Project 1 Report

Anthony Albanese
Binghamton University
Binghamton, New York, USA
aalbanese6@binghamton.edu

Kyle Enriquez
Binghamton University
Binghamton, New York, USA
kenriqu1@binghamton.edu

Ameen Kunnathu
Binghamton University
Binghamton, New York, USA
akunnat1@binghamton.edu

Andy Zheng
Binghamton University
Binghamton, New York, USA
azheng74@binghamton.edu

ACM Reference Format:
Anthony Albanese, Kyle Enriquez, Ameen Kunnathu, and Andy Zheng. 2018. Project 1 Report. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

For our project, we aimed to analyze data from different sources to determine general trends and correlations in fashion apparel. To do so, we closely analyzed Reddit and eBay as sources. For Reddit, we analyzed the subreddits r/findfashion and r/FashionReps to analyze discussion and trending topics in fashion. As for eBay, we focused on sales on apparel and clothing being auctioned recently to view their prices and categories. To gather the necessary information from these sources, we need to develop continuously functioning web crawlers which authorize access to these sites and then parse and formulate the data into a useful form. From there, the data needs to be inserted into a stable database to allow for its storage and viewing. During this process, various challenges and accommodations from our initial proposal will be taken to arrive at our ultimate goal of data analysis. In the end, through the data we gather, we can effectively visualize our findings and make preliminary remarks as we project the data that will be gathered over time.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

2 IMPLEMENTATION

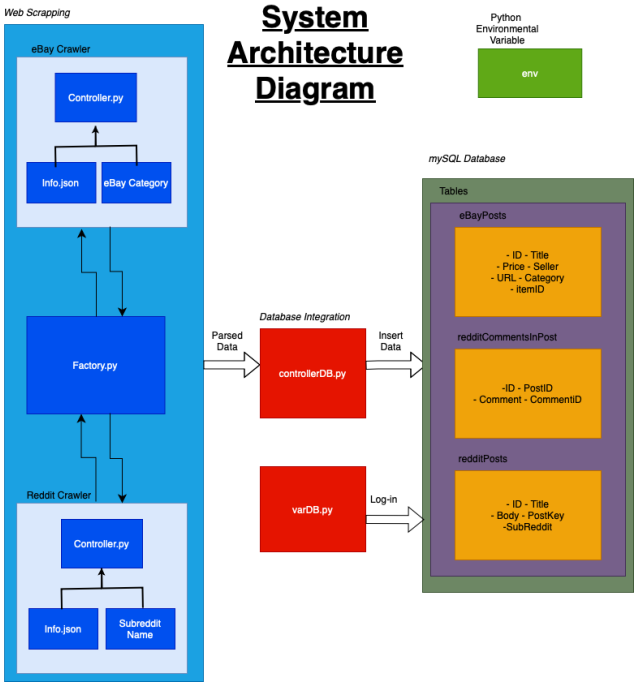

Figure 1: System Architecture Diagram

In order to effectively scrap data from different sites, several different resources were utilized. To begin, we predominantly utilized Python to build our web crawlers for Reddit and eBay. In the programs, we pass user credentials into the web scrapper, allowing us to access posts from respective sites and abstract and parse the appropriate data. To do so, we imported the request module to allow us to submit login credentials, which are passed in through a JSON file. In addition, we utilized distinct tools between the Reddit and eBay crawlers to access the respective sites. For Reddit, we implemented a method to gain authentication into the Reddit API, and from there we created another method to gather comments and posts from Reddit based on specified subreddit name and post limit parameters. As for eBay, we imported the ebaySDK library and set it to gather data from a specific category based on a keyword parameter. From there, we utilized the HTML client Beautiful

Soup as a means to parse the data we fetch from the eBay API. To supplement these web crawlers, we implemented a Faktory file to allow the data to be continuously and asynchronously collected in the background. After the data is collected, the file controllerDB.py is used to insert the data into their respective databases. To do so, we imported the mysql.Connector and myDB modules to establish a connection from the web crawlers to the databases and allow us to insert and categorize data in the tables based on specific characteristics, such as ID values and keys, descriptions, and prices. From there, the data gathered from the websites are then integrated into our databases, where we created tables to store eBay postings, Reddit comments, and Reddit posts. To implement these, we built the tables using AWS and MySQL software, allowing us to have a continuous means to store and display the data we gathered in an organized and categorized format. Through this, we can develop a comprehensive interface to view our findings, with samples of our data being able to be seen below:

id	postID	comment	commentID
887	302	Especially the first two pants lol	k5zwt2x
888	302	first two pants are fantasy pieces and they are ass cheeks	k6009hp
889	303	w2c: https://www.pandabuy.com/product?url=https://detail.tmall.com/item.htm?id=614260353450 https://…	k5yrs0q
890	303	Yk where I can get the fear of god essentials knit pull over hoodie ?	k62lm08
891	305	Cuseng ich küss dein Herz, guess Tasche für Mama <3	k61nc0c
892	305	https://weidian.com/item.html?itemID=2234722451&spider_token=4572 (Stussy Cap) https://detail.1…	k5z7ep6
893	311	https://walkers.style/shop/235023/sanctamuerte/sanctamuerte_jacket_st235023 https://www.amazon.com/…	k63x62k
894	313	Looks exactly like this one from My Mum Made It; https://www.chodenionek.com/outlet/womens-patent-sho…	k63rl8i
895	314	Looks similar to this one but different colors; https://images.app.goo.gl/ubWNm45eL8YpXNP47	k63s5fo
896	315	Could be this one; http://api-shein.shein.com/h5/sharejump/appjump?link=VzVIZ9onaP4&localcountry…	k63sf22
897	331	[The real Prada one](https://www.farfetch.com/it/shopping/women/prada-top-con-ricamo-item-18393597.as…	k632mmq
898	334	Here are some similar options - I can't find anything direct to what you were looking for however! [Ellesse…	k62w8on
899	335	Here's some options I found - Pretty Little Thing seem to be a popular brand that sells these types of things…	k62wm3i
900	335	[asos](https://www.asos.com/us/missguided/missguided-ribbed-roll-neck-midaxi-dress-in-black-black/prd/2…	k62y6fl
901	335	[many optins](https://www.amazon.com/gp/product/B01MXQF70O/)	k637dj0
902	336	https://poshmark.ca/listing/Jeffrey-Campbell-Platform-Heels-6530410b83cbec7f3e44daa2 looks like a 6.5…	k62vqs4
903	338	Here are some similar options I found: [Mocha - Satin Maxi Dress](https://www.pinterest.co.uk/pin/195836…	k62slb6
904	338	Budget up to $350-450 for the right dress!!	k62q95p
905	340	I notice there are some white pattern designs near the shoulders if that gives a clue.	k62mmpo
906	342	Try looking up "footed blazer with contrast piping"	k62vz7k
907	342	Lol *fitted. Good luck! 😅	k636kmp
908	343	I've seen a few similar styles - I'll attach below. I'd never heard of Asymmetrical Belt's before so this was fu…	k62tdo0
909	345	a lot of results just searching vintage jagermeister shirt on google images. some of the logo sizes are differ…	k62w01x
910	346	That jacket lokes like it's from Alice and the pirates, one brand part a deep rabbit hole of a fashion out of Ja…	k63cav5
911	347	[Tan Wool Coat](https://www.pinterest.co.uk/pin/754353625024925 73/) [Tan Wool Coat](https://www.pinter…	k62vnph
912	350	Try google "golf shirt no buttons"	k62xfph

Figure 2: redditComments

ID	Title	Body	PostKey	subReddit
296	[review] 10kg haul arrived today (balenciaga,vet…		17dyvc8	FashionReps
297	Any Flaws On Off-White OOO?		17dkoga	FashionReps
298	slight lil 14kg haul 😎 (rates?) low karma user	VT Dunks: https://www.pandabuy.com/share?ke…	17dkpyv	FashionReps
299	First rep haul QC! TS Reverse Mocha PK 4.0, D…		17dos10	FashionReps
300	Gym Haul Review for My Fellow Gymrats with P…	Waddup repfam, since I got a ton of DMs reque…	17dp455	FashionReps
301	HAUL (tops, pants, dress, socks, dunks)		17dpc9n	FashionReps
302	HI, I am ordering my first haul and i need some…		17drwzp	FashionReps
303	4.2 kg haul	Yes, i got essentials. No, i don't care you think it'…	17dtoi2	FashionReps
304	Qc two tshirts from 33trente (travis scott, hellstar)	travis- https://m.intl.taobao.com/detail/detail.htm…	17du4pa	FashionReps
305	8.2kg Haul to Germany (links in comments)	just a common haul.. Nike, Stussy, Soccer Jers…	17dw57q	FashionReps
306	Igor's sweater from Curb Your Enthusiasm	I really like his cardigan, can anyone find this sw…	17emtv4	findfashion
307	Blue Dress - Wolf of Wall Street - Margot Robbie	Please help me find a dress similar to Margot R…	17em79h	findfashion
308	Looking for Bob Iger's Masterclass shirt!	I've been searching for this on and off for longer…	17em5ym	findfashion
309	Looking for these earrings from Love is Blind		17elyoh	findfashion
310	This shirt	When I pressed the text below the image it does…	17elw0o	findfashion
311	I've had amazing luck in this sub so far. Please…		17el1j4	findfashion
312	Distressed leather loafers?	hello, i'm wondering where are these square-toe…	17ekqdc	findfashion
313	Can anyone help ID this purse?	I saw it in a YouTube video but I can't quite mak…	17ejr7g	findfashion
314	ID on this beanie?		17ejo9h	findfashion
315	Dying to find this swimsuit!	I know it's a little crazy but I wanted to surprise…	17ejmze	findfashion
316	Anyone know where can I find this ringer tee?		17ejk9i	findfashion
317	Help me find a similar corset?		17ejfg4	findfashion
318	Help me find this red necklace	Looking for the shorter necklace on a red string,…	17ej9sk	findfashion
319	Please help me find this T-shirt that Nick Jonas i…		17ej0bx	findfashion
320	Please help me find these sunglasses. I droppe…		17eiwil	findfashion
321	Does anyone what are these types of shoes, an…	I tried Timberland and Clarks but couldn't really…	17einqs	findfashion

Figure 3: redditPosts

ID	Title	Price	Seller	URL	Category	itemid
443	river island womens beige acrylic/polyester j	25	reskinned	https://www.ebay.com/itm/river-island-womens-	dresses	314912009719
444	nwot zara olive green cutout mini dress size	85	christytreasur…	https://www.ebay.com/itm/nwot-zara-olive-gree	dresses	126154135506
445	cos black dress $150 fit and flare cute kawai	20	libelula74	https://www.ebay.com/itm/cos-black-dress-150-	dresses	305227614881
446	zara black frilly bust sheer flowing sexy lon	5	kazaboper	https://www.ebay.com/itm/zara-black-frilly-bu	nightwear	175987869864
447	banana republic button down dress navy floral	15	wallyfiveo	https://www.ebay.com/itm/banana-republic-butt	dresses	394958126543
448	river island womens black tea jacquard dress	28	reskinned	https://www.ebay.com/itm/river-island-womens-	dresses	314912009532
449	$670 zegna men's white black striped trecapi	537	truehabit	https://www.ebay.com/itm/670-zegna-mens-white	dress s…	335088013170
450	junk food clothing coke-cola graphic white bo	25	dana_tdc	https://www.ebay.com/itm/junk-food-clothing-c	tops	175987869791
451	anne klein ladies dress sz 14 red floral zip	22	elct60	https://www.ebay.com/itm/anne-klein-ladies-dr	dresses	186135641289
452	old navy women's size large black/mustard flo	14	nycskyline	https://www.ebay.com/itm/old-navy-womens-size	dresses	386226191606
453	new look cheesecloth dress uk medium	3	2003jeh	https://www.ebay.com/itm/new-look-cheesecloth	dresses	126153523608
454	shein womens windowpane print belted wool-mix	30	mickintexas	https://www.ebay.com/itm/shein-womens-windo…	dresses	235269568739
455	1978 retro sexy women nautical dresses long l	11	westgateauct…	https://www.ebay.com/itm/1978-retro-sexy-wome	1950-1…	225835763758
456	women sexy dress bandage micro mini dress tig	10	wangjiacheng	https://www.ebay.com/itm/women-sexy-dress-ban	dresses	175984529977
457	reiss marisa lotus dress lotus coral red / uk	45	torg73	https://www.ebay.com/itm/reiss-marisa-lotus-d	dresses	175987869709
458	colorful child print fabric with children, wr	4	myhands_72	https://www.ebay.com/itm/colorful-child-print	fabric	314912009338
459	zara green dress women's large l holiday new	27	bekind2onea…	https://www.ebay.com/itm/zara-green-dress-wom	dresses	335088013083
460	used umbro soccer pants youth boys size l lar	10	5odessa	https://www.ebay.com/itm/used-umbro-soccer-pa	youth	276121871530
461	river island womens black shift sequin dress	38	reskinned	https://www.ebay.com/itm/river-island-womens-	dresses	314912009311
462	lot of 4 girls size 14 clothes	15	rylemcgear_0	https://www.ebay.com/itm/lot-4-girls-size-14-	mixed lots	314912009265
463	bratz doll lot mga clothing welcome to fabulo	11	peacedud	https://www.ebay.com/itm/bratz-doll-lot-mga-c	dolls &…	394958126279
464	catherine's sleeveless maxi dress lined gorge	25	kimkayaus	https://www.ebay.com/itm/catherines-sleeveles	dresses	166396519264
465	ll bean snowpants youth size 10 boys girls un	19	remishops	https://www.ebay.com/itm/ll-bean-snowpants-yo	snow p…	134782094782
466	ecco men's dress shoes size eu 45 us 11 leath	45	emmyrose842	https://www.ebay.com/itm/ecco-mens-dress-shoe	dress s…	225835763560
467	river island womens pink polyester sun dress	23	reskinned	https://www.ebay.com/itm/river-island-womens-	dresses	314912009175
468	danner women's gray leather trail 2650 mid at	60	venalli.b	https://www.ebay.com/itm/danner-womens-gray-l	women's	355136110235

Figure 4: eBayPosts

All of these features combine to allow a continuous abstraction, organization, and storage of data to be analyzed in a useful way.

3 PRELIMINARY EXPLORATION OF DATA

From the data we gathered, there are various noteworthy trends and facts we can determine. To begin, there is a clear correlation between big sales and fashion trends. For example, observations showed that the Bullseye deals section in eBay acted as a major source of sales in our data set. This is supplemented by several discussions in our targeted subreddits revolving around finding specific clothing items for a desired price. In addition, almost all of the listings on eBay fall at or below 100 USD, with a vast majority ranging between 5 USD and 30 USD. The data for this observation can be visualized from the graph below, which indicates the volume of eBay listings collected over time along with the prices associated with the listings:

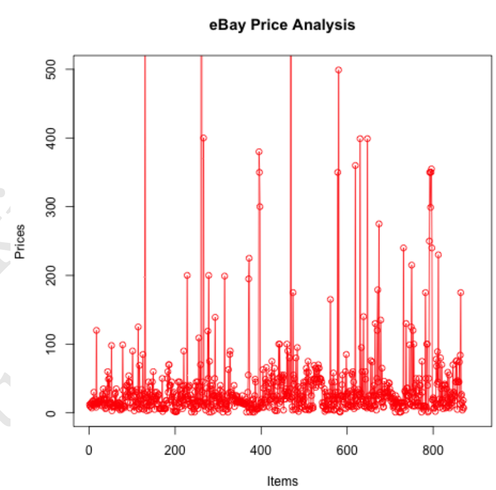

Figure 5: eBay Line Graph

In addition, we determined a noticeable disparity between trends from the different sources. For eBay, there appears to be a strong emphasis on women's clothing. In particular, a vast majority of the sales on eBay indicate dresses to be the category of the items, with approximately half of our dataset falling under this category. On the contrary, the Reddit posts prove to be extremely diversified, with most of the discourse revolving around people inquiring about the source of specific and distinct clothing items. As a result, this potentially indicates a difference in viewership when it comes to these sources, with women being a predominant user of the apparel section on eBay while Reddit consists of a varied user base. As we continue to collect data throughout the semester, we will continue to increase our sample size to assert or counter these preliminary trends and potentially spark further discussion about our findings.

4 CHANGES FROM PROPOSAL

In general, the premise of our project remained the same as our proposal. From the proposal, we analyzed data from the same sources using the same means of data extraction. Through this, we were able to work towards our goal from the proposal, which was to gather sufficient data from the sources to determine current trends

and correlations in fashion. To further solidify our project, we made a dedicated effort to incorporate the comments from the feedback to expand upon the original proposal. To begin, we expanded upon the details we included on the use of the Reddit API, ensuring we fully understood the endpoints we are using from the site. In addition, we made sure to have a detailed system architecture diagram that visuals the various components of the system and shows how they communicate and interact with each other, allowing us to maintain a sense of organization and structure for our project. Through these enhancements, we were able to develop a firmer grasp of the implementation and sources we are handling and properly handle the data we gather.

5 CHALLENGES

Throughout the project, there were a myriad of problems that we faced while collecting data. To begin, during our plan phase, we attempted to experiment with various APIs to accomplish the different objectives we had in mind. However, we soon realized that a vast majority of the APIs we viewed required special user credentials or a paid membership to utilize. As a result, this severely limited the scope of our project. In general, much of the issues stem from this project being a predominantly new experience, as web crawling is seldom used in the computer science curriculum. As a result, the project required a significant amount of experimentation and research to complete. In particular, when we successfully arrived at a project idea with compatible APIs, there were several issues we encountered regarding our implementation. To begin, we had several issues utilizing the virtual machine for our project. In particular, we periodically had issues with the remote virtual machines failing to run and connect. This halted the development of our project in various parts, particularly when it came to making the Faktory file. In addition, we also had to figure out how to implement external libraries into the remote servers to accomplish the functions of our project. Eventually, we were able to resolve this issue through the use of a Python environment variable, which allows us to focus on a specified scope rather than the system-wide version of Python. Along with this, while we were able to affirm a quick understanding of the Reddit API, eBay caused issues in the earlier implementation stage. In particular, we initially struggled to find a reliable and affordable API that can web scrap eBay and its content. Thankfully, we were eventually able to find that enrolling in the eBay Developer Program allows us to have access to data from eBay for the extent of the project's scope, allowing us to import the eBaySDK library module to utilize its functions. Finally, after we developed a means for gathering data, it was difficult to integrate the data into a running database that can consistently support the flow of data it would receive. Eventually, we overcame this issue by utilizing mySQL Workbench, which provided a stable platform and interface to store and view our data. Through these solutions, we were able to resolve the overarching concern for our project, which was being able to collect a sufficient number of data to use for analysis. As such, we were ultimately able to overcome the various challenges we encountered to complete our objective for this project.

6 PROJECTION OF DATA

From the data we collected, we can project the volume of data that will be gathered over time. To begin, for Reddit we are projected to receive 100-200 subreddit posts per day, with about 5-10 comments per post on average. As for eBay, we are projected to have about 5000 API calls per day, which would pull around 88 posts each API call. After eliminating duplicate listings, this has proven to pull a comparable amount of listings to the number of subreddit posts per day. As a result, by the end of the semester, we can predict to have around 7000 to 14000 database items for redditPosts and corresponding comments to go along with them, as well as a comparable number of eBay listings as well. As for the storage, we initiated the AWS database to have 20 GB of total storage. Due to the predominance of text data over video or image media, we predict that this will be sufficient to collect data throughout the project length. This can be shown through our most recent storage update, which indicates that we have about 18.5 GB of storage left after all the data we've collected as seen here:

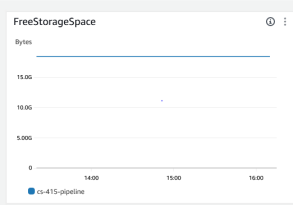

Figure 6: Database Storage Metric

Through these calculated projections, we gauge that we will gather a highly sufficient number of samples to be analyzed, as well as the storage capabilities to support the influx of data.