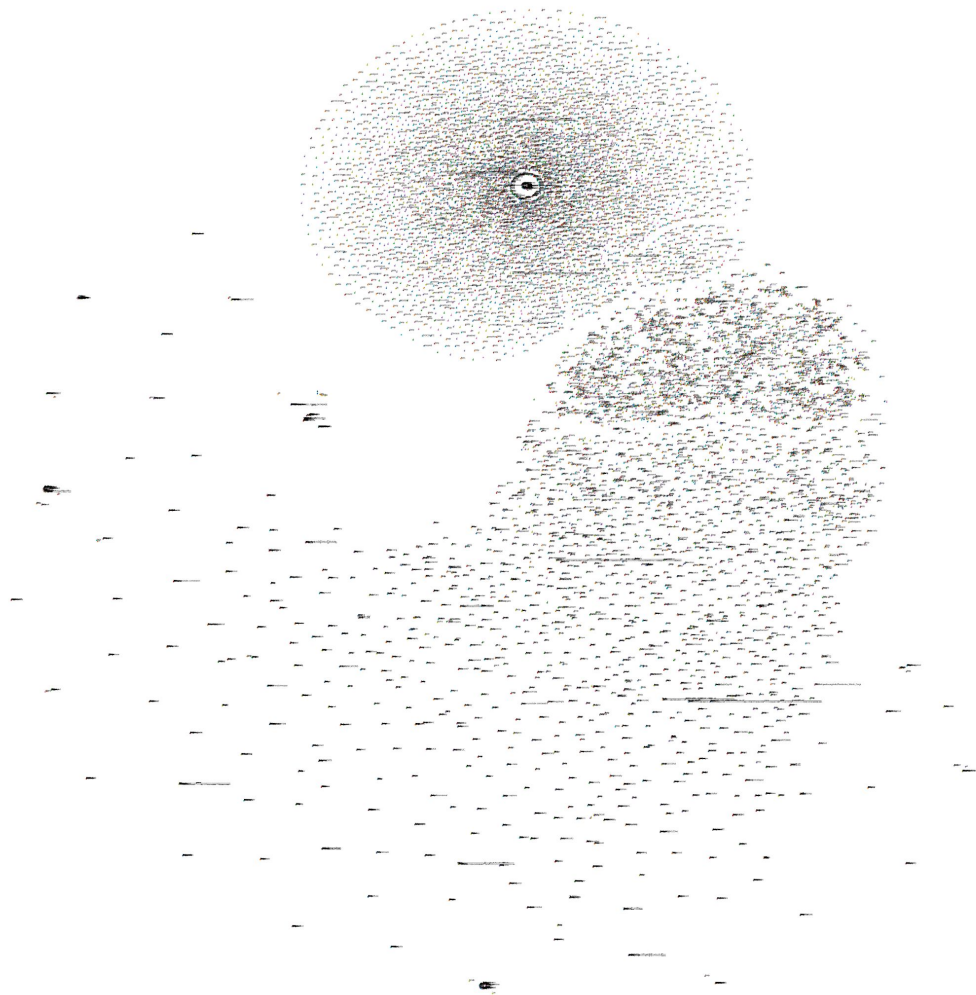


Visualization

- Distribution of words (histogram)
- Word association map (t-distributed Stochastic Neighbor Embedding)

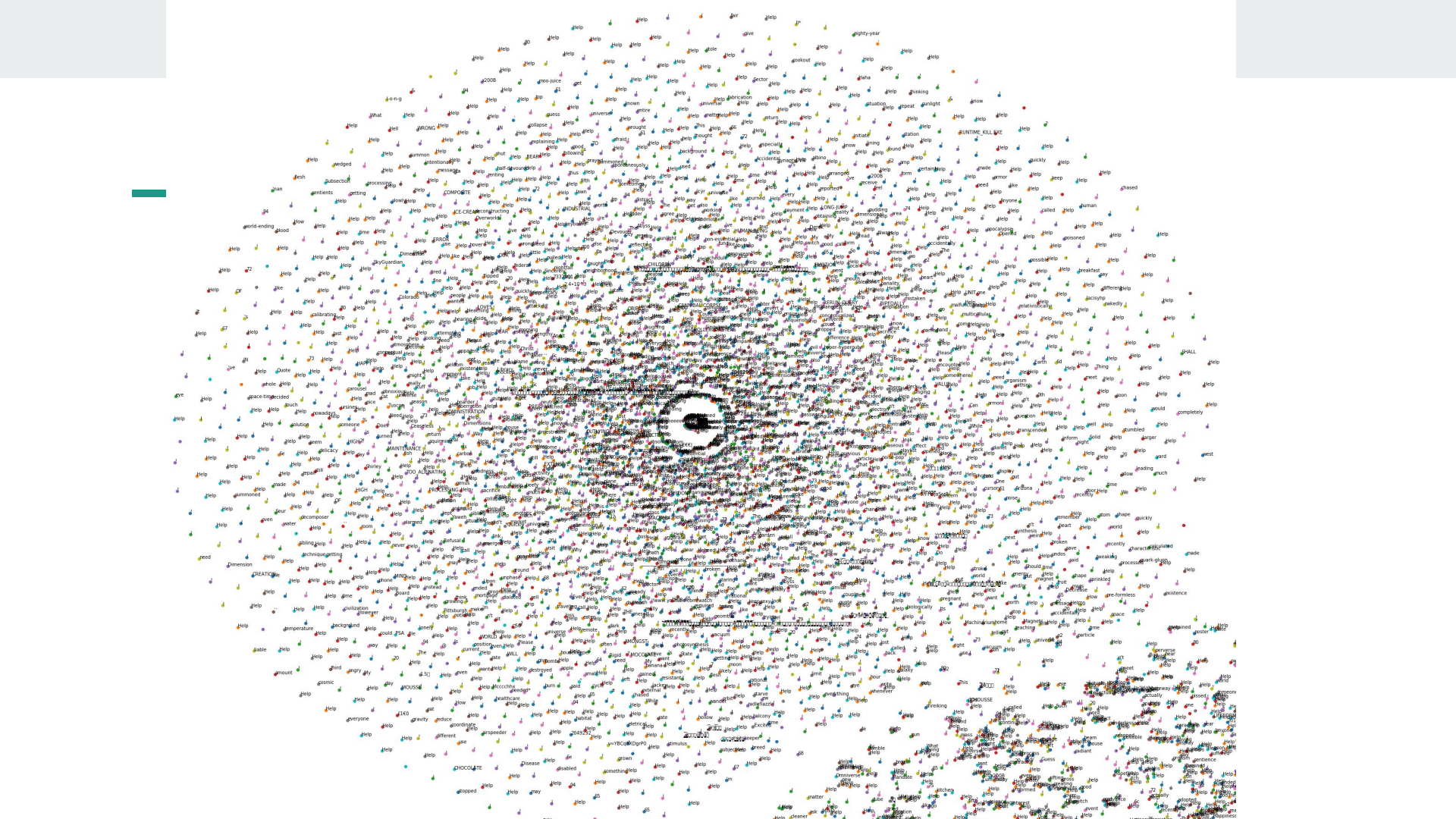
Distribution of WOrd Occurance





T-SNE

- Show word association
- Used to visualize high dimensional data
- Check out <https://distill.pub/2016/misread-t-sne/>



Holdings

49 [REDACTED]_LONGITUDE

Engineering

QUERY

piece of land

94

REMOVED FROM DATABASE



University of

Malpized

bioRxiv preprint doi: <https://doi.org/10.1101/000000>; this version posted January 1, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[Help](#) [Help out](#)

 Springer

napisz



Modeling

- Many models tried (full list in Github repo)
- Grid search and train test split used to select and evaluate models
- Logistic regression classifier won out with 89% accuracy



Potential Uses

- Filtering out insincere complaints in a social media site, for example.
- Evaluating novel text for entertainment