Gaurab Baral and Aditya Khanal

Professor Seth Adjei

DSC-200

Lab 7: Web-Scraping Report

1. **Websites Chosen:**
   1. Google: <https://www.google.com>
   2. Amazon: <https://www.amazon.com>
2. **Description of the Website:**
   1. Google: Google is a multinational technology company known for its popular search engine, online advertising, software, and services (like Android and Google Drive), hardware products, cloud services, and AI innovation. In this report, we are looking at google finance which is a web-based platform that provides financial information, stock quotes, news, interactive charts, and portfolio tracking tools for users to monitor and analyze financial markets and investments. The link to this is <https://www.google.com/finance>
   2. Amazon: Amazon is a major technology and e-commerce company based in Seattle. It offers online shopping, cloud computing (AWS), subscription services (Amazon Prime), hardware devices, digital content, and more. It's known for its customer-focused approach and impact on various industries. In this report, we are examining the price and ratings of men kurtas, providing insights on these aspects. The link to this is <https://www.amazon.com/s?k=men+kurtha&crid=18QO0YPRW3RLF&sprefix=men+kurtha%2Caps%2C98&ref=nb_sb_noss_2>
3. **Reason for Website Selection:**
   1. Google: Google Finance is selected because it is a platform that provides financial information, making it relevant for analyzing financial markets and investments. Since it provides real-time changes in the stock market price, it offers a crucial advantage in tracking and reacting to market fluctuations, which is essential for accurate financial market analysis.
   2. Amazon: We chose Amazon because it's a popular e-commerce platform where we can find a wide range of products, including kurtas. This choice is particularly relevant as we aim to track the prices of kurtas during the festive season in Nepal, a period when many people purchase these garments. Amazon's vast product range and user ratings make it an ideal source for our price and rating analysis during this festive season.

Both Google Finance and Amazon have user-friendly interfaces, making it easy to access and extract the necessary data for our report.

1. **Summary of the Terms of Service(TOS):**

Both Google and Amazon permit us to conduct web scraping activities on their websites, as long as we adhere to their respective requirements, which, as outlined in our report, appear to be within our compliance. As our scraping activities do not cause any harm, we can proceed with our website scraping endeavors.

Link to Google TOS: <https://policies.google.com/terms?hl=en&fg=1>

Link to Amazon TOS: <https://www.amazon.com/gp/help/customer/display.html?nodeId=508088&ref_=footer_cou>

1. **Terms of Service:**
   1. Google: If any of your content (1) breaches these terms, service-specific additional terms or policies, (2) violates applicable law, or (3) could harm our users, third parties, or Google, then we reserve the right to take down some or all of that content in accordance with applicable law. Examples include child pornography, content that facilitates human trafficking or harassment, terrorist content, and content that infringes someone else’s intellectual property rights. Google reserves the right to suspend or terminate your access to the services or delete your Google Account if any of these things happen:
      1. you materially or repeatedly breach these terms, service-specific additional terms, or policies.
      2. we’re required to do so to comply with a legal requirement or a court order.
      3. your conduct causes harm or liability to a user, third party, or Google — for example, by hacking, phishing, harassing, spamming, misleading others, or scraping content that doesn’t belong to you.
   2. Amazon: Subject to your compliance with these Conditions of Use and any Service Terms, and your payment of any applicable fees, Amazon or its content providers grant you a limited, non-exclusive, non-transferable, non-sublicensable license to access and make personal and non-commercial use of the Amazon Services. This license does not include any resale or commercial use of any Amazon Service, or its contents; any collection and use of any product listings, descriptions, or prices; any derivative use of any Amazon Service or its contents; any downloading, copying, or other use of account information for the benefit of any third party; or any use of data mining, robots, or similar data gathering and extraction tools. All rights not expressly granted to you in these Conditions of Use, or any Service Terms are reserved and retained by Amazon or its licensors, suppliers, publishers, rightsholders, or other content providers. No Amazon Service, nor any part of any Amazon Service, may be reproduced, duplicated, copied, sold, resold, visited, or otherwise exploited for any commercial purpose without express written consent of Amazon. You may not frame or utilize framing techniques to enclose any trademark, logo, or other proprietary information (including images, text, page layout, or form) of Amazon without express written consent. You may not use any meta tags or any other "hidden text" utilizing Amazon's name or trademarks without the express written consent of Amazon. You may not misuse the Amazon Services. You may use Amazon Services only as permitted by law. The licenses granted by Amazon terminate if you do not comply with these Conditions of Use or any Service Terms.
2. **Robots.txt:**
   1. Google: Allow: /finance
   2. Amazon: Allow: /-/en$

(See more of robots.txt at Section 6(Appendix).

1. **Appendix:**
   1. Google: <https://www.google.com/robots.txt>

User-agent: \* Disallow: /search Allow: /search/about Allow: /search/static Allow: /search/howsearchworks Disallow: /sdch Disallow: /groups Disallow: /index.html? Disallow: /? Allow: /?hl= Disallow: /?hl=\*& Allow: /?hl=\*&gws\_rd=ssl$ Disallow: /?hl=\*&\*&gws\_rd=ssl Allow: /?gws\_rd=ssl$ Allow: /?pt1=true$ Disallow: /imgres Disallow: /u/ Disallow: /preferences Disallow: /setprefs Disallow: /default Disallow: /m? Disallow: /m**/ Allow: /m/finance** Disallow: /wml? Disallow: /wml/? Disallow: /wml/search? Disallow: /xhtml? Disallow: /xhtml/? Disallow: /xhtml/search? Disallow: /xml? Disallow: /imode? Disallow: /imode/? Disallow: /imode/search? Disallow: /jsky? Disallow: /jsky/? Disallow: /jsky/search? Disallow: /pda? Disallow: /pda/? Disallow: /pda/search? Disallow: /sprint\_xhtml Disallow: /sprint\_wml Disallow: /pqa Disallow: /palm Disallow: /gwt/ Disallow: /purchases Disallow: /local? Disallow: /local\_url Disallow: /shihui? Disallow: /shihui/ Disallow: /products? Disallow: /product\_ Disallow: /products\_ Disallow: /products; Disallow: /print Disallow: /books/ Disallow: /bkshp?\*q=\* Disallow: /books?\*q=\* Disallow: /books?\*output=\* Disallow: /books?\*pg=\* Disallow: /books?\*jtp=\* Disallow: /books?\*jscmd=\* Disallow: /books?\*buy=\* Disallow: /books?\*zoom=\* Allow: /books?\*q=related Allow: /books?\*q=subject:\* Allow: /books/about Allow: /booksrightsholders Allow: /books?\*zoom=1\* Allow: /books?\*zoom=5\* Allow: /books/content?\*zoom=1\* Allow: /books/content?\*zoom=5\* Disallow: /ebooks/ Disallow: /ebooks?\*q=\* Disallow: /ebooks?\*output=\* Disallow: /ebooks?\*pg=\* Disallow: /ebooks?\*jscmd=\* Disallow: /ebooks?\*buy=\* Disallow: /ebooks?\*zoom=\* Allow: /ebooks?\*q=related:\* Allow: /ebooks?\*q=editions:\* Allow: /ebooks?\*q=subject:\* Allow: /ebooks?\*zoom=1\* Allow: /ebooks?\*zoom=5\* Disallow: /patents? Disallow: /patents/download/ Disallow: /patents/pdf/ Disallow: /patents/related/ Disallow: /scholar Disallow: /citations? Allow: /citations?user= Disallow: /citations?\*cstart= Allow: /citations?view\_op=new\_profile Allow: /citations?view\_op=top\_venues Allow: /scholar\_share Disallow: /s? Disallow: /maps? Allow: /maps?\*output=classic\* Allow: /maps?\*file= Disallow: /mapstt? Disallow: /mapslt? Disallow: /mapabcpoi? Disallow: /maphp? Disallow: /mapprint? Disallow: /maps/ Allow: /maps/search/ Allow: /maps/dir/ Allow: /maps/d/ Allow: /maps/reserve Allow: /maps/about Allow: /maps/match Disallow: /maps/api/js/ Allow: /maps/api/js Disallow: /mld? Disallow: /staticmap? Disallow: /help/maps/streetview/partners/welcome/ Disallow: /help/maps/indoormaps/partners/ Disallow: /lochp? Disallow: /center Disallow: /ie? Disallow: /blogsearch/ Disallow: /blogsearch\_feeds Disallow: /advanced\_blog\_search Disallow: /uds/ Disallow: /chart? Disallow: /transit? Allow: /calendar$ Allow: /calendar/about/ Disallow: /calendar Disallow: /cl2/feeds/ Disallow: /cl2/ical/ Disallow: /coop/directory Disallow: /coop/manage Disallow: /trends? Disallow: /trends/music? Disallow: /trends/hottrends? Disallow: /trends/viz? Disallow: /trends/embed.js? Disallow: /trends/fetchComponent? Disallow: /trends/beta Disallow: /trends/topics Disallow: /musica Disallow: /musicad Disallow: /musicas Disallow: /musicl Disallow: /musics Disallow: /musicsearch Disallow: /musicsp Disallow: /musiclp Disallow: /urchin\_test/ Disallow: /movies? Disallow: /wapsearch? Allow: /safebrowsing/diagnostic Allow: /safebrowsing/report\_badware/ Allow: /safebrowsing/report\_error/ Allow: /safebrowsing/report\_phish/ Disallow: /reviews/search? Disallow: /orkut/albums Disallow: /cbk Disallow: /recharge/dashboard/car Disallow: /recharge/dashboard/static/ Disallow: /profiles/me Allow: /profiles Disallow: /s2/profiles/me Allow: /s2/profiles Allow: /s2/oz Allow: /s2/photos Allow: /s2/search/social Allow: /s2/static Disallow: /s2 Disallow: /transconsole/portal/ Disallow: /gcc/ Disallow: /aclk Disallow: /cse? Disallow: /cse/home Disallow: /cse/panel Disallow: /cse/manage Disallow: /tbproxy/ Disallow: /imesync/ Disallow: /shenghuo/search? Disallow: /support/forum/search? Disallow: /reviews/polls/ Disallow: /hosted/images/ Disallow: /ppob/? Disallow: /ppob? Disallow: /accounts/ClientLogin Disallow: /accounts/ClientAuth Disallow: /accounts/o8 Allow: /accounts/o8/id Disallow: /topicsearch?q= Disallow: /xfx7/ Disallow: /squared/api Disallow: /squared/search Disallow: /squared/table Disallow: /qnasearch? Disallow: /app/updates Disallow: /sidewiki/entry/ Disallow: /quality\_form? Disallow: /labs/popgadget/search Disallow: /buzz/post Disallow: /compressiontest/ Disallow: /analytics/feeds/ Disallow: /analytics/partners/comments/ Disallow: /analytics/portal/ Disallow: /analytics/uploads/ Allow: /alerts/manage Allow: /alerts/remove Disallow: /alerts/ Allow: /alerts/$ Disallow: /ads/search? Disallow: /ads/plan/action\_plan? Disallow: /ads/plan/api/ Disallow: /ads/hotels/partners Disallow: /phone/compare/? Disallow: /travel/clk Disallow: /travel/flights/s/ Disallow: /hotelfinder/rpc Disallow: /hotels/rpc Disallow: /commercesearch/services/ Disallow: /evaluation/ Disallow: /chrome/browser/mobile/tour Disallow: /compare/\*/apply\* Disallow: /forms/perks/ Disallow: /shopping/suppliers/search Disallow: /ct/ Disallow: /edu/cs4hs/ Disallow: /trustedstores/s/ Disallow: /trustedstores/tm2 Disallow: /trustedstores/verify Disallow: /adwords/proposal Disallow: /shopping?\* Disallow: /shopping/product/ Disallow: /shopping/seller Disallow: /shopping/ratings/account/metrics Disallow: /shopping/ratings/merchant/immersivedetails Disallow: /shopping/reviewer Allow: /about/careers/applications/ Allow: /about/careers/applications-a/ Allow: /about/careers/applications/teams/ Allow: /about/careers/applications-a/teams/ Allow: /about/careers/applications/locations/ Allow: /about/careers/applications-a/locations/ Allow: /about/careers/applications/benefits/ Allow: /about/careers/applications-a/benefits/ Allow: /about/careers/applications/students/ Allow: /about/careers/applications-a/students/ Allow: /about/careers/applications/jobs/results/$ Allow: /about/careers/applications-a/jobs/results/$ Allow: /about/careers/applications/cloud/ Allow: /about/careers/applications-a/cloud/ Allow: /about/careers/applications/hardware/ Allow: /about/careers/applications-a/hardware/ Allow: /about/careers/applications/eeo/ Allow: /about/careers/applications-a/eeo/ Allow: /about/careers/applications/how-we-hire/ Allow: /about/careers/applications-a/how-we-hire/ Allow: /about/careers/applications/interview-tips/ Allow: /about/careers/applications-a/interview-tips/ Disallow: /about/careers/applications/ 11/5/23, 11:24 PM google.com/robots.txt https://www.google.com/robots.txt 5/6 Disallow: /about/careers/applications-a/ Disallow: /landing/signout.html Disallow: /webmasters/sitemaps/ping? Disallow: /ping? Disallow: /gallery/ Disallow: /landing/now/ontap/ Allow: /searchhistory/ Allow: /maps/reserve Allow: /maps/reserve/partners Disallow: /maps/reserve/api/ Disallow: /maps/reserve/search Disallow: /maps/reserve/bookings Disallow: /maps/reserve/settings Disallow: /maps/reserve/manage Disallow: /maps/reserve/payment Disallow: /maps/reserve/receipt Disallow: /maps/reserve/sellersignup Disallow: /maps/reserve/payments Disallow: /maps/reserve/feedback Disallow: /maps/reserve/terms Disallow: /maps/reserve/m/ Disallow: /maps/reserve/b/ Disallow: /maps/reserve/partner-dashboard Disallow: /about/views/ Disallow: /intl/\*/about/views/ Disallow: /local/cars Disallow: /local/cars/ Disallow: /local/dealership/ Disallow: /local/dining/ Disallow: /local/place/products/ Disallow: /local/place/reviews/ Disallow: /local/place/rap/ Disallow: /local/tab/ Disallow: /localservices/\* Allow: /finance Allow: /js/ Disallow: /nonprofits/account/ Disallow: /fbx Disallow: /uviewer Disallow: /landing/cmsnext-root/ # AdsBot User-agent: AdsBot-Google Disallow: /maps/api/js/ Allow: /maps/api/js Disallow: /maps/api/place/js/ Disallow: /maps/api/staticmap Disallow: /maps/api/streetview # Crawlers of certain social media sites are allowed to access page markup when google.com/imgres\* links are shared. To learn more, please contact images-robots-allowlist@google.com. User-agent: Twitterbot Allow: /imgres Allow: /search Disallow: /groups Disallow: /hosted/images/ Disallow: /m/ User-agent: facebookexternalhit Allow: /imgres Allow: /search Disallow: /groups Disallow: /hosted/images/ Disallow: /m/ 11/5/23, 11:24 PM google.com/robots.txt https://www.google.com/robots.txt 6/6 Sitemap: https://www.google.com/sitemap.xml

* 1. Amazon: <https://www.amazon.com/robots.txt>

User-agent: \* Disallow: /exec/obidos/account-access-login Disallow: /exec/obidos/change-style Disallow: /exec/obidos/flex-sign-in Disallow: /exec/obidos/handle-buy-box Disallow: /exec/obidos/tg/cm/member/ Disallow: /gp/aw/help/id=sss Disallow: /gp/cart Disallow: /gp/flex Disallow: /gp/product/e-mail-friend Disallow: /gp/product/product-availability Disallow: /gp/product/rate-this-item Disallow: /gp/sign-in Disallow: /gp/reader Disallow: /gp/sitbv3/reader Disallow: /gp/richpub/syltguides/create Disallow: /gp/gfix Disallow: /gp/associations/wizard.html Disallow: /gp/dmusic/order Disallow: /gp/legacy-handle-buy-box.html Disallow: /gp/aws/ssop Disallow: /gp/yourstore Disallow: /gp/gift-central/organizer/add-wishlist Disallow: /gp/vote Disallow: /gp/voting/ Disallow: /gp/music/wma-pop-up Disallow: /gp/customer-images Disallow: /gp/richpub/listmania/createpipeline Disallow: /gp/content-form Disallow: /gp/pdp/invitation/invite Disallow: /gp/customer-reviews/common/du Disallow: /gp/customer-reviews/write-a-review.html Disallow: /gp/associations/wizard.html Disallow: /gp/music/clipserve Disallow: /gp/customer-media/upload Disallow: /gp/history Disallow: /gp/item-dispatch Disallow: /gp/dmusic/order/handle-buy-box.html Disallow: /gp/recsradio Disallow: /gp/slredirect Disallow: /dp/shipping/ Disallow: /dp/twister-update/ Disallow: /dp/manual-submit/ Disallow: /dp/e-mail-friend/ Disallow: /dp/product-availability/ Disallow: /dp/rate-this-item/ Disallow: /gp/registry/wishlist/\*/reserve Disallow: /gp/structured-ratings/actions/get-experience.html Disallow: /gp/twitter/ Disallow: /ap/signin Disallow: /gp/registry/wishlist/ Disallow: /wishlist/ Allow: /wishlist/universal\* Allow: /wishlist/vendor-button\* Allow: /wishlist/get-button\* Disallow: /gp/wishlist/ Allow: /gp/wishlist/universal\* Allow: /gp/wishlist/vendor-button\* Allow: /gp/wishlist/ipad-install\* Disallow: /registry/wishlist/ Disallow: /review/common/du Disallow: /gp/registry/search.html Disallow: /product-reviews/B0069IY63Y Disallow: /gp/orc/rml/ Disallow: \*/gcrnsts Disallow: /gp/gc/widget Disallow: /gp/dmusic/mp3/player Disallow: /gp/entity-alert/external Disallow: /gp/customer-reviews/dynamic/sims-box Disallow: /review/dynamic/sims-box Disallow: /gp/redirect.html Disallow: /gp/twister/ajaxv2 Disallow: /ss/twister/ajax Disallow: /b?\*node=7454917011 Disallow: /b?\*node=7454927011 Disallow: /b?\*node=7454939011 Disallow: /b?\*node=7454898011 Disallow: /gp/customer-media/actions/delete/ Disallow: /gp/customer-media/actions/edit-caption/ Disallow: /gp/dmusic/ Allow: /gp/dmusic/promotions/PrimeMusic Allow: /gp/dmusic/promotions/AmazonMusicUnlimited Disallow: /gp/offer-listing/ Disallow: /b?\*node=9052533011 Disallow: /lm/R1XIHQVKXSKBNJ Disallow: /lm/R3HQ5WJSZK6QSO Disallow: /surprise/ Disallow: /local/ajax/ Disallow: \*/B00M3E1NYI Disallow: \*/B00M3E1Q5Y Disallow: \*/B00M3E1TOM Disallow: \*/B00M3E1WYO Disallow: \*/B00M3E204K Disallow: \*/B00M3E236A Disallow: \*/B00M3E260I Disallow: \*/B00M3E28WO Disallow: \*/B00M3E2BC6 Disallow: \*/B00M3E2DPQ Disallow: \*/B00M3E2GU8 Disallow: \*/B00M3E2J14 Disallow: \*/B00M3E2LOE Disallow: \*/B00M3E1HJY Disallow: /gp/socialmedia/giveaways Disallow: /gp/b2b-rd Disallow: /gp/aw/so.html Disallow: /gp/rentallist Disallow: /gp/video/dvd-rental/settings Disallow: /gp/rl/settings Disallow: /gp/video/settings Disallow: /gp/video/library Disallow: /gp/video/watchlist Disallow: /reviews/iframe Disallow: /gp/switch-language Disallow: /ga/p/ Disallow: /gp/profile/ Disallow: /giveaway/host/setup/ Disallow: /ss/customer-reviews/lighthouse/ Disallow: /ospublishing/story/\* Disallow: /gp/aw/ol/ Disallow: /gp/promotion/ Disallow: /hz/leaderboard/top-reviewers/ Disallow: /creatorhub Disallow: /creatorhub/\* Disallow: /slp/s$ Disallow: /-/ Allow: /-/es/ **Allow: /-/en$** Allow: /-/zh\_TW/ Allow: /-/zh\_TW$ 11/5/23, 11:25 PM amazon.com/robots.txt https://www.amazon.com/robots.txt 3/3 Allow: /-/he/ Allow: /-/he$ Disallow: /hz/help/contact/\*/message/$ Disallow: /gp/aw/shoppingAids/ Disallow: /rss/people/\*/reviews Disallow: /gp/pdp/rss/\*/reviews Disallow: /gp/cdp/member-reviews/ Disallow: /gp/aw/cr/ Disallow: \*/sim/B001132UEE Allow: /gp/offer-listing/B000 Allow: /gp/offer-listing/9000 Disallow: /gp/aag Allow: /gp/aag/main?\*seller=ABVFEJU8LS620 Disallow: /gp/pdp/profile/ Disallow: /gp/help/customer/express/c2c/ Disallow: /slp/\*/b$ Disallow: /hz/contact-us/ajax/initiate-trusted-contact/ Disallow: /gp/video/api Disallow: /hp/video/api Disallow: /gp/video/mystuff Disallow: /hp/video/mystuff Disallow: /gp/video/profiles Disallow: /hp/video/profiles User-agent: EtaoSpider Disallow: / User-agent: GPTBot Disallow: / User-agent: CCBot Disallow: /

1. **Conclusion:**

In conclusion, the chosen websites and their terms of service align with our goals for web scraping and data analysis. We can proceed with our scraping activities, keeping in mind the guidelines and policies outlined in this report. This will enable us to collect valuable data for our analysis while ensuring compliance with the terms and conditions of these websites.