



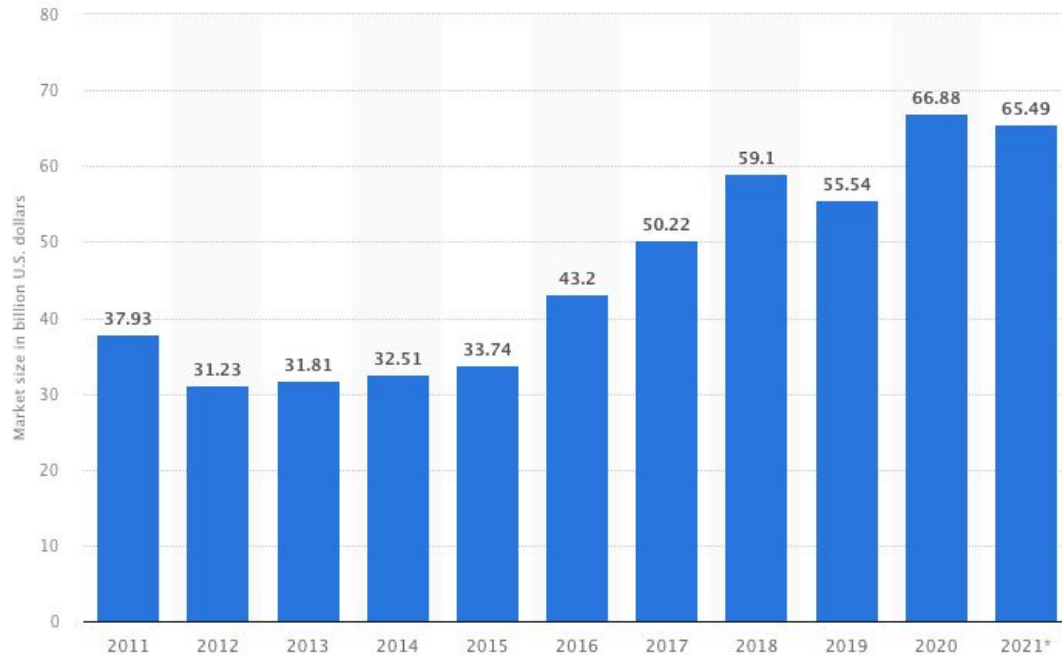
# Video Games Sentiment Analysis

---

A recommendation system for marketing team to analyze reviews

By Jason W

# Video Games Industry Market Size



Video Games Industry has been rapidly growth over the past decade and almost double the market size from 2011 to 2021

# PROBLEM



## Pricing

Too many similar products on the market with competitive pricing

Inconsistencies, glitches, bugs, bad quality assurance

## Quality



## Lifecycle

Some products are designed for only short term

# GOAL

Build a **Recommendation System** to helps marketing team



Figure out what users most **important needs** for each product category



**Competitive** product features advantage among competitors



Longevity of product **lifecycle** by picking the right materials

# DATA & PRE-PROCESSING



## Dataset

~2M Video Games  
reviews from Amazon  
(picking 10k reviews  
randomly from the  
dataset)



## Tools

Python  
Pandas  
Numpy  
NLTK  
Scikit-learn



## Text Processing

CountVectorizer  
Stemming  
Lemmatization



## Topic Model

NMF  
LDA



## Eval & Vis

Matplotlib  
Seaborn  
WordCloud  
pyLDAvis

# TOPIC MODEL

## TOPIC MODEL

15 TOPICS in total

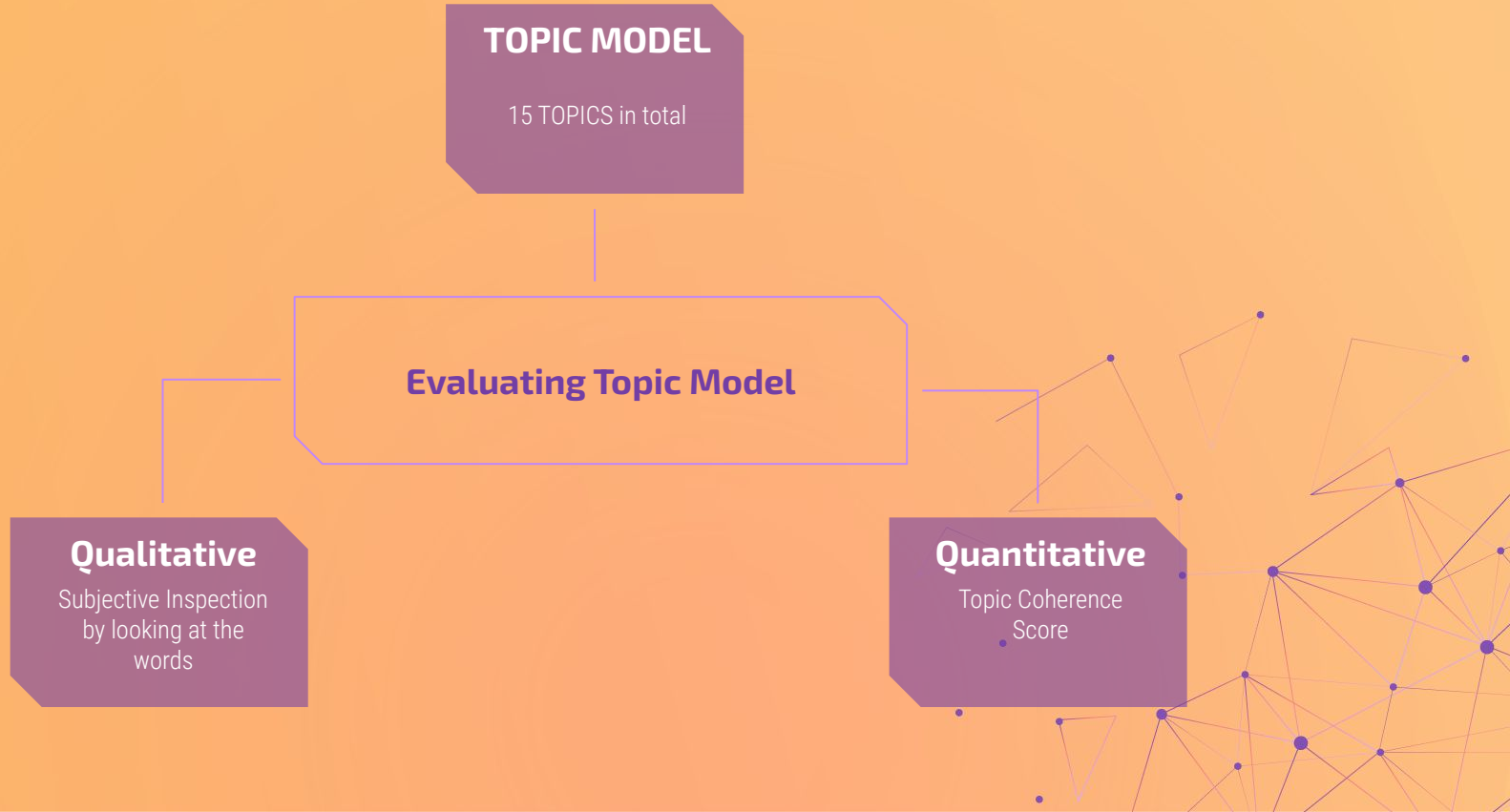
## Evaluating Topic Model

### Qualitative

Subjective Inspection  
by looking at the  
words

### Quantitative

Topic Coherence  
Score



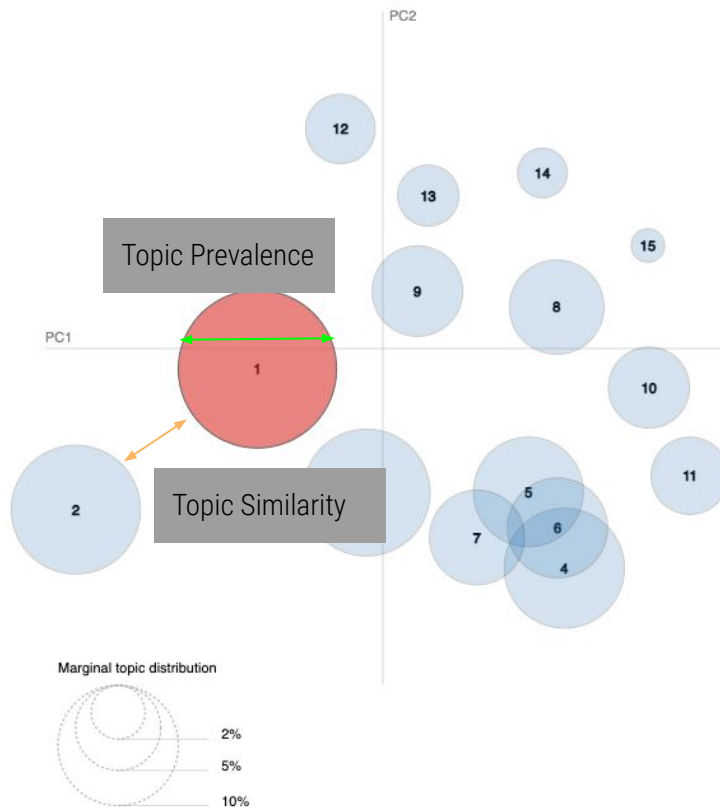
Selected Topic:

Slide to adjust relevance metric:(2)

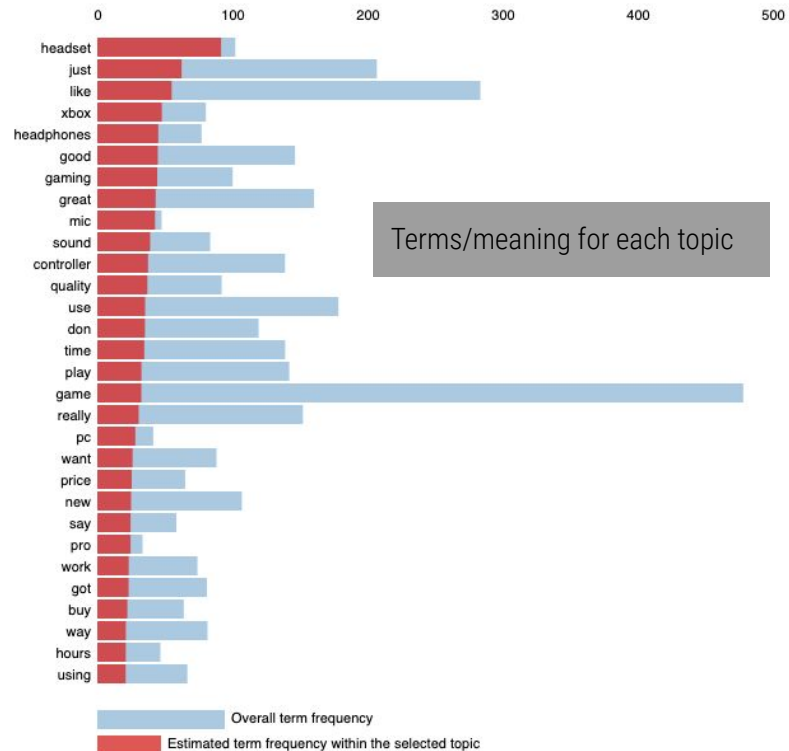
$\lambda = 1$

0.0 0.2 0.4 0.6 0.8 1.0

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 1 (17.2% of tokens)



1.  $\text{saliency}(\text{term } w) = \text{frequency}(w) * [\sum_t p(t | w) * \log(p(t | w)/p(t))]$  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)

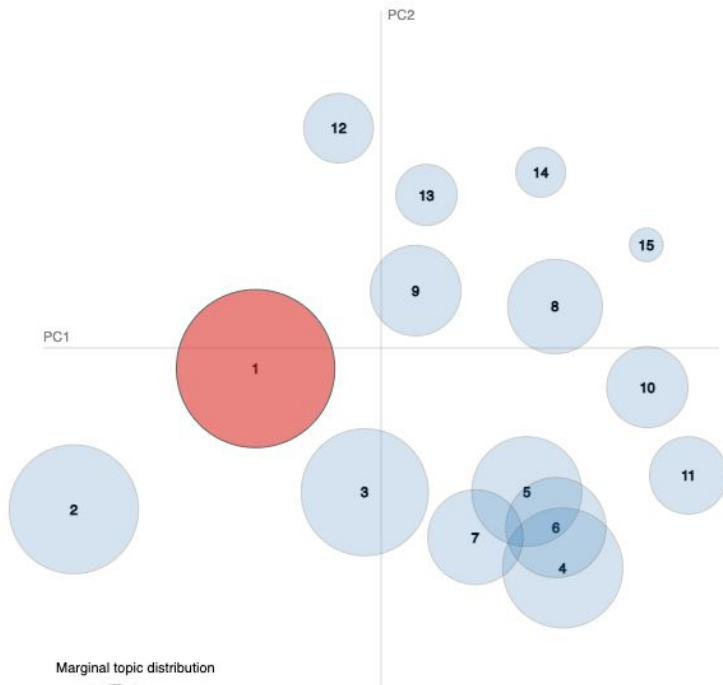
Selected Topic:

Slide to adjust relevance metric:<sup>(2)</sup>

$\lambda = 1$

0.0 0.2 0.4 0.6 0.8 1.0

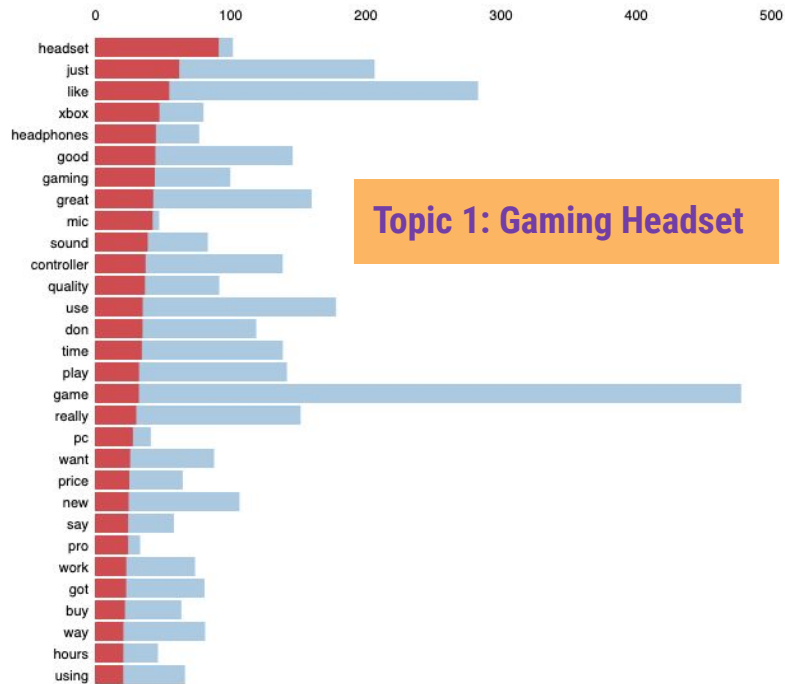
Intertopic Distance Map (via multidimensional scaling)



Marginal topic distribution



Top-30 Most Relevant Terms for Topic 1 (17.2% of tokens)



Topic 1: Gaming Headset

Overall term frequency  
Estimated term frequency within the selected topic

1.  $\text{saliency}(\text{term } w) = \text{frequency}(w) * [\sum_t p(t | w) * \log(p(t | w)/p(t))]$  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)



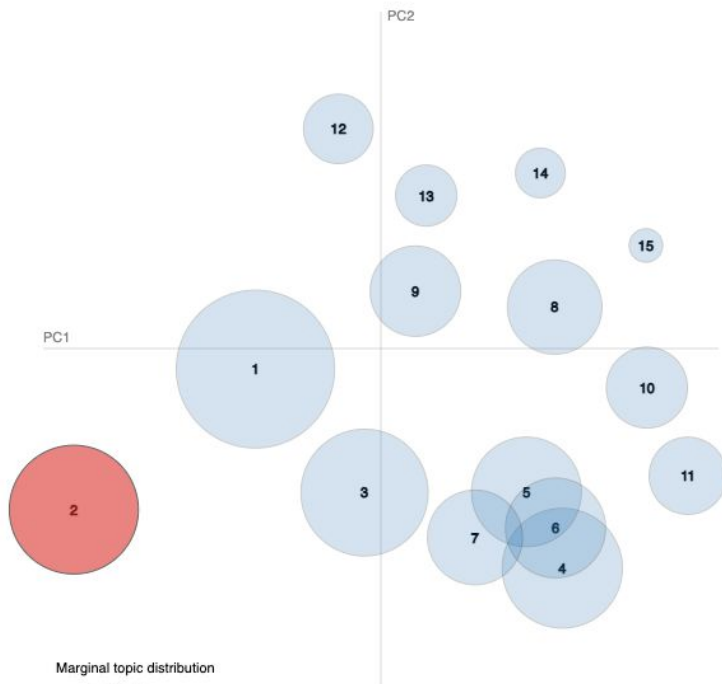
Selected Topic:

Slide to adjust relevance metric:<sup>(2)</sup>

$\lambda = 1$



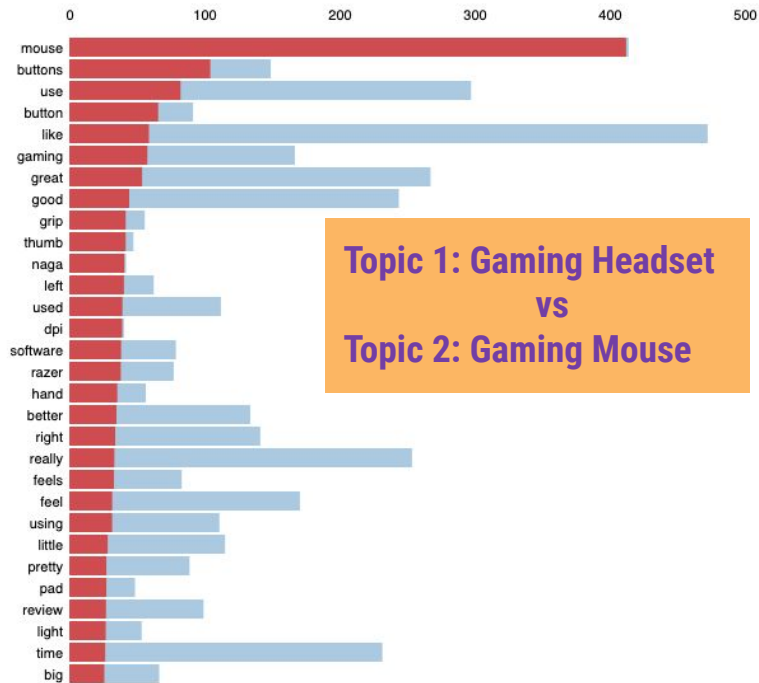
Intertopic Distance Map (via multidimensional scaling)



Marginal topic distribution



Top-30 Most Relevant Terms for Topic 2 (11.5% of tokens)



**Topic 1: Gaming Headset**  
**VS**  
**Topic 2: Gaming Mouse**

Overall term frequency  
Estimated term frequency within the selected topic

1.  $\text{saliency}(\text{term } w) = \text{frequency}(w) * [\sum_t p(t|w) * \log(p(t|w)/p(t))]$  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)

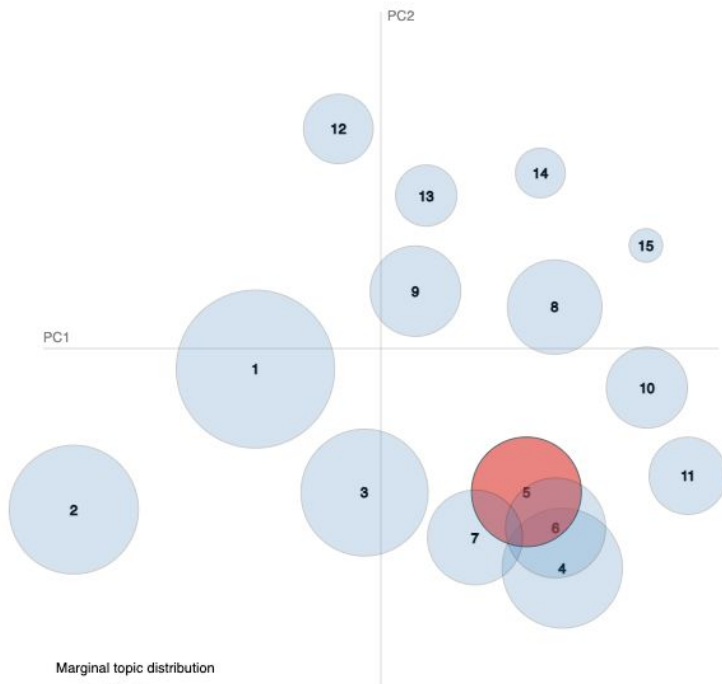
Selected Topic:

Slide to adjust relevance metric:<sup>(2)</sup>

$\lambda = 1$

0.0 0.2 0.4 0.6 0.8 1.0

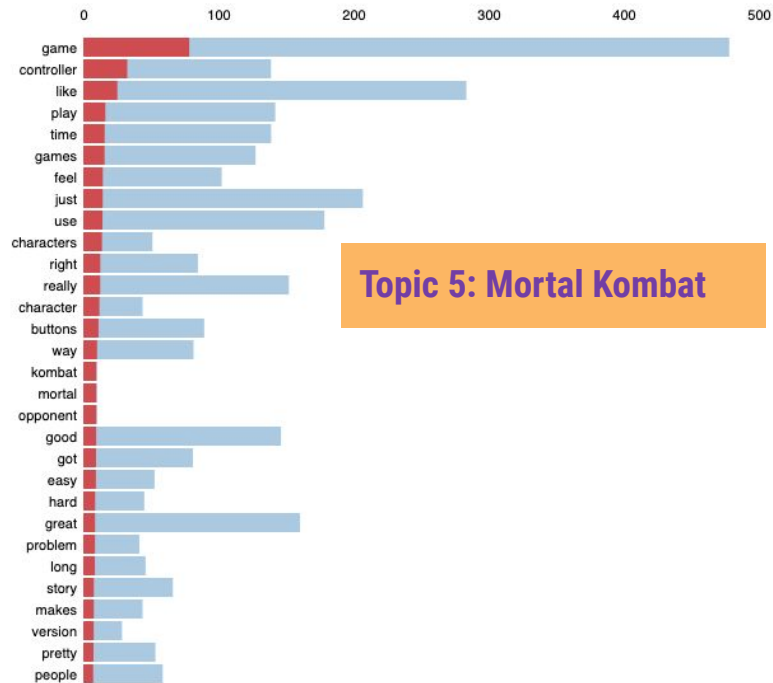
Intertopic Distance Map (via multidimensional scaling)



Marginal topic distribution



Top-30 Most Relevant Terms for Topic 5 (8.3% of tokens)



Topic 5: Mortal Kombat

Overall term frequency  
Estimated term frequency within the selected topic

1.  $\text{saliency}(\text{term } w) = \text{frequency}(w) * [\sum_t p(t|w) * \log(p(t|w)/p(t))]$  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)

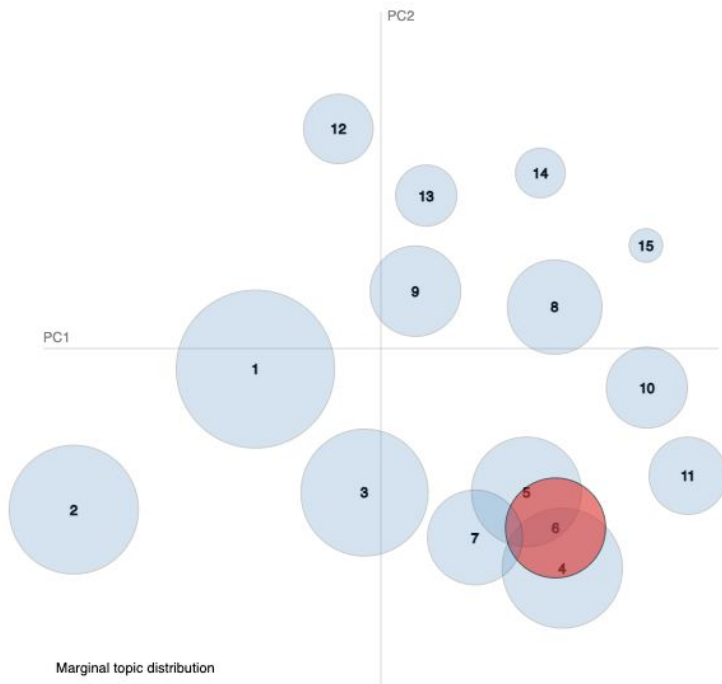
Selected Topic:

Slide to adjust relevance metric:<sup>(2)</sup>

$\lambda = 1$

0.0 0.2 0.4 0.6 0.8 1.0

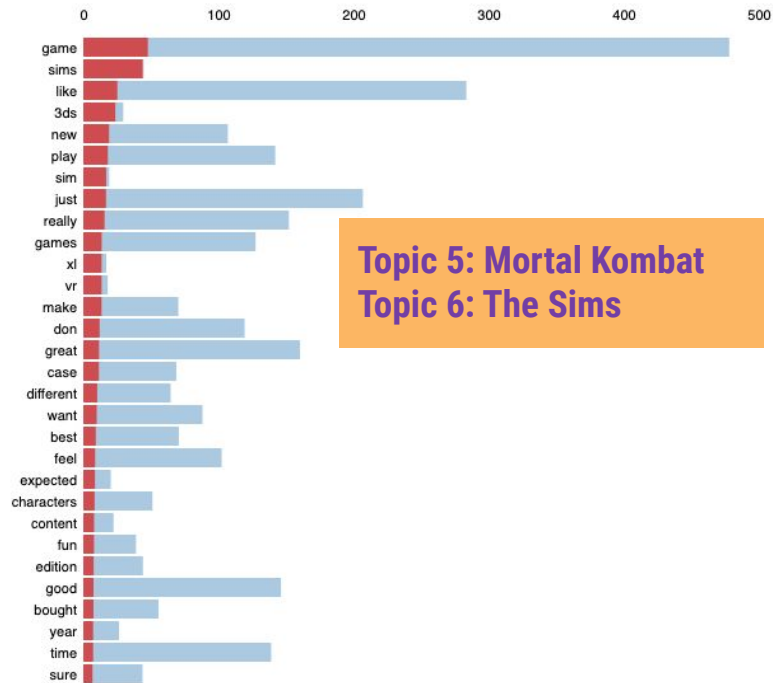
Intertopic Distance Map (via multidimensional scaling)



Marginal topic distribution



Top-30 Most Relevant Terms for Topic 6 (6.9% of tokens)



Topic 5: Mortal Kombat  
Topic 6: The Sims

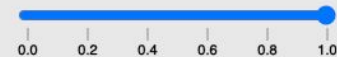
Overall term frequency  
Estimated term frequency within the selected topic

1.  $\text{saliency}(\text{term } w) = \text{frequency}(w) * [\sum_t p(t | w) * \log(p(t | w)/p(t))]$  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)

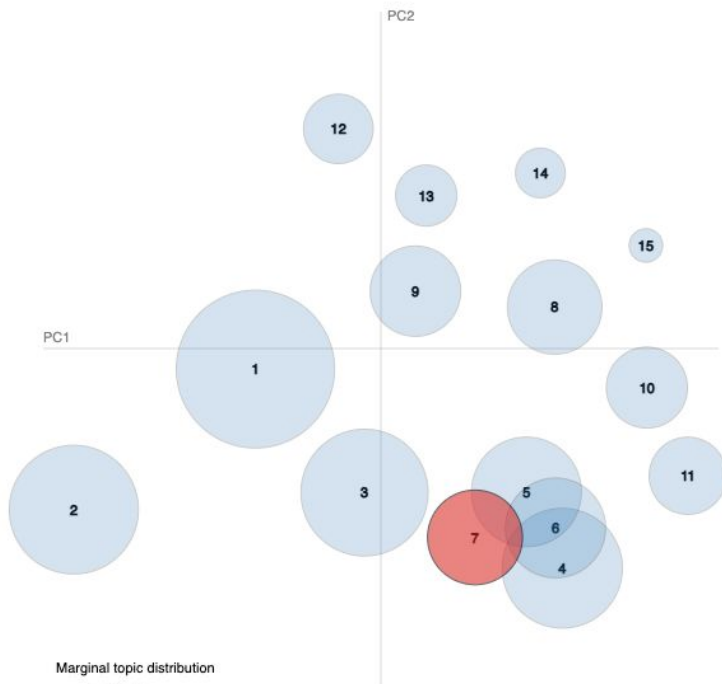
Selected Topic:

Slide to adjust relevance metric:<sup>(2)</sup>

$\lambda = 1$



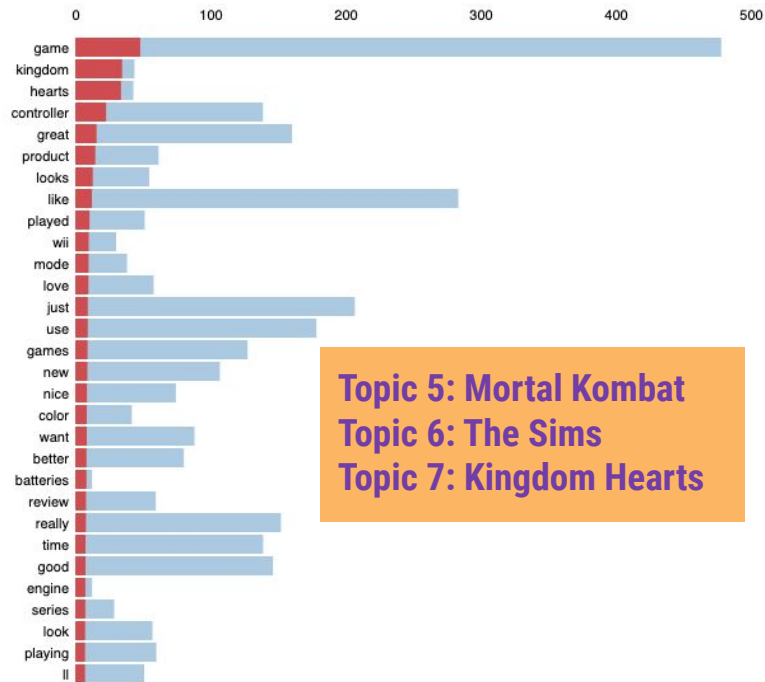
Intertopic Distance Map (via multidimensional scaling)



Marginal topic distribution



Top-30 Most Relevant Terms for Topic 7 (6.2% of tokens)

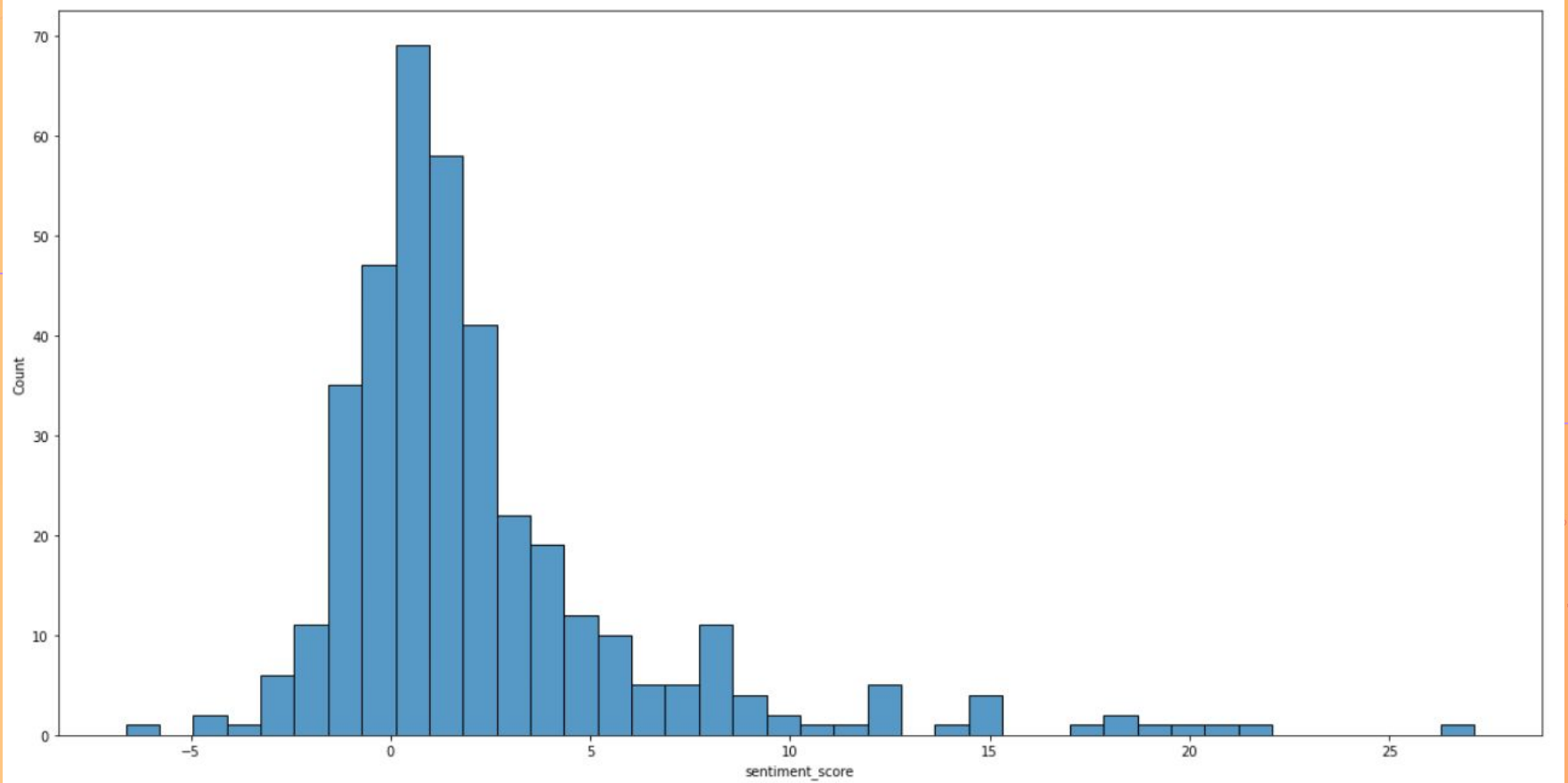


**Topic 5: Mortal Kombat**  
**Topic 6: The Sims**  
**Topic 7: Kingdom Hearts**

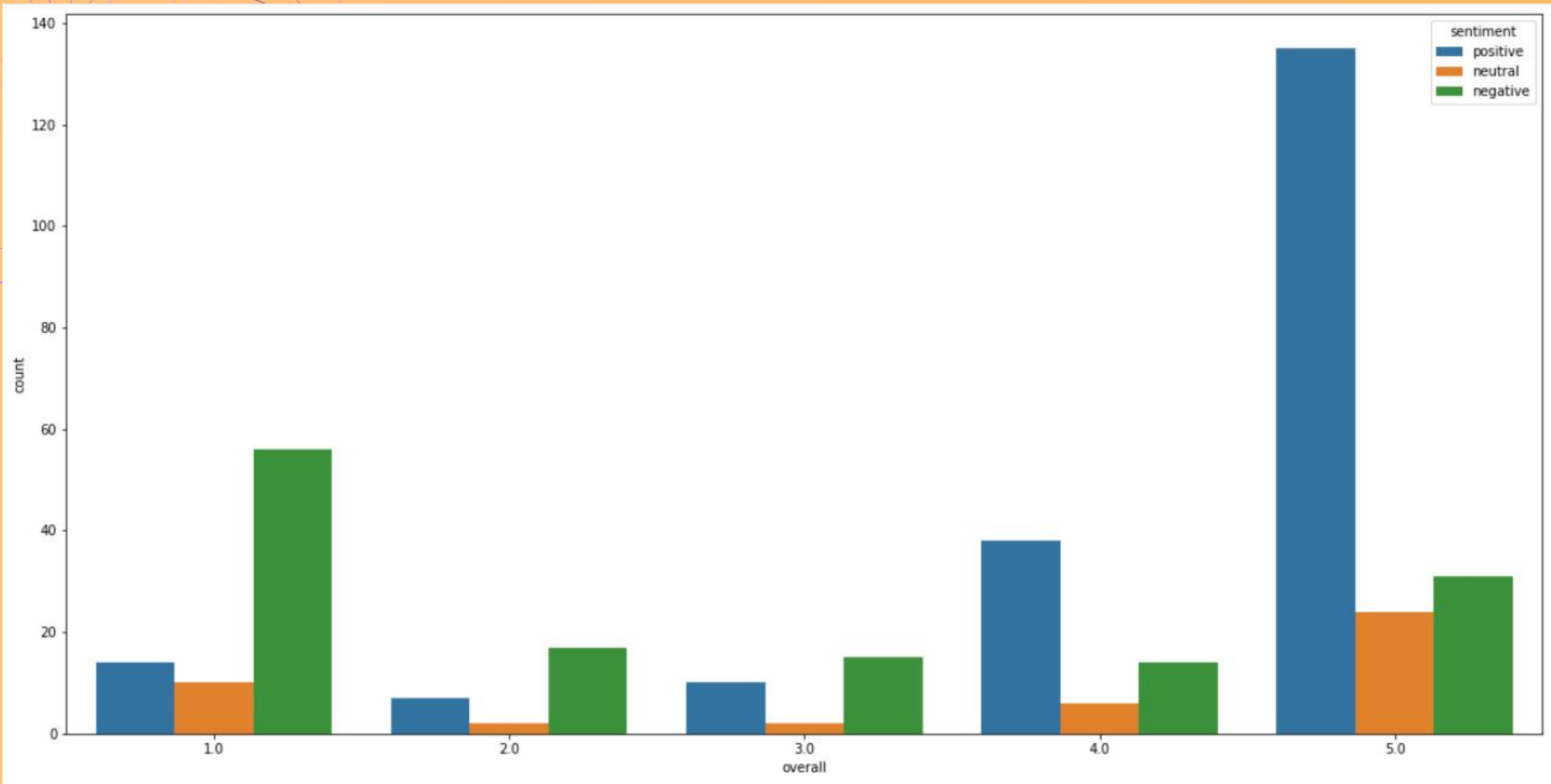
Overall term frequency  
 Estimated term frequency within the selected topic

1.  $\text{saliency}(\text{term } w) = \text{frequency}(w) * [\sum_t p(t|w) * \log(p(t|w)/p(t))]$  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)

# SENTIMENT



# SENTIMENT



# FUTURE ANALYSIS

## NLP PROCESSING

More hyperparameters

## MODEL OPTIMIZATION

Different model, # of topics

## SEARCH & RECOMMENDATION SYSTEM

Query search and recommendation





# THANKS



**GameStop**  
POWER TO THE PLAYERS®