What's there to a review?

...

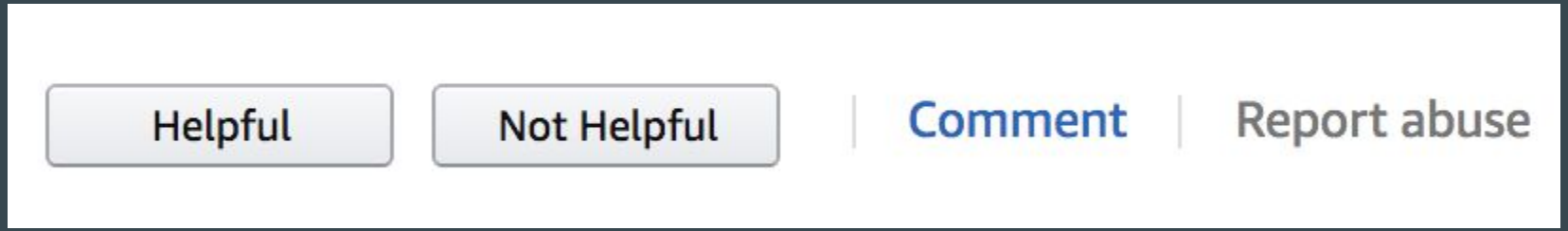
Iris Borkovsky

Action Plan

- Project scope
- Data summary
- Models
- Findings
- Appendix

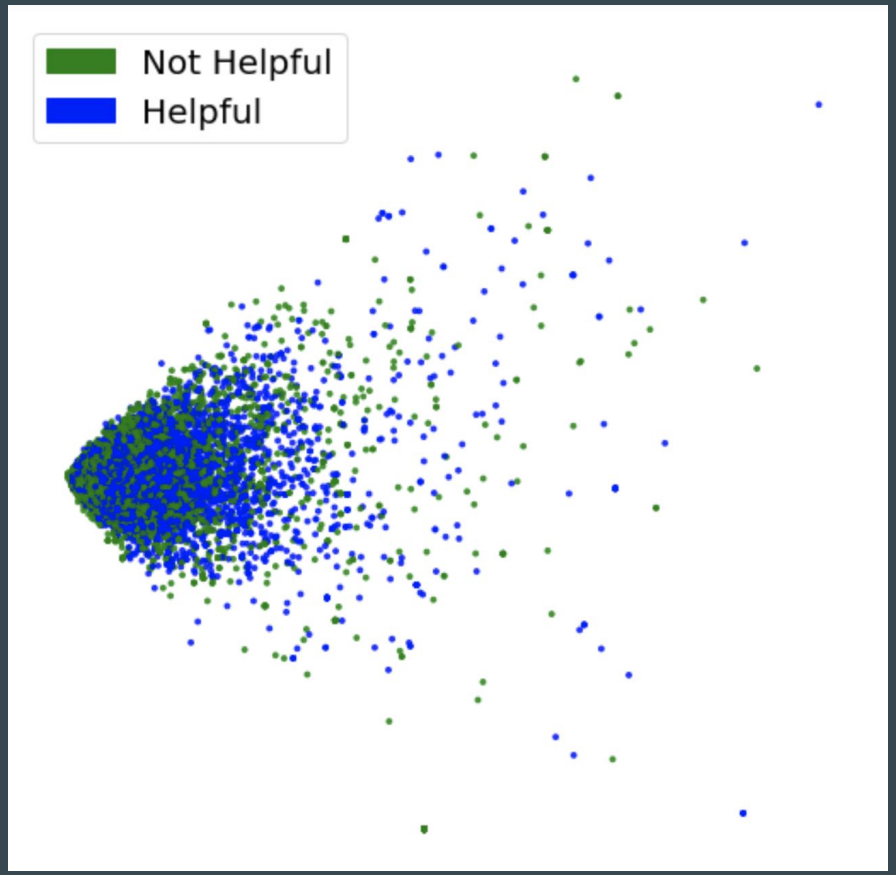
Project scope

- Can we predict if other users will find a particular review helpful?
- Topic modeling from reviews



Data

- Working data set of ~20k, from ~500k Amazon food reviews
- Focus on helpfulness numerator and denominator to craft a classifier

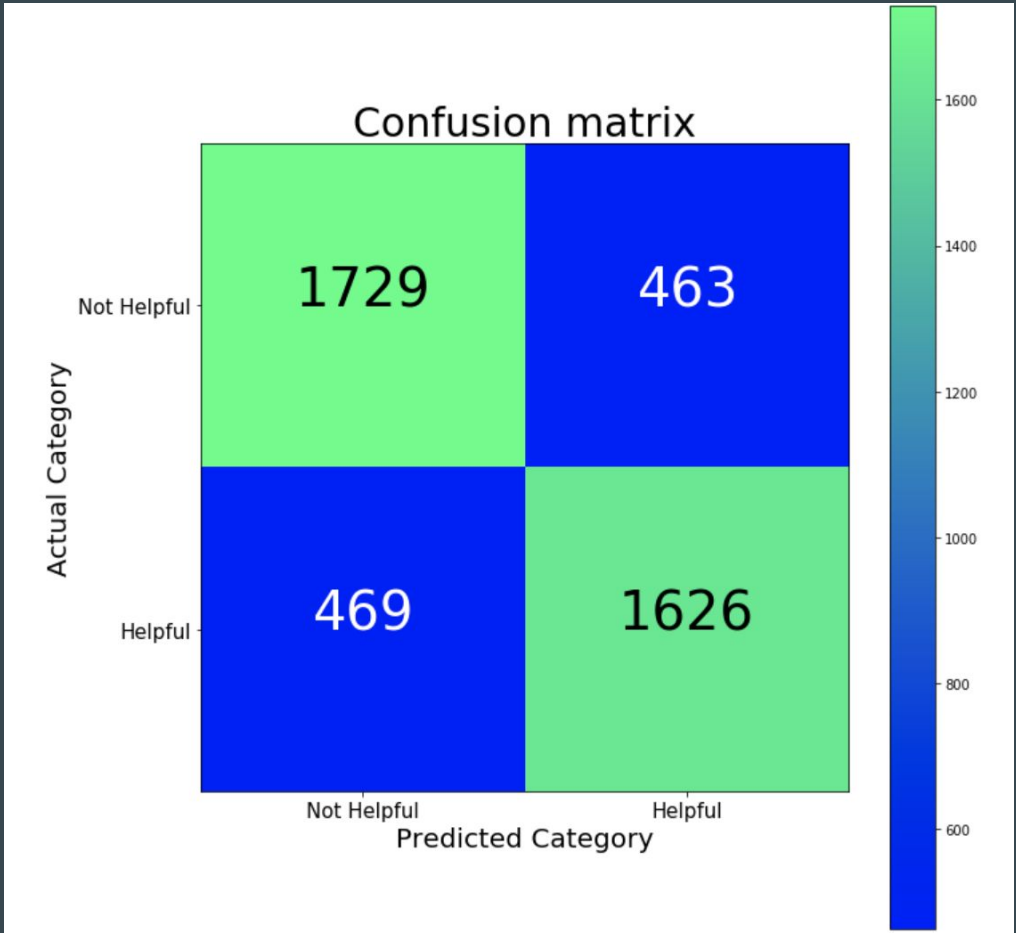

Not Helpful
Helpful

Models

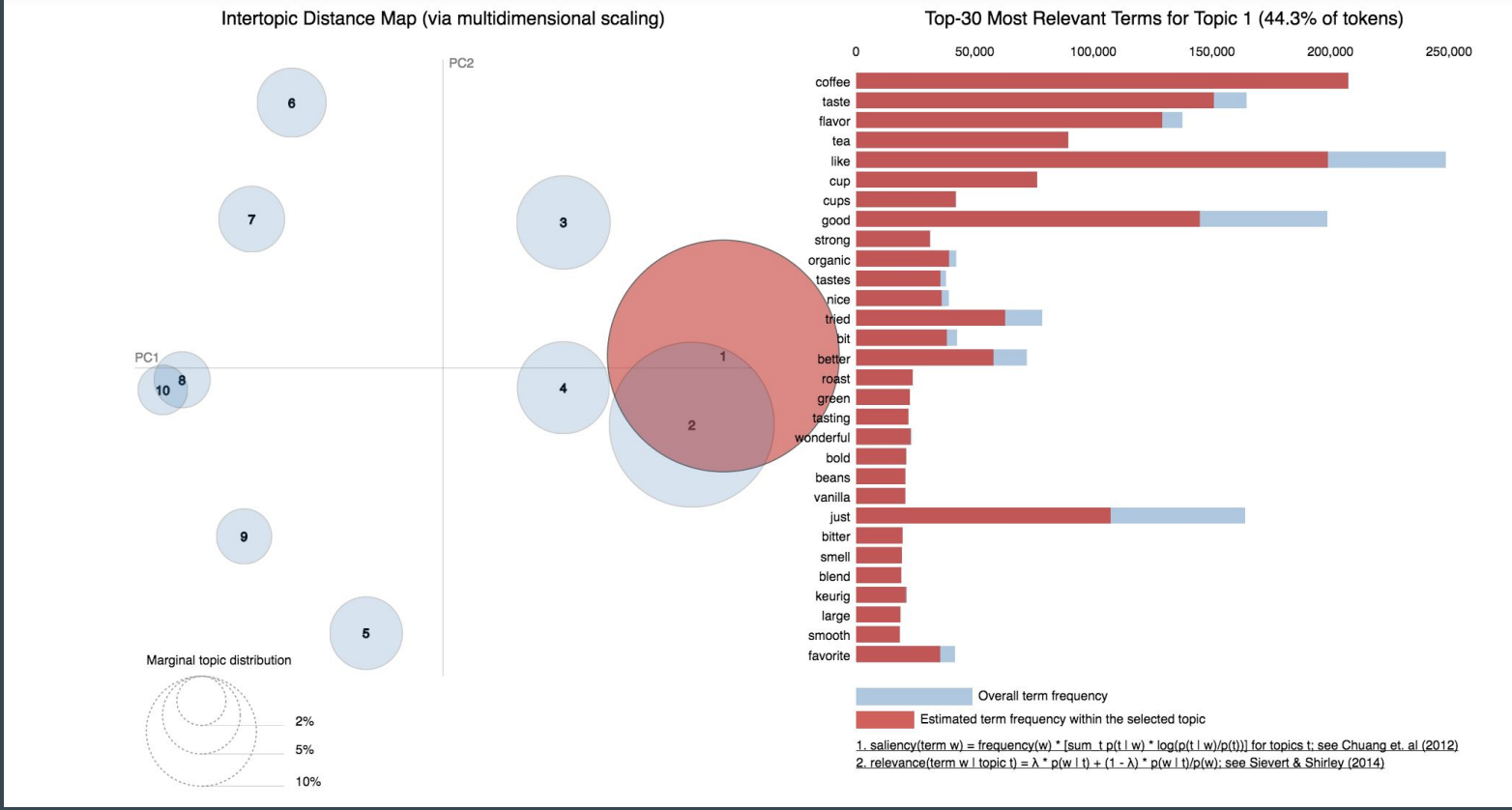
- Logistic Regression
- Latent Dirichlet Allocation
- K-means Clustering

Findings

- Regression
  Accuracy score: ~.78



Underlying categories in the data



Next steps

- Semantic meaning of words (Word2Vec)
- Additional approaches to creating the original classifier (Helpful/ Not Helpful)
- How intent relates to helpfulness

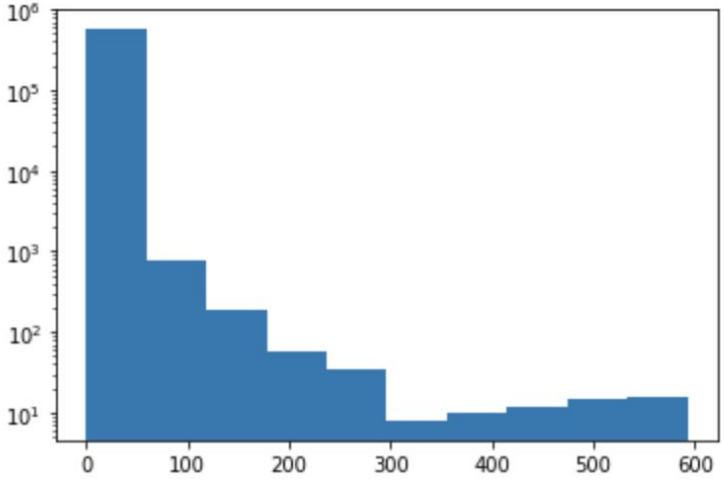
Thank you!

...

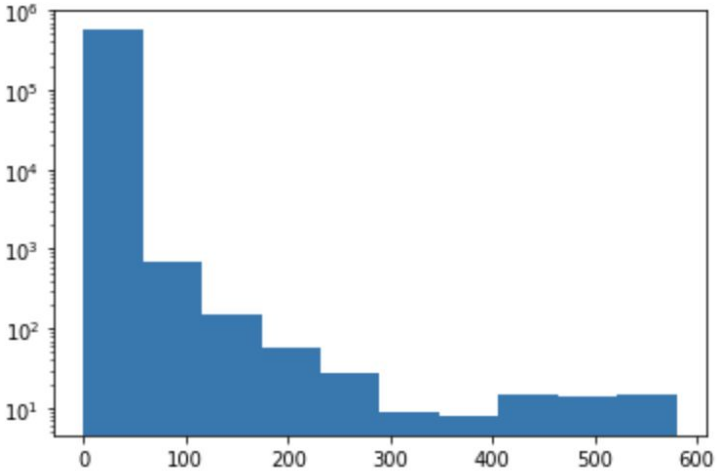
Iris Borkovsky
LinkedIn: https://www.linkedin.com/in/iris-borkovsky
GitHub: https://github.com/borkovsky
Blog: https://medium.com/@borkovsky

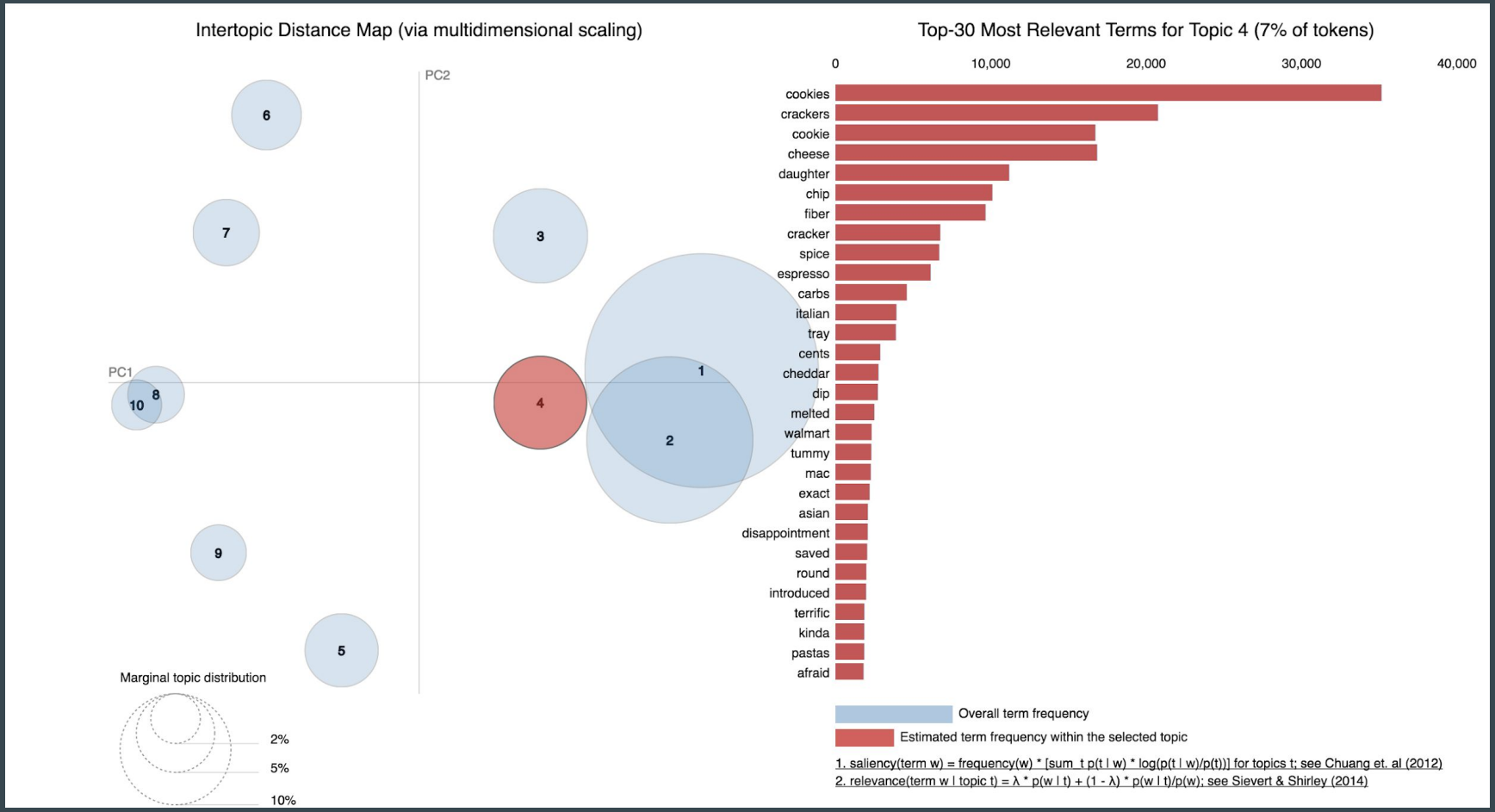
Appendix

Helpfulness Denominator without ourliers



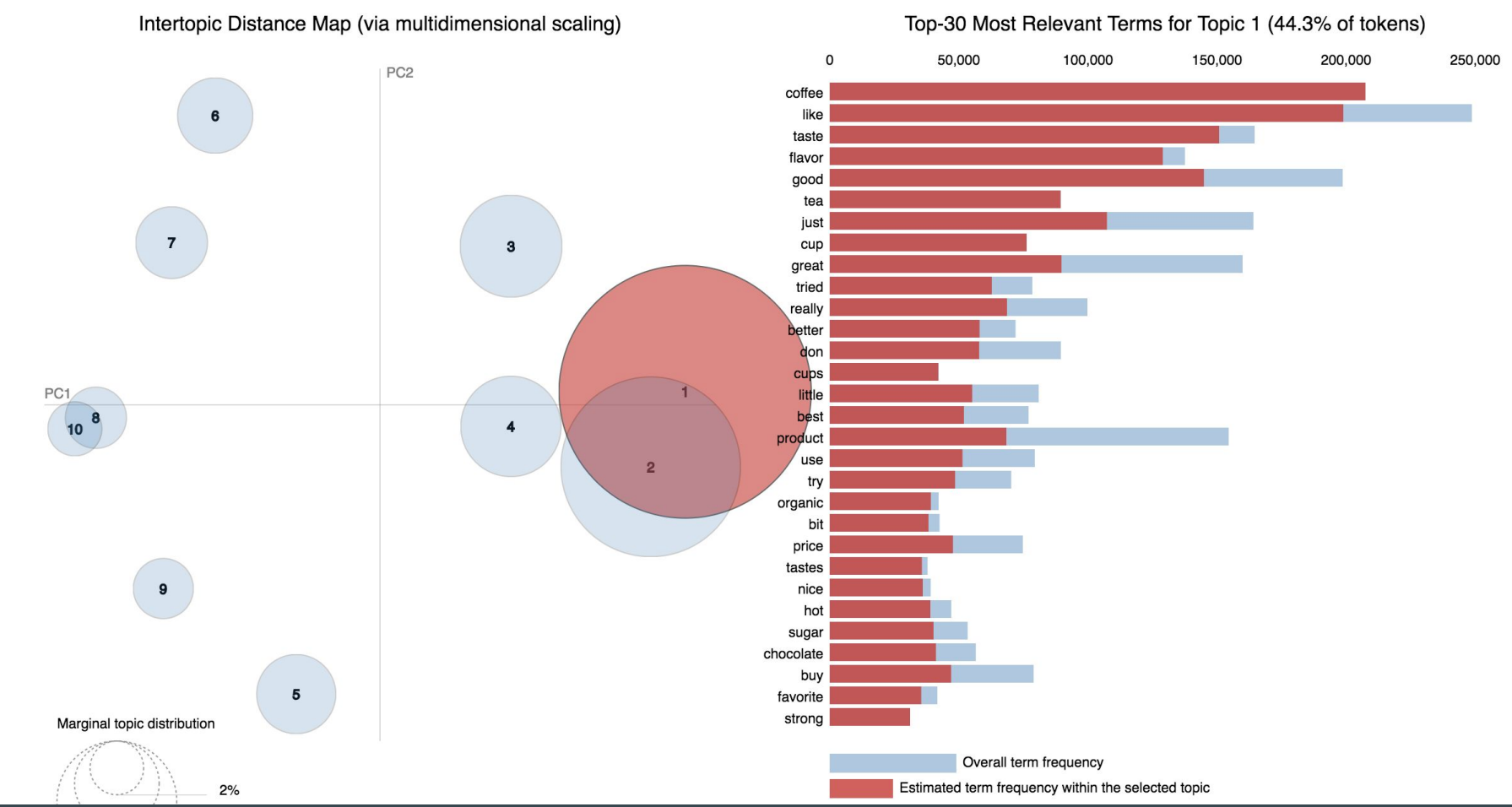
Helpfulness Numerator without ourliers




Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
6
7
3
1
8
10
4
2
9
5
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 4 (7% of tokens)
0
10,000
20,000
30,000
40,000
cookies
crackers
cookie
cheese
daughter
chip
fiber
cracker
spice
espresso
carbs
italian
tray
cents
cheddar
dip
melted
walmart
tummy
mac
exact
asian
disappointment
saved
round
introduced
terrific
kinda
pastas
afraid
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Selected Topic: 1 | Previous Topic | Next Topic | Clear Topic

Slide to adjust relevance metric:(2)
λ = 0.58
0.0 0.2 0.4 0.6 0.8 1



- There are 3,175,298 words total
- Vocabulary: 49,311