Do People Love Their Pets More Than Their Children?

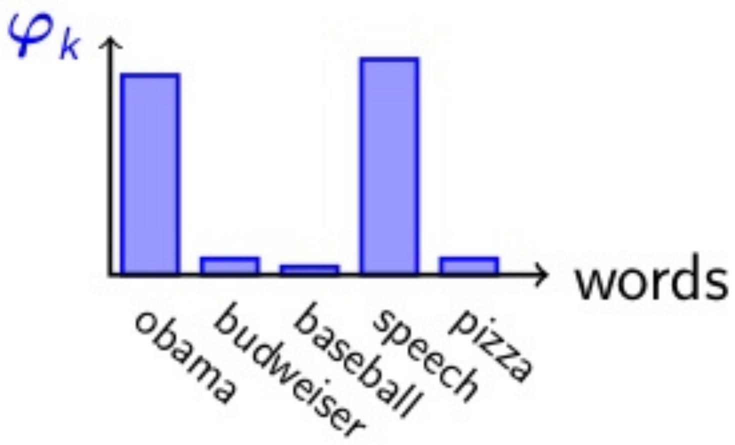
A Topic Analysis of Amazon Reviews

Latent Dirichlet Allocation

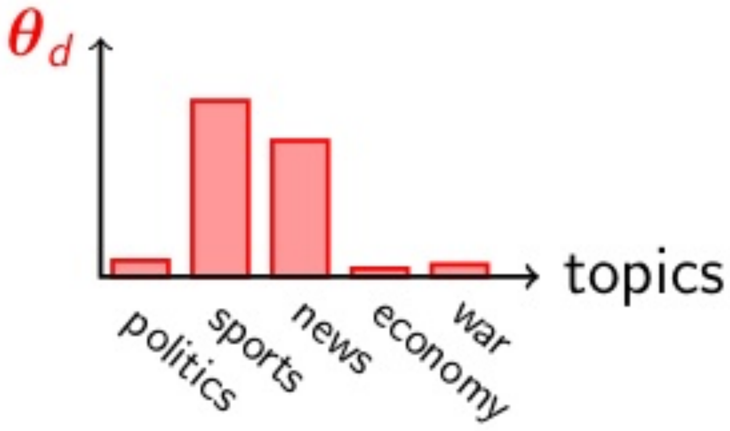
LDA discovers topics into a collection of documents.

LDA tags each document with topics.

Topic k

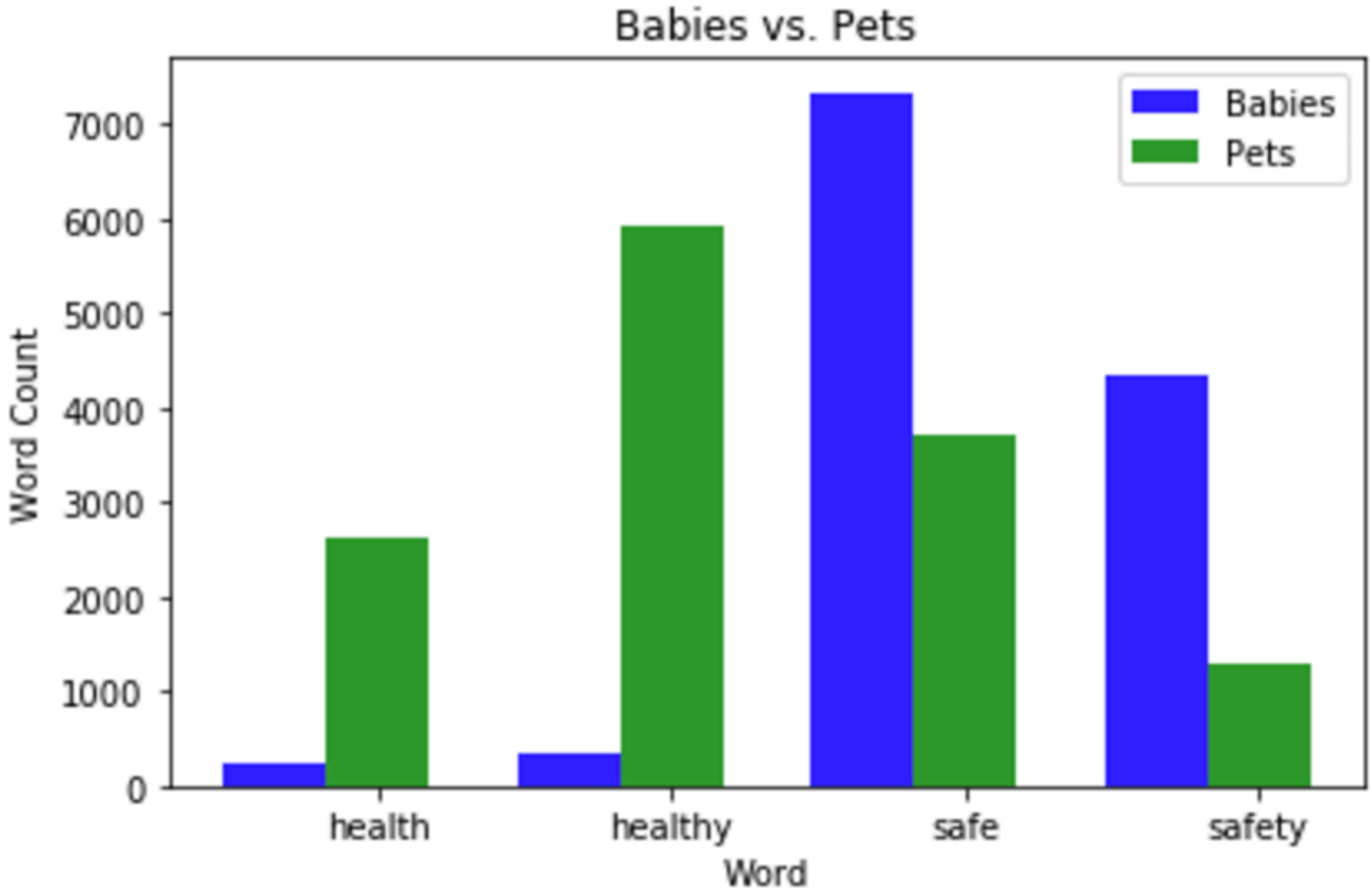


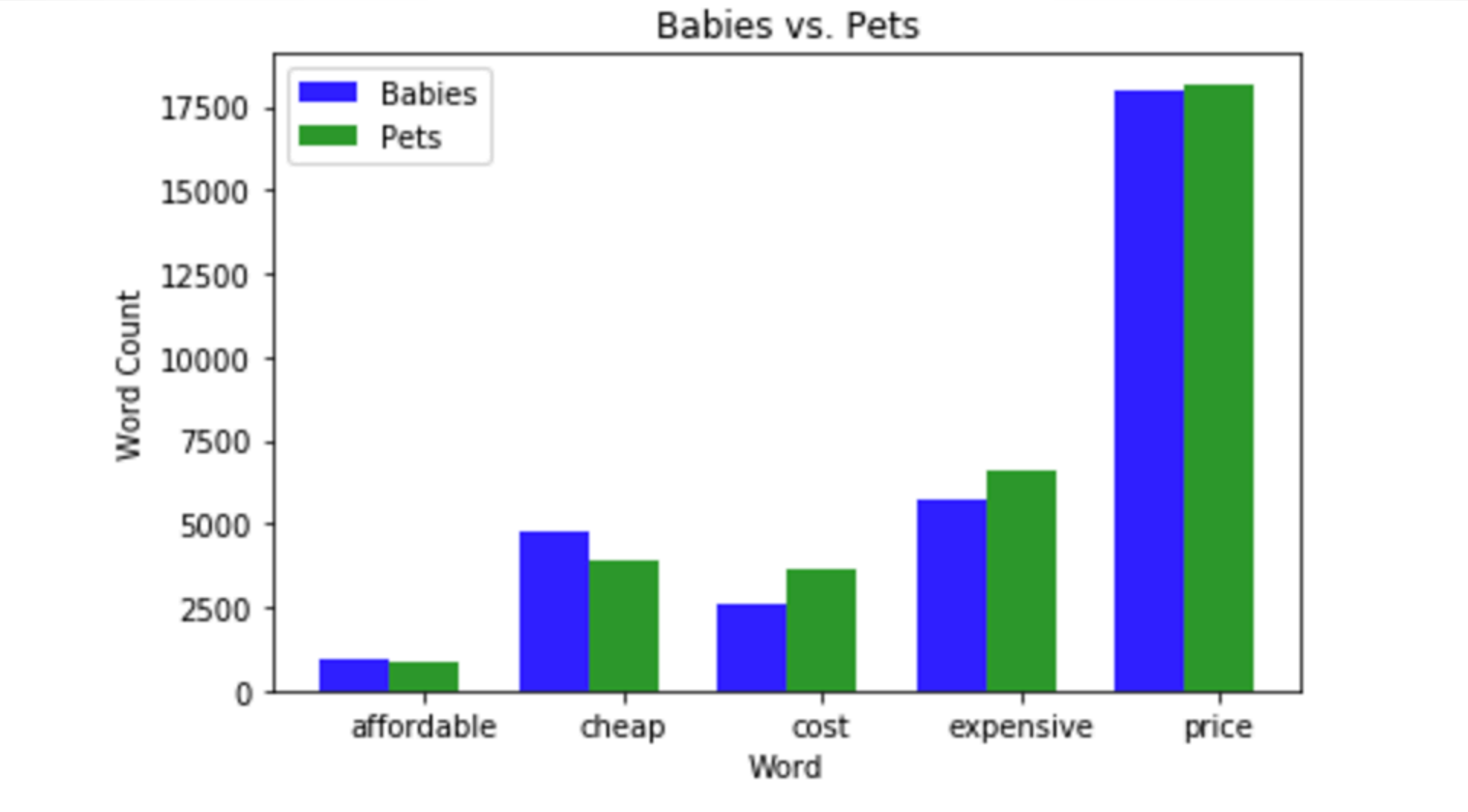
Document d



https://cdn.rawgit.com/davidlichacz/HackOnData2017/7f5414d0/
baby_lda_vis.html#topic=0&lambda=1&term=

What is love?


Babies vs. Pets
Babies
Pets
Word Count
0
1000
2000
3000
4000
5000
6000
7000
health
healthy
safe
safety
Word


Babies vs. Pets
Babies
Pets
Word Count
17500
15000
12500
10000
7500
5000
2500
0
affordable
cheap
cost
expensive
price
Word

Steps

- Tokenize

- Remove stop words

- Lemmatize: Lemmatization is the process of grouping together the inflected forms of a word so they can be analysed as a single item identified by the word's dictionary form (lemma)

Steps

- Create a document-term matrix: Describes the frequency of terms that occur in a collection of documents
- Fit the LDA model
- Query the model
- Determine the most common topics

Top 5 Review Topics

Babies	Pets
Generic	Pet Food
Toys/Teethers	Generic
Gates	Chew Toys
Baby Bottles	Fish Tanks
High Chairs	Leashes/Harnesses

Next steps

- More complete dataset
- Parameter tuning
- Bigrams
- More complex queries

Productionalizing

Productionalizing

- Part of a recommender system

- Social media analysis

- Email filtering

https://cdn.rawgit.com/davidlichacz/HackOnData2017/7f5414d0/pets_lda_vis.html#topic=0&lambda=1&term=