VARSHA PATTIPATI, USC ID: 8244565758

Information Integration on the Web(CSCI 548) – Homework1

TASK1:

The website I chose to crawl is ebay.

WEBSITE: www.ebay.com

Description: ebay is essentially a web platform for people to buy and sell goods. It is a place where sellers advertise their products(both old and new) and buyers can buy and also bid for products they want to purchase.

TASK2:

The query I chose is as follows:

Extract webpages with the following filter

Samsung Cell Phones and Smartphones

Price range:

- Minimum: $0
- Maximum: $400

Sample webpages:

a.) http://www.ebay.com/itm/Samsung-Galaxy-S-4-SGH-M919-BRAND-NEW-T-Mobile-/252074961574?hash=item3ab0d6aaa6 (Figure 1)

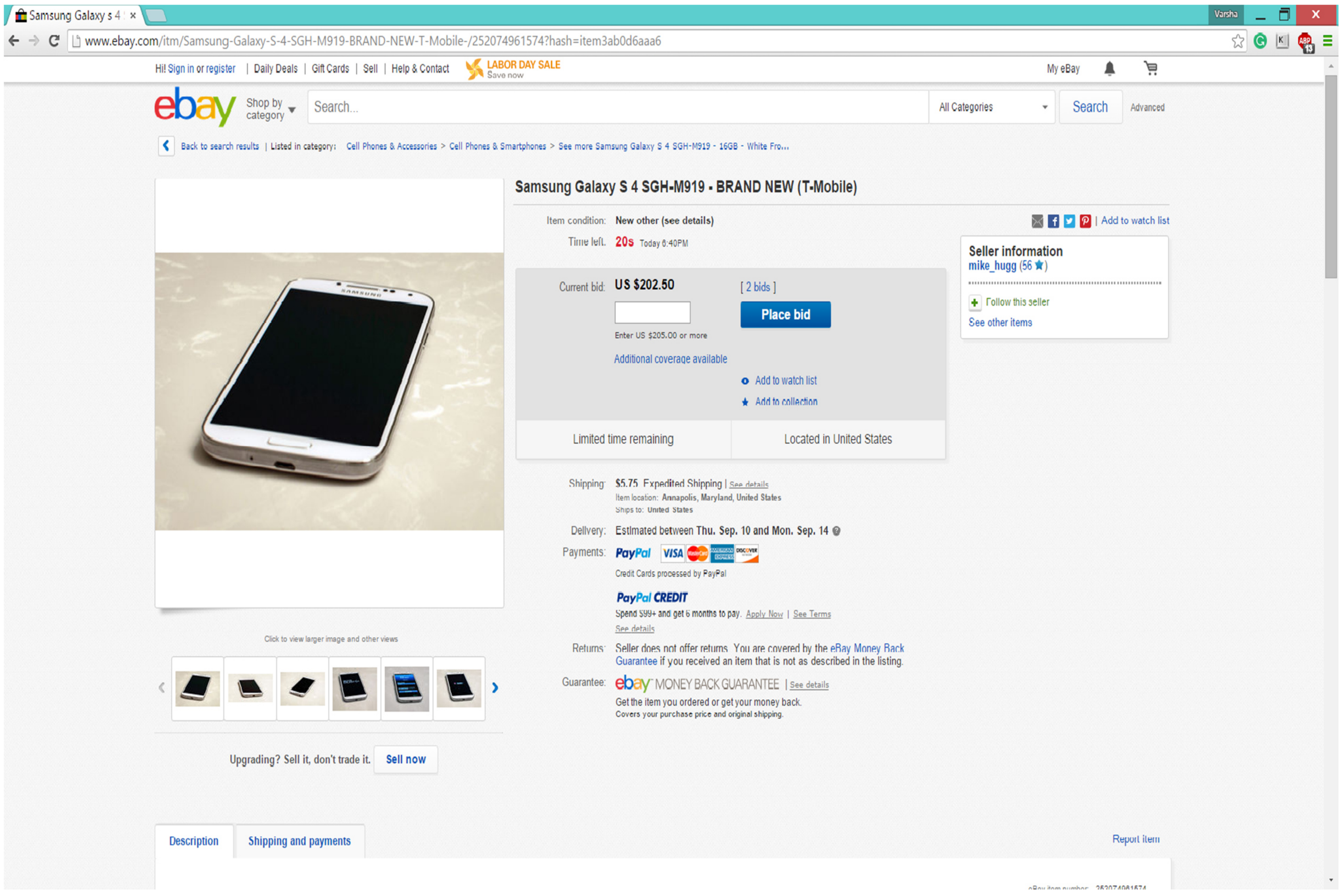
b.) http://www.ebay.com/itm/Samsung-Galaxy-S5-SM-G900A-16GB-AT-T-Unlocked-Smartphone-Black-White-/311370709654?var=&hash=item487f23fa96 (Figure 2)

TASK3:

QUESTION 1: Describe the crawler you used to get the results. If you wrote your own crawler, explain which technologies you used to develop it.

Answer: I had previously used Nutch for crawling websites in some projects I did last semester and hence very familiar with how it works. But I preferred creating one on my own in Java so that I could I could learn the language. I used jsoup parser library and imported the JAR file into my project.

Technologies used: Java, jsoup.



www.ebay.com/itm/Samsung-Galaxy-S-4-SGH-M919-BRAND-NEW-T-Mobile-/252074961574?hash=item3ab0d6aaa6

Back to search results | Listed in category: Cell Phones & Accessories > Cell Phones & Smartphones > See more Samsung Galaxy S 4 SGH-M919 - 16GB - White Fro...

Samsung Galaxy S 4 SGH-M919 - BRAND NEW (T-Mobile)

Item condition: New other (see details)

Time left: 20s Today 6:40PM

Current bid: US $202.50 [2 bids]

Enter US $205.00 or more

Additional coverage available

Limited time remaining | Located in United States

Seller information
mike_hugg (56 ★)

Shipping: $5.75 Expedited Shipping | See details
Item location: Annapolis, Maryland, United States
Ships to: United States

Delivery: Estimated between Thu. Sep. 10 and Mon. Sep. 14

Payments: PayPal VISA MasterCard AMERICAN EXPRESS DISCOVER
Credit Cards processed by PayPal

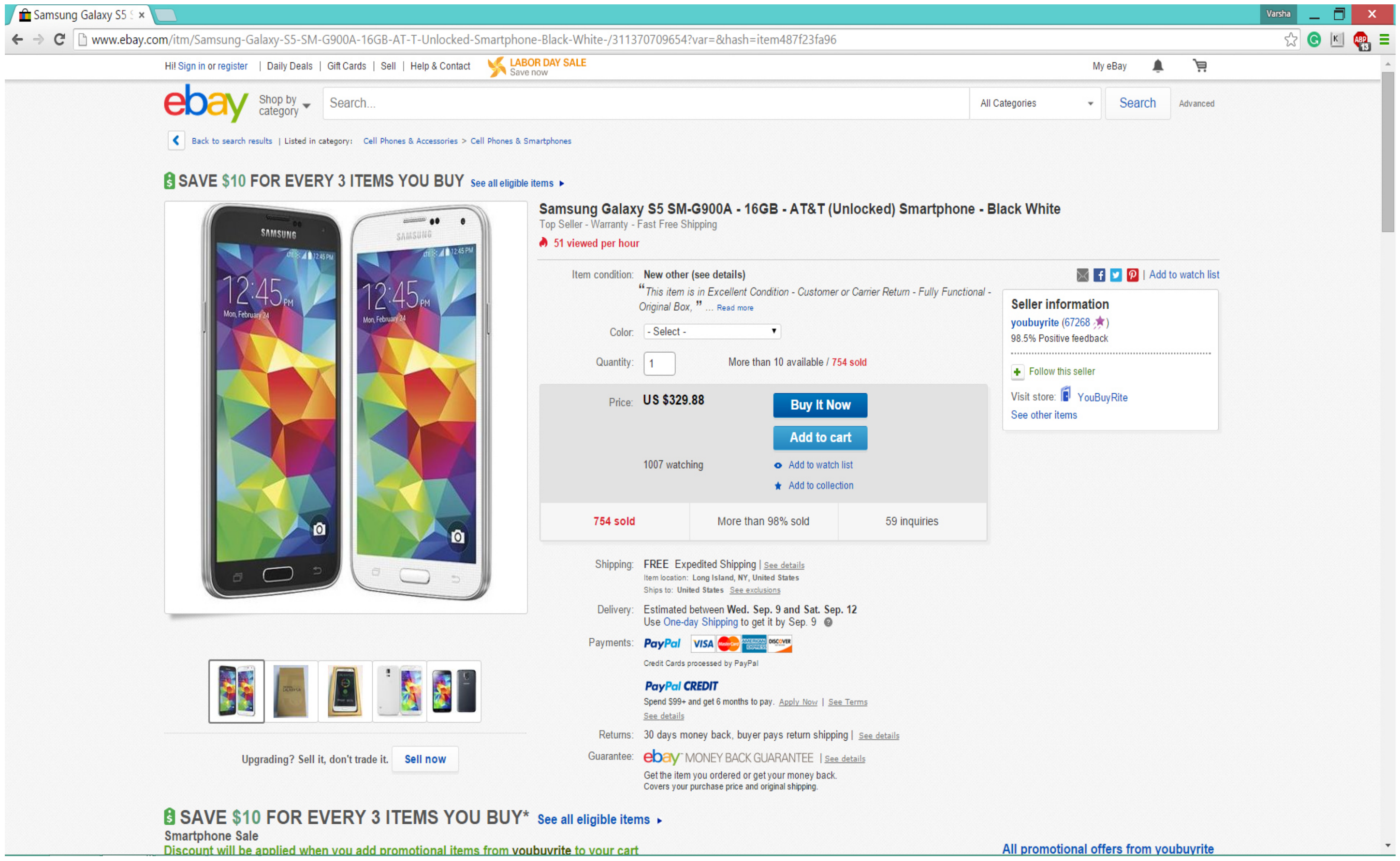
PayPal CREDIT
Spend $99+ and get 6 months to pay. Apply Now | See Terms
See details

Returns: Seller does not offer returns. You are covered by the eBay Money Back Guarantee if you received an item that is not as described in the listing.

Guarantee: ebay MONEY BACK GUARANTEE | See details
Get the item you ordered or get your money back.
Covers your purchase price and original shipping.

Upgrading? Sell it, don't trade it.

Figure1. A webpage corresponding to the query I selected


www.ebay.com/itm/Samsung-Galaxy-S5-SM-G900A-16GB-AT-T-Unlocked-Smartphone-Black-White-/311370709654?var=&hash=item487f23fa96

Hi! Sign in or register | Daily Deals | Gift Cards | Sell | Help & Contact LABOR DAY SALE Save now

Back to search results | Listed in category: Cell Phones & Accessories > Cell Phones & Smartphones

SAVE $10 FOR EVERY 3 ITEMS YOU BUY See all eligible items

Samsung Galaxy S5 SM-G900A - 16GB - AT&T (Unlocked) Smartphone - Black White
Top Seller - Warranty - Fast Free Shipping
51 viewed per hour

Item condition: New other (see details)
"This item is in Excellent Condition - Customer or Carrier Return - Fully Functional - Original Box, " ... Read more

Color: - Select -

Quantity: 1 More than 10 available / 754 sold

Price: US $329.88

1007 watching

754 sold	More than 98% sold	59 inquiries

Seller information
youbuyrite (67268)
98.5% Positive feedback
Follow this seller
Visit store: YouBuyRite
See other items

Shipping: FREE Expedited Shipping | See details
Item location: Long Island, NY, United States
Ships to: United States See exclusions

Delivery: Estimated between Wed. Sep. 9 and Sat. Sep. 12
Use One-day Shipping to get it by Sep. 9

Payments: Credit Cards processed by PayPal
PayPal CREDIT
Spend $99+ and get 6 months to pay. Apply Now | See Terms
See details

Returns: 30 days money back, buyer pays return shipping | See details

Guarantee: ebay MONEY BACK GUARANTEE | See details
Get the item you ordered or get your money back.
Covers your purchase price and original shipping.

Upgrading? Sell it, don't trade it. Sell now

SAVE $10 FOR EVERY 3 ITEMS YOU BUY* See all eligible items
Smartphone Sale
Discount will be applied when you add promotional items from youbuyrite to your cart

All promotional offers from youbuyrite

Figure2. A second webpage corresponding to the query

QUESTION 2: How does your crawler avoid a crawler trap (infinite loop)?

Answer:

- I started with some seed URLs and kept adding outgoing links from those pages to a HashSet, say 'pagesToVisit'.
- I also add pages already crawled to another HashSet, say 'pagesVisited'.
- I wrote a function to retrieve the next URL to be crawled from the HashSet('pagesToVisit') but made sure it is not crawled already by checking if 'pagesVisited' already contains the URL.
- So I never visit/crawl a webpage that has already been crawled. This is how I avoided *crawler trap*(Infinite Loop)

QUESTION 3: What is the seed URL(s)?

Answer: The seed URL is: http://www.ebay.com/sch/samsung-galaxy

QUESTION 4: How did you manage to only collect the webpages respecting the template(s) in Task 2? How did you discard irrelevant pages?

Answer:

- I observed that ebay has a way/pattern in naming the file and also storing it. All webpages that are about a particular product instead of a list of items are from a particular folder called "itm" and all URLs of webpages that match the query usually have the words "Samsung" or "Galaxy" in them.
- I used Jsoup to connect to the webpages and extract HTML.
- The price information of the products is usually embedded in the content attribute of a meta tag with a property attribute having the value "og:Description".
- I made sure that the webpages are from the "itm" folder and also the products are Samsung cell phones by parsing the URL.
- I then parsed the content attribute value of the meta tag to make sure that the price of the product complies with the query.
- I also made sure that the crawler does not go out of the domain and visit webpages on sites other than ebay by checking the domain name of each URL using the getHost() method in the URI class(Java)

QUESTION 5: If you were not able to collect 1000 pages, explain what the problem was.

Answer: I was able to collect 1000 webpages.

***ADDITIONAL:**

- I have saved the webpages as 1.html, 2.html and so on.
- This does not provide much information about the URLs of the webpages. So I saved the URLs of the pages to a file named “urlList.txt”.
- I am including this file along with the other folders I am supposed to submit. (This is saved in the source folder)
- The name of the source code file is Specific.java
  There is another file called trial.java in the source folder. Please ignore it.