Tracking Down What's Trending Now

James Denney, Steven Masschelin,
Ben Shaver, Chukwudi Uraih

Project Dashboard

- Problem and Objectives
- Methodology
- Early Returns
- Tackling Topics
- Message Recognition
- Image Text Recognition
- Recap

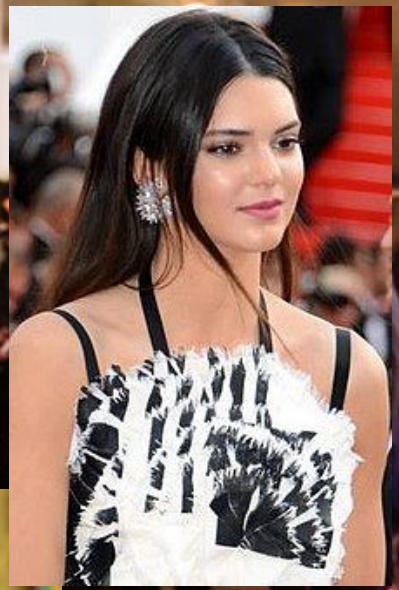
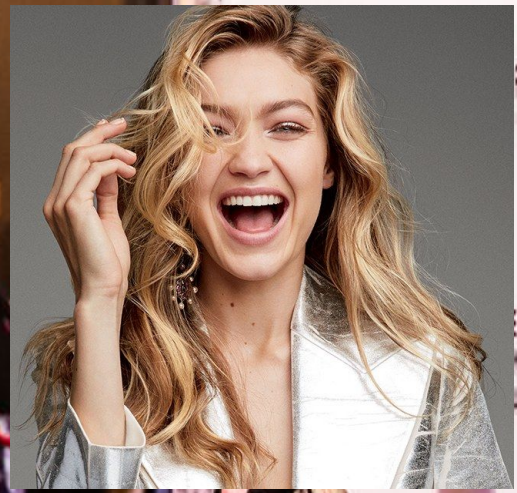
Project Dashboard

- Problem and Objectives
- Methodology
- Early Returns
- Tackling Topics
- Message Recognition
- Image Text Recognition
- Recap

Social media is a dominant presence for many company bottom lines

“How can we drive clicks?”

Better question: "***What*** drives clicks?"



Better question: "***What*** drives clicks?"

Engagement

What is it? What drives it? And how do we quantify it?

GOAL:

Identify Top Topics on Facebook and Instagram

Project Dashboard

- Problem and Objectives
- Methodology
- Early Returns
- Tackling Topics
- Message Recognition
- Image Text Recognition
- Recap

Our Methodology

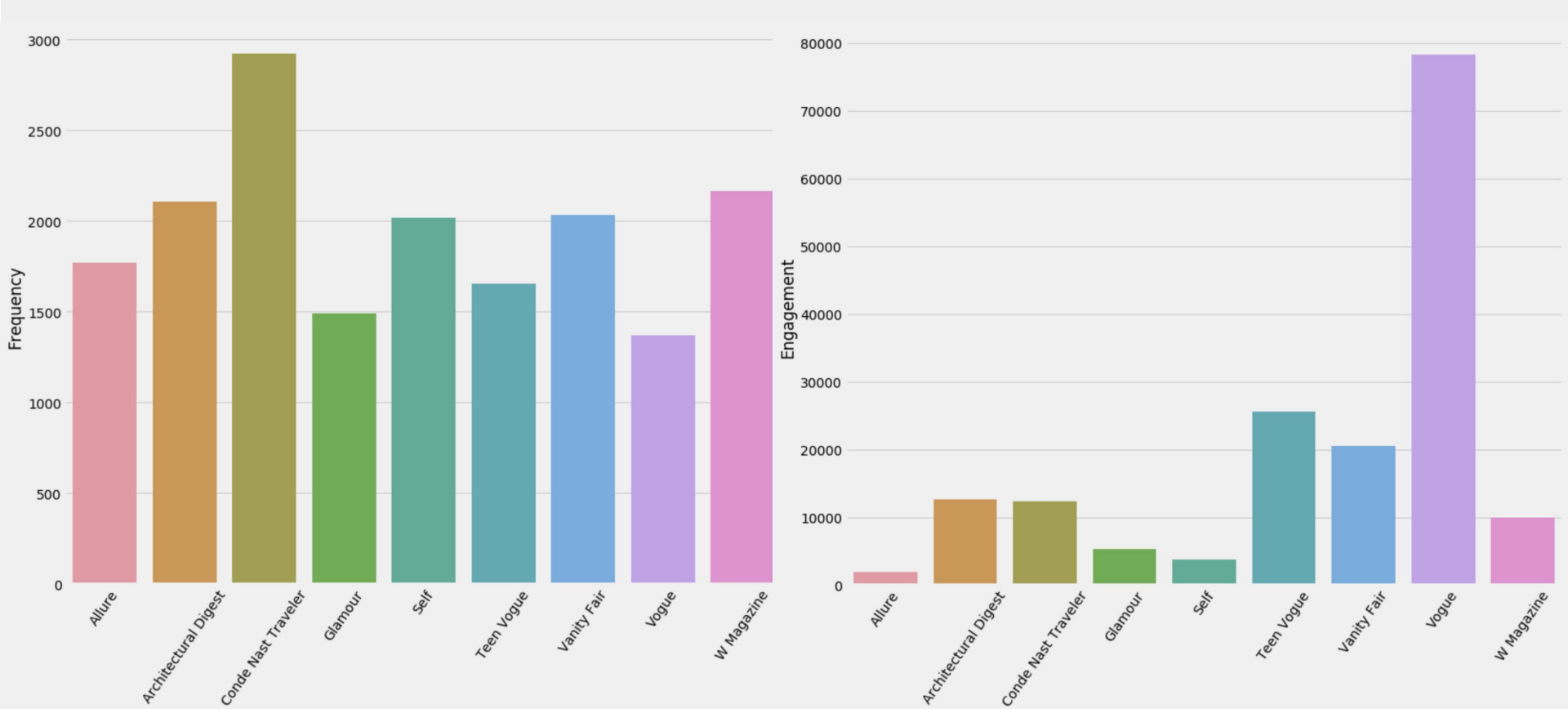
We used a three-fold approach to identify what drives engagement

- Neuro-Linguistic Programming
- Topic Modeling with Latent Dirichlet Allocation
- Image Recognition using Amazon Web Services

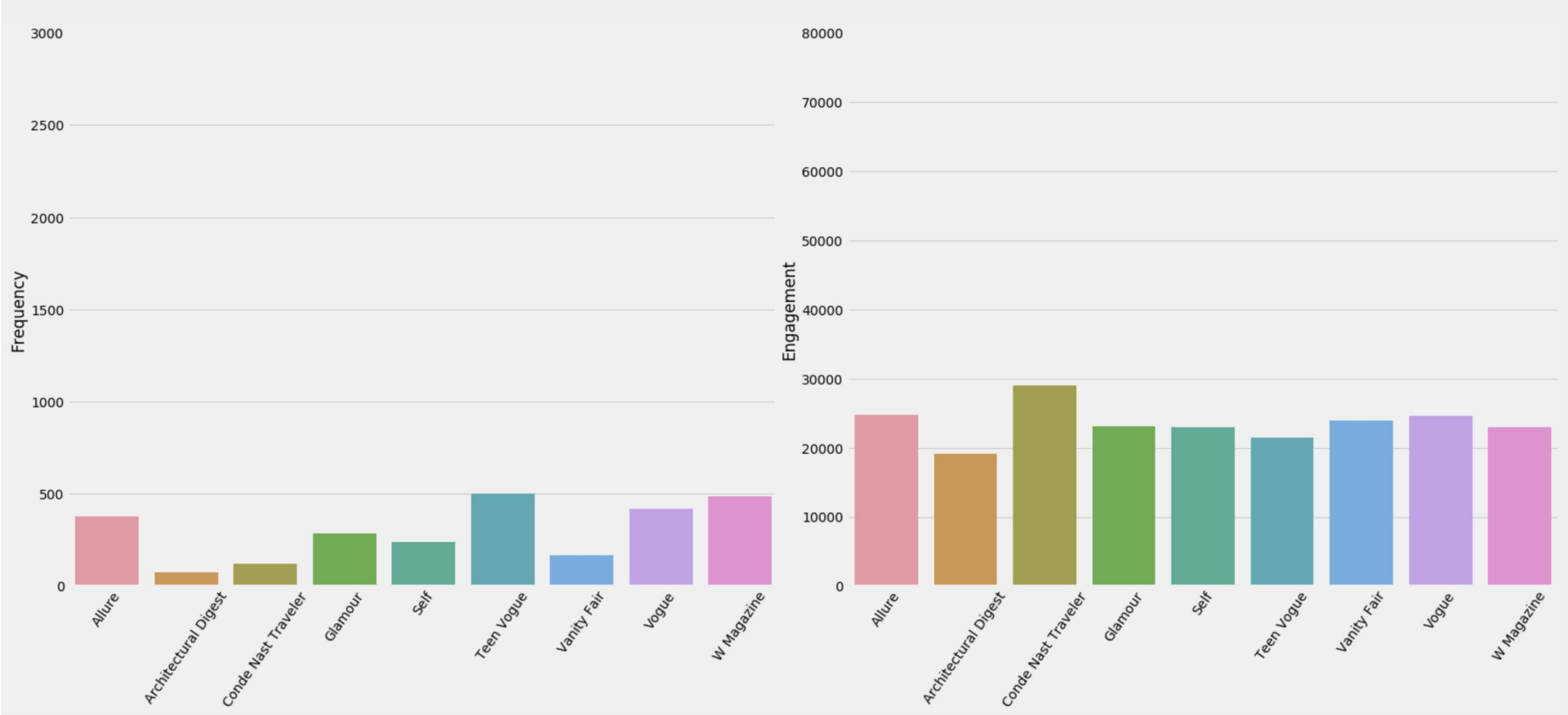
Project Dashboard

- Problem and Objectives
- Methodology
- Early Returns
- Tackling Topics
- Message Recognition
- Image Text Recognition
- Recap

Frequency and Mean Engagement: Instagram Pictures



Frequency and Mean Engagement: Instagram Videos



Project Dashboard

- Problem and Objectives
- Methodology
- Early Returns
- Tackling Topics
- Message Recognition
- Image Text Recognition
- Recap

Topic Modeling with Latent Dirichlet Allocation

- Bag of Words model -- ignores grammar and syntax
- Assumes documents are generated by a hidden (“latent”) process
- Each document is a (distributed?) mix, [or distribution,] of topics, and...
- Each topic is a distribution of words

- LDA approximates the latent distribution of
  - **P(*word*|*topic*)**
  - **P(*topic*|*document*)**
- Thus, each word/phrase is a random variable that occurs with probability
  - **P(*word*|*document*)**

Why Do Topic Modeling?

- To uncover and categorize previously unnoticed patterns
- For our purposes, LDA can be used to reduce the complexity of text data for visualization and modeling purposes.
- But:
  - Topics meanings aren't easily ascertained
  - Number of topics needs to be chosen before LDA model training

Teen Vogue Topics with *num_topics = 5*

- 'trump littl just pretti donald will movi obama live liar'
- 'taylor girl swift know harri thing need just colleg will'
- 'hadid gigi hair bella zayn look just dress prom malik'
- 'selena gomez photo instagram fashion teen beauti just vogu look'
- 'jenner kendal kyli justin bieber look beauti makeup just make'

First: More Helpful Topic Names

- Each topic is a distribution of words
- Even the top 10 words don't fully capture the topic's meaning
- Celebrity names obscure true topic meaning
- Automatically assigning topic names is an area of active research
- One avenue is to examine *Term Frequency -- Inverse Document Frequency* across all 2-to-3 word phrases in a topic
- Group by topic and concatenate all text observations in the text field
- TF-IDF returns the most distinctive phrases for each topic

'Dress shoes runway new york milan wintour'	High Fashion / Fashion Week

First: More Helpful Topic Names

Top 10 Words	Distinctive Ngram	Human-interpreted Title
trump littl just pretti donald will movi obama live liar	This is	Politics
taylor girl swift know harri thing need just colleg will	Need to	“Things you need to know”
hadid gigi hair bella zayn look just dress prom malik	Kendall Jenner	Celebrity Gossip
selena gomez photo instagram fashion teen beauti just vogu look	Selena Gomez	Photos/fashion/instagram
jenner kendal kyli justin bieber look beauti makeup just make	Kylie Jenner	Makeup

Second: Better Topics

- Too many topics yield overly-specific, uninterpretable results
- Too few topics are too vague and not helpful
- *Topic Coherence* is a quantitative measure for topic quality
  - Scores generally summarize a set of pairwise comparisons between words in a topic, such as how well a common word predicts a less common word within a topic.
  - Iterating over values of *num_topics* for Teen Vogue maximizes topic coherence with *num_topics = 3*

Second: Better Topics

Teen Vogue Topics with *num_topics* = 2	Teen Vogue Topics with *num_topics* = 3
• 'girl know just will make trump need thing teen year' • 'jenner just look hadid kendal gigi kyli selena taylor gomez'	• 'taylor swift trump justin bieber just year school peopl will' • 'jenner hadid kendal look gigi just kyli selena gomez fashion' • 'girl just know instagram beauti need thing will best love'

How to Model with LDA

- LDA returns probabilities of belonging to each topic, not a categorical variable.
- Probabilities ‘play nicely’ with a variety of modeling methods.
- With *num_topics* numerical columns, we can integrate text data into a model without massively over-complicating things

	Topic 1	Topic 2	Topic 3	impact
2	0.435002	0.057862	0.507136	0.008185
5	0.666474	0.166758	0.166768	0.098939
6	0.445626	0.443240	0.111134	0.038389
7	0.171792	0.180911	0.647296	10.236497
9	0.901071	0.048610	0.050319	0.068004
11	0.192044	0.255549	0.552407	0.043913
12	0.227028	0.208510	0.564462	0.072138
13	0.117703	0.768663	0.113634	0.108336
14	0.058009	0.279908	0.662083	0.048282
16	0.754082	0.126551	0.119367	0.007540

How to Model with LDA: Logistic Regression

- LDA's numerical output make it well-suited for classic statistical models like multiple linear regression or logistic regression
- If predicting absolute level of impact is infeasible, instead we can predict "high impact" vs "low impact" posts
- Logistic regression estimates the linear effect of topic mixture on likelihood of belonging to the high impact category
- A unit increase in percentage of the document attributed to a certain topic will increase the probability of a document being high impact by a certain amount

How to Model with LDA: Logistic Regression

- Using *only* 3 topics trained on a small corpus, logistic regression's performance is weak, indicating many other variables need to be taken into account, but:
- The direction of influence can still be inferred:
- The greater the proportion of a post is attributed to 'Topic 2,' the higher the likelihood that post is high impact. Topic 1 and Topic 3 have negative effects, given the proportion of topics must sum to one.
- In this case,
    - *'jenner hadid kendal look gigi just kyli selena gomez fashion'* outperforms the other two topics.

LDA: Limitations and Next Steps

- Train on a larger corpus:
  - Train on headlines, post content, and captions from other fashion or lifestyle magazines
  - Scrape the raw text from each link in order to increase corpus size
  - Use a pre-trained model or train LDA on a standard corpus (Wikipedia)
- Provide multiple ngram titles:
  - e.g. *"Seen out with / Rumors going"* is clearly celebrity gossip
- Streamline pipeline for selecting number of topics based on topic coherence:
  - Vanity Fair may cover a wider range of distinct topics than Teen Vogue

Text Recognition

- Employed Natural Language Processing
- Facebook messages and Instagram captions and hashtags
- Which words occurred most frequently?
- Methods employed
  - Count Vectorizer
  - TF-IDF

Facebook Messages, Vogue

Phrase	Frequency (Raw)
Red carpet	426
Fashion week	424
New York	338
Met Gala	327
Kendall Jenner	295
Gigi Hadid	271
Street style	259
Kim Kardashian	235
Selena Gomez	224
Bella Hadid	212

Facebook Messages, Teen Vogue

Phrase	Frequency (Raw)
Kendall Jenner	674
Gigi Hadid	575
Kylie Jenner	559
Selena Gomez	554
Taylor Swift	462
Justin Bieber	339
Donald Trump	302
Ariana Grande	203
Bella Hadid	162
Miley Cyrus	161

Captions, Vogue

Word	Frequency (Raw)
Photographed	700
Vogue	461
Styled	177
2017	173
Fashion	152

Captions, Teen Vogue

Word	Frequency (Raw)
TeenVogue	145
Happy	143
Today	134
Us	130
One	124

Captions, Vogue (TF-IDF)

Word	Frequency (Percent)
Photographed	43.103
Vogue	37.929
Styled	22.377
2017	20.696
Fashion	17.976

Captions, Teen Vogue (TF-IDF)

Word	Frequency (Percent)
Happy	40.573
Birthday	30.737
Today	19.167
MondayMotivation	18.748
TeenVogue	18.492

Hashtags, Vogue

Word	Frequency (Raw)
Vogue125	45
Metgala	34
AnnieLeibovitz	33
IrvingPenn	26
NYFW	21

Hashtags, Teen Vogue

Word	Frequency (Raw)
GirlGaze	40
MondayMotivation	39
ShineTheory	30
TBT	27
NYFW	27

Hashtags, Vogue (TF-IDF)

Word	Frequency (Percent)
Vogue125	31.652
AnnieLeibovitz	26.833
IrvingPenn	24.152
MetGala	22.720
PFW	20.049

Hashtags, Teen Vogue (TF-IDF)

Word	Frequency (Percent)
MondayMotivation	37.735
GirlGaze	27.786
ShineTheory	27.304
NYFW	24.078
TBT	22.469

Project Dashboard

- Problem and Objectives
- Methodology
- Early Returns
- Tackling Topics
- Message Recognition
- Image Text Recognition
- Recap

What's in a Picture?

- Sent Instagram pictures through Amazon Web Services image processing
- Identified, pulled, and analyzed results

Results

Clown	83.3 %
Human	83.3 %
Performer	83.3 %
Person	83.3 %
Carnival	59.3 %
Crowd	59.3 %

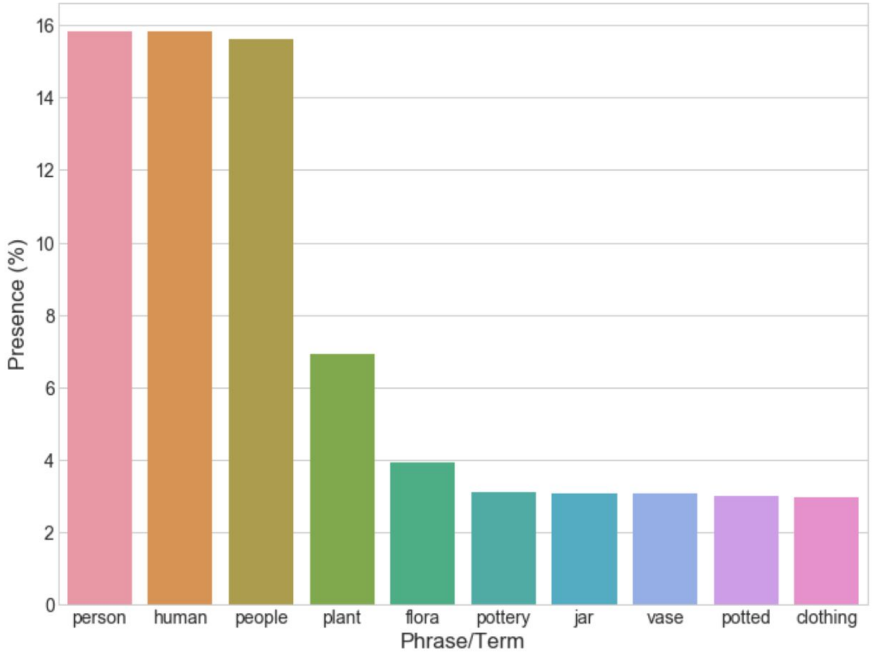
Show more

Request

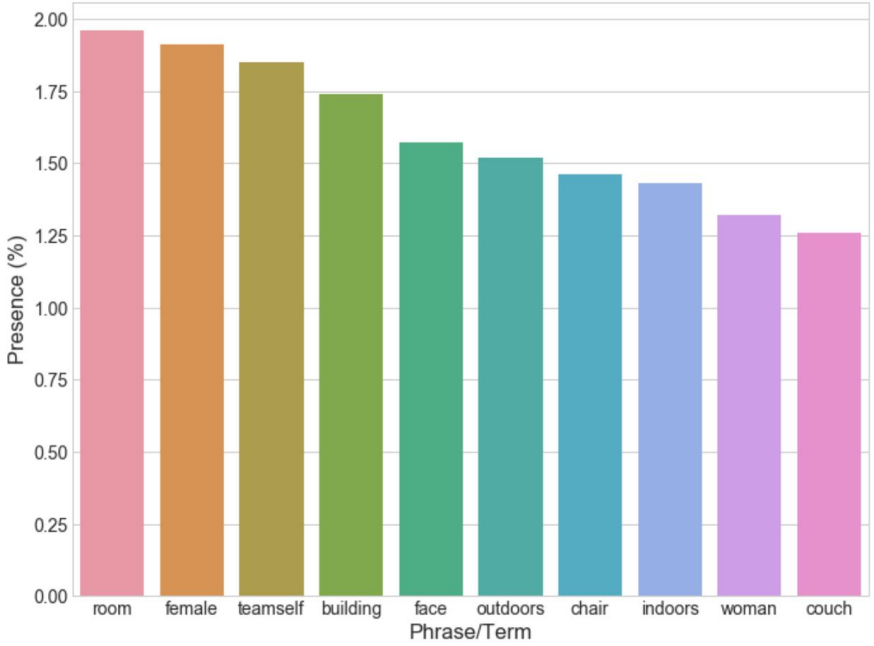
Response

Teen Vogue Instagram Pictures: What Do We See?

Ten Most Frequent Phrases



Ten Least Frequent Phrases



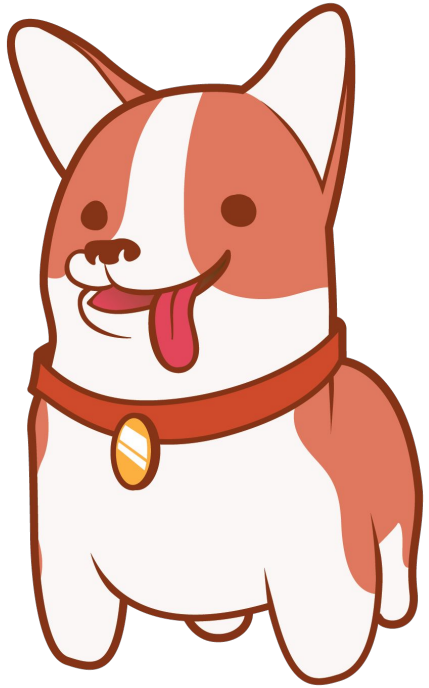
Project Dashboard

- Problem and Objectives
- Methodology
- Early Returns
- Tackling Topics
- Message Recognition
- Image Text Recognition
- Recap

Into the Future

- Analyzed topics and image text features of nine Condé Nast publications
- Focus on Instagram and Facebook channels
- Created tools to identify most frequent subjects and highest-engaging topic for each brand
- Amazon Web Services pulls text features from Instagram images

Depending on a brand's desired metric, it is possible to model and predict engagement in the long run using these tools!



TrackMaven