reddit

search engine

spring 2022 - information retrieval

ALOYSIUS ARNO WIPUTRA	1244139
ASHLEY VALDEZ	1257891
ANDREW MOLINA	1255187
DEVAAUM SHAH	1310191

Chapter 1

project background

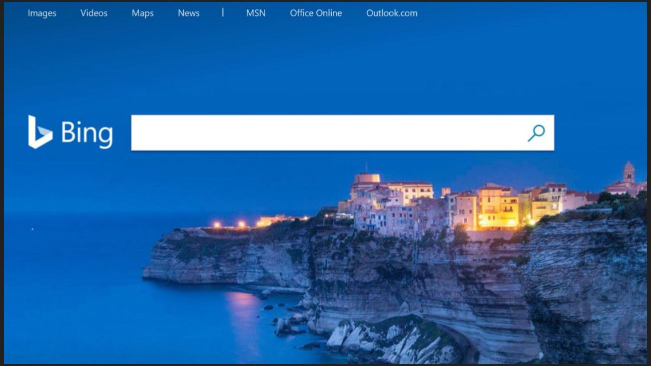
Reddit Overview

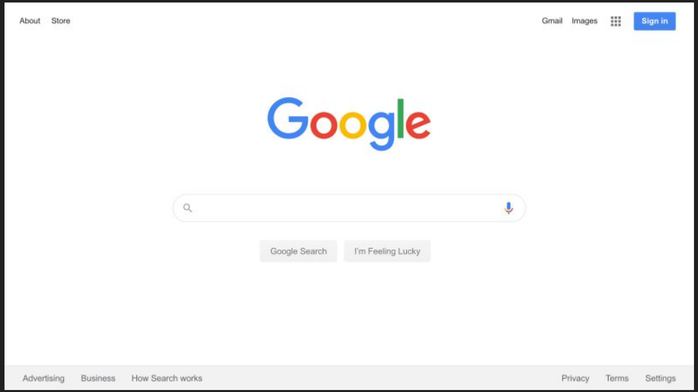
- Collection of forums that are shared and searched
  - Text posts
  - News/information
  - Images/videos
- Users search for communities (subreddits) that interest them
  - ex: video games or sports
- Users can upvote or downvote on posts in the subreddit which makes that particular post more popular on that specific subreddit
  - allows interesting posts to be at the top of sorting list and uninteresting posts at the bottom

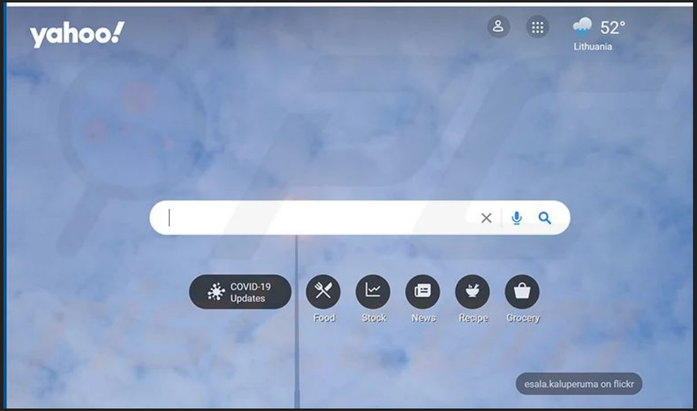
Other Search Engines

1. Google (Known as #1 search engine)
2. Bing
3. Yahoo

Competitors - more on the same level

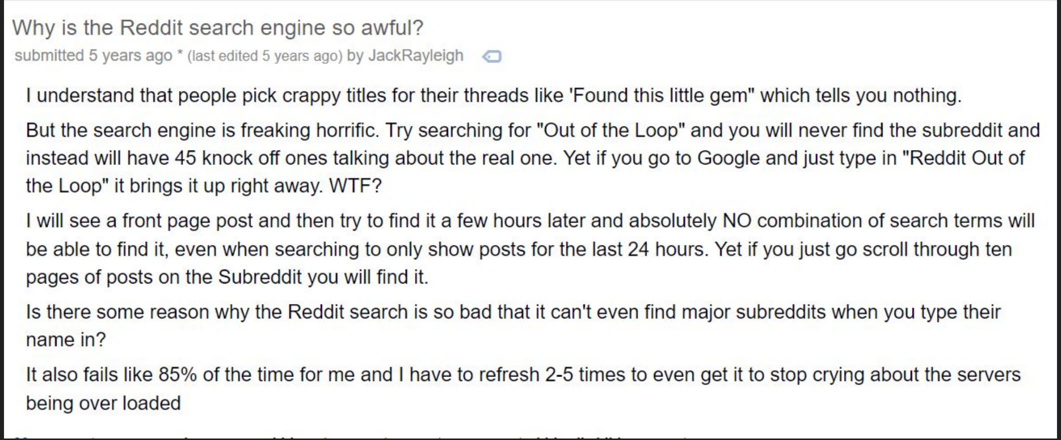






Issues with Reddit Search Engine

- posts that you search just don't show up
    - people rely on google instead
        - Using parameter "site:reddit.com"



Why is the Reddit search engine so awful?

submitted 5 years ago * (last edited 5 years ago) by JackRayleigh

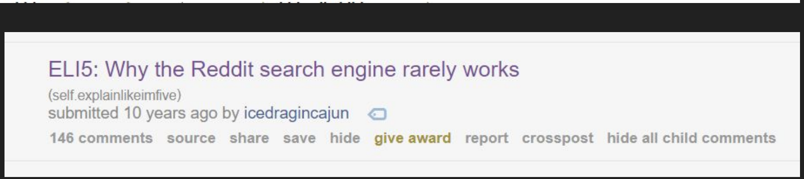
I understand that people pick crappy titles for their threads like 'Found this little gem" which tells you nothing.

But the search engine is freaking horrific. Try searching for "Out of the Loop" and you will never find the subreddit and instead will have 45 knock off ones talking about the real one. Yet if you go to Google and just type in "Reddit Out of the Loop" it brings it up right away. WTF?

I will see a front page post and then try to find it a few hours later and absolutely NO combination of search terms will be able to find it, even when searching to only show posts for the last 24 hours. Yet if you just go scroll through ten pages of posts on the Subreddit you will find it.

Is there some reason why the Reddit search is so bad that it can't even find major subreddits when you type their name in?

It also fails like 85% of the time for me and I have to refresh 2-5 times to even get it to stop crying about the servers being over loaded

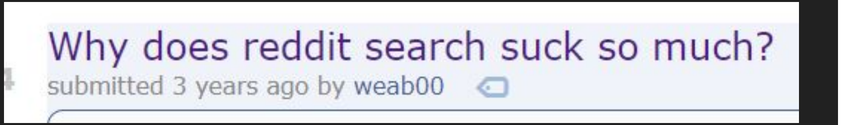


ELI5: Why the Reddit search engine rarely works

(self.explainlikeimfive)

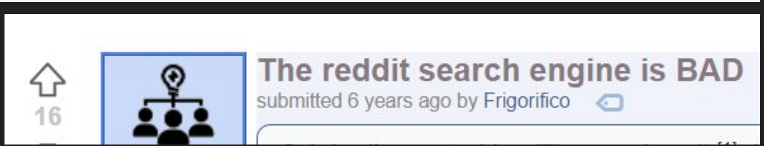
submitted 10 years ago by icedragincajun

146 comments source share save hide give award report crosspost hide all child comments



Why does reddit search suck so much?

submitted 3 years ago by weab00



The reddit search engine is BAD

submitted 6 years ago by Frigorifico

The Problem

- Our problem was optimizing the result where we obtain useful information from reddit
- We tackled this problem by creating a search engine that helps locate useful information on reddit

Methods

- Methods such as **lemmatization, stemming, tokenization,** and **stopword removal** are vital components to natural language processing and are key to the functioning of our search engine

Definitions

- **Tokenization** is essentially splitting raw text into individual words or terms called tokens
- **Stemming** identifies the root form of a word
- **Lemmatization** groups different forms of the same word
- **Stopword removal** breaks down the size of the raw text by removing commonly used words

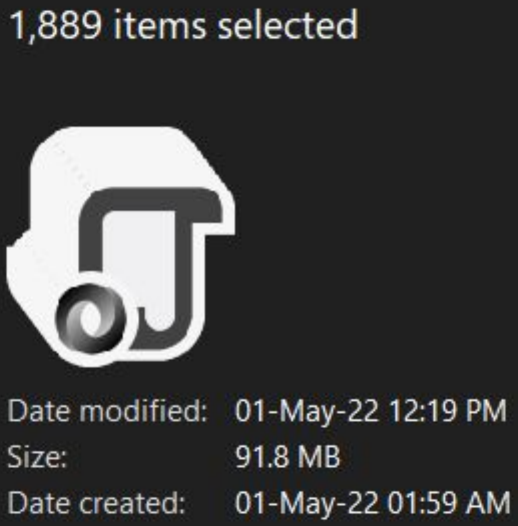
Chapter 2

data description

data description

Description

- 1889 json files
- 91.8 MB

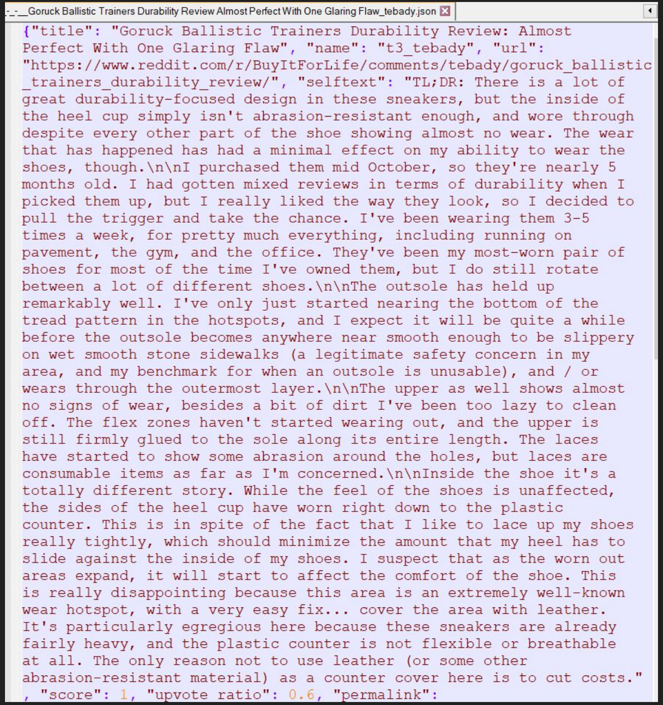


1,889 items selected

Date modified: 01-May-22 12:19 PM

Size: 91.8 MB

Date created: 01-May-22 01:59 AM

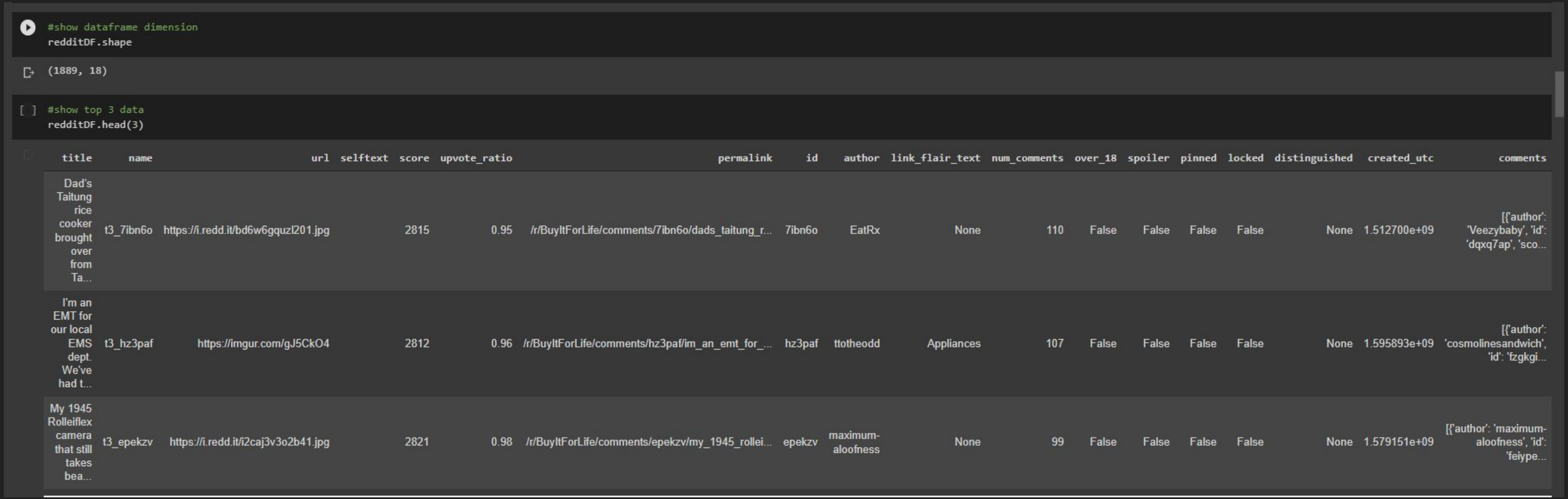


-_-__Goruck Ballistic Trainers Durability Review Almost Perfect With One Glaring Flaw_tebady.json

```
{"title": "Goruck Ballistic Trainers Durability Review: Almost
Perfect With One Glaring Flaw", "name": "t3_tebady", "url":
"https://www.reddit.com/r/BuyItForLife/comments/tebady/goruck_ballistic
_trainers_durability_review/", "selftext": "TL;DR: There is a lot of
great durability-focused design in these sneakers, but the inside of
the heel cup simply isn't abrasion-resistant enough, and wore through
despite every other part of the shoe showing almost no wear. The wear
that has happened has had a minimal effect on my ability to wear the
shoes, though.\n\nI purchased them mid October, so they're nearly 5
months old. I had gotten mixed reviews in terms of durability when I
picked them up, but I really liked the way they look, so I decided to
pull the trigger and take the chance. I've been wearing them 3-5
times a week, for pretty much everything, including running on
pavement, the gym, and the office. They've been my most-worn pair of
shoes for most of the time I've owned them, but I do still rotate
between a lot of different shoes.\n\nThe outsole has held up
remarkably well. I've only just started nearing the bottom of the
tread pattern in the hotspots, and I expect it will be quite a while
before the outsole becomes anywhere near smooth enough to be slippery
on wet smooth stone sidewalks (a legitimate safety concern in my
area, and my benchmark for when an outsole is unusable), and / or
wears through the outermost layer.\n\nThe upper as well shows almost
no signs of wear, besides a bit of dirt I've been too lazy to clean
off. The flex zones haven't started wearing out, and the upper is
still firmly glued to the sole along its entire length. The laces
have started to show some abrasion around the holes, but laces are
consumable items as far as I'm concerned.\n\nInside the shoe it's a
totally different story. While the feel of the shoes is unaffected,
the sides of the heel cup have worn right down to the plastic
counter. This is in spite of the fact that I like to lace up my shoes
really tightly, which should minimize the amount that my heel has to
slide against the inside of my shoes. I suspect that as the worn out
areas expand, it will start to affect the comfort of the shoe. This
is really disappointing because this area is an extremely well-known
wear hotspot, with a very easy fix... cover the area with leather.
It's particularly egregious here because these sneakers are already
fairly heavy, and the plastic counter is not flexible or breathable
at all. The only reason not to use leather (or some other
abrasion-resistant material) as a counter cover here is to cut costs."
, "score": 1, "upvote ratio": 0.6, "permalink":
```

Description

1889 entries, 18 columns



```
#show dataframe dimension
redditDF.shape
```

```
(1889, 18)
```

```
#show top 3 data
redditDF.head(3)
```

title	name	url	selftext	score	upvote_ratio	permalink	id	author	link_flair_text	num_comments	over_18	spoiler	pinned	locked	distinguished	created_utc	comments
Dad's Taitung rice cooker brought over from Ta...	t3_7ibn6o	https://i.redd.it/bd6w6gquzl201.jpg		2815	0.95	/r/BuyItForLife/comments/7ibn6o/dads_taitung_r...	7ibn6o	EatRx	None	110	False	False	False	False	None	1.512700e+09	[{'author': 'Veezybaby', 'id': 'dqxq7ap', 'sco...
I'm an EMT for our local EMS dept. We've had t...	t3_hz3paf	https://imgur.com/gJ5CkO4		2812	0.96	/r/BuyItForLife/comments/hz3paf/im_an_emt_for_...	hz3paf	ttotheodd	Appliances	107	False	False	False	False	None	1.595893e+09	[{'author': 'cosmolinesandwich', 'id': 'fzgkgi...
My 1945 Rolleiflex camera that still takes bea...	t3_epekzv	https://i.redd.it/i2caj3v3o2b41.jpg		2821	0.98	/r/BuyItForLife/comments/epekzv/my_1945_rollei...	epekzv	maximum-aloofness	None	99	False	False	False	False	None	1.579151e+09	[{'author': 'maximum-aloofness', 'id': 'feiype...

Description

All the column types and description

title	: post name	**link_flair_text**	: categories
name	: identifier	**num_comments**	: reply count
url	: content link	**over_18**	: NSFW?
selftext	: body	**spoiler**	: spoilers?
score	: total points	**pinned**	: sticky
upvote_ratio	: upvote/downvote	**locked**	: no replies
permalink	: link to thread	**distinguished**	: mod highlight
id	: post identifier	**created_utc**	: time made
author	: user who posted	**comments**	: content

data description

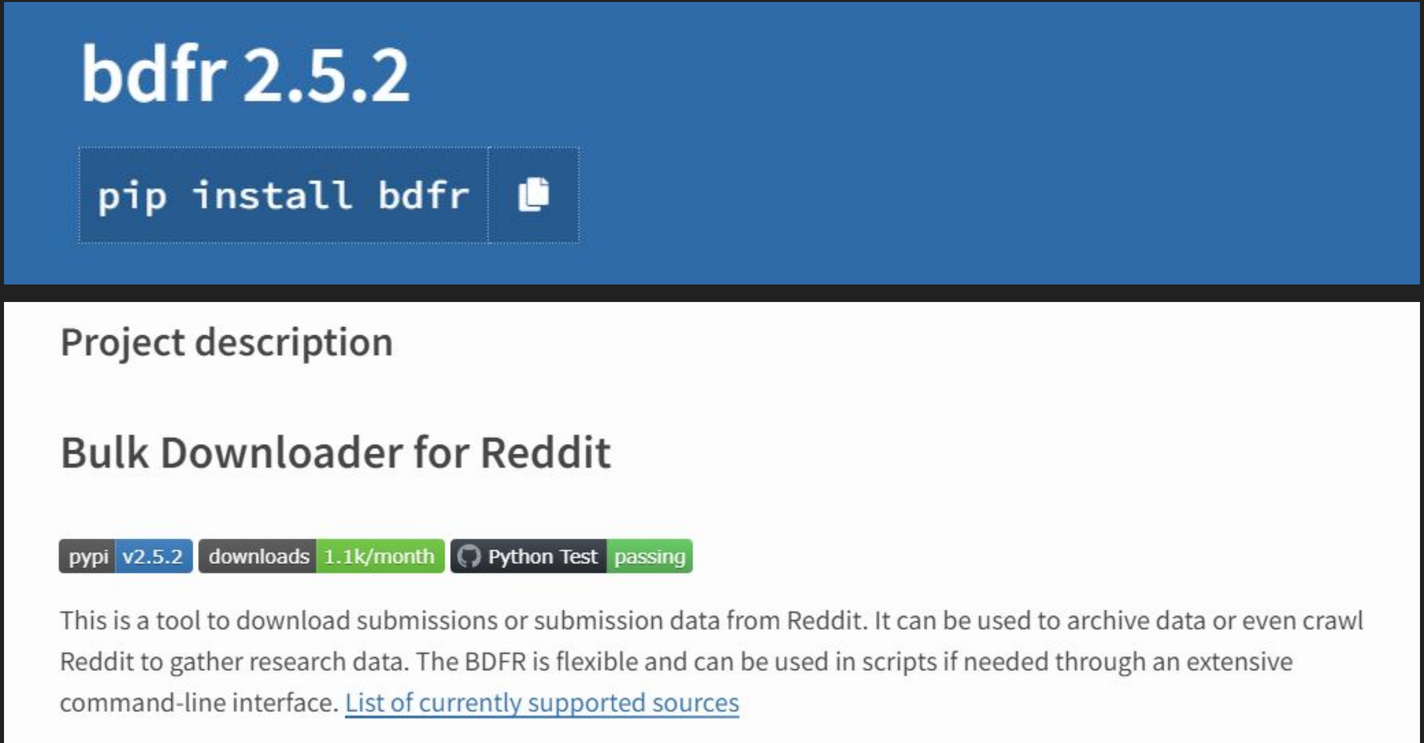
Methodology

Ali Parlakçı
aliparlakci

I don't know what I would do if Computer Science didn't exist.

28 followers · 6 following

Sabancı University



bdfr 2.5.2

```
pip install bdfr
```

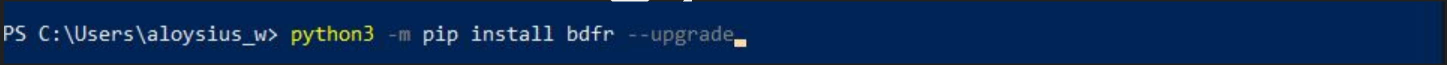
Project description

Bulk Downloader for Reddit

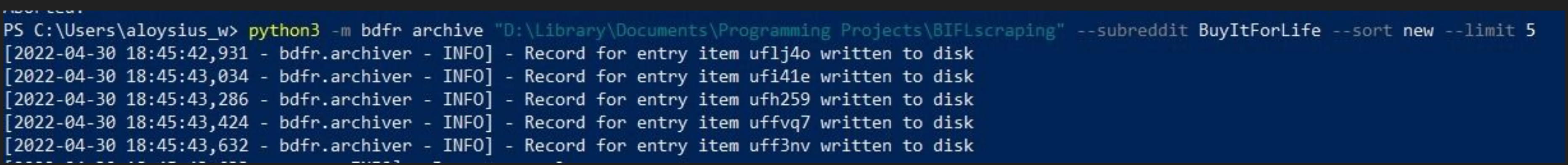
This is a tool to download submissions or submission data from Reddit. It can be used to archive data or even crawl Reddit to gather research data. The BDFR is flexible and can be used in scripts if needed through an extensive command-line interface. List of currently supported sources

data description

Methodology



```
PS C:\Users\aloysius_w> python3 -m pip install bdfr --upgrade
```



```
PS C:\Users\aloysius_w> python3 -m bdfr archive "D:\Library\Documents\Programming Projects\BIFLscraping" --subreddit BuyItForLife --sort new --limit 5
[2022-04-30 18:45:42,931 - bdfr.archiver - INFO] - Record for entry item uflj4o written to disk
[2022-04-30 18:45:43,034 - bdfr.archiver - INFO] - Record for entry item ufi41e written to disk
[2022-04-30 18:45:43,286 - bdfr.archiver - INFO] - Record for entry item ufh259 written to disk
[2022-04-30 18:45:43,424 - bdfr.archiver - INFO] - Record for entry item uffvq7 written to disk
[2022-04-30 18:45:43,632 - bdfr.archiver - INFO] - Record for entry item uff3nv written to disk
```

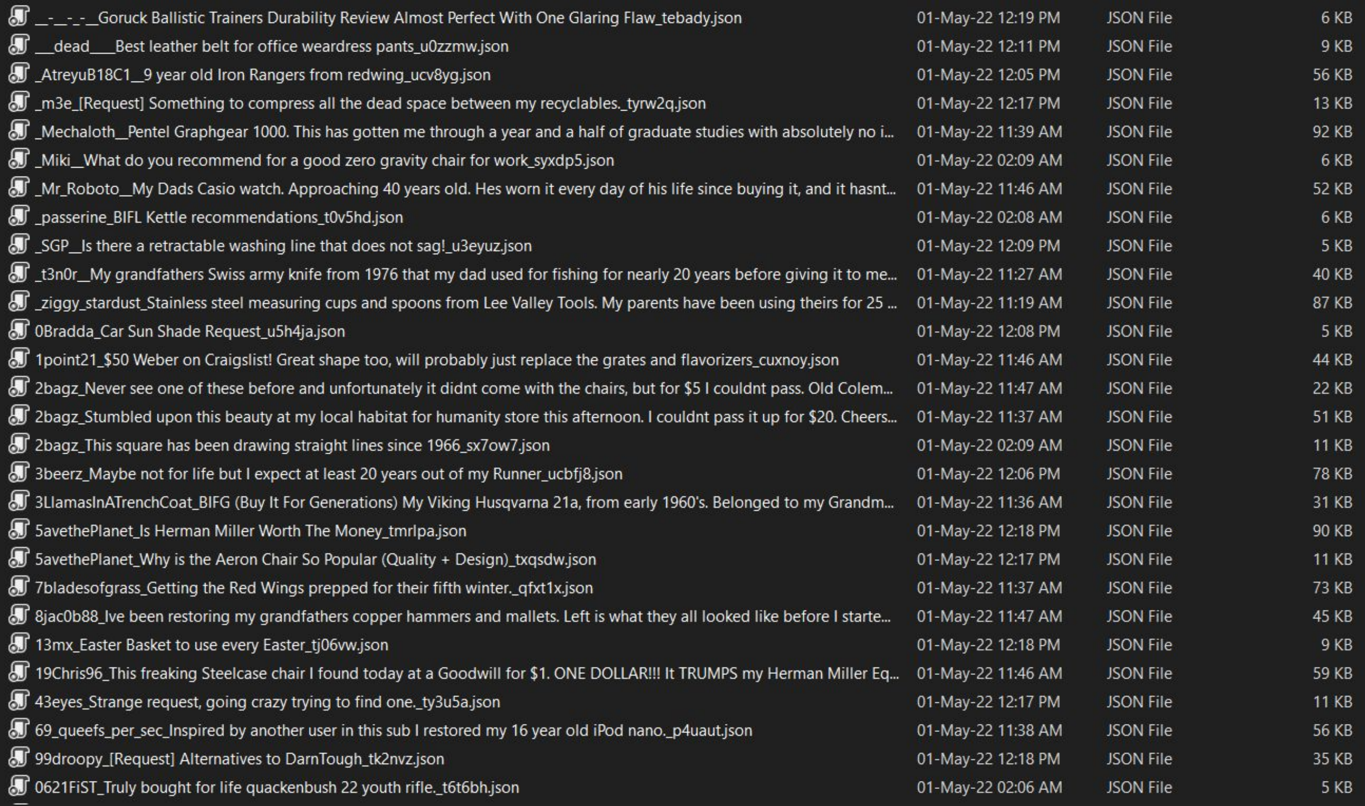
- install and run with powershell and python3
- with the following arguments
  - archive
    - other options are **clone** and **download**
  - --subreddit
  - --sort
    - by default, sorts by new

Issues Encountered

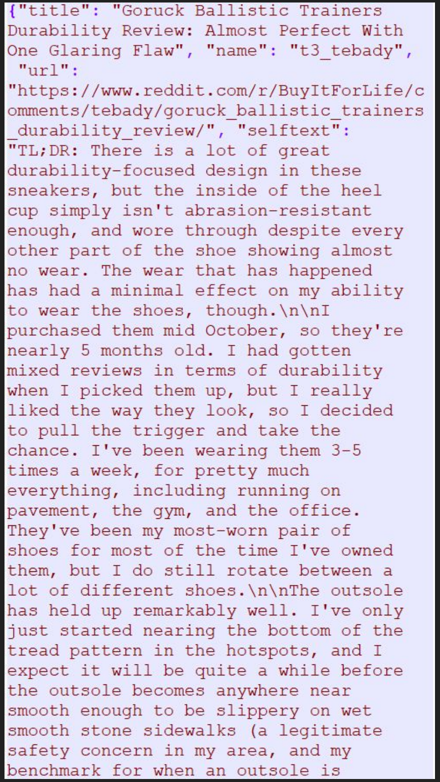
- script ran for a few hours
- documentation said there are no time limits
- did not make a complete archive of the subreddit
    - ran multiple times with different arguments to get more data

data description

Sample



__-__-_-__Goruck Ballistic Trainers Durability Review Almost Perfect With One Glaring Flaw_tebady.json	01-May-22 12:19 PM	JSON File	6 KB
__dead___Best leather belt for office weardress pants_u0zzmw.json	01-May-22 12:11 PM	JSON File	9 KB
_AtreyuB18C1__9 year old Iron Rangers from redwing_ucv8yg.json	01-May-22 12:05 PM	JSON File	56 KB
_m3e_[Request] Something to compress all the dead space between my recyclables._tyrw2q.json	01-May-22 12:17 PM	JSON File	13 KB
_Mechaloth__Pentel Graphgear 1000. This has gotten me through a year and a half of graduate studies with absolutely no i...	01-May-22 11:39 AM	JSON File	92 KB
_Miki__What do you recommend for a good zero gravity chair for work_syxdp5.json	01-May-22 02:09 AM	JSON File	6 KB
_Mr_Roboto__My Dads Casio watch. Approaching 40 years old. Hes worn it every day of his life since buying it, and it hasnt...	01-May-22 11:46 AM	JSON File	52 KB
_passerine_BIFL Kettle recommendations_t0v5hd.json	01-May-22 02:08 AM	JSON File	6 KB
_SGP__Is there a retractable washing line that does not sag!_u3eyuz.json	01-May-22 12:09 PM	JSON File	5 KB
_t3n0r__My grandfathers Swiss army knife from 1976 that my dad used for fishing for nearly 20 years before giving it to me...	01-May-22 11:27 AM	JSON File	40 KB
_ziggy_stardust_Stainless steel measuring cups and spoons from Lee Valley Tools. My parents have been using theirs for 25 ...	01-May-22 11:19 AM	JSON File	87 KB
0Bradda_Car Sun Shade Request_u5h4ja.json	01-May-22 12:08 PM	JSON File	5 KB
1point21_$50 Weber on Craigslist! Great shape too, will probably just replace the grates and flavorizers_cuxnoy.json	01-May-22 11:46 AM	JSON File	44 KB
2bagz_Never see one of these before and unfortunately it didnt come with the chairs, but for $5 I couldnt pass. Old Colem...	01-May-22 11:47 AM	JSON File	22 KB
2bagz_Stumbled upon this beauty at my local habitat for humanity store this afternoon. I couldnt pass it up for $20. Cheers...	01-May-22 11:37 AM	JSON File	51 KB
2bagz_This square has been drawing straight lines since 1966_sx7ow7.json	01-May-22 02:09 AM	JSON File	11 KB
3beerz_Maybe not for life but I expect at least 20 years out of my Runner_ucbfj8.json	01-May-22 12:06 PM	JSON File	78 KB
3LlamasInATrenchCoat_BIFG (Buy It For Generations) My Viking Husqvarna 21a, from early 1960's. Belonged to my Grandm...	01-May-22 11:36 AM	JSON File	31 KB
5avethePlanet_Is Herman Miller Worth The Money_tmrlpa.json	01-May-22 12:18 PM	JSON File	90 KB
5avethePlanet_Why is the Aeron Chair So Popular (Quality + Design)_txqsdw.json	01-May-22 12:17 PM	JSON File	11 KB
7bladesofgrass_Getting the Red Wings prepped for their fifth winter._qfxt1x.json	01-May-22 11:37 AM	JSON File	73 KB
8jac0b88_Ive been restoring my grandfathers copper hammers and mallets. Left is what they all looked like before I starte...	01-May-22 11:47 AM	JSON File	45 KB
13mx_Easter Basket to use every Easter_tj06vw.json	01-May-22 12:18 PM	JSON File	9 KB
19Chris96_This freaking Steelcase chair I found today at a Goodwill for $1. ONE DOLLAR!!! It TRUMPS my Herman Miller Eq...	01-May-22 11:46 AM	JSON File	59 KB
43eyes_Strange request, going crazy trying to find one._ty3u5a.json	01-May-22 12:17 PM	JSON File	11 KB
69_queefs_per_sec_Inspired by another user in this sub I restored my 16 year old iPod nano._p4uaut.json	01-May-22 11:38 AM	JSON File	56 KB
99droopy_[Request] Alternatives to DarnTough_tk2nvz.json	01-May-22 12:18 PM	JSON File	35 KB
0621FiST_Truly bought for life quackenbush 22 youth rifle._t6t6bh.json	01-May-22 02:06 AM	JSON File	5 KB



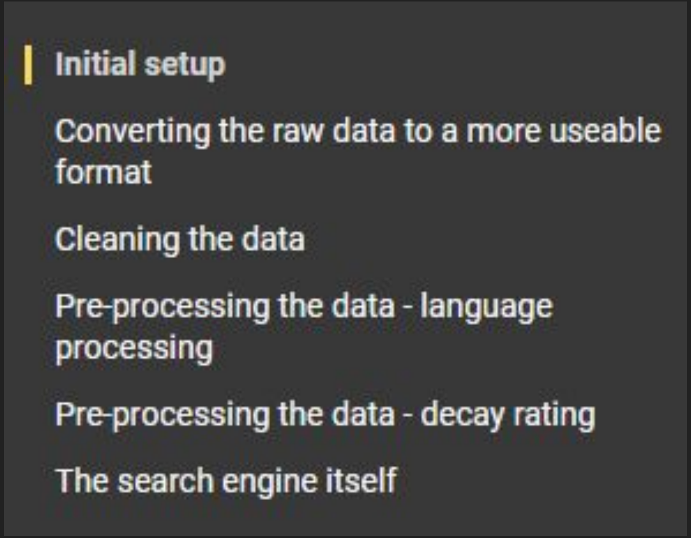
{"title": "Goruck Ballistic Trainers Durability Review: Almost Perfect With One Glaring Flaw", "name": "t3_tebady", "url": "https://www.reddit.com/r/BuyItForLife/comments/tebady/goruck_ballistic_trainers_durability_review/", "selftext": "TL;DR: There is a lot of great durability-focused design in these sneakers, but the inside of the heel cup simply isn't abrasion-resistant enough, and wore through despite every other part of the shoe showing almost no wear. The wear that has happened has had a minimal effect on my ability to wear the shoes, though.\n\nI purchased them mid October, so they're nearly 5 months old. I had gotten mixed reviews in terms of durability when I picked them up, but I really liked the way they look, so I decided to pull the trigger and take the chance. I've been wearing them 3-5 times a week, for pretty much everything, including running on pavement, the gym, and the office. They've been my most-worn pair of shoes for most of the time I've owned them, but I do still rotate between a lot of different shoes.\n\nThe outsole has held up remarkably well. I've only just started nearing the bottom of the tread pattern in the hotspots, and I expect it will be quite a while before the outsole becomes anywhere near smooth enough to be slippery on wet smooth stone sidewalks (a legitimate safety concern in my area, and my benchmark for when an outsole is

Chapter 3

approach

our approach

Table of Contents



Initial setup

Converting the raw data to a more useable format

Cleaning the data

Pre-processing the data - language processing

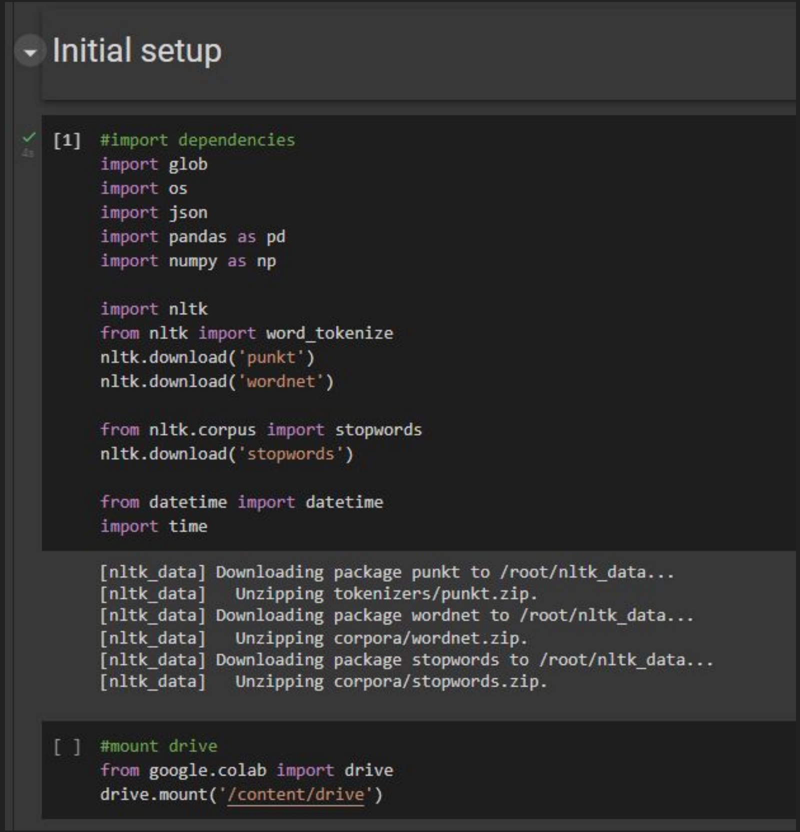
Pre-processing the data - decay rating

The search engine itself

- organized by categories

our approach

Initial Setup



Initial setup

```
[1] #import dependencies
    import glob
    import os
    import json
    import pandas as pd
    import numpy as np

    import nltk
    from nltk import word_tokenize
    nltk.download('punkt')
    nltk.download('wordnet')

    from nltk.corpus import stopwords
    nltk.download('stopwords')

    from datetime import datetime
    import time
```

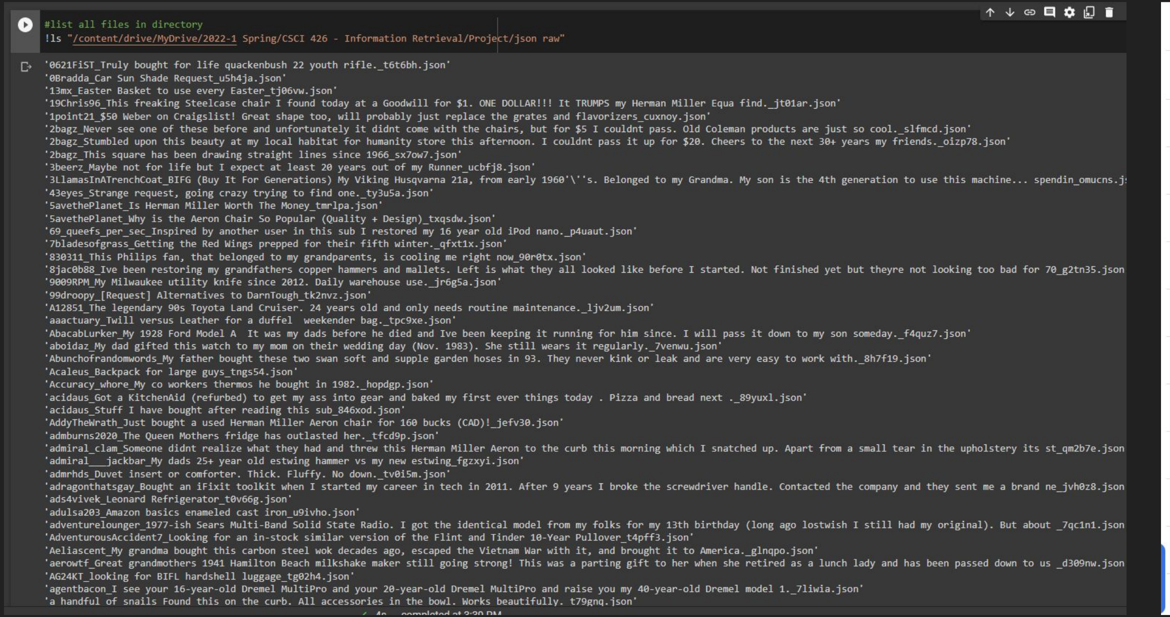
```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Unzipping corpora/wordnet.zip.
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

```
[ ] #mount drive
    from google.colab import drive
    drive.mount('/content/drive')
```

- glob, os, json
  - data imports
- pandas, numpy
  - processing and calculation
- nltk
  - natural language processing
    - tokenization
    - stopword removal
    - lemmatization
- datetime
  - decay factor
- drive
  - data storage

our approach

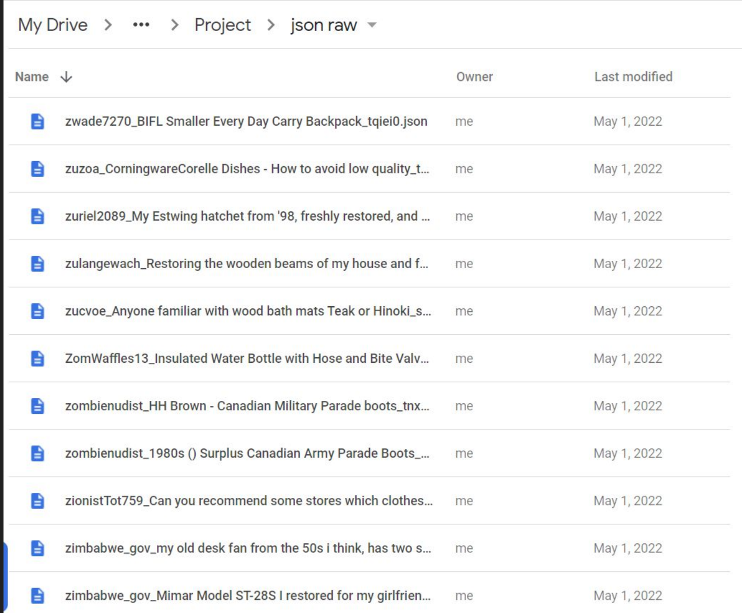
Data Conversion



```
#list all files in directory
!ls "/content/drive/MyDrive/2022-1 Spring/CSCI 426 - Information Retrieval/Project/json raw"
```

```
'0621FiST_Truly bought for life quackenbush 22 youth rifle._t6t6bh.json'
'0Bradda_Car Sun Shade Request_u5h4ja.json'
'13mx_Easter Basket to use every Easter_tj06vw.json'
'19Chris96_This freaking Steelcase chair I found today at a Goodwill for $1. ONE DOLLAR!!! It TRUMPS my Herman Miller Equa find._jt01ar.json'
'1point21_$50 Weber on Craigslist! Great shape too, will probably just replace the grates and flavorizers_cuxnoy.json'
'2bagz_Never see one of these before and unfortunately it didnt come with the chairs, but for $5 I couldnt pass. Old Coleman products are just so cool._slfmcd.json'
'2bagz_Stumbled upon this beauty at my local habitat for humanity store this afternoon. I couldnt pass it up for $20. Cheers to the next 30+ years my friends._oizp78.json'
'2bagz_This square has been drawing straight lines since 1966_sx7ow7.json'
'3beerz_Maybe not for life but I expect at least 20 years out of my Runner_ucbfj8.json'
'3LlamasInATrenchCoat_BIFG (Buy It For Generations) My Viking Husqvarna 21a, from early 1960'\''s. Belonged to my Grandma. My son is the 4th generation to use this machine... spendin_omucns.j
'43eyes_Strange request, going crazy trying to find one._ty3u5a.json'
'5avethePlanet_Is Herman Miller Worth The Money_tmrlpa.json'
'5avethePlanet_Why is the Aeron Chair So Popular (Quality + Design)_txqsdw.json'
'69_queefs_per_sec_Inspired by another user in this sub I restored my 16 year old iPod nano._p4uaut.json'
'7bladesofgrass_Getting the Red Wings prepped for their fifth winter._qfxt1x.json'
'830311_This Philips fan, that belonged to my grandparents, is cooling me right now_90r0tx.json'
'8jac0b88_Ive been restoring my grandfathers copper hammers and mallets. Left is what they all looked like before I started. Not finished yet but theyre not looking too bad for 70_g2tn35.json
'9009RPM_My Milwaukee utility knife since 2012. Daily warehouse use._jr6g5a.json'
'99droopy_[Request] Alternatives to DarnTough_tk2nvz.json'
'A12851_The legendary 90s Toyota Land Cruiser. 24 years old and only needs routine maintenance._ljv2um.json'
'aaactuary_Twill versus Leather for a duffel  weekender bag._tpc9xe.json'
'AbacabLurker_My 1928 Ford Model A  It was my dads before he died and Ive been keeping it running for him since. I will pass it down to my son someday._f4quz7.json'
'aboidaz_My dad gifted this watch to my mom on their wedding day (Nov. 1983). She still wears it regularly._7venwu.json'
'Abunchofrandomwords_My father bought these two swan soft and supple garden hoses in 93. They never kink or leak and are very easy to work with._8h7f19.json'
'Acaleus_Backpack for large guys_tngs54.json'
'Accuracy_whore_My co workers thermos he bought in 1982._hopdgp.json'
'acidaus_Got a KitchenAid (refurbed) to get my ass into gear and baked my first ever things today . Pizza and bread next ._89yuxl.json'
'acidaus_Stuff I have bought after reading this sub_846xod.json'
'AddyTheWrath_Just bought a used Herman Miller Aeron chair for 160 bucks (CAD)!_jefv30.json'
'admburns2020_The Queen Mothers fridge has outlasted her._tfcd9p.json'
'admiral_clam_Someone didnt realize what they had and threw this Herman Miller Aeron to the curb this morning which I snatched up. Apart from a small tear in the upholstery its st_qm2b7e.json
'admiral___jackbar_My dads 25+ year old estwing hammer vs my new estwing_fgzxyi.json'
'admrhds_Duvet insert or comforter. Thick. Fluffy. No down._tv0i5m.json'
'adragonthatsgay_Bought an iFixit toolkit when I started my career in tech in 2011. After 9 years I broke the screwdriver handle. Contacted the company and they sent me a brand ne_jvh0z8.json
'ads4vivek_Leonard Refrigerator_t0v66g.json'
'adulsa203_Amazon basics enameled cast iron_u9ivho.json'
'adventurelounger_1977-ish Sears Multi-Band Solid State Radio. I got the identical model from my folks for my 13th birthday (long ago lostwish I still had my original). But about _7qc1n1.json
'AdventurousAccident7_Looking for an in-stock similar version of the Flint and Tinder 10-Year Pullover_t4pff3.json'
'Aeliascent_My grandma bought this carbon steel wok decades ago, escaped the Vietnam War with it, and brought it to America._glnqpo.json'
'aerowtf_Great grandmothers 1941 Hamilton Beach milkshake maker still going strong! This was a parting gift to her when she retired as a lunch lady and has been passed down to us _d309nw.json
'AG24KT_looking for BIFL hardshell luggage_tg02h4.json'
'agentbacon_I see your 16-year-old Dremel MultiPro and your 20-year-old Dremel MultiPro and raise you my 40-year-old Dremel model 1._7liwia.json'
'a handful of snails Found this on the curb. All accessories in the bowl. Works beautifully._t79gnq.json'
```

✓ 4s completed at 3:30 PM

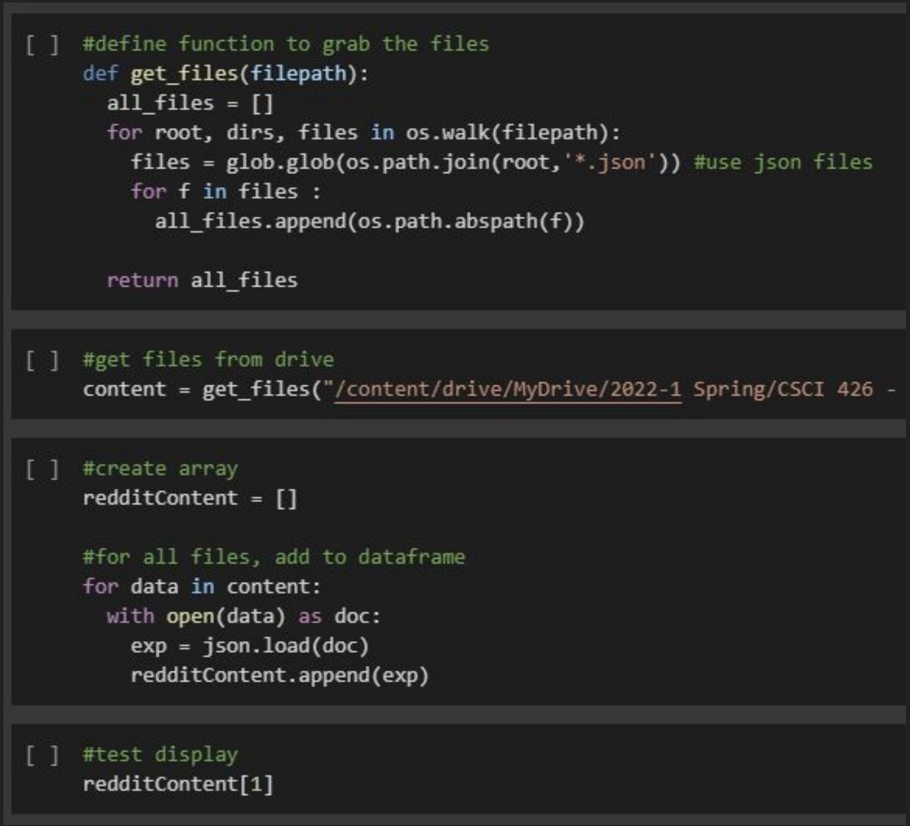


My Drive > ••• > Project > json raw

Name ↓	Owner	Last modified
zwade7270_BIFL Smaller Every Day Carry Backpack_tqiei0.json	me	May 1, 2022
zuzoa_CorningwareCorelle Dishes - How to avoid low quality_t...	me	May 1, 2022
zuriel2089_My Estwing hatchet from '98, freshly restored, and ...	me	May 1, 2022
zulangewach_Restoring the wooden beams of my house and f...	me	May 1, 2022
zucvoe_Anyone familiar with wood bath mats Teak or Hinoki_s...	me	May 1, 2022
ZomWaffles13_Insulated Water Bottle with Hose and Bite Valv...	me	May 1, 2022
zombienudist_HH Brown - Canadian Military Parade boots_tnx...	me	May 1, 2022
zombienudist_1980s () Surplus Canadian Army Parade Boots_...	me	May 1, 2022
zionistTot759_Can you recommend some stores which clothes...	me	May 1, 2022
zimbabwe_gov_my old desk fan from the 50s i think, has two s...	me	May 1, 2022
zimbabwe_gov_Mimar Model ST-28S I restored for my girlfrien...	me	May 1, 2022

our approach

Data Conversion



```
[ ]  #define function to grab the files
     def get_files(filepath):
       all_files = []
       for root, dirs, files in os.walk(filepath):
         files = glob.glob(os.path.join(root,'*.json')) #use json files
         for f in files :
           all_files.append(os.path.abspath(f))

       return all_files
```

```
[ ]  #get files from drive
     content = get_files("/content/drive/MyDrive/2022-1 Spring/CSCI 426 -
```

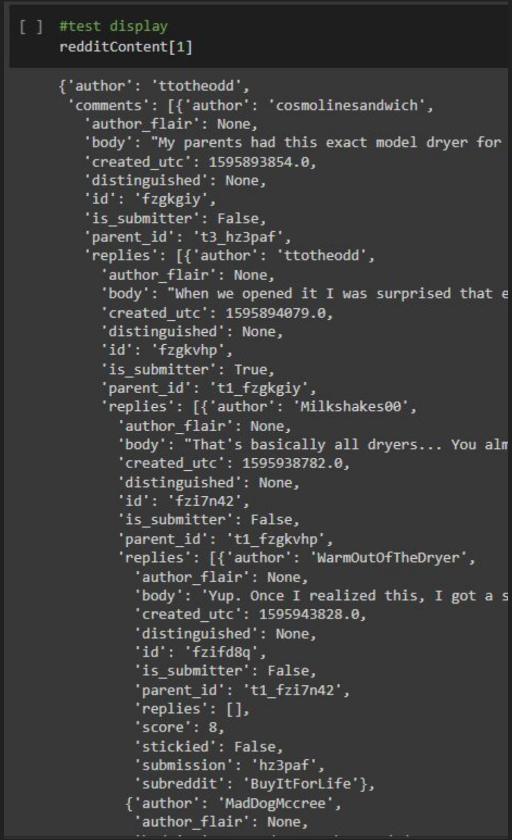
```
[ ]  #create array
     redditContent = []

     #for all files, add to dataframe
     for data in content:
       with open(data) as doc:
         exp = json.load(doc)
         redditContent.append(exp)
```

```
[ ]  #test display
     redditContent[1]
```

- function to **get all the .json files** in a folder
- **fetch the raw files** from drive
- **append all files** to an array

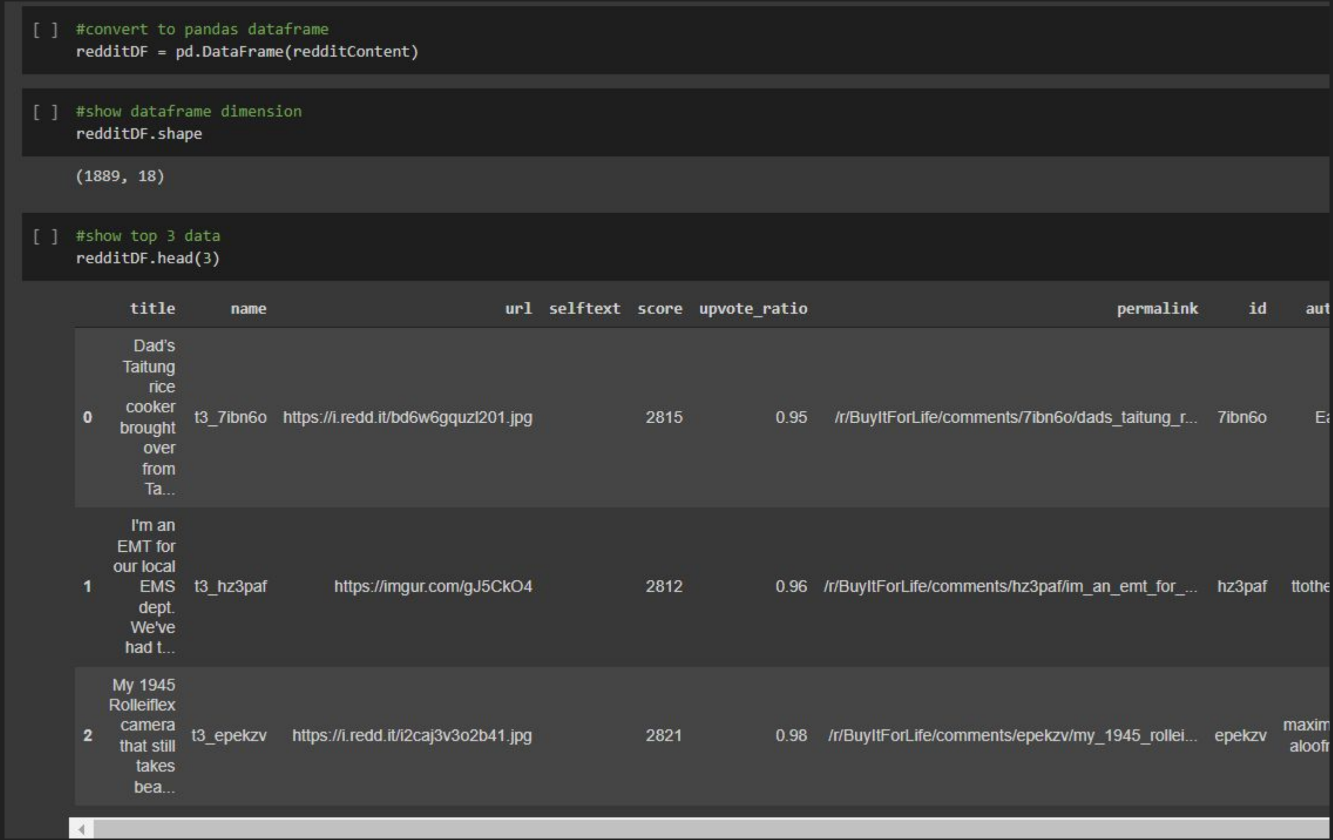
our approach

Data Conversion



```
[ ]  #test display
     redditContent[1]
```

```
{'author': 'ttotheodd',
 'comments': [{'author': 'cosmolinesandwich',
   'author_flair': None,
   'body': "My parents had this exact model dryer for
   'created_utc': 1595893854.0,
   'distinguished': None,
   'id': 'fzgkgiy',
   'is_submitter': False,
   'parent_id': 't3_hz3paf',
   'replies': [{'author': 'ttotheodd',
     'author_flair': None,
     'body': "When we opened it I was surprised that e
     'created_utc': 1595894079.0,
     'distinguished': None,
     'id': 'fzgkvhp',
     'is_submitter': True,
     'parent_id': 't1_fzgkgiy',
     'replies': [{'author': 'Milkshakes00',
       'author_flair': None,
       'body': "That's basically all dryers... You alm
       'created_utc': 1595938782.0,
       'distinguished': None,
       'id': 'fzi7n42',
       'is_submitter': False,
       'parent_id': 't1_fzgkvhp',
       'replies': [{'author': 'WarmOutOfTheDryer',
         'author_flair': None,
         'body': 'Yup. Once I realized this, I got a s
         'created_utc': 1595943828.0,
         'distinguished': None,
         'id': 'fzifd8q',
         'is_submitter': False,
         'parent_id': 't1_fzi7n42',
         'replies': [],
         'score': 8,
         'stickied': False,
         'submission': 'hz3paf',
         'subreddit': 'BuyItForLife'},
        {'author': 'MadDogMccree',
         'author_flair': None,
```



```
[ ]  #convert to pandas dataframe
     redditDF = pd.DataFrame(redditContent)
```

```
[ ]  #show dataframe dimension
     redditDF.shape
```

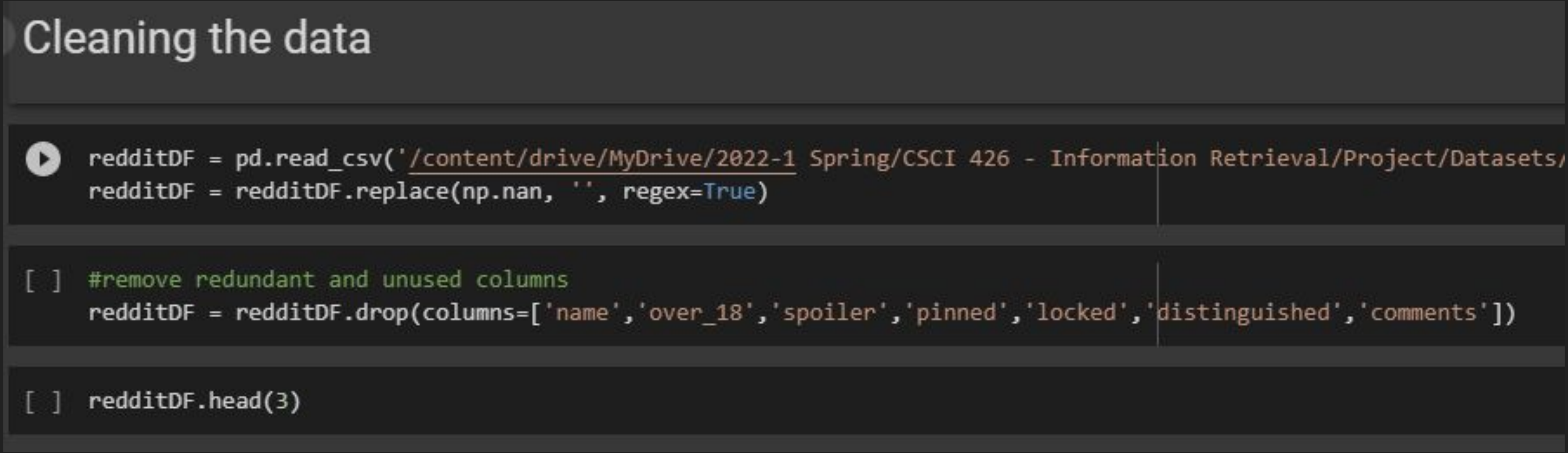
```
(1889, 18)
```

```
[ ]  #show top 3 data
     redditDF.head(3)
```

	title	name	url	selftext	score	upvote_ratio	permalink	id	aut
0	Dad's Taitung rice cooker brought over from Ta...	t3_7ibn6o	https://i.redd.it/bd6w6gquzl201.jpg		2815	0.95	/r/BuyItForLife/comments/7ibn6o/dads_taitung_r...	7ibn6o	Ea
1	I'm an EMT for our local EMS dept. We've had t...	t3_hz3paf	https://imgur.com/gJ5CkO4		2812	0.96	/r/BuyItForLife/comments/hz3paf/im_an_emt_for_...	hz3paf	ttothe
2	My 1945 Rolleiflex camera that still takes bea...	t3_epekzv	https://i.redd.it/i2caj3v3o2b41.jpg		2821	0.98	/r/BuyItForLife/comments/epekzv/my_1945_rollei...	epekzv	maxim aloofr

our approach

Cleaning the Data



Cleaning the data

```
redditDF = pd.read_csv('/content/drive/MyDrive/2022-1 Spring/CSCI 426 - Information Retrieval/Project/Datasets/
redditDF = redditDF.replace(np.nan, '', regex=True)
```

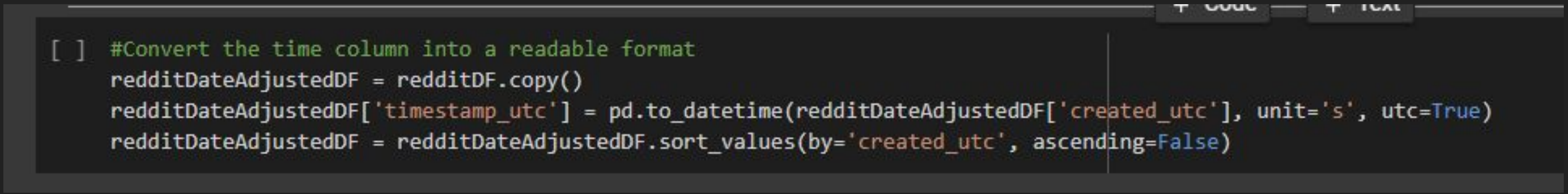
```
#remove redundant and unused columns
redditDF = redditDF.drop(columns=['name','over_18','spoiler','pinned','locked','distinguished','comments'])
```

```
redditDF.head(3)
```

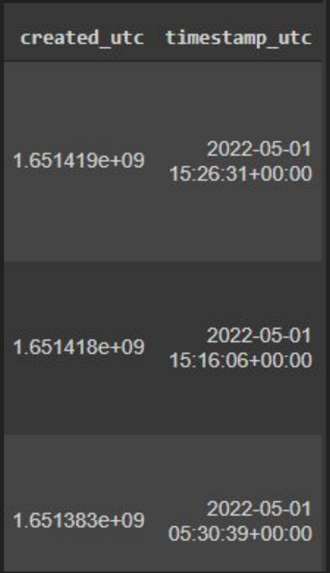
drop NAN entries, drop unused columns

our approach

Cleaning the Data



```
#Convert the time column into a readable format
redditDateAdjustedDF = redditDF.copy()
redditDateAdjustedDF['timestamp_utc'] = pd.to_datetime(redditDateAdjustedDF['created_utc'], unit='s', utc=True)
redditDateAdjustedDF = redditDateAdjustedDF.sort_values(by='created_utc', ascending=False)
```

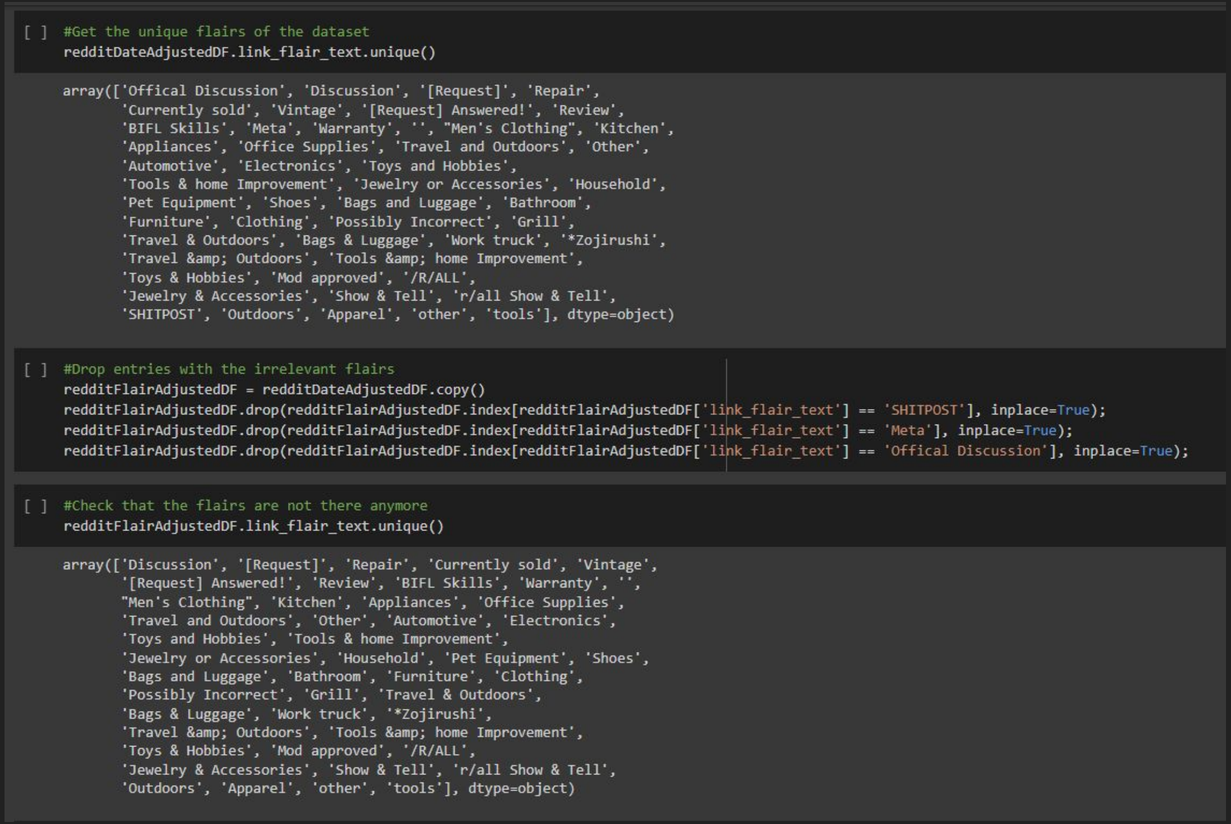


created_utc	timestamp_utc
1.651419e+09	2022-05-01 15:26:31+00:00
1.651418e+09	2022-05-01 15:16:06+00:00
1.651383e+09	2022-05-01 05:30:39+00:00

- pandas function to convert unix timestamp to a general format

our approach

Cleaning the Data

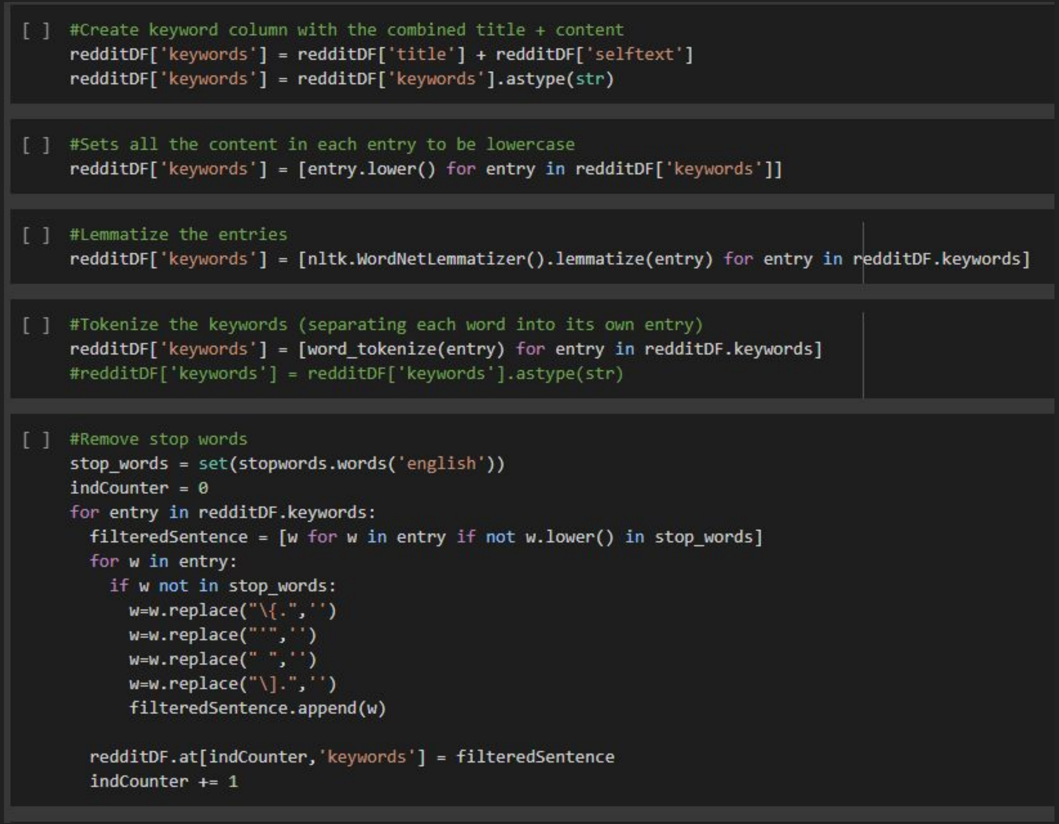


```
[ ]  #Get the unique flairs of the dataset
     redditDateAdjustedDF.link_flair_text.unique()

     array(['Offical Discussion', 'Discussion', '[Request]', 'Repair',
            'Currently sold', 'Vintage', '[Request] Answered!', 'Review',
            'BIFL Skills', 'Meta', 'Warranty', '', "Men's Clothing", 'Kitchen',
            'Appliances', 'Office Supplies', 'Travel and Outdoors', 'Other',
            'Automotive', 'Electronics', 'Toys and Hobbies',
            'Tools & home Improvement', 'Jewelry or Accessories', 'Household',
            'Pet Equipment', 'Shoes', 'Bags and Luggage', 'Bathroom',
            'Furniture', 'Clothing', 'Possibly Incorrect', 'Grill',
            'Travel & Outdoors', 'Bags & Luggage', 'Work truck', '*Zojirushi',
            'Travel &amp; Outdoors', 'Tools &amp; home Improvement',
            'Toys & Hobbies', 'Mod approved', '/R/ALL',
            'Jewelry & Accessories', 'Show & Tell', 'r/all Show & Tell',
            'SHITPOST', 'Outdoors', 'Apparel', 'other', 'tools'], dtype=object)

[ ]  #Drop entries with the irrelevant flairs
     redditFlairAdjustedDF = redditDateAdjustedDF.copy()
     redditFlairAdjustedDF.drop(redditFlairAdjustedDF.index[redditFlairAdjustedDF['link_flair_text'] == 'SHITPOST'], inplace=True);
     redditFlairAdjustedDF.drop(redditFlairAdjustedDF.index[redditFlairAdjustedDF['link_flair_text'] == 'Meta'], inplace=True);
     redditFlairAdjustedDF.drop(redditFlairAdjustedDF.index[redditFlairAdjustedDF['link_flair_text'] == 'Offical Discussion'], inplace=True);

[ ]  #Check that the flairs are not there anymore
     redditFlairAdjustedDF.link_flair_text.unique()

     array(['Discussion', '[Request]', 'Repair', 'Currently sold', 'Vintage',
            '[Request] Answered!', 'Review', 'BIFL Skills', 'Warranty', '',
            "Men's Clothing", 'Kitchen', 'Appliances', 'Office Supplies',
            'Travel and Outdoors', 'Other', 'Automotive', 'Electronics',
            'Toys and Hobbies', 'Tools & home Improvement',
            'Jewelry or Accessories', 'Household', 'Pet Equipment', 'Shoes',
            'Bags and Luggage', 'Bathroom', 'Furniture', 'Clothing',
            'Possibly Incorrect', 'Grill', 'Travel & Outdoors',
            'Bags & Luggage', 'Work truck', '*Zojirushi',
            'Travel &amp; Outdoors', 'Tools &amp; home Improvement',
            'Toys & Hobbies', 'Mod approved', '/R/ALL',
            'Jewelry & Accessories', 'Show & Tell', 'r/all Show & Tell',
            'Outdoors', 'Apparel', 'other', 'tools'], dtype=object)
```

- Removing irrelevant values based on flair
  - meme posts
  - meta posts
  - official discussions

our approach

Language pre-processing



```
[ ]  #Create keyword column with the combined title + content
     redditDF['keywords'] = redditDF['title'] + redditDF['selftext']
     redditDF['keywords'] = redditDF['keywords'].astype(str)
```

```
[ ]  #Sets all the content in each entry to be lowercase
     redditDF['keywords'] = [entry.lower() for entry in redditDF['keywords']]
```

```
[ ]  #Lemmatize the entries
     redditDF['keywords'] = [nltk.WordNetLemmatizer().lemmatize(entry) for entry in redditDF.keywords]
```

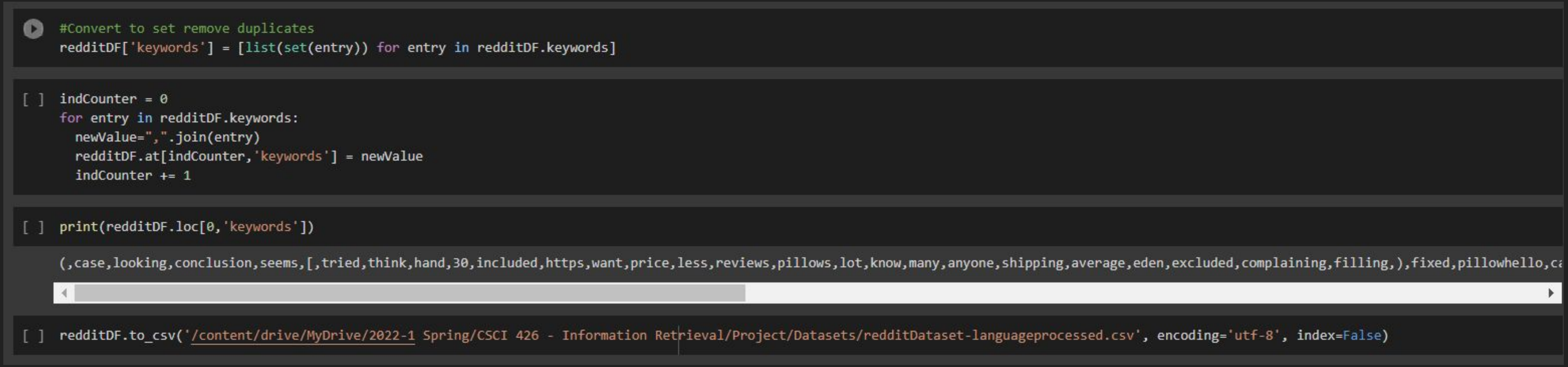
```
[ ]  #Tokenize the keywords (separating each word into its own entry)
     redditDF['keywords'] = [word_tokenize(entry) for entry in redditDF.keywords]
     #redditDF['keywords'] = redditDF['keywords'].astype(str)
```

```
[ ]  #Remove stop words
     stop_words = set(stopwords.words('english'))
     indCounter = 0
     for entry in redditDF.keywords:
       filteredSentence = [w for w in entry if not w.lower() in stop_words]
       for w in entry:
         if w not in stop_words:
           w=w.replace("\{.",'')
           w=w.replace("'",'')
           w=w.replace(" ",'')
           w=w.replace("\].",'')
           filteredSentence.append(w)

       redditDF.at[indCounter,'keywords'] = filteredSentence
       indCounter += 1
```

- Create a **new keyword column**
  - use **title + body**
- **Lowercase all values** in the new column
- **Stem** entries
  - back to **word root**
- **Tokenize**
  - words in **separate entries**
- Remove **stopwords**

our approach

Language pre-processing



```
#Convert to set remove duplicates
redditDF['keywords'] = [list(set(entry)) for entry in redditDF.keywords]
```

```
indCounter = 0
for entry in redditDF.keywords:
  newValue=",".join(entry)
  redditDF.at[indCounter,'keywords'] = newValue
  indCounter += 1
```

```
print(redditDF.loc[0,'keywords'])
```

```
(,case,looking,conclusion,seems,[,tried,think,hand,30,included,https,want,price,less,reviews,pillows,lot,know,many,anyone,shipping,average,eden,excluded,complaining,filling,),fixed,pillowhello,ca
```

```
redditDF.to_csv('/content/drive/MyDrive/2022-1 Spring/CSCI 426 - Information Retrieval/Project/Datasets/redditDataset-languageprocessed.csv', encoding='utf-8', index=False)
```

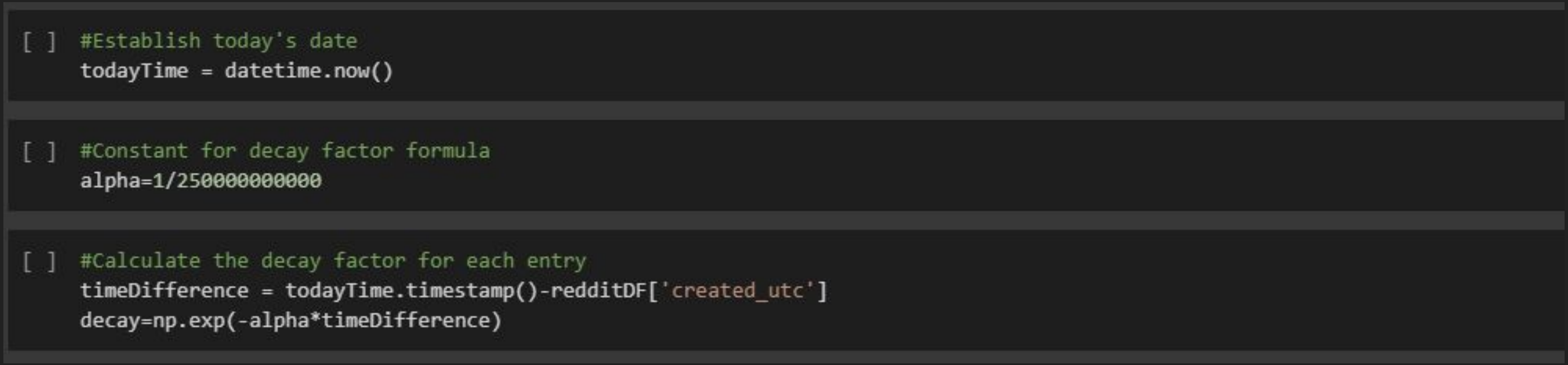
- Remove duplicate entries
    - use natural property of a set
- Create a function to combine words into something easier to parse
    - prevents later headache

our approach

Decay factor

$$e^{-a\,(current\ time\ -\ rating\ time)}$$

e = rating
a = constant = 1/25000000000



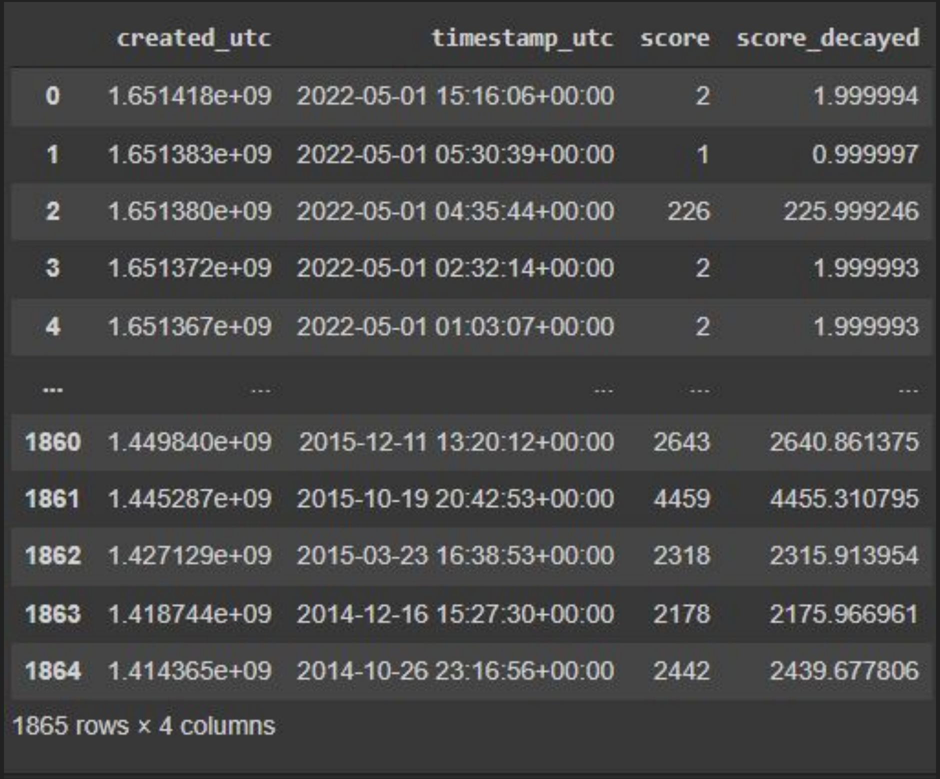
```
[ ]  #Establish today's date
     todayTime = datetime.now()
```

```
[ ]  #Constant for decay factor formula
     alpha=1/250000000000
```

```
[ ]  #Calculate the decay factor for each entry
     timeDifference = todayTime.timestamp()-redditDF['created_utc']
     decay=np.exp(-alpha*timeDifference)
```

our approach

Decay factor



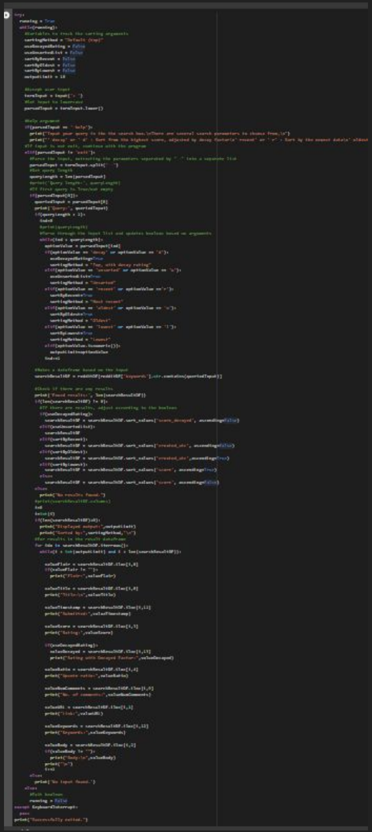
	created_utc	timestamp_utc	score	score_decayed
0	1.651418e+09	2022-05-01 15:16:06+00:00	2	1.999994
1	1.651383e+09	2022-05-01 05:30:39+00:00	1	0.999997
2	1.651380e+09	2022-05-01 04:35:44+00:00	226	225.999246
3	1.651372e+09	2022-05-01 02:32:14+00:00	2	1.999993
4	1.651367e+09	2022-05-01 01:03:07+00:00	2	1.999993
...	...	...	...	...
1860	1.449840e+09	2015-12-11 13:20:12+00:00	2643	2640.861375
1861	1.445287e+09	2015-10-19 20:42:53+00:00	4459	4455.310795
1862	1.427129e+09	2015-03-23 16:38:53+00:00	2318	2315.913954
1863	1.418744e+09	2014-12-16 15:27:30+00:00	2178	2175.966961
1864	1.414365e+09	2014-10-26 23:16:56+00:00	2442	2439.677806

1865 rows × 4 columns

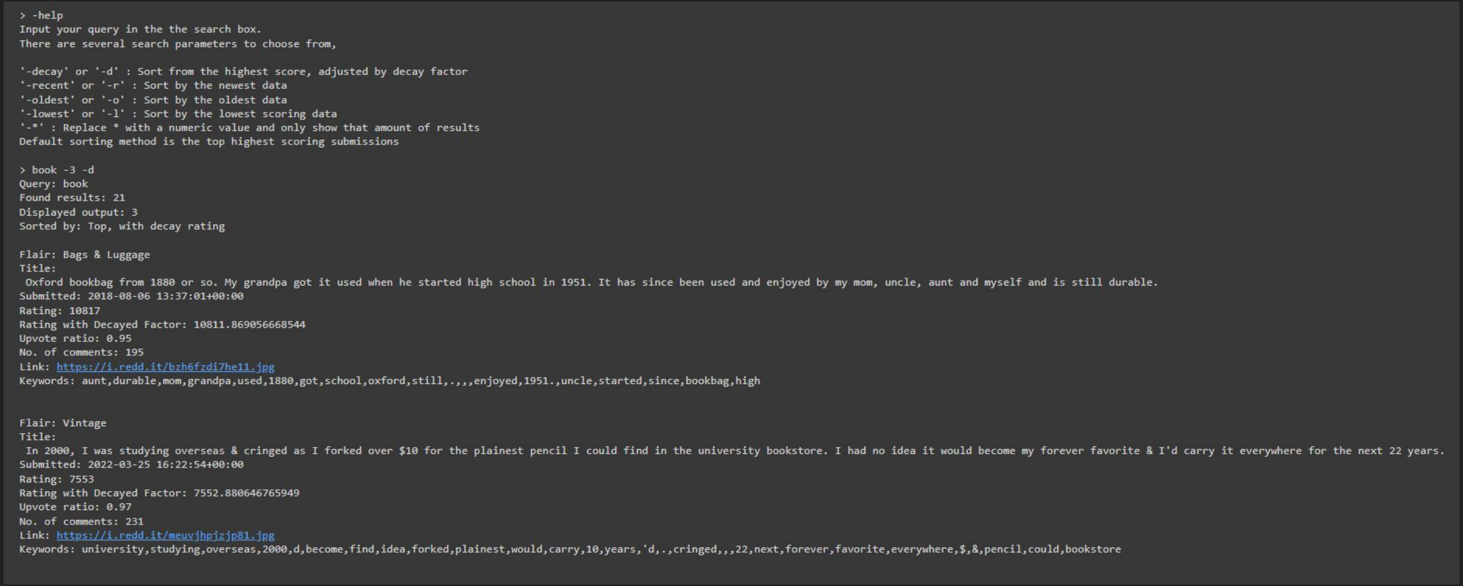
- Example result
  - the older the post, the more the rating gets adjusted

our approach

The search engine itself



```
try:
  running = True
  while(running):
    #Variables to track the sorting arguments
    sortingMethod = "Default (top)"
    useDecayedRating = False
    useUnsortedList = False
    sortByRecent = False
    sortByOldest = False
    sortByLowest = False
    outputLimit = 10

    #Accept user input
    termInput = input("> ")
    #Set input to lowercase
    parsedInput = termInput.lower()

    #Help argument
    if(parsedInput == '-help'):
      print("Input your query in the the search box.\nThere are several search parameters to choose from,\n")
      print("'-decay' or '-d' : Sort from the highest score, adjusted by decay factor\n'-recent' or '-r' : Sort by the newest data\n'-oldest
    #If input is not exit, continue with the program
    elif(parsedInput != 'exit'):
      #Parse the input, extracting the parameters separated by " -" into a separate list
      parsedInput = termInput.split(' -')
      #Get query length
      queryLength = len(parsedInput)
      #print('Query length:', queryLength)
      #If first query is True/not empty
      if(parsedInput[0]):
        queriedInput = parsedInput[0]
        print('Query:', queriedInput)
        if(queryLength > 1):
          ind=0
          #print(queryLength)
          #Parse through the input list and updates boolean based on arguments
          while(ind < queryLength):
            optionValue = parsedInput[ind]
            if(optionValue == 'decay' or optionValue == 'd'):
              useDecayedRating=True
              sortingMethod = "Top, with decay rating"
            elif(optionValue == 'unsorted' or optionValue == 'u'):
              useUnsortedList=True
              sortingMethod = "Unsorted"
            elif(optionValue == 'recent' or optionValue =='r'):
              sortByRecent=True
              sortingMethod = "Most recent"
            elif(optionValue == 'oldest' or optionValue == 'o'):
              sortByOldest=True
              sortingMethod = "Oldest"
            elif(optionValue == 'lowest' or optionValue == 'l'):
              sortByLowest=True
              sortingMethod = "Lowest"
            elif(optionValue.isnumeric()):
              outputLimit=optionValue
            ind+=1

        #Makes a dataframe based on the input
        searchResultDF = redditDF[redditDF['keywords'].str.contains(queriedInput)]

        #Check if there are any results
        print('Found results:', len(searchResultDF))
        if(len(searchResultDF) != 0):
          #If there are results, adjust according to the boolean
          if(useDecayedRating):
            searchResultDF = searchResultDF.sort_values(['score_decayed'], ascending=False)
          elif(useUnsortedList):
            searchResultDF
          elif(sortByRecent):
            searchResultDF = searchResultDF.sort_values(['created_utc'], ascending=False)
          elif(sortByOldest):
            searchResultDF = searchResultDF.sort_values(['created_utc'],ascending=True)
          elif(sortByLowest):
            searchResultDF = searchResultDF.sort_values(['score'], ascending=True)
          else:
            searchResultDF = searchResultDF.sort_values(['score'], ascending=False)
        else:
          print("No results found.")
        #print(searchResultDF.columns)
        i=0
        i=int(i)
        if(len(searchResultDF)>0):
          print("Displayed output:",outputLimit)
          print("Sorted by:",sortingMethod,"\n")
        #For results in the result dataframe
        for ids in searchResultDF.iterrows():
          while(i < int(outputLimit) and i < len(searchResultDF)):

            valueFlair = searchResultDF.iloc[i,8]
            if(valueFlair != ""):
              print("Flair:",valueFlair)

            valueTitle = searchResultDF.iloc[i,0]
            print("Title:\n",valueTitle)

            valueTimestamp = searchResultDF.iloc[i,12]
            print("Submitted:",valueTimestamp)

            valueScore = searchResultDF.iloc[i,1]
            print("Rating:",valueScore)

            if(useDecayedRating):
              valueDecayed = searchResultDF.iloc[i,13]
              print("Rating with Decayed Factor:",valueDecayed)

            valueRatio = searchResultDF.iloc[i,4]
            print("Upvote ratio:",valueRatio)

            valueNumComments = searchResultDF.iloc[i,6]
            print("No. of comments:",valueNumComments)

            valueURL = searchResultDF.iloc[i,3]
            print("Link:",valueURL)

            valueKeywords = searchResultDF.iloc[i,11]
            print("Keywords:",valueKeywords)

            valueBody = searchResultDF.iloc[i,2]
            if(valueBody != ""):
              print("Body:\n",valueBody)
            print("\n")
            i+=1
      else:
        print('No input found.')
    else:
      #Exit boolean
      running = False
except KeyboardInterrupt:
  pass
print("Successfully exited.")
```

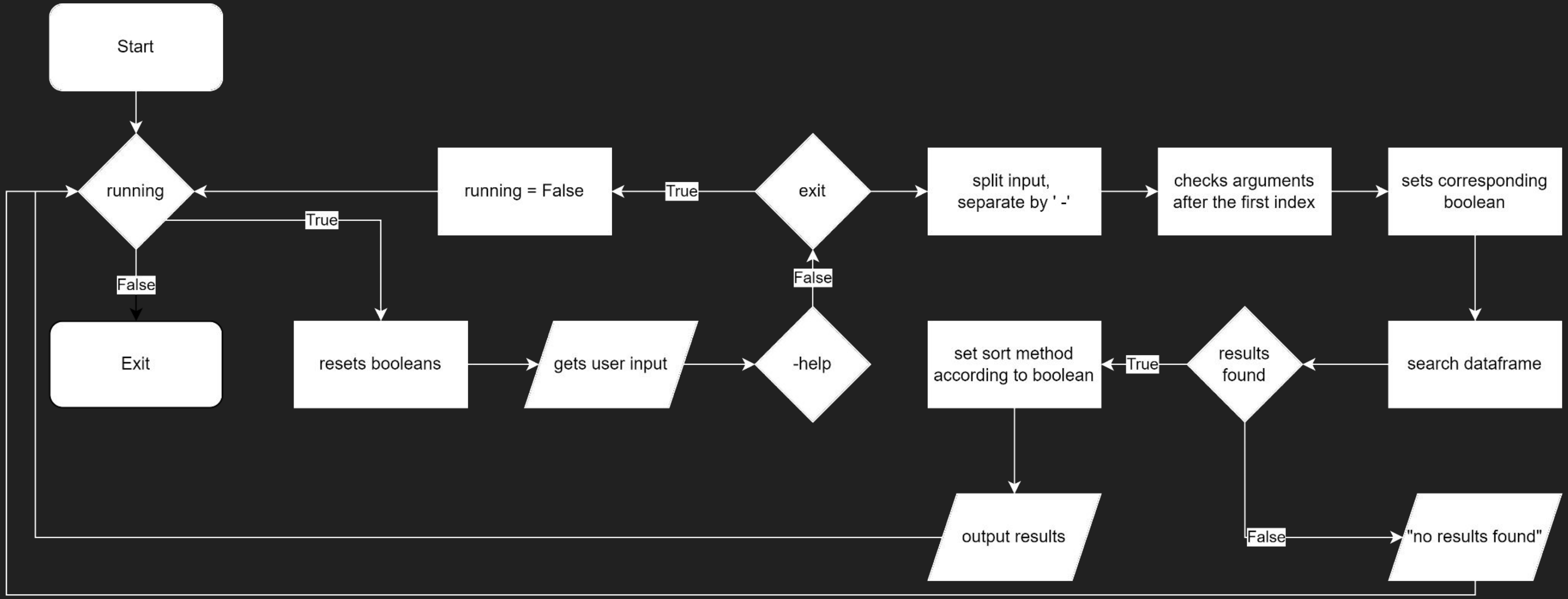


```
> -help
Input your query in the the search box.
There are several search parameters to choose from,

'-decay' or '-d' : Sort from the highest score, adjusted by decay factor
'-recent' or '-r' : Sort by the newest data
'-oldest' or '-o' : Sort by the oldest data
'-lowest' or '-l' : Sort by the lowest scoring data
'-*' : Replace * with a numeric value and only show that amount of results
Default sorting method is the top highest scoring submissions

> book -3 -d
Query: book
Found results: 21
Displayed output: 3
Sorted by: Top, with decay rating

Flair: Bags & Luggage
Title:
 Oxford bookbag from 1880 or so. My grandpa got it used when he started high school in 1951. It has since been used and enjoyed by my mom, uncle, aunt and myself and is still durable.
Submitted: 2018-08-06 13:37:01+00:00
Rating: 10817
Rating with Decayed Factor: 10811.869056668544
Upvote ratio: 0.95
No. of comments: 195
Link: https://i.redd.it/bzh6fzdi7he11.jpg
Keywords: aunt,durable,mom,grandpa,used,1880,got,school,oxford,still,.,,,enjoyed,1951.,uncle,started,since,bookbag,high


Flair: Vintage
Title:
 In 2000, I was studying overseas & cringed as I forked over $10 for the plainest pencil I could find in the university bookstore. I had no idea it would become my forever favorite & I'd carry it everywhere for the next 22 years.
Submitted: 2022-03-25 16:22:54+00:00
Rating: 7553
Rating with Decayed Factor: 7552.880646765949
Upvote ratio: 0.97
No. of comments: 231
Link: https://i.redd.it/meuvjhpjzjp81.jpg
Keywords: university,studying,overseas,2000,d,become,find,idea,forked,plainest,would,carry,10,years,'d,.,cringed,,,22,next,forever,favorite,everywhere,$,&,pencil,could,bookstore
```

- keyword search
  - most of the work goes into parsing

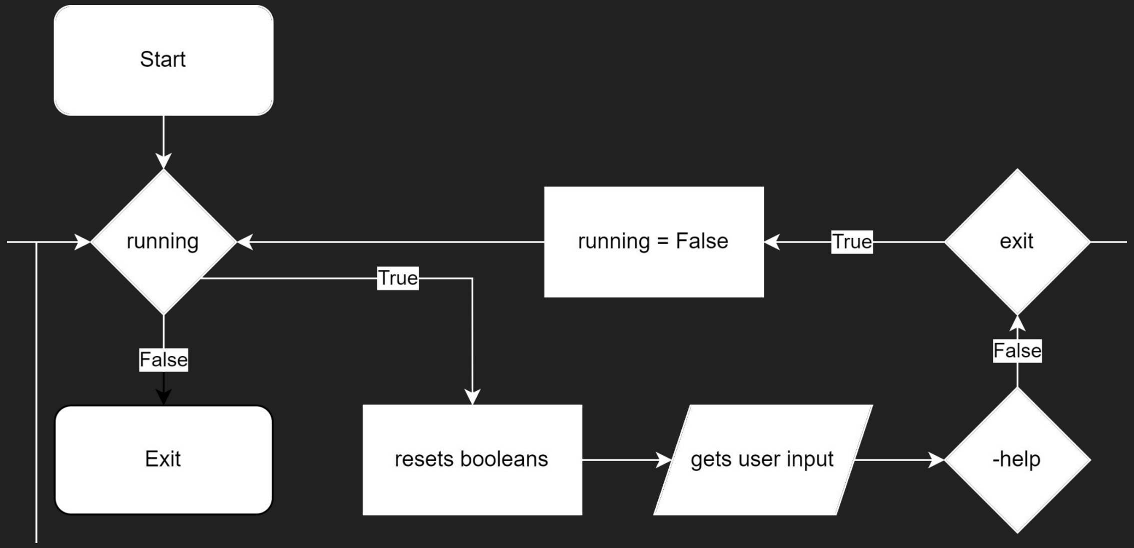
our approach

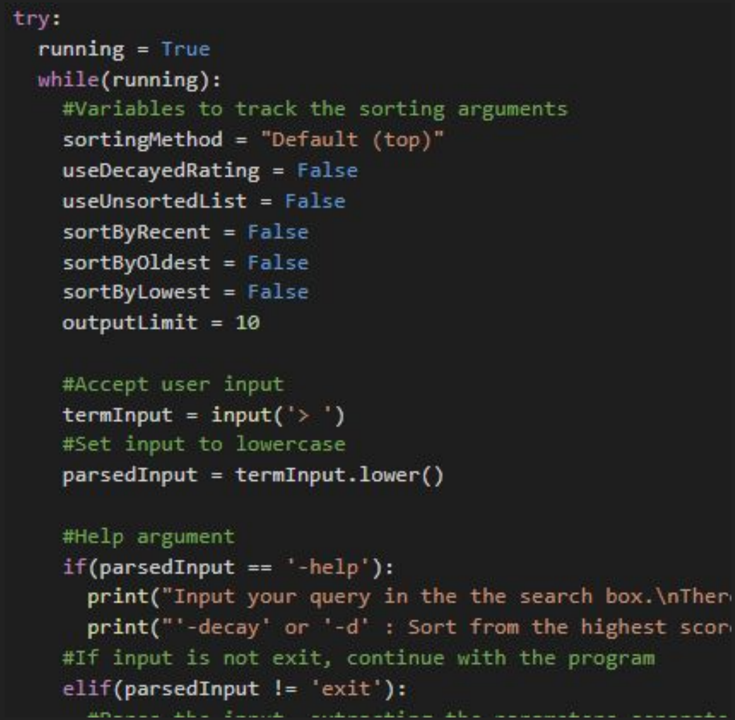
The search engine itself



our approach

The search engine itself

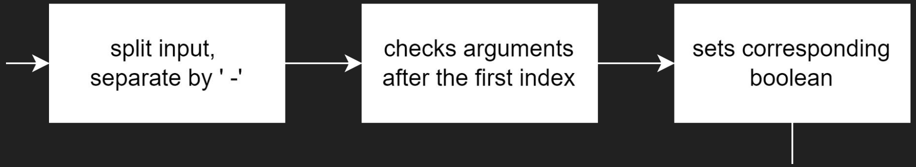




```
try:
  running = True
  while(running):
    #Variables to track the sorting arguments
    sortingMethod = "Default (top)"
    useDecayedRating = False
    useUnsortedList = False
    sortByRecent = False
    sortByOldest = False
    sortByLowest = False
    outputLimit = 10

    #Accept user input
    termInput = input('> ')
    #Set input to lowercase
    parsedInput = termInput.lower()

    #Help argument
    if(parsedInput == '-help'):
      print("Input your query in the the search box.\nTher
      print("'-decay' or '-d' : Sort from the highest scor
    #If input is not exit, continue with the program
    elif(parsedInput != 'exit'):
```

our approach

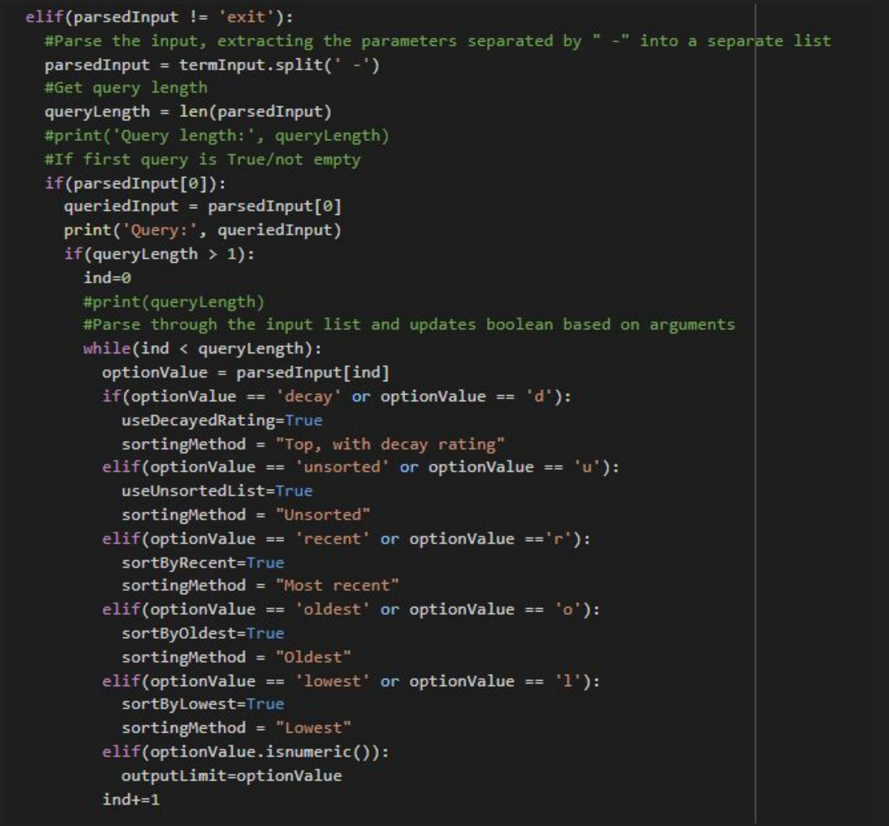
The search engine itself



split input, separate by ' -'

checks arguments after the first index

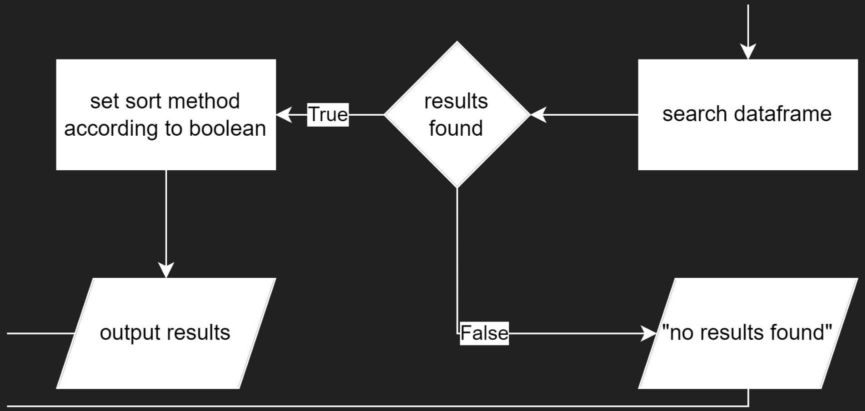
sets corresponding boolean

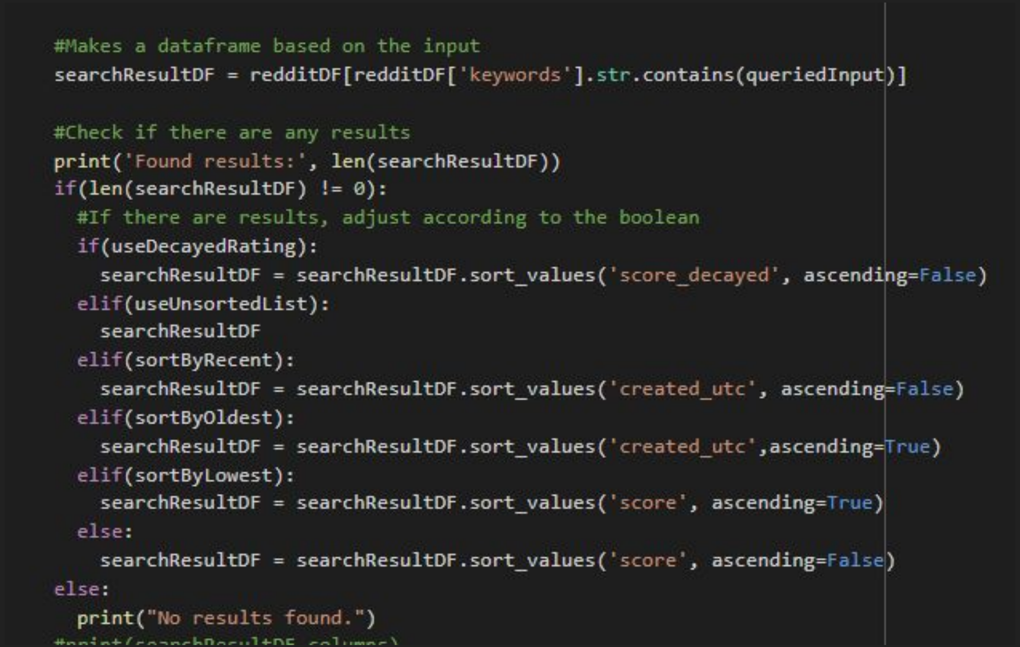


```
elif(parsedInput != 'exit'):
  #Parse the input, extracting the parameters separated by " -" into a separate list
  parsedInput = termInput.split(' -')
  #Get query length
  queryLength = len(parsedInput)
  #print('Query length:', queryLength)
  #If first query is True/not empty
  if(parsedInput[0]):
    queriedInput = parsedInput[0]
    print('Query:', queriedInput)
    if(queryLength > 1):
      ind=0
      #print(queryLength)
      #Parse through the input list and updates boolean based on arguments
      while(ind < queryLength):
        optionValue = parsedInput[ind]
        if(optionValue == 'decay' or optionValue == 'd'):
          useDecayedRating=True
          sortingMethod = "Top, with decay rating"
        elif(optionValue == 'unsorted' or optionValue == 'u'):
          useUnsortedList=True
          sortingMethod = "Unsorted"
        elif(optionValue == 'recent' or optionValue =='r'):
          sortByRecent=True
          sortingMethod = "Most recent"
        elif(optionValue == 'oldest' or optionValue == 'o'):
          sortByOldest=True
          sortingMethod = "Oldest"
        elif(optionValue == 'lowest' or optionValue == 'l'):
          sortByLowest=True
          sortingMethod = "Lowest"
        elif(optionValue.isnumeric()):
          outputLimit=optionValue
        ind+=1
```

our approach

The search engine itself

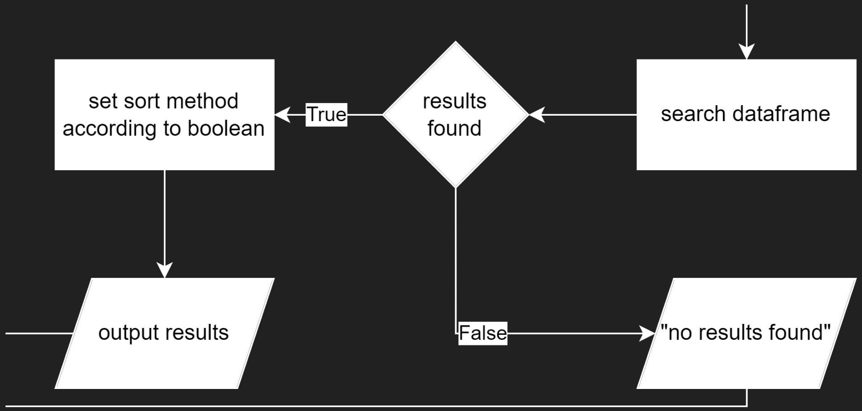


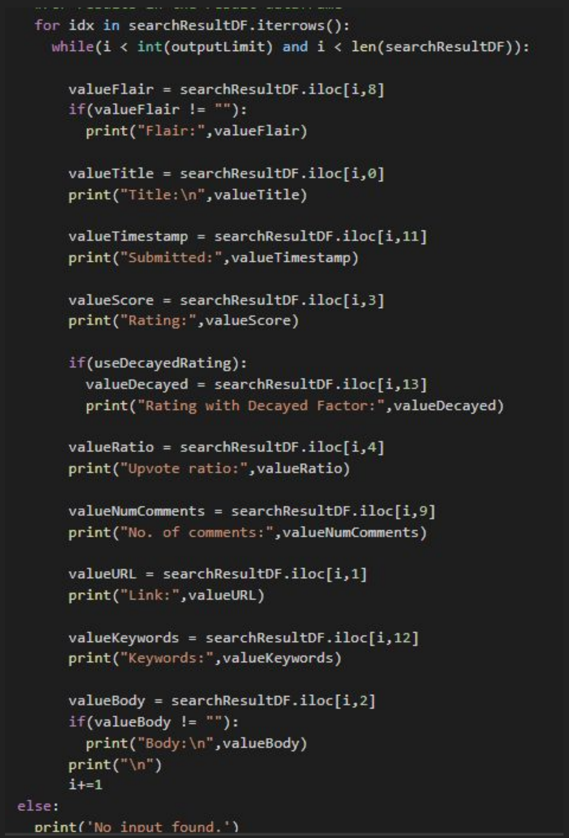


```
#Makes a dataframe based on the input
searchResultDF = redditDF[redditDF['keywords'].str.contains(queriedInput)]

#Check if there are any results
print('Found results:', len(searchResultDF))
if(len(searchResultDF) != 0):
  #If there are results, adjust according to the boolean
  if(useDecayedRating):
    searchResultDF = searchResultDF.sort_values('score_decayed', ascending=False)
  elif(useUnsortedList):
    searchResultDF
  elif(sortByRecent):
    searchResultDF = searchResultDF.sort_values('created_utc', ascending=False)
  elif(sortByOldest):
    searchResultDF = searchResultDF.sort_values('created_utc',ascending=True)
  elif(sortByLowest):
    searchResultDF = searchResultDF.sort_values('score', ascending=True)
  else:
    searchResultDF = searchResultDF.sort_values('score', ascending=False)
else:
  print("No results found.")
#print(searchResultDF.columns)
```

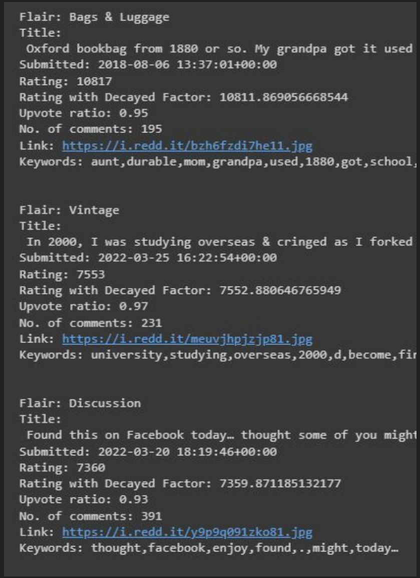
our approach

The search engine itself





```
  for idx in searchResultDF.iterrows():
    while(i < int(outputLimit) and i < len(searchResultDF)):

      valueFlair = searchResultDF.iloc[i,8]
      if(valueFlair != ""):
        print("Flair:",valueFlair)

      valueTitle = searchResultDF.iloc[i,0]
      print("Title:\n",valueTitle)

      valueTimestamp = searchResultDF.iloc[i,11]
      print("Submitted:",valueTimestamp)

      valueScore = searchResultDF.iloc[i,3]
      print("Rating:",valueScore)

      if(useDecayedRating):
        valueDecayed = searchResultDF.iloc[i,13]
        print("Rating with Decayed Factor:",valueDecayed)

      valueRatio = searchResultDF.iloc[i,4]
      print("Upvote ratio:",valueRatio)

      valueNumComments = searchResultDF.iloc[i,9]
      print("No. of comments:",valueNumComments)

      valueURL = searchResultDF.iloc[i,1]
      print("Link:",valueURL)

      valueKeywords = searchResultDF.iloc[i,12]
      print("Keywords:",valueKeywords)

      valueBody = searchResultDF.iloc[i,2]
      if(valueBody != ""):
        print("Body:\n",valueBody)
      print("\n")
      i+=1
else:
  print('No input found.')
```



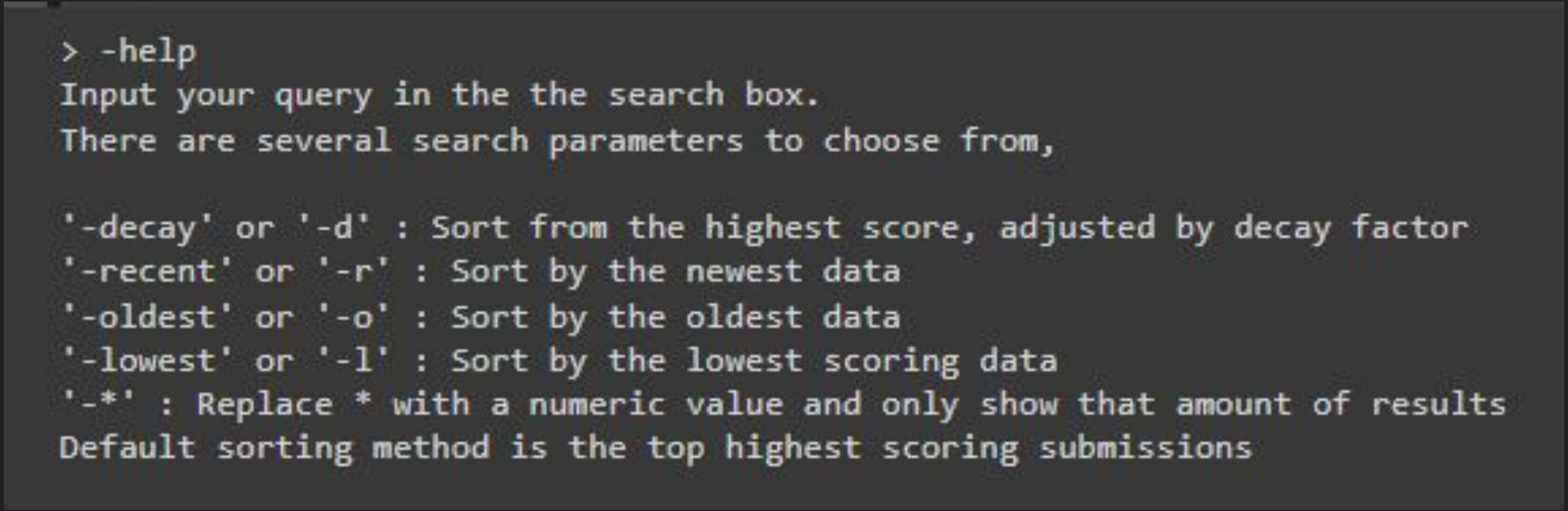
```
Flair: Bags & Luggage
Title:
 Oxford bookbag from 1880 or so. My grandpa got it used
Submitted: 2018-08-06 13:37:01+00:00
Rating: 10817
Rating with Decayed Factor: 10811.869056668544
Upvote ratio: 0.95
No. of comments: 195
Link: https://i.redd.it/bzh6fzdi7he11.jpg
Keywords: aunt,durable,mom,grandpa,used,1880,got,school,


Flair: Vintage
Title:
 In 2000, I was studying overseas & cringed as I forked
Submitted: 2022-03-25 16:22:54+00:00
Rating: 7553
Rating with Decayed Factor: 7552.880646765949
Upvote ratio: 0.97
No. of comments: 231
Link: https://i.redd.it/meuvjhpjzjp81.jpg
Keywords: university,studying,overseas,2000,d,become,fir


Flair: Discussion
Title:
 Found this on Facebook today… thought some of you might
Submitted: 2022-03-20 18:19:46+00:00
Rating: 7360
Rating with Decayed Factor: 7359.871185132177
Upvote ratio: 0.93
No. of comments: 391
Link: https://i.redd.it/y9p9q091zko81.jpg
Keywords: thought,facebook,enjoy,found,.,might,today…
```

Chapter 4

experiments and results

experiment and results

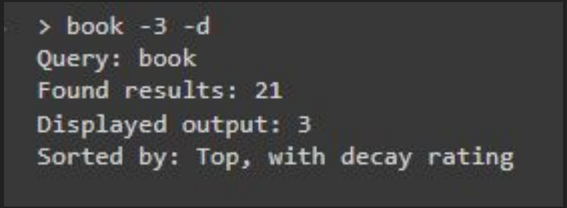
Sample output



```
> -help
Input your query in the the search box.
There are several search parameters to choose from,

'-decay' or '-d' : Sort from the highest score, adjusted by decay factor
'-recent' or '-r' : Sort by the newest data
'-oldest' or '-o' : Sort by the oldest data
'-lowest' or '-l' : Sort by the lowest scoring data
'-*' : Replace * with a numeric value and only show that amount of results
Default sorting method is the top highest scoring submissions
```

experiment and results

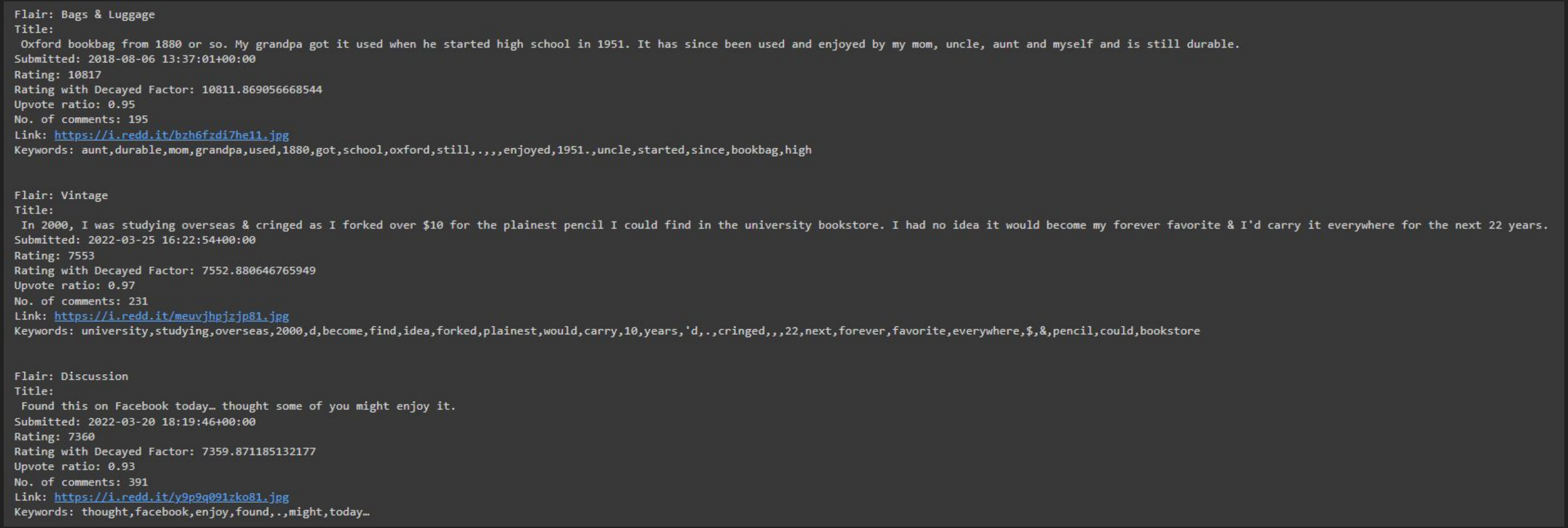
Sample output



```
> book -3 -d
Query: book
Found results: 21
Displayed output: 3
Sorted by: Top, with decay rating
```

‘book -3 -d’

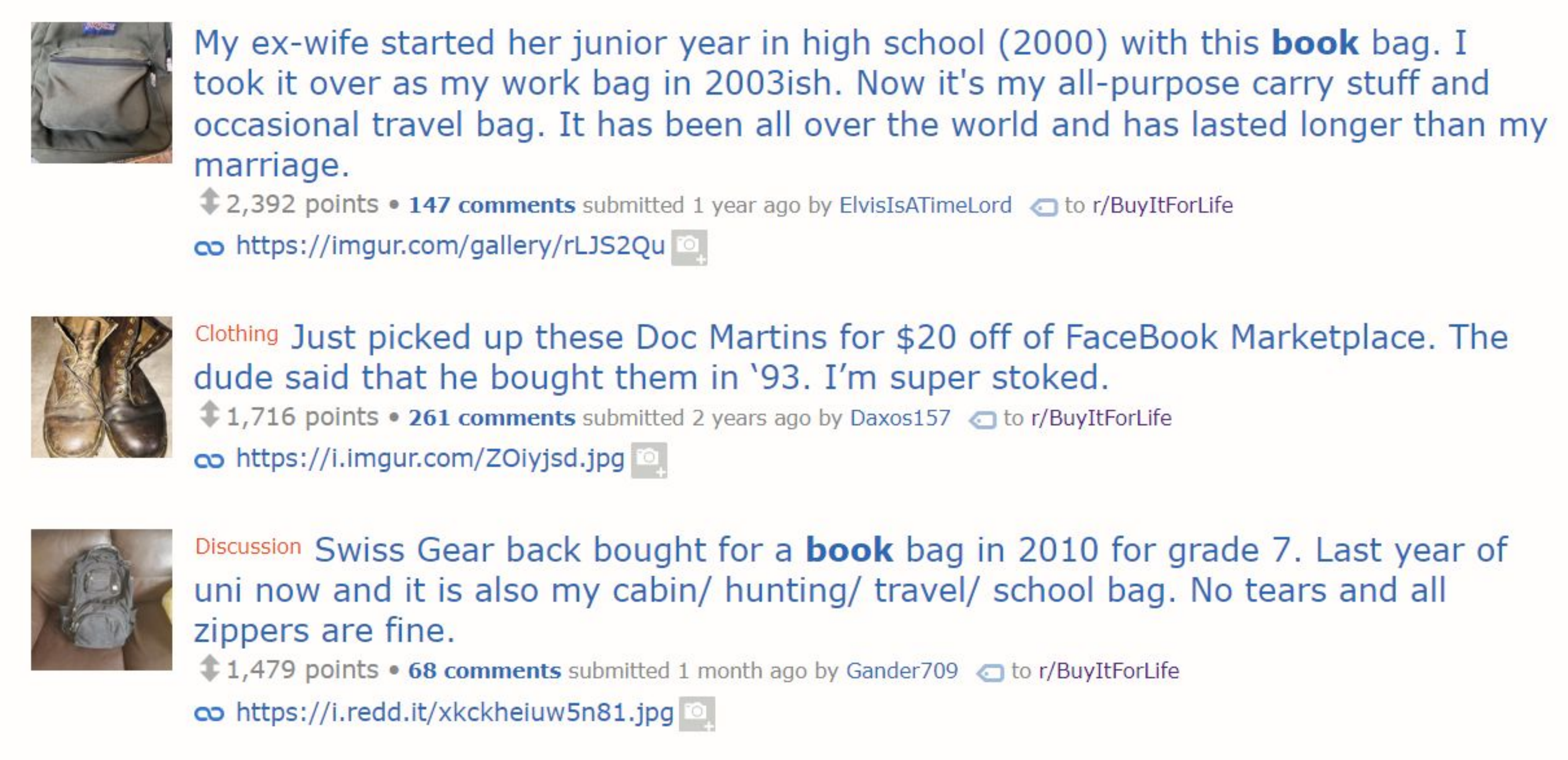
book : searches “book”
-3 : outputs 3 results
-d : sort by top with decayed factor



```
Flair: Bags & Luggage
Title:
 Oxford bookbag from 1880 or so. My grandpa got it used when he started high school in 1951. It has since been used and enjoyed by my mom, uncle, aunt and myself and is still durable.
Submitted: 2018-08-06 13:37:01+00:00
Rating: 10817
Rating with Decayed Factor: 10811.869056668544
Upvote ratio: 0.95
No. of comments: 195
Link: https://i.redd.it/bzh6fzdi7he11.jpg
Keywords: aunt,durable,mom,grandpa,used,1880,got,school,oxford,still,.,,,enjoyed,1951.,uncle,started,since,bookbag,high


Flair: Vintage
Title:
 In 2000, I was studying overseas & cringed as I forked over $10 for the plainest pencil I could find in the university bookstore. I had no idea it would become my forever favorite & I'd carry it everywhere for the next 22 years.
Submitted: 2022-03-25 16:22:54+00:00
Rating: 7553
Rating with Decayed Factor: 7552.880646765949
Upvote ratio: 0.97
No. of comments: 231
Link: https://i.redd.it/meuvjhpjzjp81.jpg
Keywords: university,studying,overseas,2000,d,become,find,idea,forked,plainest,would,carry,10,years,'d,.,cringed,,,22,next,forever,favorite,everywhere,$,&,pencil,could,bookstore


Flair: Discussion
Title:
 Found this on Facebook today… thought some of you might enjoy it.
Submitted: 2022-03-20 18:19:46+00:00
Rating: 7360
Rating with Decayed Factor: 7359.871185132177
Upvote ratio: 0.93
No. of comments: 391
Link: https://i.redd.it/y9p9q091zko81.jpg
Keywords: thought,facebook,enjoy,found,.,might,today…
```

experiment and results

Comparisons



My ex-wife started her junior year in high school (2000) with this **book** bag. I took it over as my work bag in 2003ish. Now it's my all-purpose carry stuff and occasional travel bag. It has been all over the world and has lasted longer than my marriage.

2,392 points • **147 comments** submitted 1 year ago by ElvisIsATimeLord to r/BuyItForLife

https://imgur.com/gallery/rLJS2Qu

Clothing Just picked up these Doc Martins for $20 off of FaceBook Marketplace. The dude said that he bought them in '93. I'm super stoked.

1,716 points • **261 comments** submitted 2 years ago by Daxos157 to r/BuyItForLife

https://i.imgur.com/ZOiyjsd.jpg

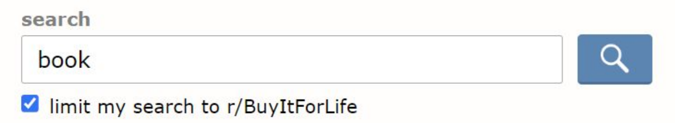
Discussion Swiss Gear back bought for a **book** bag in 2010 for grade 7. Last year of uni now and it is also my cabin/ hunting/ travel/ school bag. No tears and all zippers are fine.

1,479 points • **68 comments** submitted 1 month ago by Gander709 to r/BuyItForLife

https://i.redd.it/xkckheiuw5n81.jpg

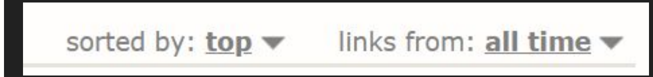
experiment and results

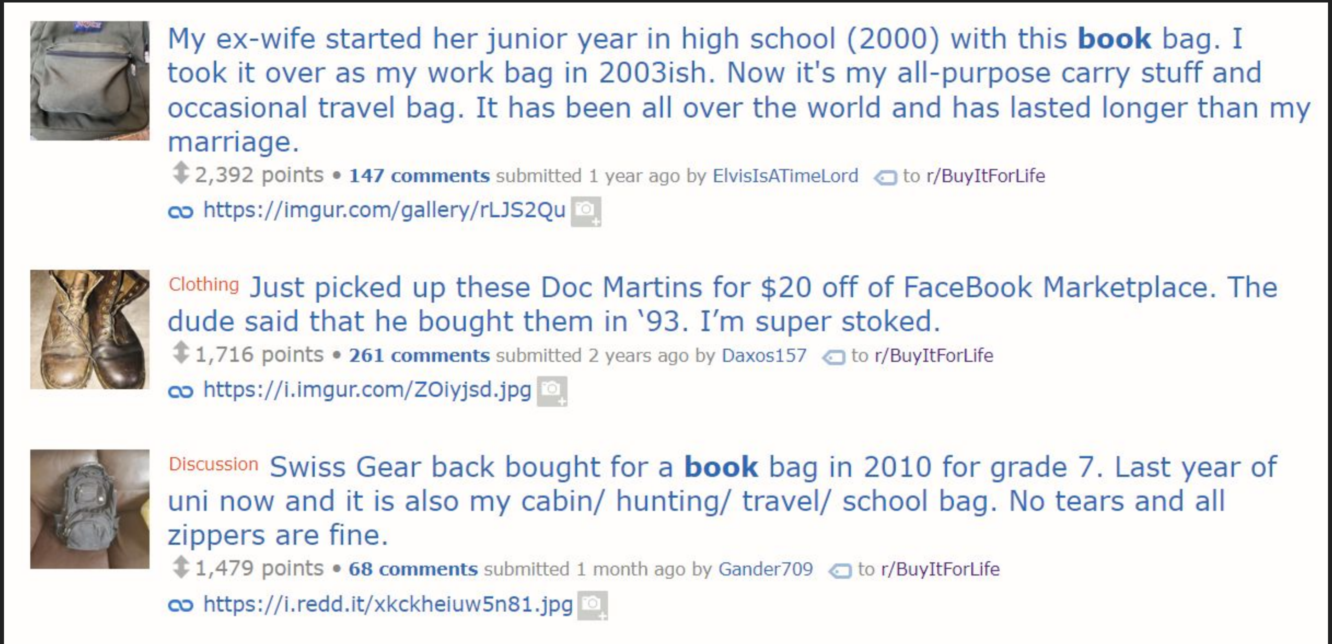
Comparisons


search

book

☑ limit my search to r/BuyItForLife


sorted by: top ▾ links from: all time ▾


My ex-wife started her junior year in high school (2000) with this **book** bag. I took it over as my work bag in 2003ish. Now it's my all-purpose carry stuff and occasional travel bag. It has been all over the world and has lasted longer than my marriage.

2,392 points • **147 comments** submitted 1 year ago by ElvisIsATimeLord to r/BuyItForLife

https://imgur.com/gallery/rLJS2Qu

Clothing Just picked up these Doc Martins for $20 off of FaceBook Marketplace. The dude said that he bought them in '93. I'm super stoked.

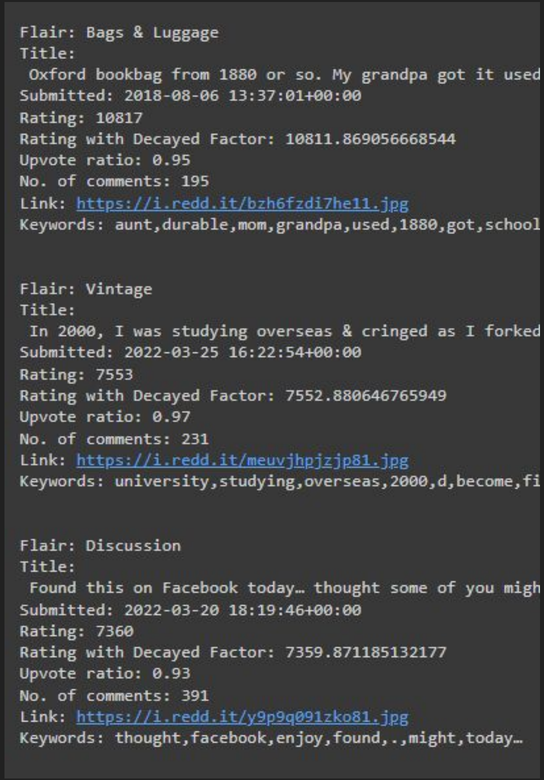
1,716 points • **261 comments** submitted 2 years ago by Daxos157 to r/BuyItForLife

https://i.imgur.com/ZOiyjsd.jpg

Discussion Swiss Gear back bought for a **book** bag in 2010 for grade 7. Last year of uni now and it is also my cabin/ hunting/ travel/ school bag. No tears and all zippers are fine.

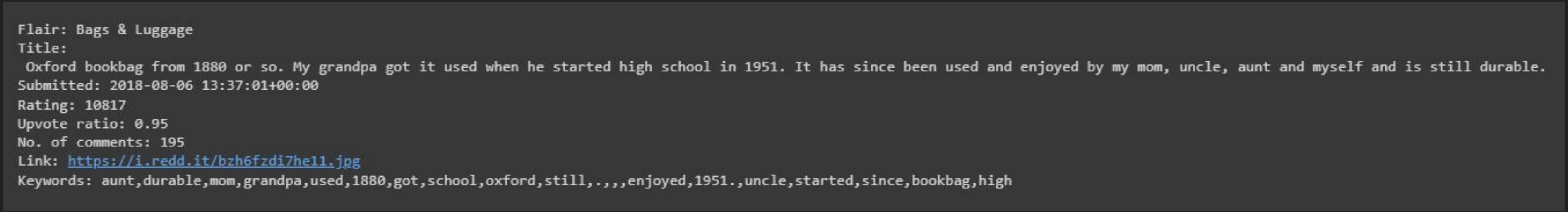
1,479 points • **68 comments** submitted 1 month ago by Gander709 to r/BuyItForLife

https://i.redd.it/xkckheiuw5n81.jpg

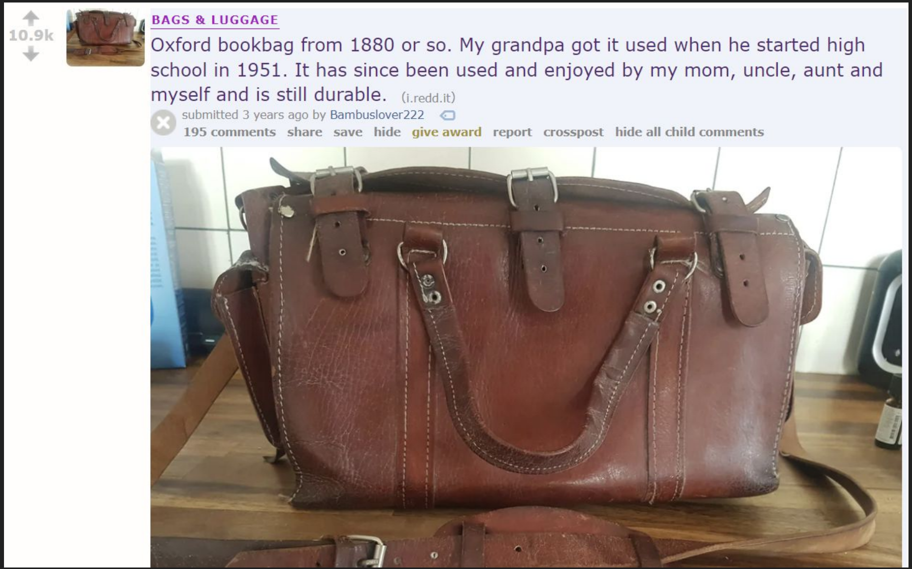

```
Flair: Bags & Luggage
Title:
 Oxford bookbag from 1880 or so. My grandpa got it used
Submitted: 2018-08-06 13:37:01+00:00
Rating: 10817
Rating with Decayed Factor: 10811.869056668544
Upvote ratio: 0.95
No. of comments: 195
Link: https://i.redd.it/bzh6fzdi7he11.jpg
Keywords: aunt,durable,mom,grandpa,used,1880,got,school


Flair: Vintage
Title:
 In 2000, I was studying overseas & cringed as I forked
Submitted: 2022-03-25 16:22:54+00:00
Rating: 7553
Rating with Decayed Factor: 7552.880646765949
Upvote ratio: 0.97
No. of comments: 231
Link: https://i.redd.it/meuvjhpjzjp81.jpg
Keywords: university,studying,overseas,2000,d,become,fi


Flair: Discussion
Title:
 Found this on Facebook today… thought some of you migh
Submitted: 2022-03-20 18:19:46+00:00
Rating: 7360
Rating with Decayed Factor: 7359.871185132177
Upvote ratio: 0.93
No. of comments: 391
Link: https://i.redd.it/y9p9q091zko81.jpg
Keywords: thought,facebook,enjoy,found,.,might,today…
```

experiment and results

Comparisons



```
Flair: Bags & Luggage
Title:
 Oxford bookbag from 1880 or so. My grandpa got it used when he started high school in 1951. It has since been used and enjoyed by my mom, uncle, aunt and myself and is still durable.
Submitted: 2018-08-06 13:37:01+00:00
Rating: 10817
Upvote ratio: 0.95
No. of comments: 195
Link: https://i.redd.it/bzh6fzdi7he11.jpg
Keywords: aunt,durable,mom,grandpa,used,1880,got,school,oxford,still,.,,,enjoyed,1951.,uncle,started,since,bookbag,high
```



BAGS & LUGGAGE

Oxford bookbag from 1880 or so. My grandpa got it used when he started high school in 1951. It has since been used and enjoyed by my mom, uncle, aunt and myself and is still durable. (i.redd.it)

submitted 3 years ago by Bambuslover222

195 comments share save hide give award report crosspost hide all child comments

https://www.reddit.com/r/BuyItForLife/comments/951ce6/oxford_bookbag_from_1880_or_so_my_grandpa_got_it/

Chapter 4

demo

Chapter 5

conclusion

Conclusion

Our thoughts

- Based on limited testing, our method to look at keywords gave possibly more accurate results based on supposed sorting
  - Top results from all time
- There are still many more improvements to be had
- Language processing is pretty powerful

Conclusion

What did not work?

- Using advanced search methods
    - Couldn’t figure it out in time
        - TF-IDF
        - Cosine similarity
- Front end implementation
    - Wanted to use Flask
        - Library to implement web apps

Conclusion

Future Work

- Expand search parameters
  - filter certain dates
  - search comments
  - etc
- Give additional search methods
  - similarity and TF-IDF metrics
  - use additional values for weighing
- More robust input parser
  - tokenize and lemmatize input, multiple queries
- Live search instead of using archived data

fin

thank you