

Amazon Reviews Analysis

Carolina Gonzalez

Business Case

- Amazon Reviews is categorized as a product in itself
- Customer reviews offer insights about:
 - Who is buying the products
 - What is important for customers
 - Active communities
 - Customer segmentation

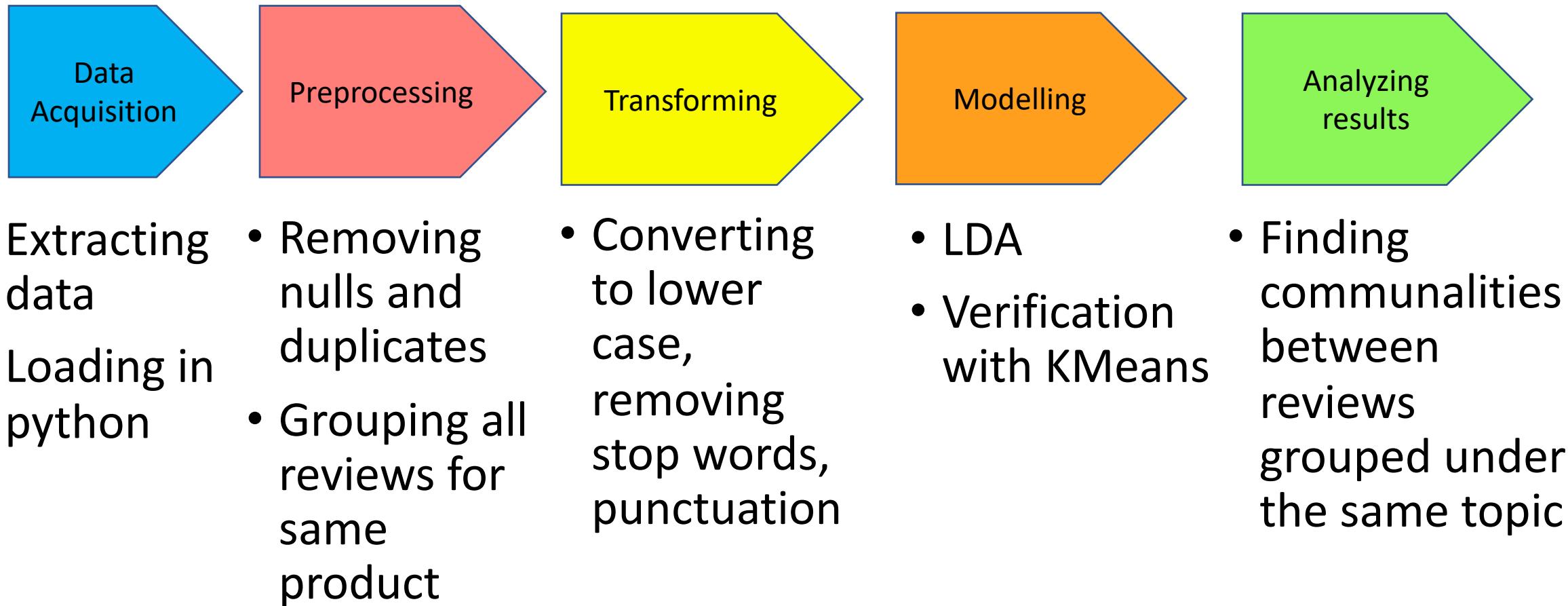


Data

- A collection of reviews written in the Amazon.com marketplace and associated metadata from 1995 until 2015
 - Amazon Customer Reviews Dataset available in S3
 - Sports category
 - 4,832,539 reviews
- Prices from a selected topic:
 - Scrapped using Requests in python



Topic pipeline



THE MODEL

- Latent Dirichlet Allocation (LDA) for topic identification
 - Topic 0: Teams and Gifts
 - Topic 1: Airsoft Guns
 - Topic 2: Clothes
 - Topic 3: Fishing
 - Topic 4: Workout routines
- Evaluating potential relationship with product price
 - Part of speech regression, vocab regression and classification, neural networks



Topic 0: Teams and Gifts

Sally the family shopper:

- She is likely married
- She does her **Christmas** shopping online
- She is likely to buy sports apparel
- She often buys **gifts** for her **husband** and **kids**
- **Arrival** time is important



Topic 1: Airsoft

John the enthusiast

- He is very is very much into **airsoft guns**
- He would write about specific features of the product.
- **Quality** is important
- **Price** is a factor



Conclusions

- Amazon reviews for the sport category were split into topics
- Five distinct topics were identified
- The language in each topics was used to identify commonalities and potential reviewers characteristics
- There is not evidence of relationship between product price and reviews

Future work:

- Further examine recommendation systems and cross selling



Tools

- Spyder
- Requests
- GENSIM
- NLTK
- KMEANS
- LDA
- LDavis
- keras.text_to_word_sequence
- sklearn.cross_validation
- sklearn.model_selection
- sklearn.preprocessing
- TfidfVectorizer
- LSA (TruncatedSVD)
- Word2Vec
- GridSearchCV
- keras.engine
- keras.layers:Dense, Embedding, Conv1D, MaxPooling1D, Flatten
- GradientBoostingClassifier
- sklearn.linear_model
- sklearn.naive_bayes

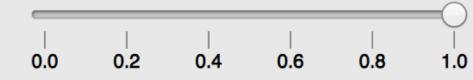




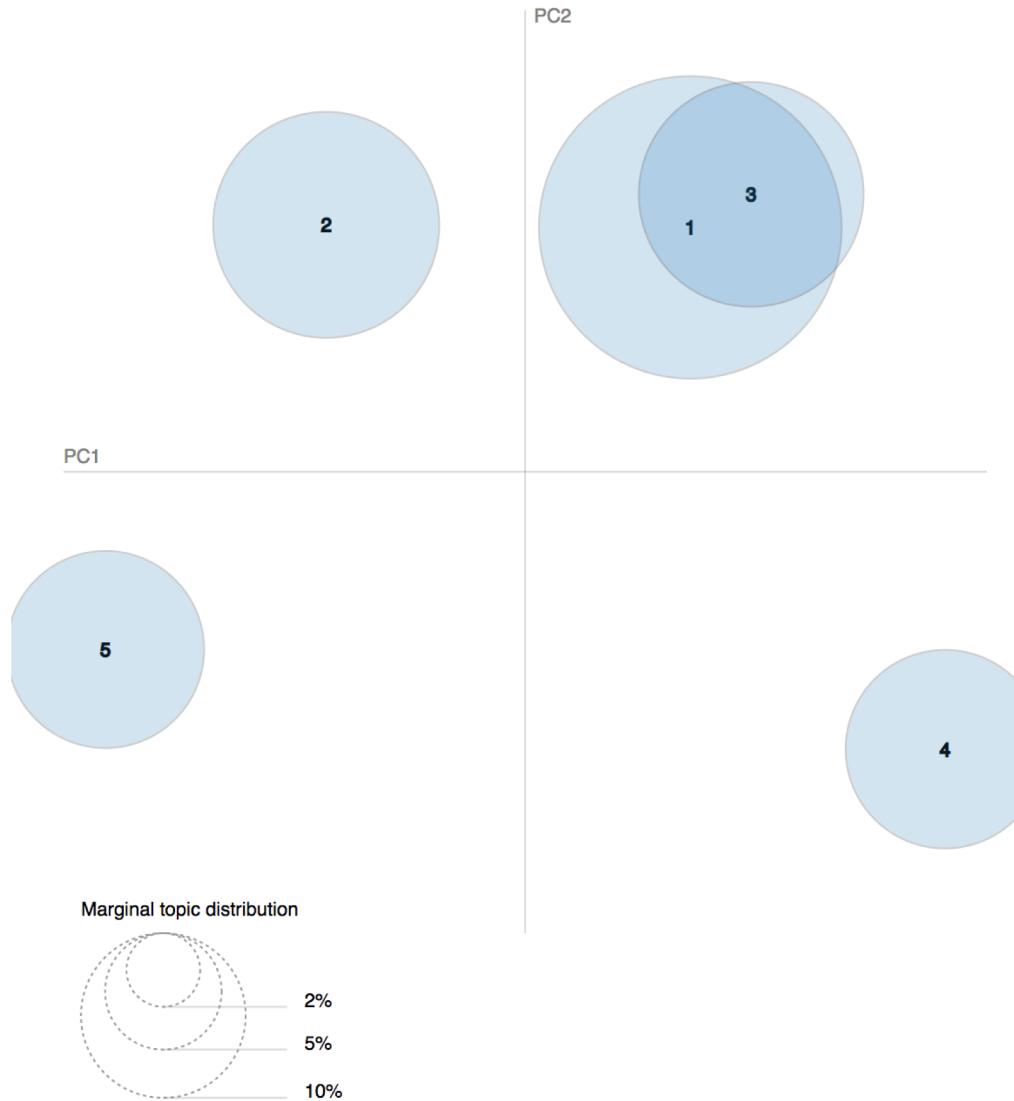
Topic 0: Teams and Gifts

Slide to adjust relevance metric:⁽²⁾

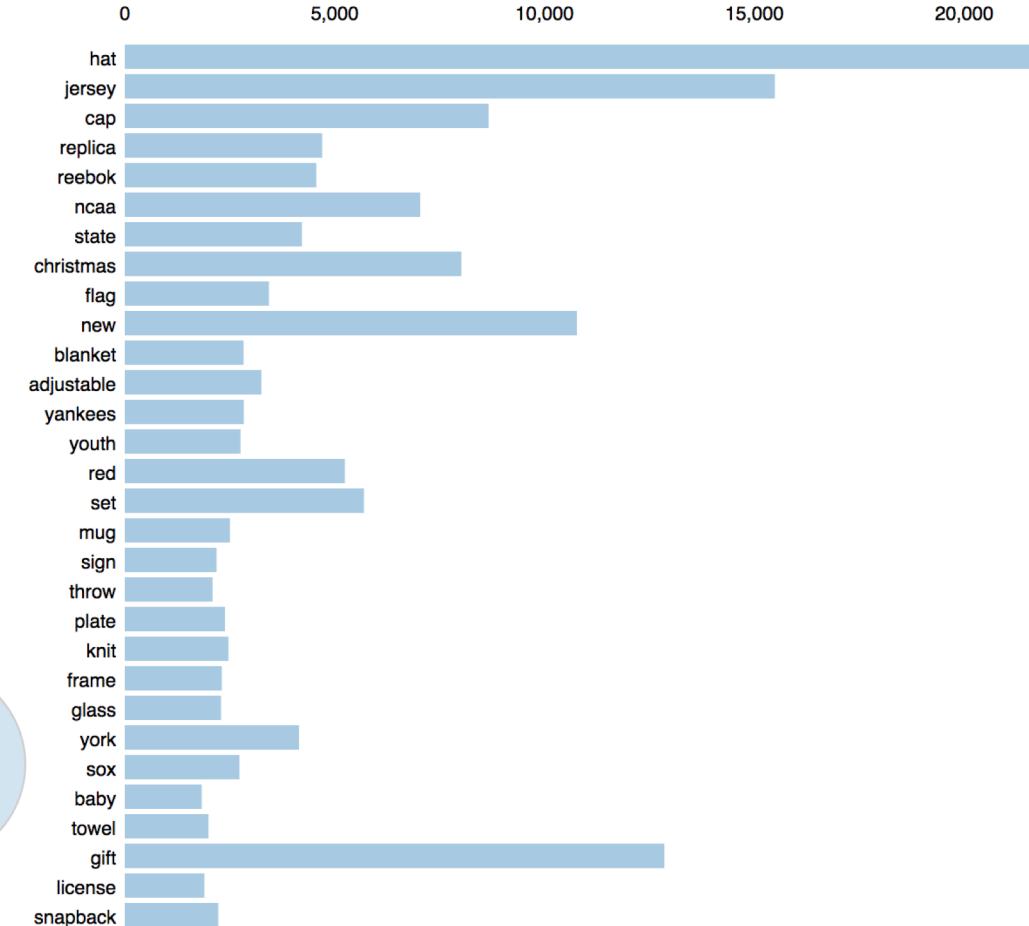
$\lambda = 1$



Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Salient Terms¹



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) = $\lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$; see Sievert & Shirley (2014)

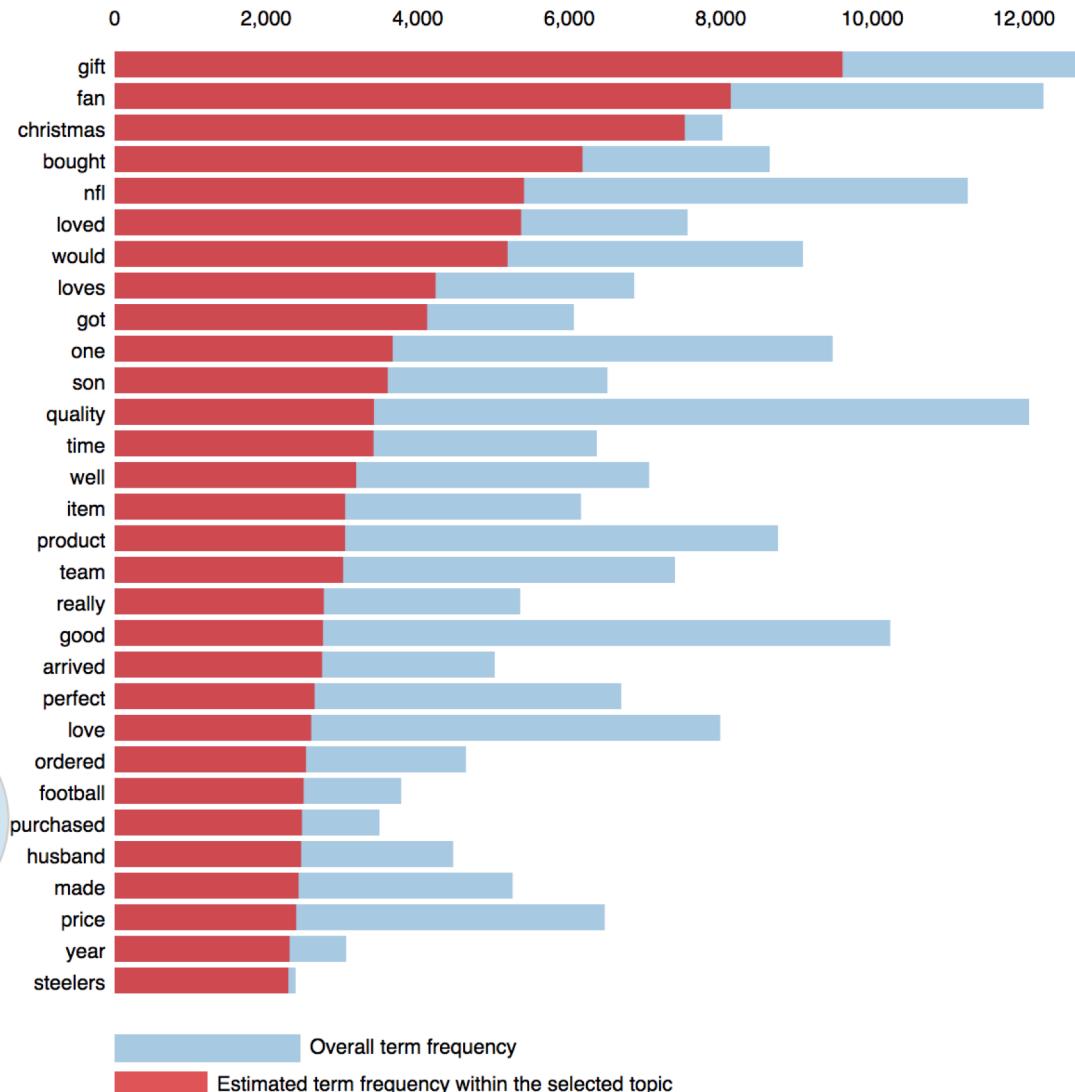
Topic 0: Teams and Gifts Subtopics

Slide to adjust relevance metric:(2)

$\lambda = 1$



Top-30 Most Relevant Terms for Topic 1 (33.7% of tokens)



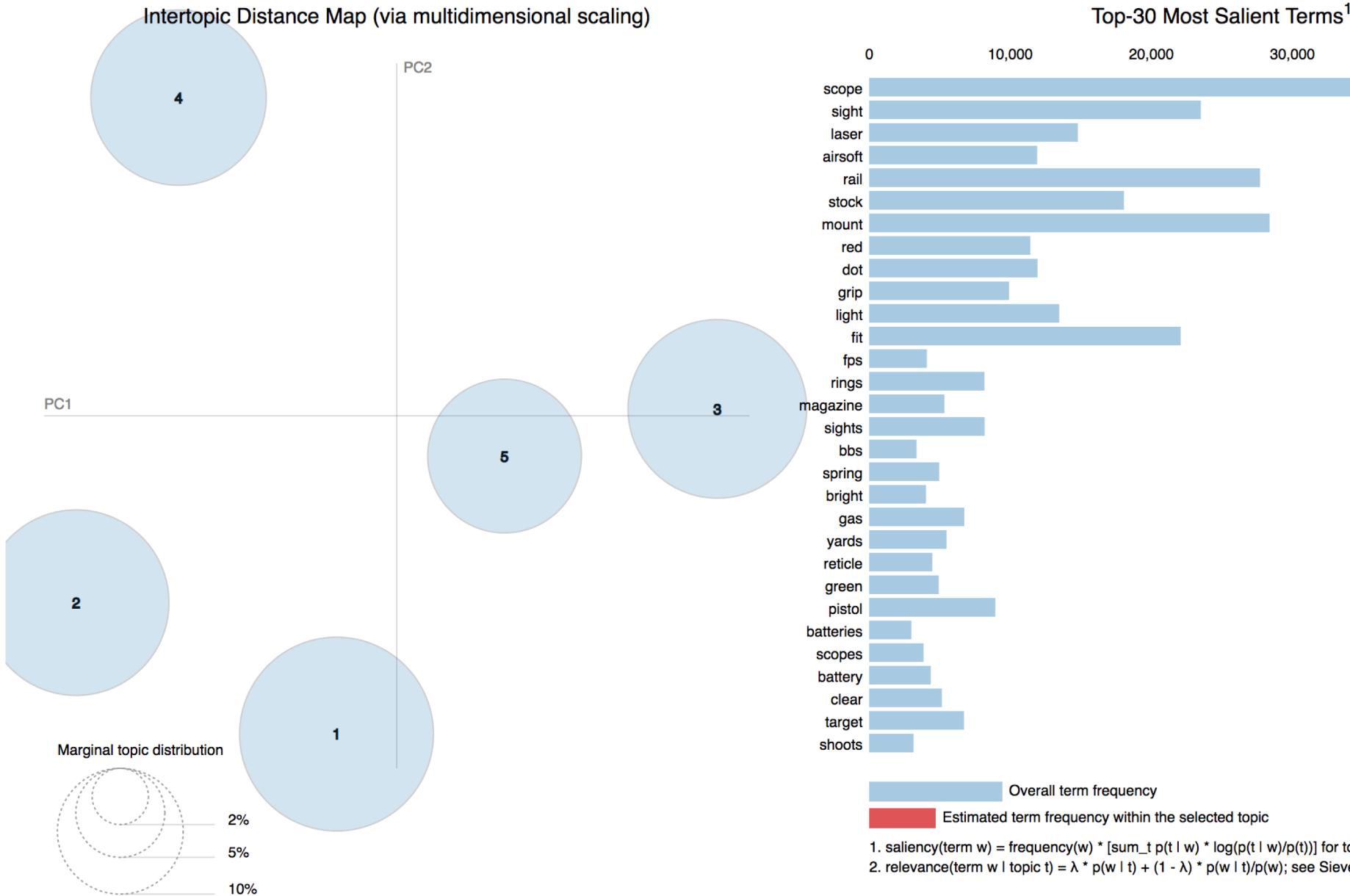
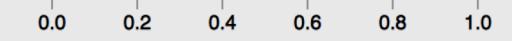
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) = $\lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$; see Sievert & Shirley (2014)

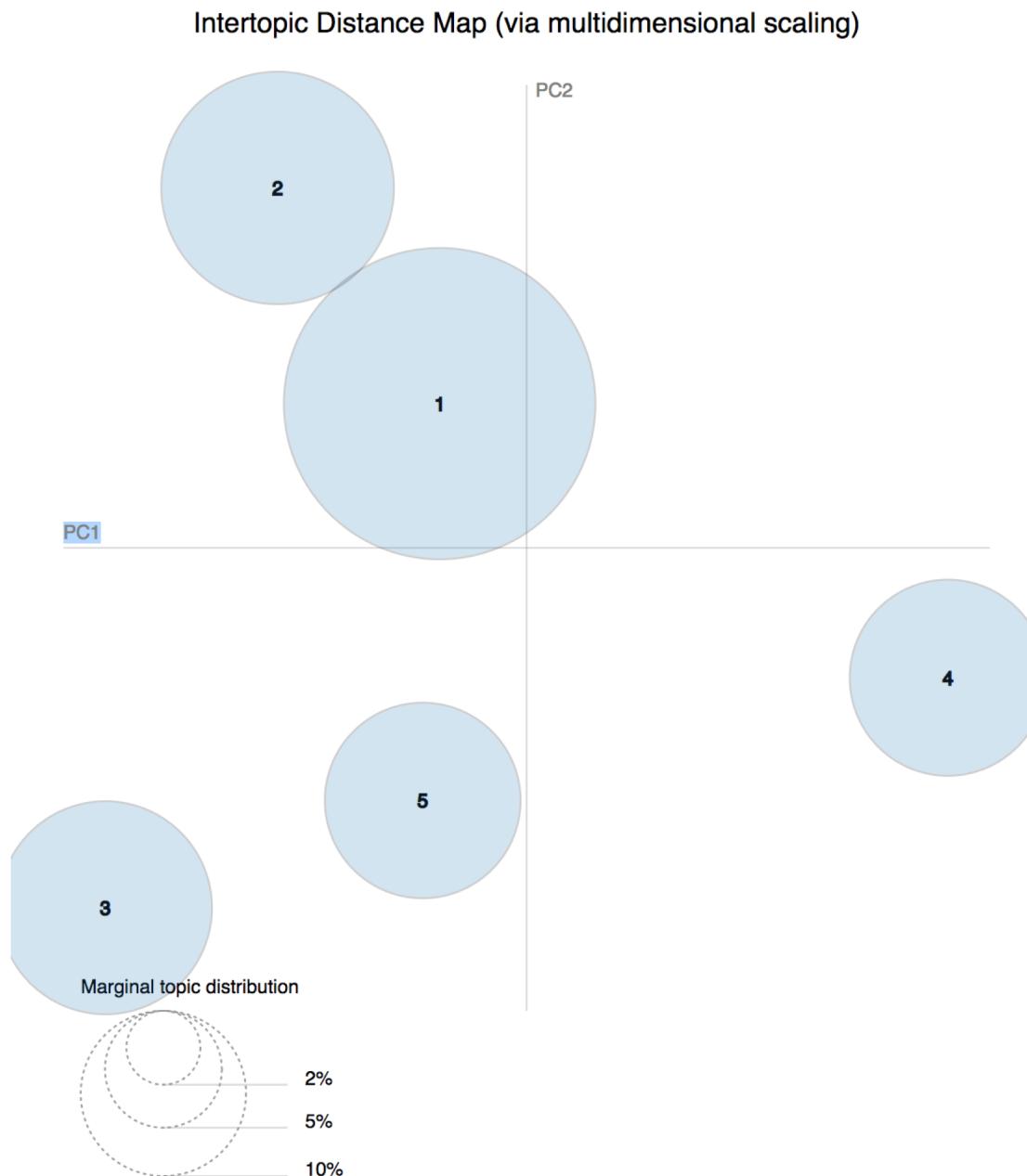
Topic 1: AirSoft Guns

Slide to adjust relevance metric:(2)

$\lambda = 1$



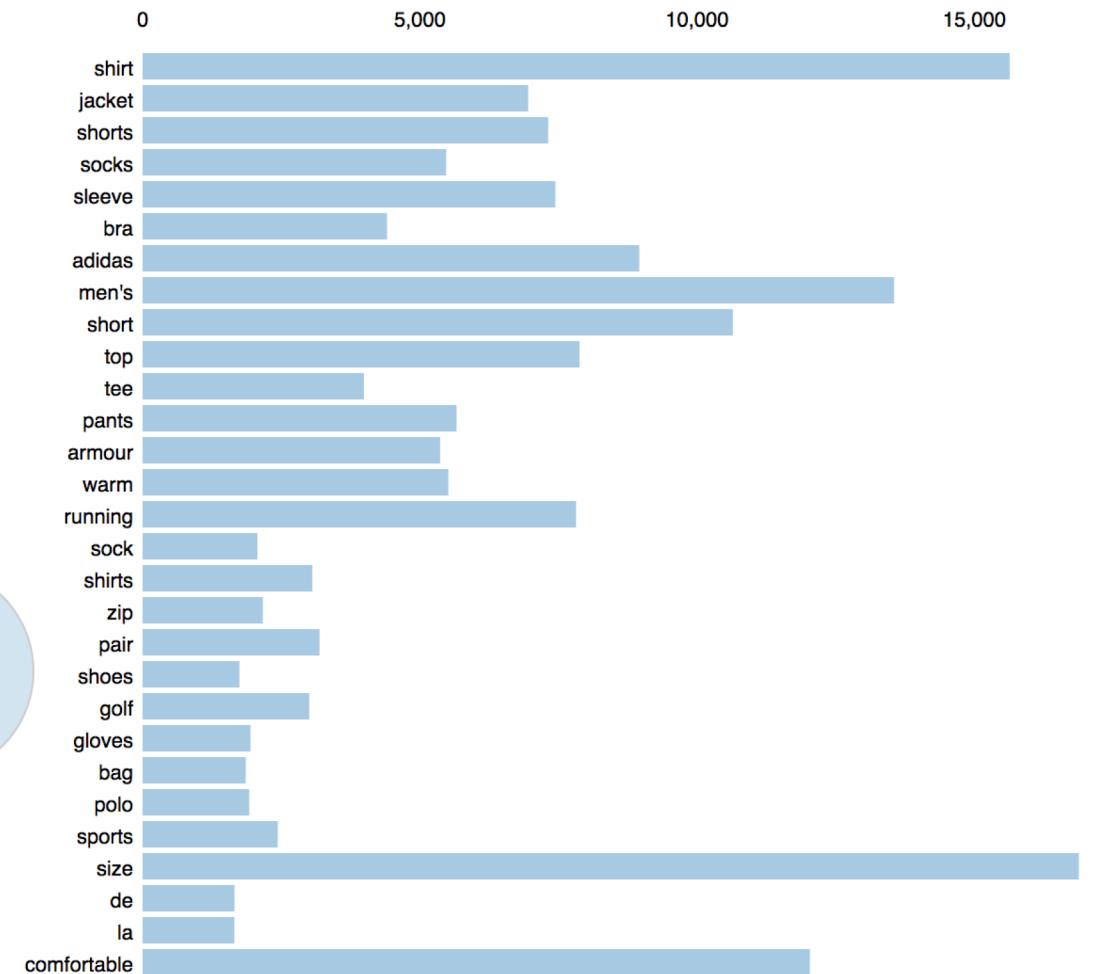
Topic 2: Clothes



Slide to adjust relevance metric:⁽²⁾

$\lambda = 1$

Top-30 Most Salient Terms⁽¹⁾



Overall term frequency

Estimated term frequency within the selected topic

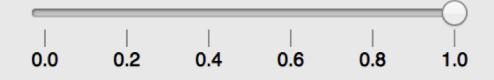
1. $\text{salency}(\text{term } w) = \text{frequency}(w) * [\sum_t p(t|w) * \log(p(t|w)/p(t))]$ for topics t ; see Chuang et. al (2012)

2. $\text{relevance}(\text{term } w | \text{topic } t) = \lambda * p(w|t) + (1 - \lambda) * p(w|t)/p(w)$; see Sievert & Shirley (2014)

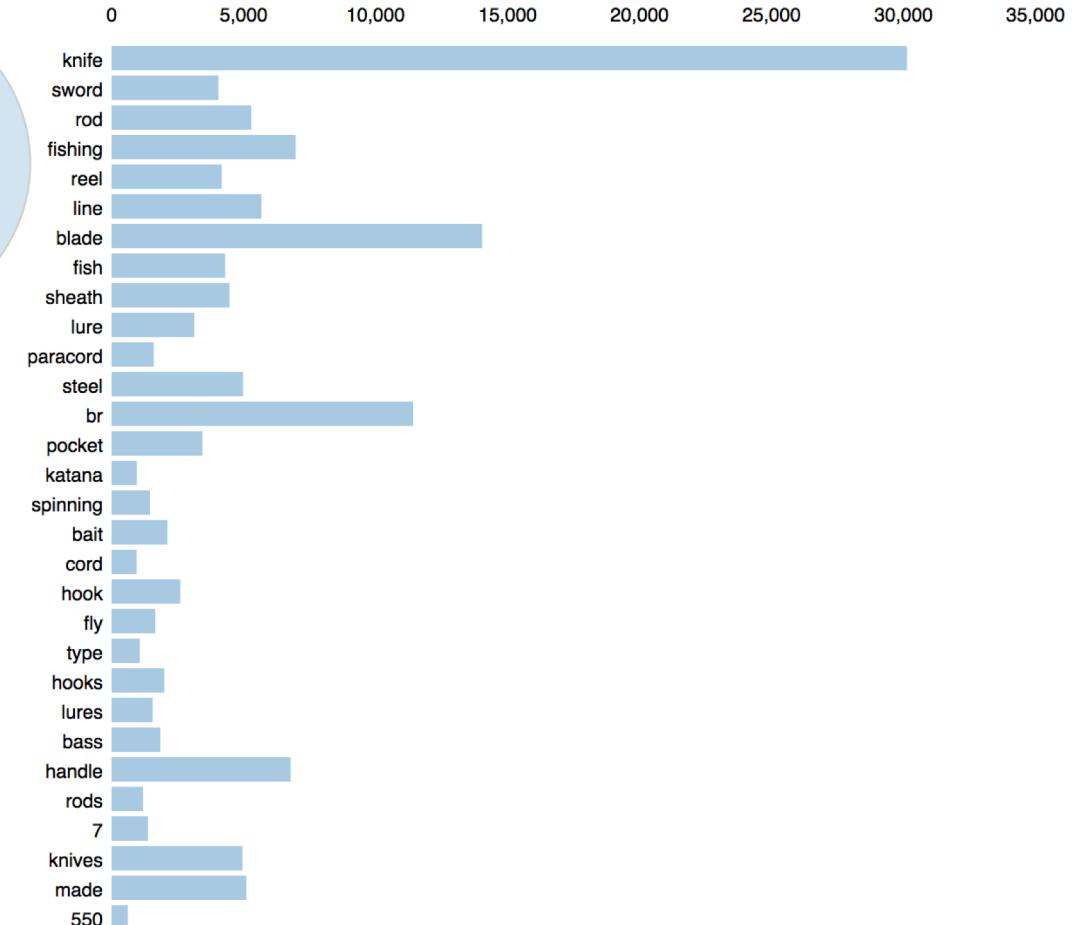
Topic 3: Fishing and Knives

Slide to adjust relevance metric:⁽²⁾

$\lambda = 1$



Top-30 Most Salient Terms¹



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) = $\lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$; see Sievert & Shirley (2014)

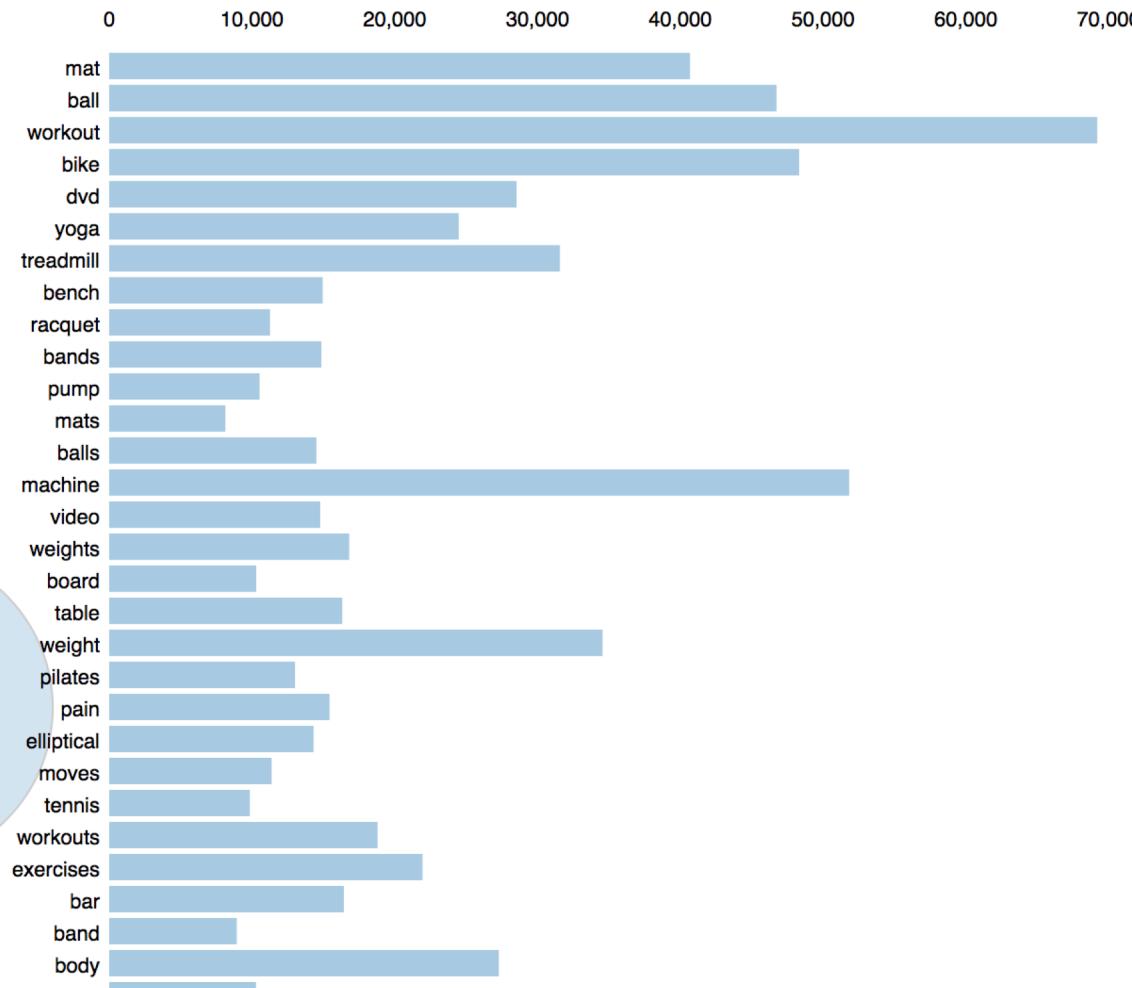
Topic 4: Workout Routines



Slide to adjust relevance metric:(2)

$\lambda = 1$

Top-30 Most Salient Terms⁽¹⁾



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) = $\lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$; see Sievert & Shirley (2014)

Topic 2: Clothes

Kathy the golfer

- The most important things about her purchase are comfort and size
- Materials are important
- She cares about staying warm
- Price is not the main concern in her clothing purchases



Topic 3: Fishing and knives

Paul the fisherman

- Quality and experience are important
- Knives are very big portion of this topic
- Pocket knives and swords are different subcategories
- Price is part of his vocabulary



Topic 4: Workout routines

Home gym Jeff

- He would buy exercise equipment
- Workout DVDs
- Play tennis and sometimes practice yoga.
- This topic could probably be subdivided

