KEATRA NESBITT

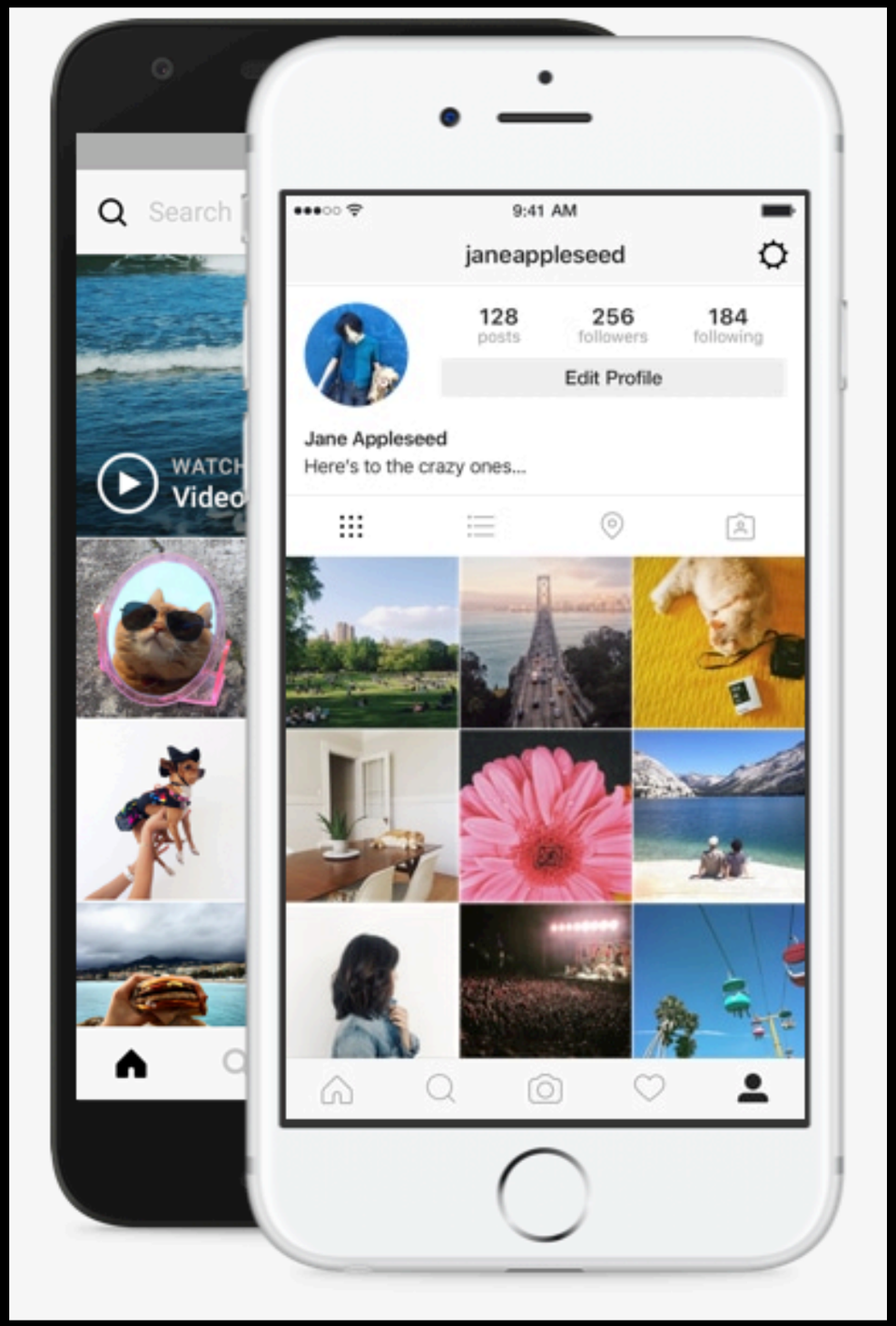
INSTAGRAM NLP

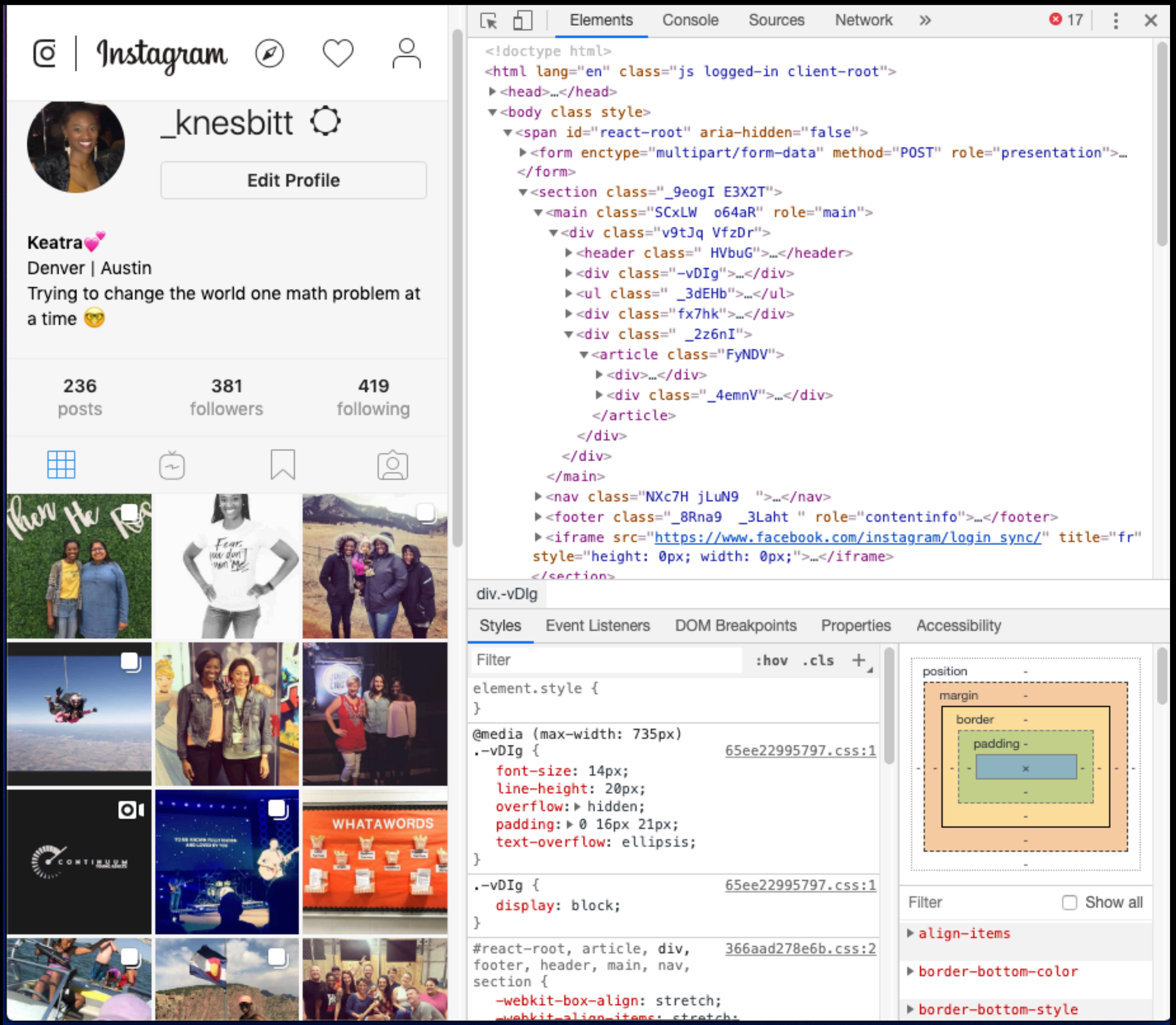
PROJECT DESCRIPTION

- **Goal:** Use Natural Language Processing to predict the number of likes on a picture

- **Data:** 17 Instagram profiles with 6806 pictures

- **Resources:** Python, Selenium, NLTK, and Sklearn,




Instagram
_knesbitt
Edit Profile
Keatra💕
Denver | Austin
Trying to change the world one math problem at a time 🤓
236
posts
381
followers
419
following
Elements
Console
Sources
Network
17
<!doctype html>
<html lang="en" class="js logged-in client-root">
<head>…</head>
<body class style>

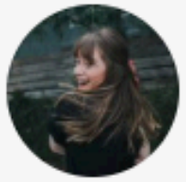
<form enctype="multipart/form-data" method="POST" role="presentation">…
</form>
<section class="_9eogI E3X2T">
<main class="SCxLW o64aR" role="main">

<header class=" HVbuG">…</header>
…
<ul class=" _3dEHb">…
…

<article class="FyNDV">
…
…
</article>

</main>
<nav class="NXc7H jLuN9 ">…</nav>
<footer class="_8Rna9 _3Laht " role="contentinfo">…</footer>
<iframe src="https://www.facebook.com/instagram/login_sync/" title="fr"
style="height: 0px; width: 0px;">…</iframe>
</section>
div.-vDIg
Styles
Event Listeners
DOM Breakpoints
Properties
Accessibility
Filter
:hov .cls +
element.style {
}
@media (max-width: 735px)
.-vDIg {
65ee22995797.css:1
font-size: 14px;
line-height: 20px;
overflow: ▸ hidden;
padding: ▸ 0 16px 21px;
text-overflow: ellipsis;
}
.-vDIg {
65ee22995797.css:1
display: block;
}
#react-root, article, div,
footer, header, main, nav,
section {
366aad278e6b.css:2
-webkit-box-align: stretch;
position
margin
border
padding
Filter
Show all
align-items
border-bottom-color
border-bottom-style

Instagram



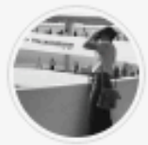
briannanmoore13 ···

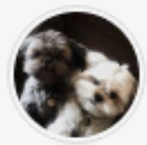
Following

Brianna
TX ➳ Photography IG @briannamoore
www.briannamoorephotography.com

Followed by cassidyjwilkinson, adriana_renee15, m_gonz17 + 33 more


my name

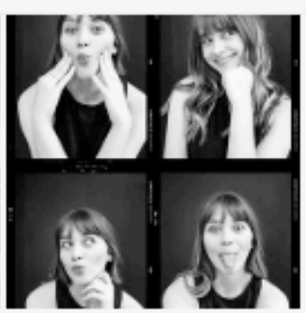

➳


cash + ju...

1,083 posts

1,550 followers

985 following



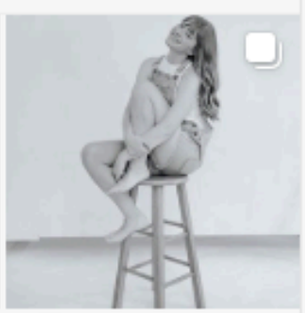
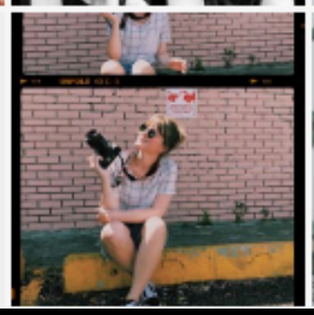
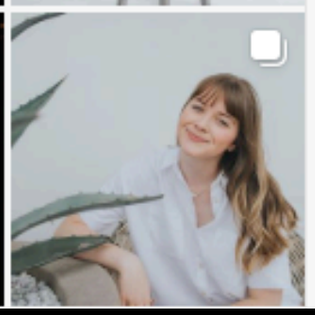


Image may contain: one or more people, people standing, shorts and indoor





Elements Console Sources Network » 321

```
<!doctype html>
<html lang="en" class="js logged-in client-root">
▶<head>…</head>
▼<body class style>
 ▼<span id="react-root" aria-hidden="false">
  ▶<form enctype="multipart/form-data" method="POST" role=
  "presentation">…</form>
  ▼<section class="_9eogI E3X2T">
   ▼<main class="SCxLW  o64aR" role="main">
    ▼<div class="v9tJq VfzDr">
     ▶<header class=" HVbuG">…</header>
     ▶<div class="-vDIg">…</div>
     ▶<div class="_4bSq7">…</div>
     ▶<ul class=" _3dEHb">…</ul>
     ▶<div class="fx7hk">…</div>
     ▼<div class=" _2z6nI">
      ▼<article class="FyNDV">
       ▼<div>
        ▶<div style="flex-direction: column; padding-bottom:
        1343.31px; padding-top: 0px;">…</div>
        </div>
       ▶<div class="_4emnV">…</div>
       </article>
      </div>
     </div>
    </main>
   ▶<nav class="NXc7H jLuN9 "> </nav>
```

div.-vDIg

Styles | Event Listeners | DOM Breakpoints | Properties | Accessibility

Filter :hov .cls +

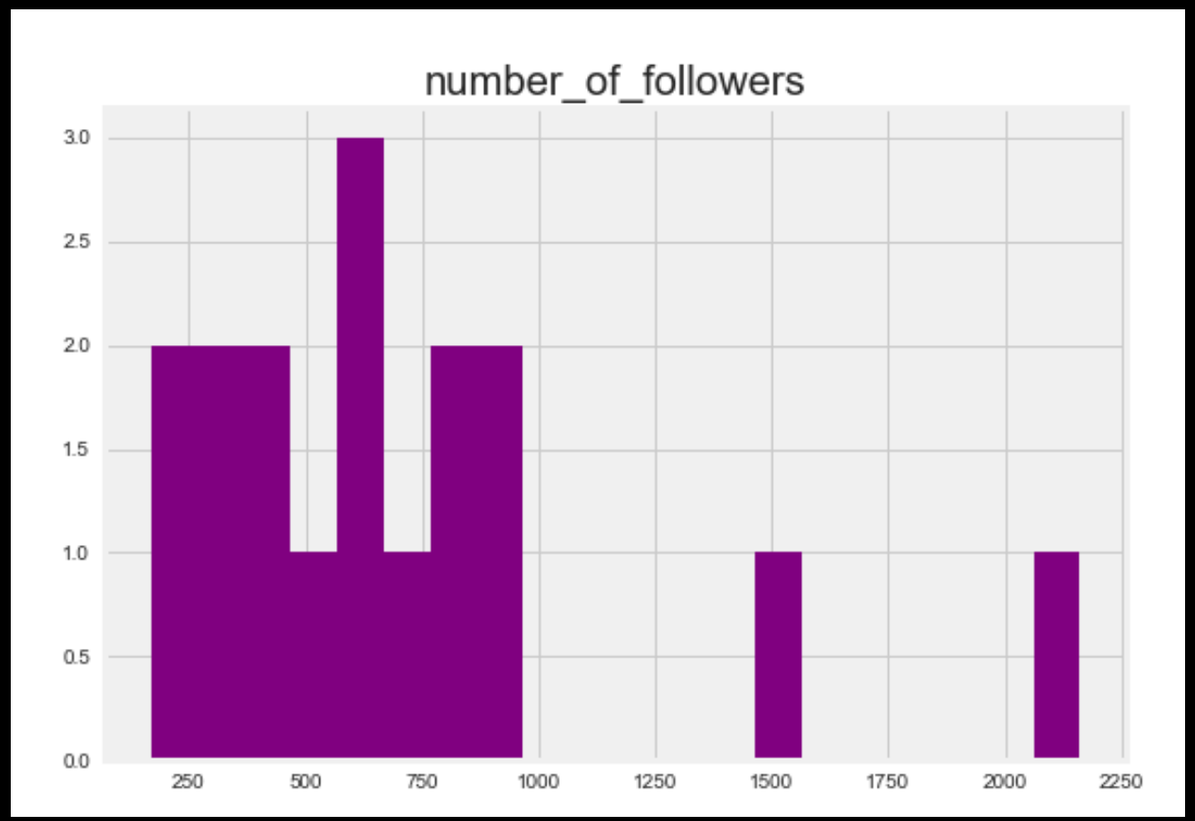
```
element.style {
}

@media (max-width: 735px)
.-vDIg {                          65ee22995797.css:1
  font-size: 14px;
  line-height: 20px;
  overflow: ▶ hidden;
  padding: ▶ 0 16px 21px;
  text-overflow: ellipsis;
}

.-vDIg {                          65ee22995797.css:1
  display: block;
}

#react-root,                      366aad278e6b.css:2
article, div,
footer, header, main, nav, section {
  -webkit-box-align: stretch;
```

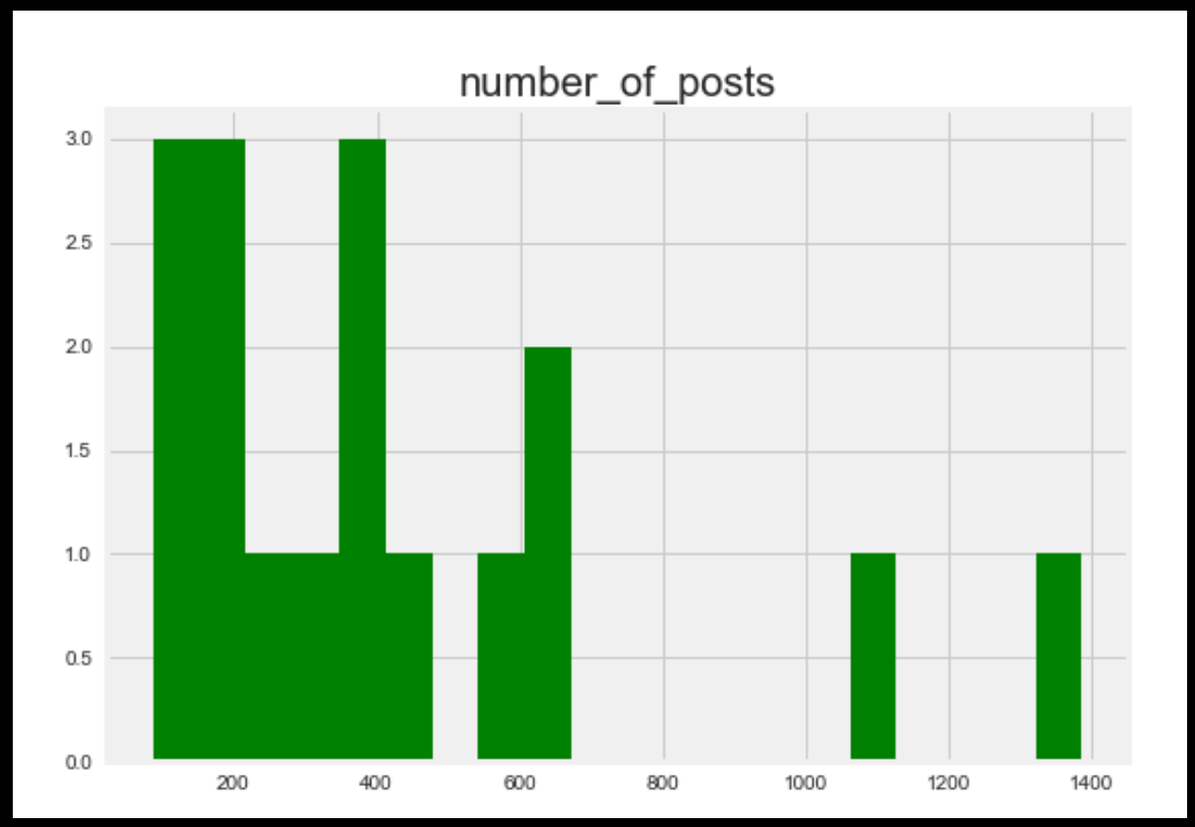
position -
margin -
border -
padding -

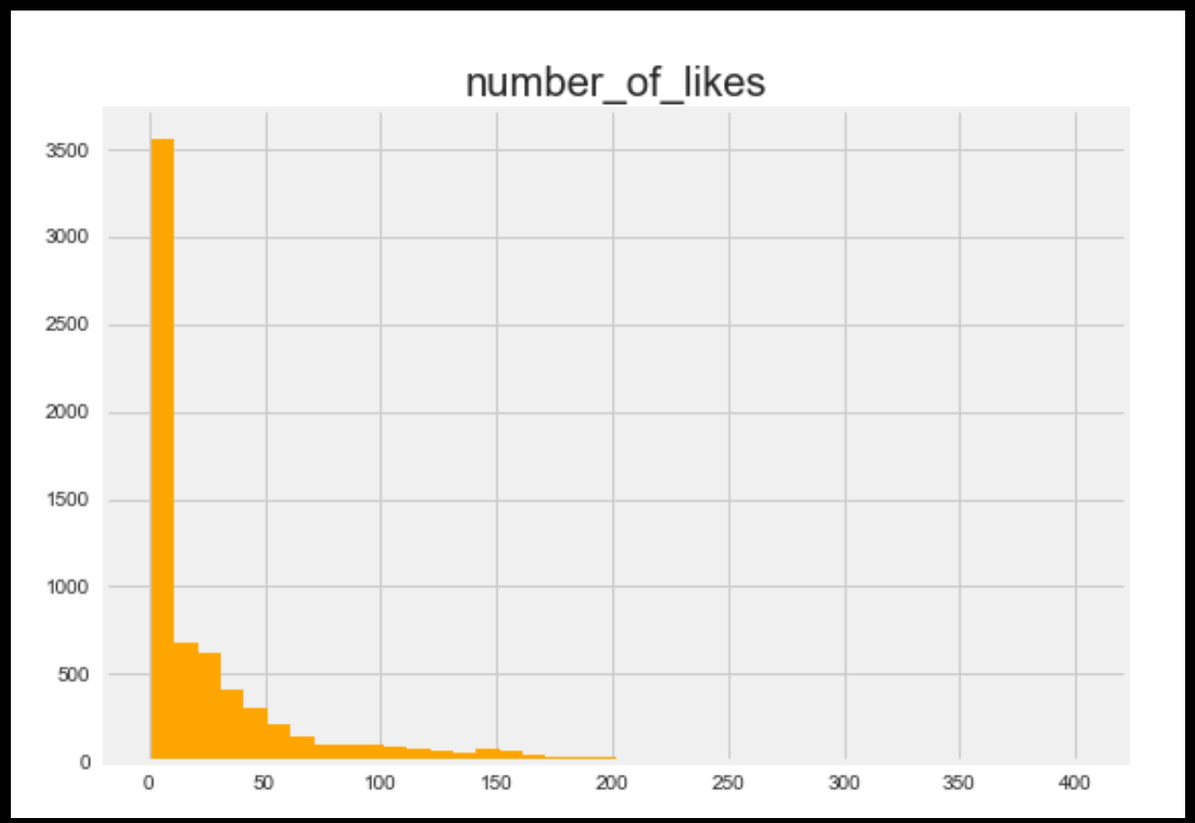
Filter Show all

▶ align-items
▶ border-bottom-color
▶ border-bottom-style

DATA COLLECTED



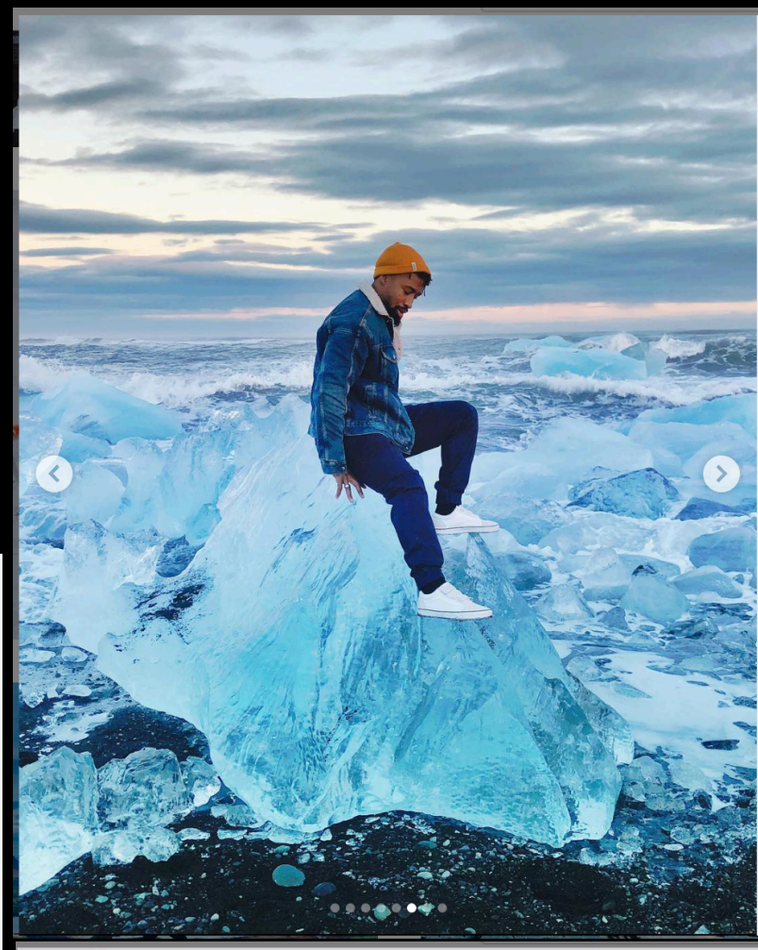


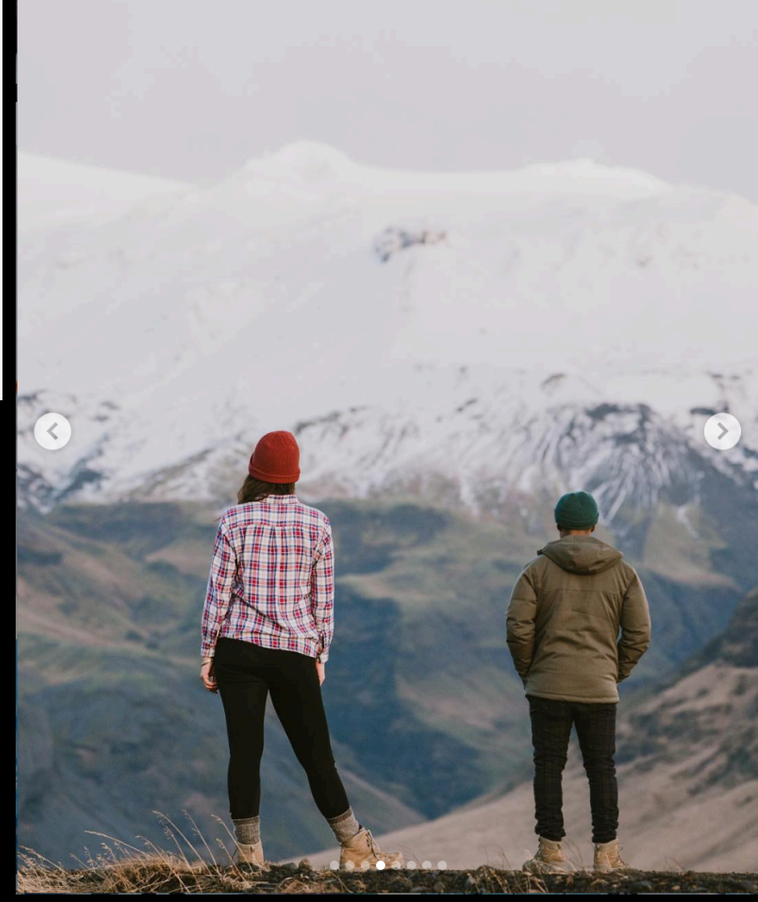


HIGHEST NUMBER OF LIKES

The caption read:

"Last day in Iceland. It's been an amazing 11 days. Got to meet and work with some amazing people, work with an amazing company/brand, and got to play model while being on some epic adventures. I will never forget this trip and I'm looking forward to the next. Iceland, it's been REAL, real! Shoutout to @jamesharnoisphoto and @hownottotravellikeabasicbitch for extending their stay with me and for these bad ass and mostly basic shots. 😂😂 *#iceland #niceland #LITlit #workaction #reykjavik #reiadventures #reishoot #bluelagoon #mountains #inclusivity #beach #hiking #whyhike #optoutside #adventures #travel #europe"*





USING THE SELENIUM WEB DRIVER

SCRAPING CODE

```
def scrape_page(webdriver, links, username):
        '''This function will go to all links provided
        and scrape each picture for the number of likes
        and the caption. It will only provide the caption if the
        identified user is the title'''
        picture_info = []

        for link in links:
                # Open new tab
                webdriver.execute_script("window.open('');")
                time.sleep(3)

                # Switch to the new window
                webdriver.switch_to.window(webdriver.window_handles[1])
                webdriver.get(link)
                time.sleep(5)
                try:
                        likes_list = webdriver.find_elements_by_class_name('zV_Nj')
                        if len(likes_list) != 0:
                                if len(likes_list) == 1:
                                        num_likes = 5

                                else:
                                        num_likes = int(likes_list[1].text.split(' ')[0]) + 1

                                try:
                                        title = webdriver.find_element_by_class_name('_6lAjh').text
                                        if title == username:
                                                caption_list = webdriver.find_elements_by_xpath("//div[@cl
                                                '''num_of_comments = len(caption_list)'''
                                                caption = caption_list[0].text
                                        else:
                                                caption = None
                                except:
                                        caption = None


                                picture_info.append([num_likes, caption])
                except:
                        num_likes = None
                webdriver.close()

                # Switch focus back to main tab
                webdriver.switch_to.window(webdriver.window_handles[0])
                time.sleep(5)
```

FREQUENCY OF WORDS

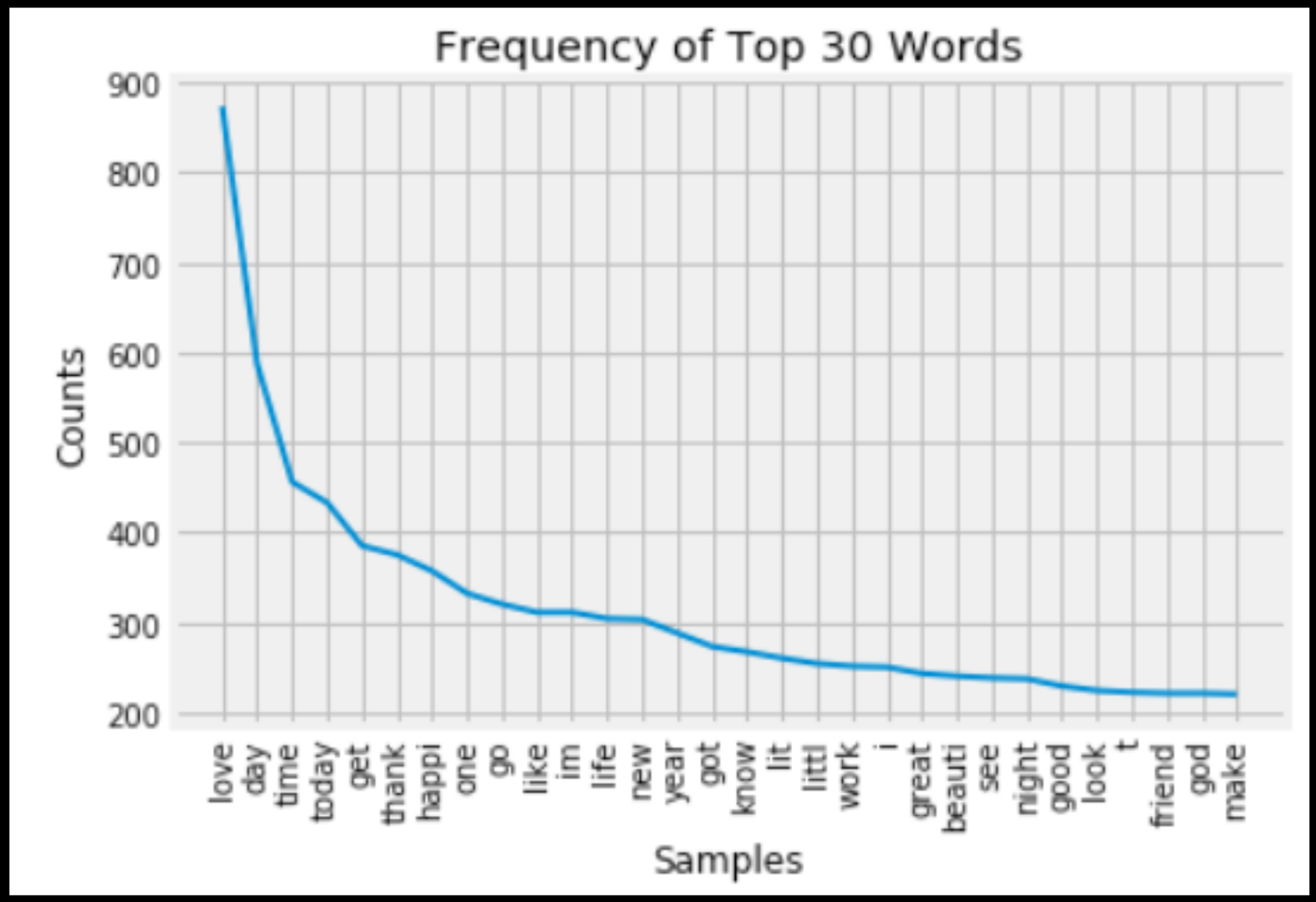
<u>Top 3%:</u>

beauti, day, friend, get, go, good, got, great, happy, im, know, life, like, **lit**, littl, love, make, new, one, see, thank, time, today, work, year.

<u>Top 5%:</u>

day, love, time today

Word used most often (10%)?



SKLEARN TF-IDF VECTOR

- Total words in corpus: 76,744
- I created a vector with a minimum document frequency of .25% which contained only 1% of the total words
- There were 10485 ignored words

```
ignored_words
```

```
{'central', 'schwinn', 'aspir', 'horatio', 'grandfath', 'suppo
'polo', 'legday', 'infinit', 'darkest', 'sperri', 'be', 'meme'
'saltlit', '5th', 'getcurious', 'weirdinterest', 'judgement',
'unclerich', 'yelp', 'trial', 'celeri', 'recognit', 'flour', '
'menwithstyl', 'warned', 'ifreakinglovehim', 'secrettosexi', '
'sweetsourmeatballsquinoaasparagushealthycleaneatsmovienightpo
'costco', 'ohcourtney16', 'hairston', 'downstair', 'handm', 'c
'workandworship', 'doubtkil', 'nightowl', 'shiner', 'amen', 's
'milleniumpark', 'champion', 'jlc_drum', 'wan', 'brownric', 'm
'hillfest2016', 'denverfashion', 'advic', 'section', 'maktub',
'turtl', 'althoughitsmonday', 'closeenough', 'outlet', 'freshc
'bridesmaid', 'myfitnessp', 'astrosopeningday', 'obnoxi', 'nug
'dreamingwithabrokenheart', 'olympics', 'gostro', 'cranberri',
'spreadthenew', 'halloween2014', 'justin_great', 'dreamgirl',
'purpl', 'tinier', '8pm', 'tgifriday', 'bartonspr', 'downtownh
'couldn', 'puls', 'jeremiah', 'itwork', 'lang1221lang', 'shawn
'cardashian', 'glimps', 'connecticut', 'forgiv', 'culturecola'
'alwaysworkinghard', 'saddest', 'skagenwatch', 'clcstudent', '
'thetruthbehindstudyabroad', 'summerof2012', 'appoint', 'fligh
'3dbqvqzmziht4', 'somehow', 'forgotten', 'bandanna', 'mywholew
'peep', 'itsbeenaday', 'mashedpotato', 'prayforpari', 'laser',
'shestooyoungforuoubro', 'quarterlyreview', 'theolog', 'ript',
'elishakelchphoto', 'hyatt', 'burinskyyi', 'cesarsa95', 'stran
'bowti', 'hash', 'snugglesforday', 'soy', 'housewarm', 'porkch
'mamafu', 'où', 'panel', 'ti', 'lovi', 'roster', 'rollersk', '
'dredhead', 'ilovemus', 'letter', 'attir', 'postthrow', 'anxie
'fomo', 'photobooth', 'pinkpant', 'gooey', 'woodland', 'showro
'mycheecksaresorefromsmil', 'goals', '1015pm', 'hohoho', 'easi
'monsterjam', 'earnit', 'separatingow', 'mybeautifulbabi', 're
```

RESULTS AND REFLECTION

- Random Forest with 100 trees score of: **-0.076**
- Linear Regression with a score of: **-0.054**
- Changing the parameters did not affect the scores in a significant way.
- My data is too sparse, there are errors in the data, and I should consider adding more features to the data set

NEXT STEPS

- Update scrape code to get accurate number of likes, collect timestamp of post,
- Add total number of followers of user per post
- Identify emoji's and use them in tokenizer