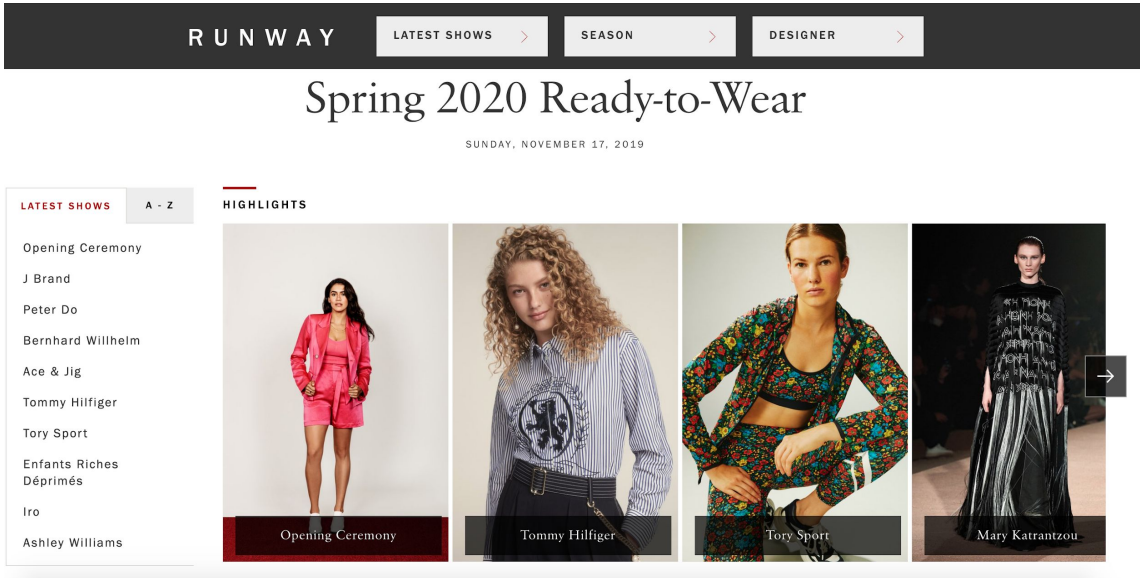
RE-CLASSIFYING VOGUE COLLECTION COVERAGE




Goal

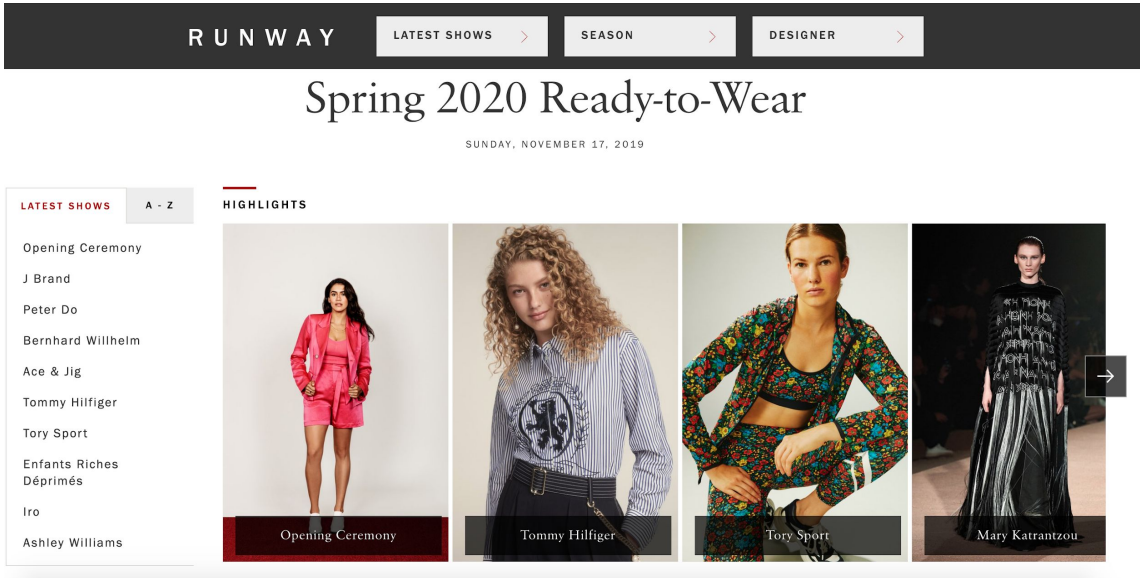
To reclassify Vogue runway show coverage from the Vogue website.

- What other categories can be created?
- How is the text for each collection related?
- Articles currently only categorized by season, year, and designer



The Data

- Gathered 3525 articles covering major collections from 2018 to 2020
- Used Selenium to scrape from the Vogue website



Bag of Words

- Created a bag of words.
- These are the most frequently used words in the corpus.

Once Common Stop Words Removed

Further Cleaning

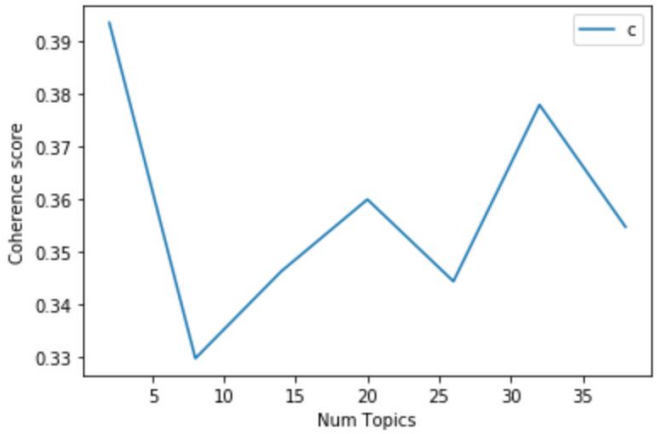
- Created a list of 100 most common words in corpus
- Used domain knowledge to pick terms (very general) to remove from corpus that would not help in modeling more specific topics (collection type, designer, year, and common terms like ‘runway’, ‘designer’, ‘collection’
- Removed numbers and tokens with a length less than 3

Next Steps

- Lemmatized the updated corpus
- Created bi-gram and tri-gram models with a minimum count of 10 and converted them into tokens
- Filtered extreemes (since trying to find more specific niche words tokens must appear 10 times and appear in no more than 10% of the documents.

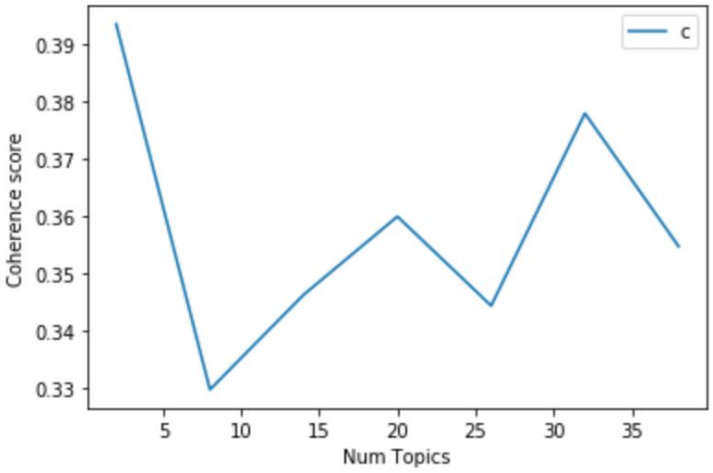
First Model

- Created an LDA model to extract topic from the textual data.
- First chosen topic count = 5
- Coherence score was .374
- To improve model created a line graph showing how coherence scores increase by topic



Choosing number of Topics

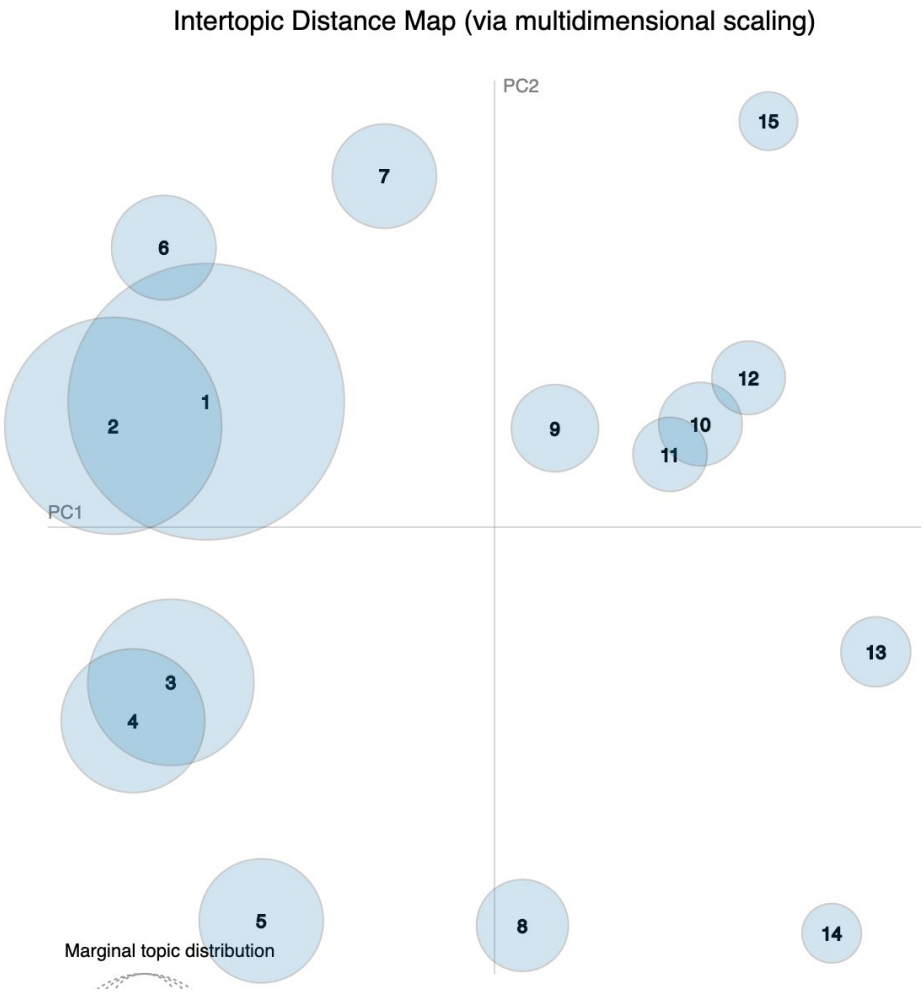
- First created 20 topics according to first peak
- Coherence score was .0.402
  - Improved by 5.5%
- Thought this might be too many topics

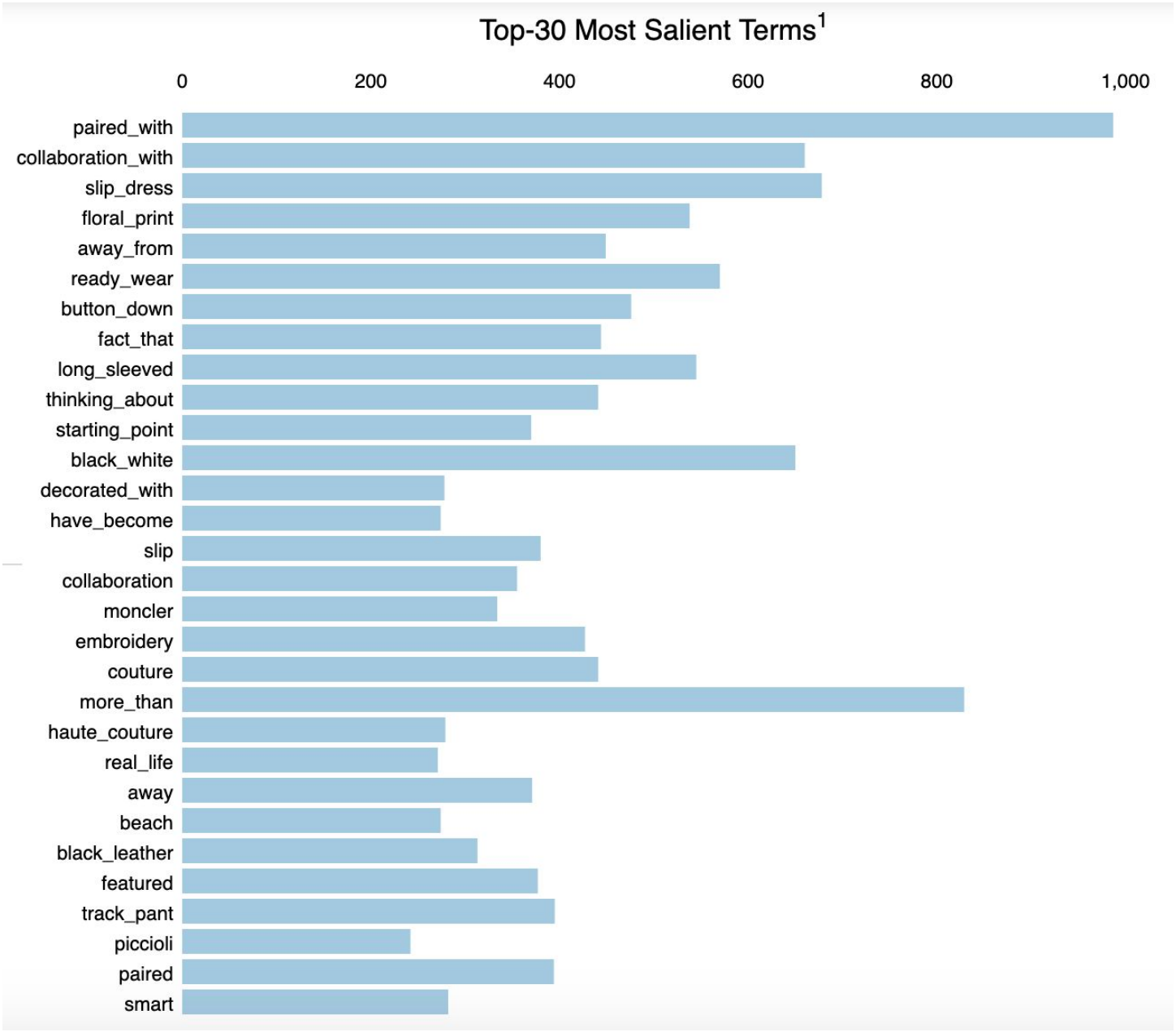


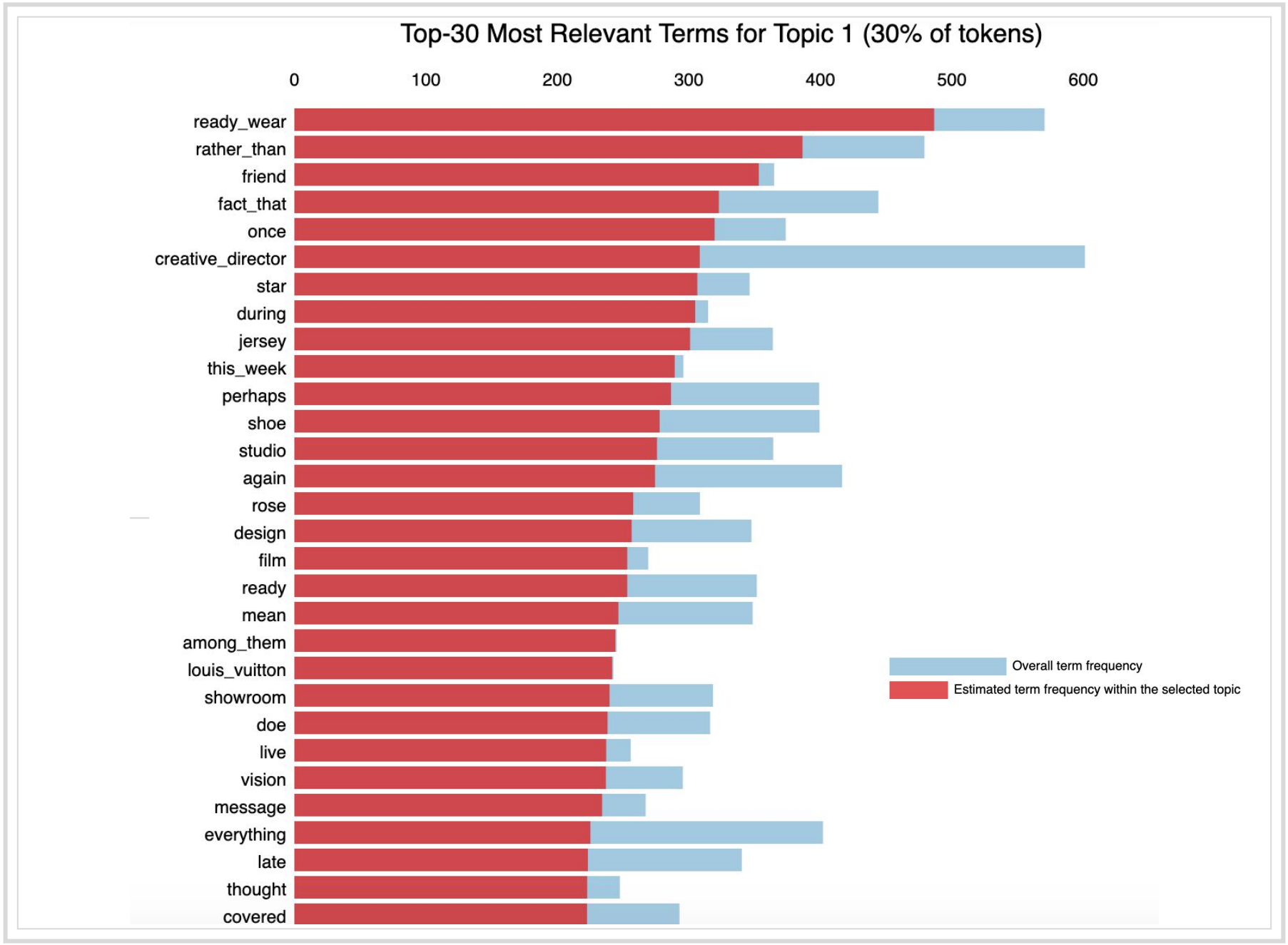
Choosing number of Topics

- Tried a few more models with various number of topics between 10 and 20
- Arrived at 15 where coherence score actually increased
  - Coherence score was 0.403
- This was chosen as the final model

Topic Relationships




Top-30 Most Salient Terms[1]
0
200
400
600
800
1,000
paired_with
collaboration_with
slip_dress
floral_print
away_from
ready_wear
button_down
fact_that
long_sleeved
thinking_about
starting_point
black_white
decorated_with
have_become
slip
collaboration
moncler
embroidery
couture
more_than
haute_couture
real_life
away
beach
black_leather
featured
track_pant
piccioli
paired
smart


Top-30 Most Relevant Terms for Topic 1 (30% of tokens)
0
100
200
300
400
500
600
ready_wear
rather_than
friend
fact_that
once
creative_director
star
during
jersey
this_week
perhaps
shoe
studio
again
rose
design
film
ready
mean
among_them
louis_vuitton
showroom
doe
live
vision
message
everything
late
thought
covered
Overall term frequency
Estimated term frequency within the selected topic

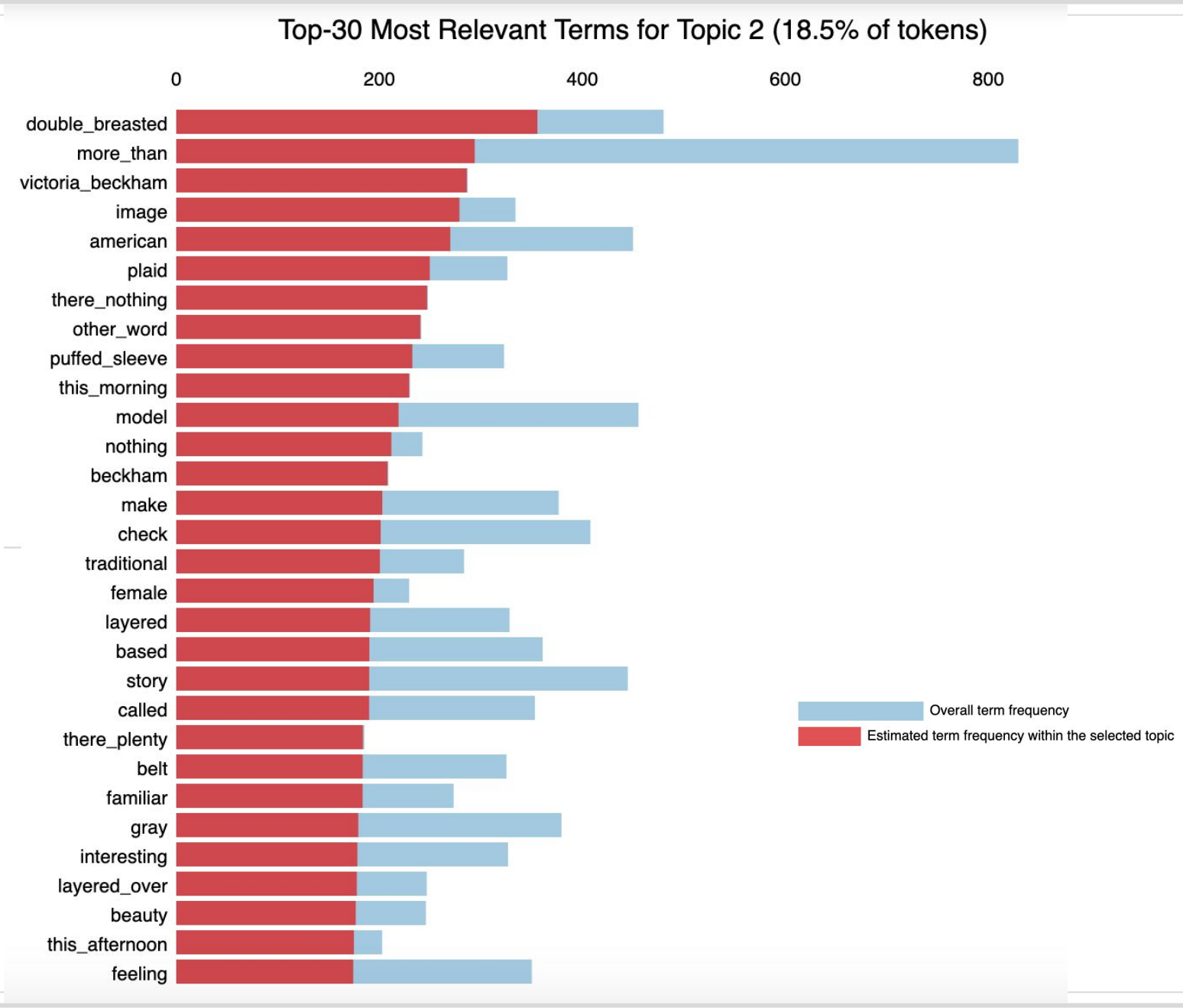
Topic 1

Jersey

Or is this Jersey?

Louis Vuitton

Has mostly language terms, and proper nouns, not specifics


Top-30 Most Relevant Terms for Topic 2 (18.5% of tokens)
0
200
400
600
800
double_breasted
more_than
victoria_beckham
image
american
plaid
there_nothing
other_word
puffed_sleeve
this_morning
model
nothing
beckham
make
check
traditional
female
layered
based
story
called
there_plenty
belt
familiar
gray
interesting
layered_over
beauty
this_afternoon
feeling
Overall term frequency
Estimated term frequency within the selected topic

Topic 2

Double Breasted

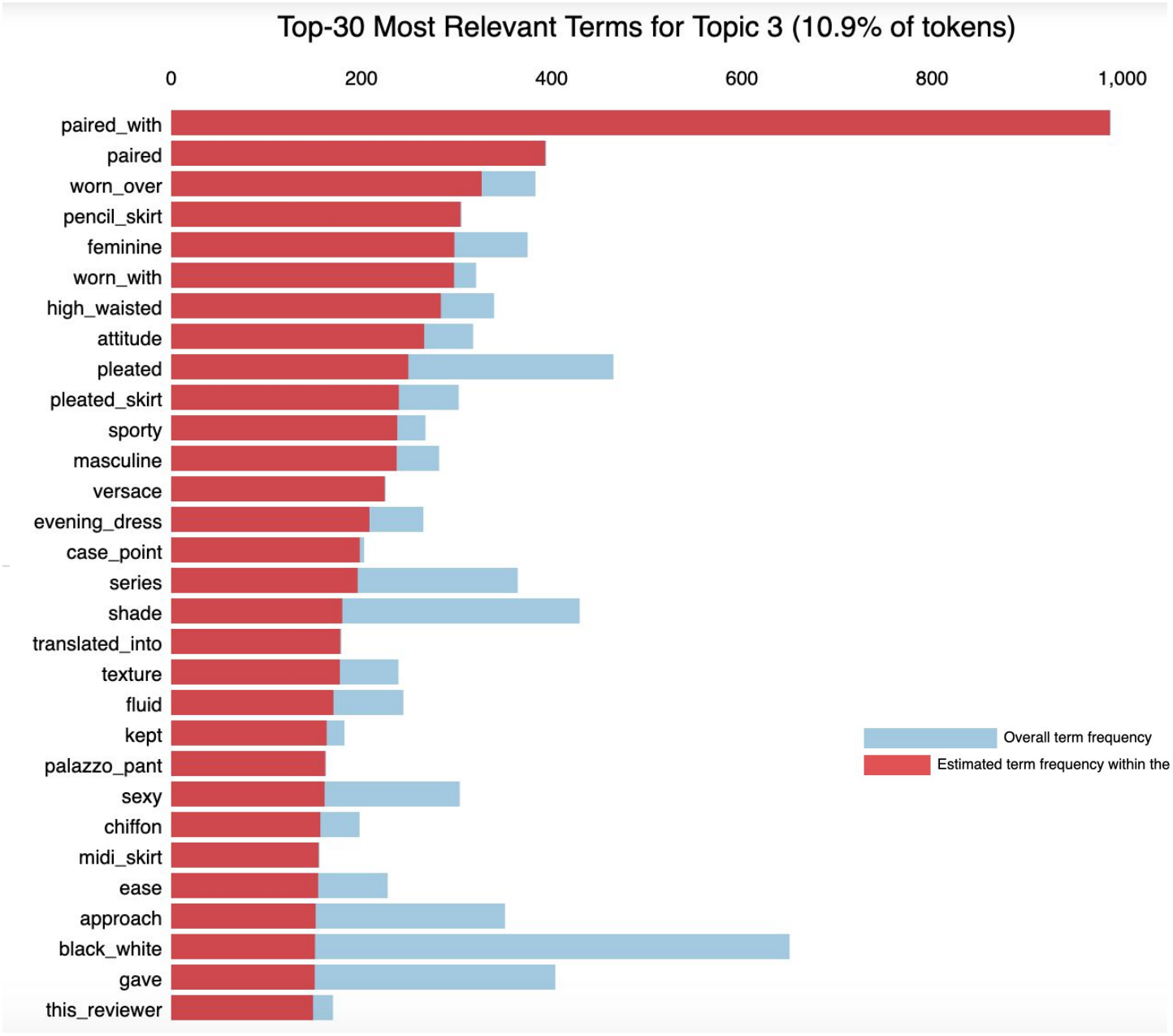
Plaid & Double Breasted

Plaid & Puff Sleeve

Puff Sleeve

Double Breasted & Puff Sleeve

Many of the trends intersect


Top-30 Most Relevant Terms for Topic 3 (10.9% of tokens)
0
200
400
600
800
1,000
paired_with
paired
worn_over
pencil_skirt
feminine
worn_with
high_waisted
attitude
pleated
pleated_skirt
sporty
masculine
versace
evening_dress
case_point
series
shade
translated_into
texture
fluid
kept
palazzo_pant
sexy
chiffon
midi_skirt
ease
approach
black_white
gave
this_reviewer
Overall term frequency
Estimated term frequency within the selected topic

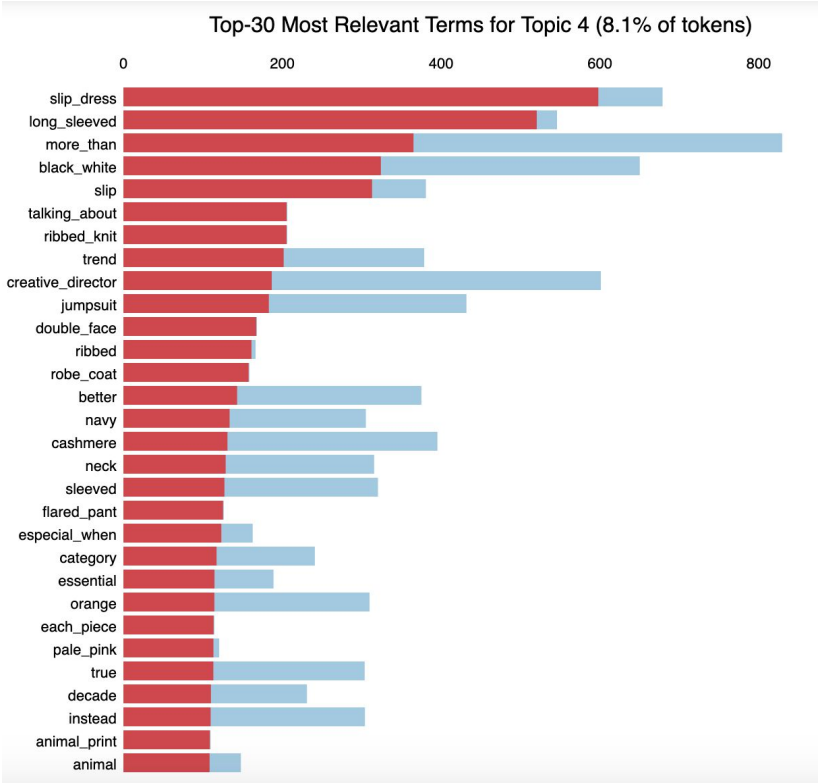
Topic 3

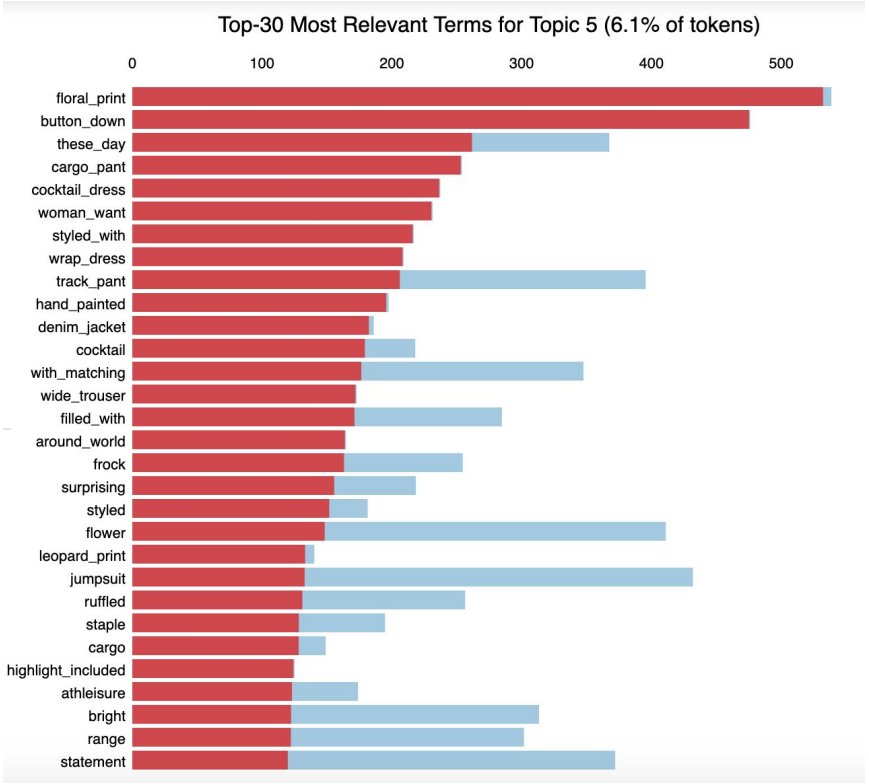
Pencil Skirt

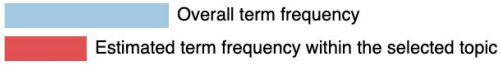
High Waisted

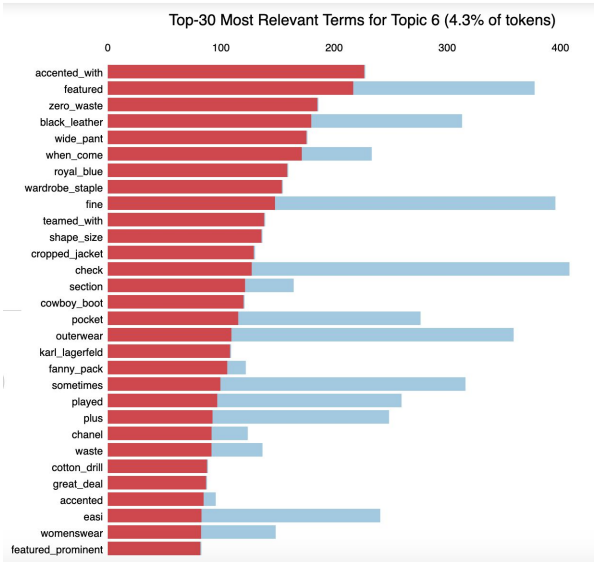
Pleated Skirt

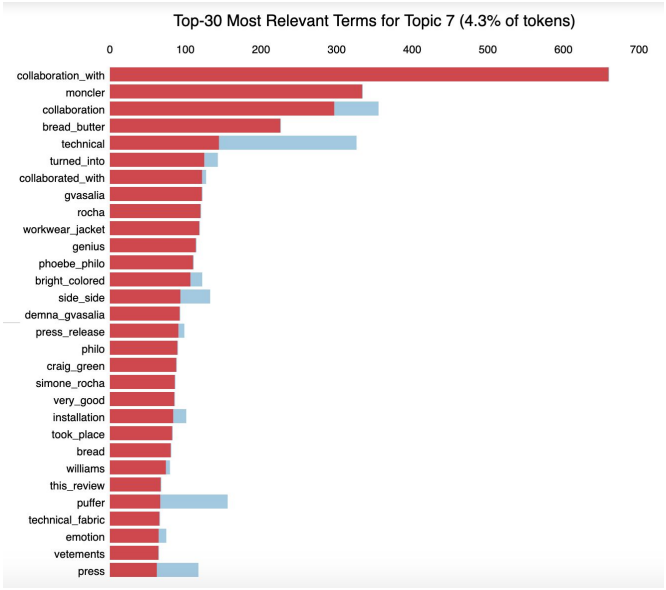
Bigrams proved important for trends

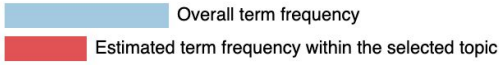

Top-30 Most Relevant Terms for Topic 4 (8.1% of tokens)
0
200
400
600
800
slip_dress
long_sleeved
more_than
black_white
slip
talking_about
ribbed_knit
trend
creative_director
jumpsuit
double_face
ribbed
robe_coat
better
navy
cashmere
neck
sleeved
flared_pant
especial_when
category
essential
orange
each_piece
pale_pink
true
decade
instead
animal_print
animal

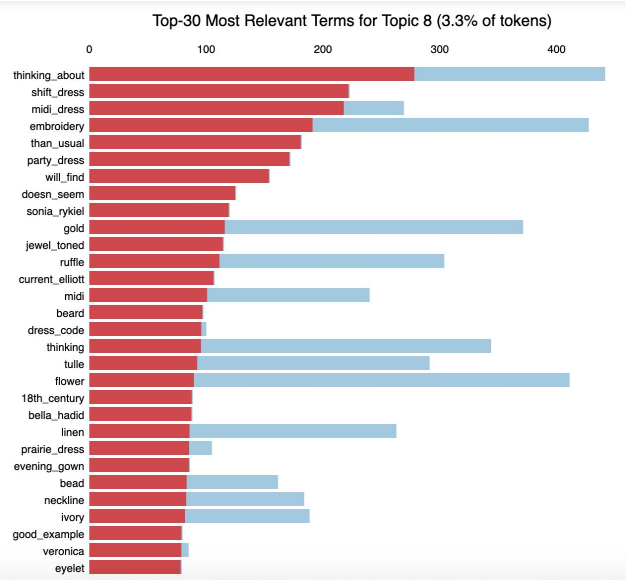

Top-30 Most Relevant Terms for Topic 5 (6.1% of tokens)
0
100
200
300
400
500
floral_print
button_down
these_day
cargo_pant
cocktail_dress
woman_want
styled_with
wrap_dress
track_pant
hand_painted
denim_jacket
cocktail
with_matching
wide_trouser
filled_with
around_world
frock
surprising
styled
flower
leopard_print
jumpsuit
ruffled
staple
cargo
highlight_included
athleisure
bright
range
statement


Overall term frequency
Estimated term frequency within the selected topic


Top-30 Most Relevant Terms for Topic 6 (4.3% of tokens)
0
100
200
300
400
accented_with
featured
zero_waste
black_leather
wide_pant
when_come
royal_blue
wardrobe_staple
fine
teamed_with
shape_size
cropped_jacket
check
section
cowboy_boot
pocket
outerwear
karl_lagerfeld
fanny_pack
sometimes
played
plus
chanel
waste
cotton_drill
great_deal
accented
easi
womenswear
featured_prominent


Top-30 Most Relevant Terms for Topic 7 (4.3% of tokens)
0
100
200
300
400
500
600
700
collaboration_with
moncler
collaboration
bread_butter
technical
turned_into
collaborated_with
gvasalia
rocha
workwear_jacket
genius
phoebe_philo
bright_colored
side_side
demna_gvasalia
press_release
philo
craig_green
simone_rocha
very_good
installation
took_place
bread
williams
this_review
puffer
technical_fabric
emotion
vetements
press

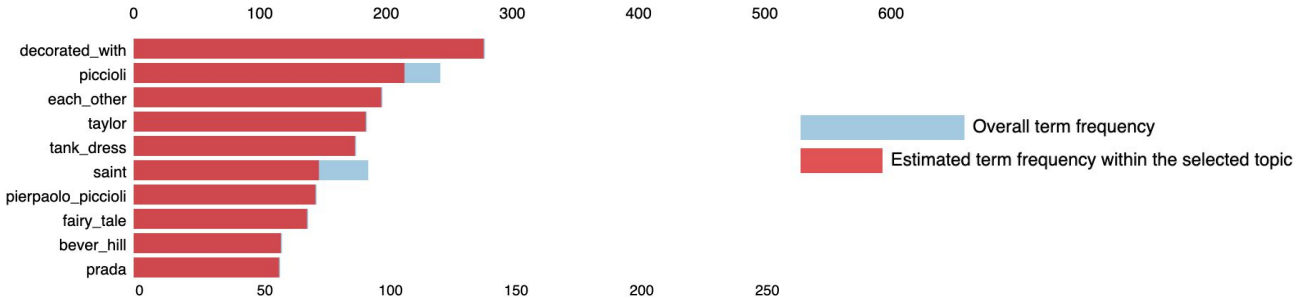

Overall term frequency
Estimated term frequency within the selected topic


Top-30 Most Relevant Terms for Topic 8 (3.3% of tokens)
0
100
200
300
400
thinking_about
shift_dress
midi_dress
embroidery
than_usual
party_dress
will_find
doesn_seem
sonia_rykiel
gold
jewel_toned
ruffle
current_elliott
midi
beard
dress_code
thinking
tulle
flower
18th_century
bella_hadid
linen
prairie_dress
evening_gown
bead
neckline
ivory
good_example
veronica
eyelet

Top 10 Most Relevant terms

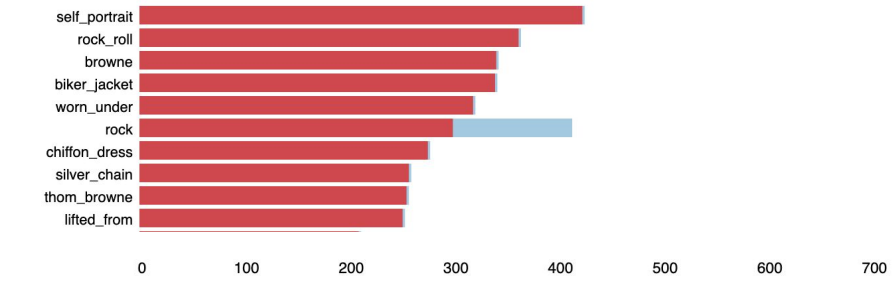
Topic 12

(2.1% of tokens)



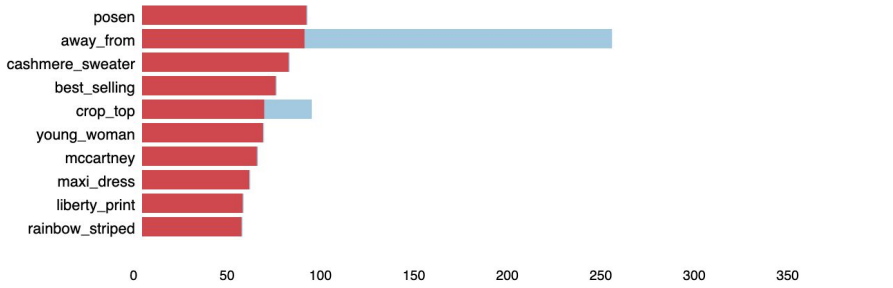
Topic 13

(1.91% of tokens)



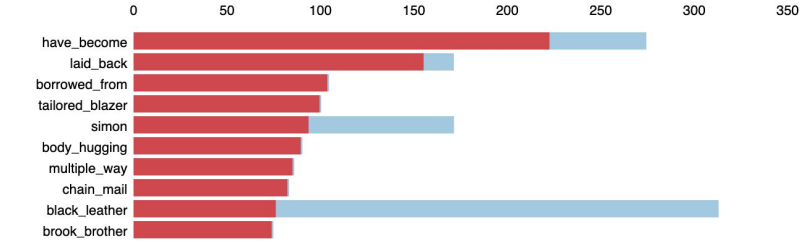
Topic 14

(1.4% of tokens)



Topic 15

(1.31% of tokens)



Future Work:

- Analyze text more for additional stop words to eliminate designer names and common words not relevant to goals of this topic modeling
- Collect even more data
- Figure out ways to further improve coherence score.
- Create bar plot visual for topic distribution
- Test these topics on other vogue runway text

thanks!

Any questions?