Analysointityökalun rakentaminen verkkoharavoinnin pohjalta

Johanna Rantanen

Haaga-Helia ammattikorkeakoulu
Tietojenkäsittelyn koulutusohjelma
Opinnäytetyö
2024

Tiivistelmä

Tekijä
Johanna Rantanen

Tutkinto
Tradenomi

Raportin/Opinnäytetyön nimi
Analysointityökalun rakentaminen verkkoharavoinnin pohjalta

Sivu- ja liitesivumäärä
30

Tiivistelmä

Tässä toiminnallisessa opinnäytetyössä rakennettiin analysointityökalu tuotteiden keskihintojen vertailuun tiedon elinkaarimallia noudattaen. Keskihinnat työkaluun haettiin verkkoharavoinnin ohjelmalla kolmelta eri verkkosivulta Python ohjelmointikieltä hyödyntäen. Työ oli rajattu tutkimaan tietyn hetken hintatietoja, sillä opinnäytetyön tavoitteena oli selvittää, voidaanko verkkoharavoinnin pohjalta rakentaa toimiva analysoinnin työkalu. Työ perustuu tekijän omaan mielenkiintoon eikä sillä ole ulkoista toimeksiantajaa. Valmis lopputulos eli verkkoharavoinnin robotti, koodin hakema CSV-tiedosto, sekä Power BI -ohjelmalla luodut analysoinnin ja visualisoinnin näkymät jaetaan GitHubissa. Verkkoharavoinnin Projekti toteutettiin kevään 2024 aikana.

Työn keskeinen tietoperusta koostui kahdesta eri kokonaisuudesta tiedon elinkaaresta ja verkkoharavoinnista. Tiedon elinkaari on iteratiivinen prosessi, joka koostuu kahdeksasta eri vaiheesta datan luomisesta, keräämisestä, käsittelystä, varastoinnista, hallinnasta, analysoinnista, visualisoinnista ja tulkitsemisesta. Verkkoharavoinnin osuudessa käsiteltiin verkkoharavoinnin prosessia ja teknisiä ratkaisuja. Lisäksi tarkasteltiin verkkoharavointiin liittyviä haasteita, sekä eettisiä ja laillisia ongelmia. Tietoperustaa hyödynnettiin seuraavassa tutkimusvaiheessa, jossa rakennettiin analysoinnin työkalu käymällä läpi kaikki tiedon elinkaaren vaiheet järjestyksessä. Verkkoharavointia hyödynnettiin prosessin datan keräämisen ja käsittelyn vaiheissa. Työn lopussa arvioitiin luodun työkalun soveltuvuutta analysoinnin perustana.

Opinnäytetyössä asetetut tavoitteet saavutettiin, sillä siinä luotiin toimiva verkkoharavoinnin ohjelma Python ohjelmointikielellä, sekä interaktiivinen analysoinnin työkalu Power BI -sovelluksessa tiedon elinkaarimallia noudattaen.

Asiasanat
Verkkoharavointi, Python, Tiedon elinkaari, Power BI

Sisällys

1. Johdanto

Nopeasti digitalisoituvassa toimintaympäristössä on tarjolla yhä enemmän tietoa. Internetin luonteen vuoksi data on kuitenkin usein huonosti jäsenneltyä ja hajallaan usealla nettisivuilla eri muodoissa. Tämän tekee tiedon hakemisesta, tallentamisesta ja analysoimisesta haastavaa. Tiedonhaun prosessia voidaan automatisoida verkkoharavointi (web scraping) tekniikalla, joka hakee tarvittavat tiedot automaattisesti nettisivuilta, muokkaa datan haluttuun muotoon ja tallentaa sen paikalliselle levylle tarkempaa analysointia varten. (Nigam & Biswas 2021.)

Haettu data tulee analysoida, jotta siitä on hyötyä. Tehokkaan data-analyysin ja tiedonhallinnan avulla yritys voi saavuttaa huomattavaa kilpailuetua muihin verrattuna. Kun organisaatiossa on tietoa nopeasti saatavilla, on helpompaa tehdä hyvin informoituja päätöksiä. (Penn LPS 2022.) Tiedonhallinnan elinkaaren avulla voidaan varmistaa, että tarvittava data on oikeassa paikassa oikeaan aikaan.

1.1 Työn tavoitteet, tausta ja rajaus

Työn tarkoituksena on selvittää, miten verkkoharavoinnin ohjelma voidaan luoda Python ohjelmointikielellä ja voidaanko sen pohjalta luoda analysoinnin työkalu tiedon elinkaarimallin avulla. Tiedon elinkaari koostuu kahdeksasta eri vaiheesta datan luomisesta, keräämisestä, käsittelystä, varastoinnista, hallinnasta, analysoinnista, visualisoinnista ja johtopäätöksistä. Näistä vaiheista datan kerääminen ja käsittely toteutetaan verkkoharavoinnin ohjelmointiprojektilla. Työkalun avulla haetaan ajantasaista hintatietoa kolmelta eri verkkosivulta ja analysoidaan, missä kaupassa on halvimmat hinnat ja näin ollen asiakkaan kannattaisi asioida. Työn tavoitteena on tehdä toimiva ohjelma, jolla pystytään hakemaan dataa automaattisesti verkkosivuilta, sekä käyttää tätä tietoa analysointityökalun luomiseen Power BI -ohjelmassa.

Hintatiedot haetaan suomalaisen kosmetiikkayrityksen Lumene Oy:n Lumo ihonhoitosarjan tuotteille. Tuotteet on valittu, sillä käytän itse kyseisiä tuotteita ja haluaisin tietää, mistä saisin ne edullisimmin. Tämän vuoksi hinnat haetaan kolmesta Suomessa toimivasta vähittäiskaupasta, jotka ovat sallineet verkkoharavoinnin verkkosivuillaan. Tutkimuksessa haetaan hinnat yhdeltä tietyltä hetkeltä ja historiatietojen hakeminen on rajattu työn ulkopuolelle, sillä tarkoituksena on keskittyä datan elinkaarimallin toimivuuden selvittämiseen.

Teoriaosuudessa tarkastellaan ensin Harvard Business Schoolin (HBO) datan elinkaarimallia, jonka pohjalta tutkimus on laadittu. Tämän jälkeen käsitellään verkkoharavoinnin prosessia, työkaluja, sekä verkkoharavoinnin eettisiä ja laillisia ongelmia. Työkalut osiossa keskitytään erityisesti Python ohjelmointikielen verkkoharavointiin tarkoitettujen kirjastojen tutkimiseen, sillä niitä

käytetään tutkimuksen kolmannessa empiria osuudessa, jossa rakennetaan haravointirobotti. Tutkimuksessa käydään kaikki elinkaarimallin vaiheet läpi toimivan työkalun luomiseksi Power BI järjestelmässä. Työ huipentuu johtopäätöksiin, jossa tarkastellaan projektin lopputulosta ja esitetään jatkotutkimuksen ehdotuksia.

1.2 Keskeiset käsitteet

CSS	CSS on tyylisivukieli, jolla kuvataan HTML-dokumentin ulkoasua. Sen avulla määritellään, miltä dokumentin elementit näyttävät ruudulla. (MDN 2024a.)
DAX	DAX (Data Analysis Expressions) on Power BI:n kirjasto, joka sisältää funktioita ja operaattoreita, jolla voidaan luoda kaavoja ja mittareita taulukkotietomallien taulukoista ja sarakkeista (Microsoft 2023).
HTML	HTML (HyperText Markup Language) eli verkkosivujen määrittelykieli kertoo verkkosivujen sisällön merkityksen ja rakenteen (MDN 2024c).
HTTP	HTTP (HyperText Transfer Protocol) on verkkoprotokolla, joka mahdollistaa hypermedia dokumenttien siirtämisen netissä selaimen ja serverin välillä niin, että ihmiset voivat lukea niitä (MDN 2024e).
Jupyter	Jupyter on avoimen lähdekoodin interaktiivinen kehitysympäristö koodille, visualisoinneille, yhtälöille ja narratiiviselle tekstille. Ohjelma toimii verkossa ja se tukee useita ohjelmointikieliä, mutta sitä käytetään pääasiassa Python, R ja Julia kielien kanssa. (Sharma 2024.)
Power BI	Power BI on liiketoimintatiedon analysointiin ja visualisointiin tarkoitettu BI-järjestelmä, joka yhdistää tietoa useasta eri lähteestä (Suramaparna 2024)
Python	Korkeatasoinen ohjelmointikieli, jota voidaan käyttää verkkoharavoinnissa. (Dhanashree 2023).
SQL	SQL (Structured Query Language) on kuvaileva tietokonekieli, jolla päivitetään, haetaan ja lasketaan tietoja taulukkopohjaisista tietokannoista (MDN 2024f).
Tiedon elinkaari	Tiedon elinkaarella tarkoitetaan tiedonhallinnan eri vaiheita datan tuottamisesta tiedon tulkintaan päätöksenteon perusteena. Elinkaari on

	iteratiivinen prosessi ja siinä on kahdeksan eri vaihetta, jotka ovat datan luominen, kerääminen, käsittely, varastointi, hallinta, analysointi, visualisointi ja tulkitseminen. (HBO.)
URL	URL (Uniform Resource Locator) on tekstinpätkä, joka kertoo missä resurssi, kuten verkkosivu, sijaitsee internetissä (MDN 2024d).
Verkkoharavointi	Tiedon automaattista keräämistä ja prosessointia verkkosivuilta, sekä sen tallentamista myöhempää analyysiä varten (Gallagher & Beveridge 2021).
Verkkotunnus	Verkkotunnus on sivuston osoite, jota käytetään, kun halutaan vierailla tietyllä sivustolla. (MDN 2024b)

2. Tietoperusta

Tietoperusta koostuu kahdesta eri osasta. Ensimmäinen osa käsittelee tiedon elinkaarta ja sen kahdeksaa eri vaihetta; datan luomista, keräämistä, käsittelyä, säilytystä, hallintaa, analysointia, visualisointia ja tulkintaa. Tiedon elinkaari muodostaa perustan tutkimuksen toteuttamiselle. Toinen tietoperustan osa kuvailee verkkoharavoinnin prosessia ja työkaluja. Verkkoharavoinnin perusteella on tutkimuksessa toteutettu tiedon elinkaaren datan keräämisen ja osittain käsittelyn vaiheet. Toisessa osiossa tarkastellaan hakurobottien toteuttamiseen liittyviä haasteita, sekä eettisiä ja laillisia ongelmia.

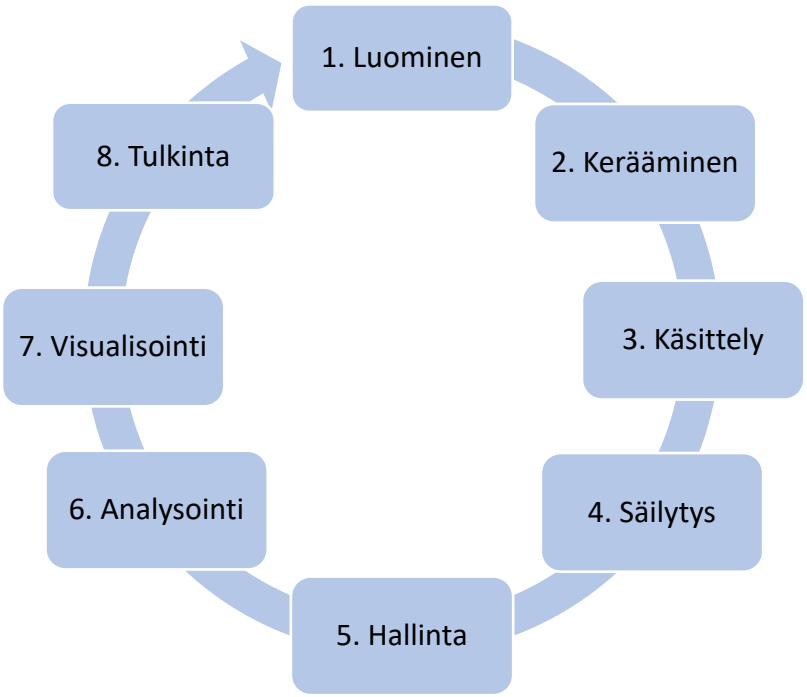
2.1 Data-analytiikka

Data-analytiikka on laaja käsite, joka viittaa prosessiin, jossa raakadataa käsitellään ja analysoidaan niin, että siitä saadaan merkityksellisiä ja hyödyllisiä oivalluksia liiketoimintapäätösten tueksi. Organisaatiot keräävät suuria määriä raakadataa, mutta se ei itsessään merkitse mitään. Data tuleekin käsitellä ja analysoida, jotta sitä voidaan hyödyntää. (Stevens 2023.)

2.2 Tiedon elinkaari

Tiedon elinkaari kuvaa tiedon tuottamisen vaiheet datan syntymisestä sen muuttamiseksi päätöksen teon taustalla oleviksi oivalluksiksi (Stobierski 2021a). Elinkaari korostaa erityisesti datan hallinnan ja käsittelyn tärkeyttä koko elinkaaren aikana niin, että tietoa on helposti saatavilla ja se on luotettavaa silloin, kun sitä tarvitaan. Näin yrityksen resursseja voidaan optimoida ja hyödyntää tehokkaasti. (Kamaly 2022.) Tiedon elinkaari varmistaa, että dataa on saatavilla data-analytiikan tarpeisiin. (HBO 2024).

Harvard Business Schoolin tiedon elinkaarimalli (Kuva 1) koostuu kahdeksasta eri vaiheesta datan luomisesta, keräämisestä, käsittelystä, säilytyksestä, hallinnasta, analysoinnista, visualisoinnista ja tulkinnasta. Prosessia kuvataan sykliksi sillä prosessin viimeinen vaihe luo uutta dataa, joka taas käynnistää uuden dataprojektin. (Stobierski 2021a.)



Kuva 1. Tiedon elinkaari

2.2.1 Datan luominen

Maailmassa syntyy päivittäin käsittämätön määrä dataa yritysten, asiakkaiden tai muiden kolmansien osapuolien toimesta. Dataa syntyy yrityksen kaikista toiminnoista kuten, myynnistä, tilauksista, työntekijöiden palkkaamisesta, kommunikaatiosta ja muusta vuorovaikutuksesta. (Stobierski 2021a.)

2.2.2 Datan kerääminen

Datan suuren määrän takia kaikkea luotua dataa ei pystytä keräämään. Yrityksen tulee määritellä tarkasti, mihin tarkoitukseen dataa kerätään ja mitä sillä pyritään saavuttamaan, jotta kerätty data on merkityksellistä. Jos yritys kerää vain mahdollisimman paljon dataa, voi kerääminen ja käsittely hidastua niin, että informaatio ei ole enää relevanttia tai kaikkea kerättyä dataa ei pystytä analysoimaan. (Stobierski 2021a.) Yrityksen tulee valita luotettavat tietolähteet, jotta data on luotettavaa, eheää ja yhdenmukaista (Javaid 2024). Dataa voidaan kerätä monella eri tavalla kuten kyselyillä, haastatteluilla tai esimerkiksi nettiharavoinnilla, josta kerron lisää työn seuraavassa osiossa (Stobierski 2021a).

2.2.3 Datan käsittely

Datan keräämisen jälkeen tulee se vielä käsitellä, jotta eri lähteistä kerätty raakadata saadaan yhtenäiseen muotoon. Data saattaa vaatia paljonkin käsittelyä, jotta siitä saadaan merkityksellistä

informaatiota. Datasetti tulee validoida luotettavuuden ja virheettömyyden takaamiseksi. Se tehdään tarkastamalla ja korjaamalla virheet, duplikaatit, sekä puutteelliset tiedot. (Urazaliev 2023.) Tämän jälkeen data muunnetaan muotoon, joka voidaan varastoida tehokkaammin ja salataan tiedot yksityisyyden suojaamiseksi (Stobierski 2021a).

2.2.4 Datan varastointi

Datan varastointi on prosessi, jossa tallennetaan kerättyä ja käsiteltyä tietoa siten, että se on helposti saatavilla ja suojattuna, vaikka raakadatan alkuperäinen lähde rikkoutuisi **(**Stobierski 2021a). Dataa voidaan varastoida usealla eri tavalla riippuen datan rakenteesta. Strukturoitua organisoitua dataa, joissa data on jaettuna selkeisiin sarakkeisiin ja riveihin, tallennetaan taulukkoina SQL (Structured Query Language) relaatiotietokantaan. Strukturoitua dataa on helppo järjestää ja yhdistellä eri tietokannoista. Yritykset keräävät myös rakenteetonta dataa, jolla ei ole ennalta määriteltyä tietomallia tai rakennetta, kuten esimerkiksi kuvat, tekstitiedostot tai äänitiedostot. Rakenteetonta dataa on vaikeampi ymmärtää perinteisten ohjelmien avulla ja sitä ei voida tallentaa relaatiotietokantaan. (Big Data Framework 2019.) Markkinoilla on kuitenkin useita muita teknologioita ja työkaluja, jotka eivät vaadi tiukkaa taulurakennetta (Segment 2023).

2.2.5 Datan hallinta

Datan hallinnalla viitataan datan organisointiin, säilytykseen ja palauttamiseen prosessin aikana. Vaikka datan hallinta on mainittuna yhtenä askeleena datan elinkaaressa kattaa se koko prosessin alusta loppuun. Siihen kuuluu muun muassa käyttöoikeuksien hallintaa, dataan tehtyjen muutosten seurantaa ja datan suojausta. (Stobierski 2021a.)

2.2.6 Datan analysointi

Datan analysointi viittaa prosessiin, jossa pyritään luomaan merkityksellisiä oivalluksia kerätystä raakadatasta, sillä data itsessään ei tarkoita mitään, jollei sitä analysoida. Analyysin tarkoituksena on vastata tiettyyn kysymykseen niin, että vastausta voidaan käyttää päätöksenteon perusteena. Analysoinnilla voidaan löytää merkityksellisiä trendejä, riippuvuuksia, vaihteluita ja malleja datasta. Dataa voidaan analysoida monella eri tavalla, riippuen siitä mitä halutaan saavuttaa ja millaista dataa on kerätty. Analyysiä on neljää päätyyppiä: kuvailevaa, diagnostista, ennakoivaa ja ohjaavaa. (Stevens 2023.)

Kuvaileva analyysi vastaa kysymykseen, mitä on tapahtunut. Se tarjoaa yksinkertaisesti tilannekatsauksen menneistä tapahtumista, mutta ei selitystä siitä, miksi niin on tapahtunut. Kuvailevassa analyysissä käytetään kahta päätekniikkaa datan kokoamista (data aggregation) ja datan louhimista (data mining). Datan kokoaminen on nimensä mukaan datan kokoamista ja sen esittämistä

tiivistetyssä muodossa niin, että käyttäjän on helppo saada yleiskatsaus tilanteesta. Datan louhimisessa etsitään datasta malleja ja trendejä, joilla voidaan ennustaa tulevaisuuden tapahtumia. (Stevens 2023.)

Diagnostinen analyysi vastaa kysymykseen, miksi jotain tapahtui. Se tarkastelee kuvailevan analyysin lukuja tarkemmin ja pyrkii tunnistamaan syitä havaittuihin poikkeamiin. Tarkoituksena on hakea lisää dataa lukujen pohjalle, jotta voidaan ymmärtää miksi tietyt mittarit ovat suoriutuneet erityisen hyvin tai huonosti. Näin voidaan ymmärtää, mitkä toimet ovat vauhdittaneet yrityksen liiketoimintaa ja mitkä haitanneet. (Stevens 2023.)

Ennakoiva analyysi vastaa kysymykseen, mitä todennäköisesti tapahtuu tulevaisuudessa. Se hyödyntää koneoppimista ennustaakseen historiallisten mallien ja trendien avulla tulevien tapahtumien todennäköisyyksiä. Analyysi tunnistaa eri muuttujien välisiä syy-seuraussuhteita, joita voidaan käyttää päätöksenteon tukena. (Stevens 2023.) Sen avulla voidaan muun muassa tunnistaa, että kesäkuukausina myydään usein enemmän jäätelöä, kuin loppuvuonna ja sopeuttaa liiketoimintaa tämän mukaan.

Ohjaileva analyysi vastaa kysymykseen, mikä on paras toimintatapa jatkossa. Sen tarkoituksena on luoda malli, joka laskee edellisten analyysien pohjalta eri ennusteiden todennäköisyyksiä. Se pyrkii ennustamaan ja luokittelemaan usean eri muuttujan muodostamia lopputuloksia niin, että voidaan ennustaa eri toimenpiteiden vaikutuksia liiketoimintaan ja näin ohjata yrityksen toimintaa. (Stevens 2023.)

2.2.7 Datan visualisointi

Datan visualisointi on datan esittämistä graafisessa muodossa, kuten kaavioilla, kuvaajilla, infografiikoilla tai jopa animaatioilla. Nämä visuaaliset näkymät esittävät monimutkaisia muuttujien välisiä suhteita yksinkertaisemmassa muodossa, jotta datasta on helpompi tehdä johtopäätöksiä ja käyttää päätöksenteossa. (IBM.)

Yritykset keräävät dataa huomattavia määriä päivässä ja visualisoinnin ongelmaksi muodostuu, miten valita olennaiset tiedot halutun johtopäätöksen välittämiseksi. Visualisointiin liittyykin tiettyjä parhaita käytäntöjä tukemaan selkeää ja hyödyllistä kommunikointia. (IBM.)

Ensimmäisenä käytänteenä on kontekstin asettaminen visualisoinnin käyttäjille. Yleisölle olisi tärkeää tarjota taustatietoa siitä, miksi tietty mittari on tärkeä. Ilman kontekstia luku ei itsessään kerro mitään vaan sitä tulisi pystyä vertaamaan johonkin konkreettiseen tietoon kuten esimerkiksi tavoitteisiin, alan keskiarvoon tai muihin keskeisiin mittareihin. (IBM.)

Toisena käytäntönä on yleisön tunteminen. Visualisointia tehdessä tulisi ymmärtää, mitä sen käyttäjät haluavat visualisoinnilla saavuttaa ja suunnitella näkymät sen mukaan (IBM). Näkymiä suunnitellessa tulee ottaa huomioon käyttäjien taustat ja kokemus ja valita käytettävät mittarit ja kaaviot sen perusteella. Esimerkiksi, jos käyttäjillä on kokemusta datan analysoinnista, voi heille laatia monimutkaisempia interaktiivisia näkymiä, joissa dataa voi suodattaa ja säätää. Jos taas käyttäjillä ei ole kokemusta datan käsittelystä, voi heille laatia yksinkertaisemman näkymän, joka esittää halutun lopputuloksen ilman lisäanalysoinnin mahdollisuutta. (Evolytics 2024.)

Kolmantena käytäntönä on tehokkaan visualisoinnin valitseminen johtopäätöksen kommunikoimiseen. Tietyt visualisoinnit soveltuvat parhaiten tietynlaisen datan esittämiseen, esimerkiksi hajontakuviot sopivat kahden muuttujan välisen suhteen osoittamiseen ja pylväsdiagrammi taas sopii hyvin aikasarjojen seuraamiseen. Vääränlainen visualisointi voi merkittävästi heikentää johtopäätösten kommunikointia. (IBM.)

Viimeisenä käytäntönä on pitää visualisointi yksinkertaisena, jotta käyttäjän huomio kiinnittyy halutun viestin välittämiseen. Visualisoinnissa tulisi käyttää selkeitä ja yksinkertaisia grafiikoita sekä miettiä tarkasti, mitä lisätietoja tarvitaan halutun johtopäätöksen kommunikoinnissa ja poistaa tiedot, jotka eivät tue sitä. Visualisointia tehtäessä tulisi pohtia tarvitaanko kaikkiin kaavion datapisteisiin otsikot vai riittääkö tietyn pisteen painottaminen. Näkymien ja mittareiden värit olisi myös hyvä pitää yksinkertaisena, sillä liiallinen värien käyttö voi vaikeuttaa visualisoinnin lukemista. Värejä tulisi käyttää harkiten korostamaan johtopäätöksen kannalta tärkeitä lukuja. (IBM.)

Datan visualisointiin on tarjolla useita eri työkaluja. Yksinkertaisimmillaan ohjelmaan syötetään itse data, jota sen jälkeen visualisoidaan (Stobierski 2021b). Yksi yleisimpiä ja yksinkertaisimpia tapoja visualisointiin on Microsoft Excel (Miller 2021). Se ei itsessään ole visualisoinnin ohjelmisto vaan laskentataulukko-ohjelma, jossa voidaan analysoida dataa ja luoda erilaisia kaavioita ja graafeja valmiista malleista. Se soveltuukin enemmän yksittäisten kuvaajien laatimiseen yksinkertaisista dataseteistä. (Stobierski 2021b.)

Monimutkaisempien kokonaisuuksien ja syy-seuraussuhteiden esittämiseen voidaan käyttää visualisointiin tarkoitettuja Business Intelligence (BI) -järjestelmiä, joihin voidaan hakea dataa useasta eri lähteestä. BI-järjestelmissä luodaan erilaisia näkymiä tiettyjen ongelmien ratkomiseksi tai kokonaisuuksien näyttämiseksi. Järjestelmissä on sisään rakennettuna erilaisia visualisointeja, joita on helppo muokata ja lisätä näkymiin tarpeen mukaan. BI-järjestelmien etuna on interaktiiviset näkymät ja kaaviot, joissa voidaan tarkastella dataa mittareiden takana ja näin syvällisemmin ymmärtää, mitä mittarit liiketoiminnasta kertovat. Toisin kuin Excel-raporttien kanssa, BI-järjestelmät näyttävät automaattisesti käyttäjille ajantasaista dataa ja niiden kanssa ei tarvitse huolehtia versionhallinnasta. Excel-raportit lähetetään usein sähköpostitse ja ne näyttävät enemmänkin tietyn valmiiksi

määritellyn hetken tilanteen. (Suramaparna 2024.) Suosittuja BI-järjestelmiä on muun muassa Microsoft Power BI, Tableau ja QlikSense.

Monet BI-järjestelmät käyttävät tekoälyä ja koneoppimista visualisoinnin ja analysoinnin tehostamisessa ja automatisoinnissa. Algoritmit tunnistavat automaattisesti malleja ja yhteyksiä monimutkaisista dataseteistä ja luo niiden perusteella ennusteita tulevista tapahtumista. Tekoäly havaitsee poikkeamia dataseteissä ja tarkastelee tarkemmin miksi nämä poikkeamat ovat syntyneet. (Insightsoftware 2023.) Tekoälyn avulla voidaan parantaa visualisointien interaktiivisuutta niin, että käyttäjä voi tutkia lukuja dynaamisesti. Järjestelmät voivat myös tarjota automaattisia ehdotuksia datan visualisoimiseksi tunnistamalla datan erityispiirteitä ja suhteita, sekä ehdottamalla sopivia näkymiä tai kaavioita niiden esittämiseksi. (Czaban 2023.) Tekoälyn avulla voidaan ymmärtää dataa syvällisemmin.

2.2.8 Datan tulkinta

Prosessin viimeisessä vaiheessa analysoitua ja visualisoitua dataa tarkastellaan oman ammattitaidon ja ymmärryksen kautta johtopäätöksen saavuttamiseksi. Tarkoituksena on ymmärtää mitä data meille kertoo ja mitä vaikutuksia sillä on. (Stobierski 2021a.)

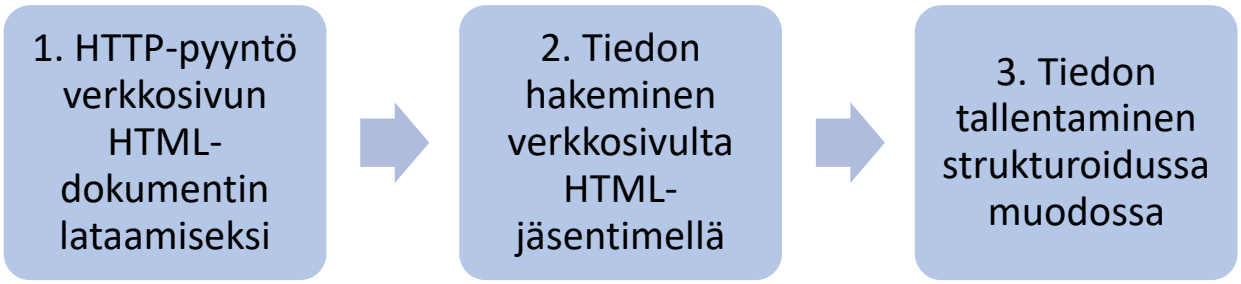
2.3 Verkkoharavointi

Verkkoharavointi (web scraping) on tiedon keräämistä ja prosessointia verkkosivuilta, sekä sen tallentamista myöhempää analyysiä varten. Sen avulla voidaan kerätä useasta eri lähteestä suuria määriä dataa kuten tekstiä, digitaalista sisältöä, käyttäjäprofiileja, sijaintitietoja ja muuta metadataa. Verkkoharavoinnin automaation aste vaihtelee manuaalisesta datan kopioinnista ja liittämisestä ohjelmointikielellä luotuun automaattiseen tiedonkeräykseen. (Gallagher & Beveridge 2021.)

Verkkoharavointi on nopea ja edullinen tapa kerätä ajantasaista tietoa yrityksen tarpeisiin. Dataa voidaan kerätä moniin eri käyttötarpeisiin, joista yksi yleisimpiä on hintavertailu, jota käytämme myös tässä työssä. Hintavertailussa tuotteiden hintoja haetaan eri sivustoilta parhaiden kauppojen löytämiseksi. Muita käyttötapoja on muun muassa markkinoiden ja trendien seuraaminen liiketoimintaympäristön muutosten havaitsemiseksi, sekä kilpailija-analyysi. Hakemalla tietoja kilpailijoiden tuotteista, hinnoista, kampanjoista ja asiakaspalautteista voidaan saavuttaa strategista etua omien liiketoimien suunnittelussa. Verkosta voidaan kerätä myös omien tuotteiden arvosteluja kuluttajien mielipiteiden kartoittamiseksi tai tehostaa myyntiä hakemalla nettisivuilta mahdollisten asiakkaiden yhteystietoja. (Besinsky 2024.) Verkkoharavoinnilla voidaan kerätä dataa koneoppimisen mallien kouluttamiseksi (Ermakovich 2023).

2.3.1 Verkkoharavoinnin prosessi

Verkkoharavoinnin prosessi koostuu kolmesta eri vaiheesta, jotka ovat tiedon saatavuus, datan kerääminen verkkosivulta ja tiedon tallentaminen (Kuva 2).



Kuva 2. Verkkoharavoinnin prosessi

Ensimmäisellä vaiheella eli tiedon saatavuudella viitataan pääsyyn verkkosivuille, josta data halutaan hakea. (Chasin et. Al 2018 p. 964.) Verkkoharavoinnin työkalu toteuttaa tämän laatimalla Hypertext Transfer Protocol (HTTP) -pyynnön kohdesivuston HTML-muotoisen sisällön saamiseksi. (Callagher & Beveridge 2021).

Toisessa vaiheessa verkkosivulta jäsennetään HTML-sisältö jäsentimellä, joka tulkkaa HTML-koodin ja luo siitä rakenteen, jota on helppo käsitellä. Sen avulla voidaan erotella erilaiset elementit toisistaan, kuten otsikot, tekstit tai linkit. Tämän jälkeen halutut elementit poimitaan HTML-rakenteesta. Kun halutut tiedot on haettu, verkkoharavoinnin ohjelma käsittelee datan ja tallentaa sen strukturoidussa muodossa CSV- tai XLS-tiedostona haluttuun tietokantaan. (ProWebScaper 2024).

2.3.2 Verkkoharavoinnin työkaluja ja kirjastoja

Verkkoharavointiin on tarjolla useita erilaisia työkaluja ja kirjastoja, kuten mukautetut skriptit, selainlaajennukset, työpöytäsovellukset ja pilvipohjaiset palvelut (Besinsky 2024).

Mukautetut skriptit ovat kehittäjien koodaamia ohjelmia, jotka hakevat ennalta määriteltyä dataa tietyiltä sivustoilta (Besinsky 2024). Niiden ohjelmoinnissa voidaan hyödyntää erilaisia ohjelmointikieliä, joista Python on yksi yleisin, sen yksinkertaisuuden, monipuolisuuden ja runsaiden verkkoharavointiin tarkoitettujen kirjastojen ansiosta. (Dhanashree 2023.) Tämän vuoksi Pythonia käytetään, myös tässä opinnäytetyössä.

Pythonin suosittuja kirjastoja verkkoharavoinnissa on muun muassa Selenium, Scrapy, Requests, Beautiful Soup, Lxml ja Pandas.

Selenium

Selenium on alkujaan tarkoitettu verkkosivujen testaamiseen, mutta sitä käytetään laajasti myös verkkoharavointiin. Kirjastossa on tyypillisiä verkkoharavoinnin kirjastoja laajemmat ominaisuudet ja toiminnallisuus. Selenium webdriverin avulla voidaan hallita verkkoselainta ja simuloida ihmisten käyttäytymistä nettisivuilla. Kirjaston avulla voidaan automatisoida toimintoja, kuten linkkien avausta, lomakkeiden täyttämistä ja muiden painikkeiden klikkausta. Selenium soveltuu erityisen hyvin dynaamisesti ladattavan sisällön ja JavaScriptin käsittelyyn. (Skakun 2024a.)

Scrapy

Scrapy on avoimen lähdekoodin verkkoharavoinnin viitekehys, jolla voidaan suorittaa kaikki verkkoharavoinnissa tarvittavat komennot. Sen avulla voidaan luoda HTTP-pyynnöt, hakea tiedot ja jäsentää ne, sekä tallentaa tiedot JSON-, CSV- tai XML-muodossa (Skakun 2024b.) Scrapy tukee CSS- ja XPath valitsimia, joiden avulla voidaan tehokkaammin paikantaa ja eristää tietoja verkkosivujen HTML-rakenteesta. Scrapya voidaan käyttää verkkosivujen automaattiseen testaamiseen ja tiedon louhintaan. (Daivi 2024.) Scrapylla voidaan luoda useita eri hakuohjelmia yhden projektin sisällä. Tämän vuoksi se sopii erityisesti suuriin ja skaalautuviin projekteihin ja onkin melko vaikeakäyttöinen aloittelijoille. (Skakun 2024b.)

Requests

Requests kirjasto on tarkoitettu HTTP-pyyntöjen luomiseen ja vastausten käsittelyyn Pythonilla. Pyyntöjen luominen on verkkoharavoinnin ensimmäinen vaihe. Requests tarjoaa helpon tavan kommunikoida suojattujen verkkosivujen kanssa. (Ronquillo 2024.)

BeautifulSoup ja LXML

BeautifulSoup on suosituin Python kirjasto verkkoharavointiin, sillä se sopii aloittelijoille koodin yksinkertaisuuden ja mukautuvuuden ansiosta. Se jäsentää HTML- ja XML- dokumentit puurakenteeseen, josta voidaan helposti etsiä ja eristää dataa, joka on verkkoharavoinnin toinen vaihe. Sen avulla voidaan etsiä ja tunnistaa sivuston rakenteesta erilaisia elementtejä, kuten otsikoita tai linkkejä, sekä poimia tekstiä HTML-tageista. (Educative 2024.) BeutifulSoup käyttää oletuksena Pythonin standardikirjaston HTML-jäsennintä, joka soveltuu virheettömien verkkosivujen parsimiseen. Virheellisten verkkosivujen parsimiseen soveltuu paremmin ulkoiset erikseen ladattavat jäsentimet, kuten Lxml. (Ramirez 2023.) Lxml on erittäin nopea HTML- ja XML-jäsennin, joka soveltuu rikkinäisten verkkosivujen parsimiseen (LXML 2024).

Pandas

Pandas kirjasto on suosittu datankäsittely ja analysoinnin työkalu, jota käytetään datan putsaamiseen ja taulukointiin. Se mahdollistaa muun muassa eri datasettien yhdistämisen, puuttuvien tietojen käsittelyn ja tietojen kirjoittamisen eri muodoissa, kuten CSV-, Microsoft Excel-, SQL- tai tekstitiedostona. (Pandas 2024.) Se mahdollistaa siis verkkoharavoinnin viimeisen vaiheen eli tietojen tallentamisen strukturoidussa muodossa. Pandas kirjaston avulla voidaan haravoida HTML-taulukoita verkkosivuilta (Ksn-developer 2023).

Erilaisten kirjastojen lisäksi verkkoharavointiin voidaan käyttää selainlaajennuksia eli selaimeen ladattavia ohjelmia. Laajennusta voidaan käyttää samalla, kun sivua selataan klikkaamalla sivuston elementtejä, joita halutaan haravoida. Verkkoharavointiin on olemassa myös erilaisia työpöydälle ladattavia ohjelmia, jotka hakevat sivustot ja halutut elementit sovelluksen kautta. Sovelluksissa on valmiina helposti käytettäviä toiminnallisuuksia, jotka mahdollistavat haun ilman ohjelmointia. (Besinsky 2024.)

Yritykset voivat ostaa verkkoharavointia myös pilvipohjaisena palveluna, jolloin datan haku voidaan ulkoistaa yritykselle, jolla on jo valmiiksi siihen tarvittava infrastruktuuri, kuten lisenssit ja henkilöstö. Tällöin yritys maksaa verkkoharavoinnista vain käytön mukaan ja data tulee valmiina toimittajalta. (Dilmegani 2024.)

2.3.3 Verkkoharavoinnin haasteet

Vaikka verkkoharavoinnin prosessi on yksinkertainen, liittyy siihen erilaisia haasteita. Yksi suurimmista ongelmista koskee verkkosivujen jatkuvasti muuttuvia rakenteita ylläpitäjien päivittäessä sivuja ja lisätessä sinne uusia tietoja. Hakurobotti etsii haettavan datan verkkosivujen elementtien perusteella ja, jos näitä muutetaan, tulee myös hakuohjelmaa muuttaa ja haettavat elementit määritellä uudestaan. Pienikin muutos verkkosivuilla voi kaataa hakurobotin tai tuoda virheellistä dataa. Tämän vuoksi hakuohjelmaa tulee päivittää säännöllisesti. (ProWebScraper 2024.)

Verkkosivujen muuttuvien rakenteiden vuoksi haetun datan laatuun voi liittyä ongelmia. Verkkosivuilta haettua dataa ei voida käyttää, jos datassa on virheitä tai puutteita. Päätöksenteon tueksi tarvitaan korkealaatuista dataa, jotta ei tehdä vääriä päätöksiä. Verkkoharavoinnissa voidaan hakea tietoa laajalla skaalalla eri verkkosivuilta ja, jos yhdessäkin datasetissä on virhe, saattaa se antaa väärän kuvan päätöksentekijälle. (Karatas 2024.)

Verkkosivujen rakenteellisia ongelmia voidaan kiertää käyttämällä hyväksi tekoälyn algoritmeja, jotka mukautuvat automaattisesti nettisivujen päivityksiin ja muutoksiin. Datan keräämiseen on nykyään tarjolla useita tekoälyä hyödyntäviä ohjelmistoja, jotka eivät vaadi teknistä osaamista.

Valmiissa ohjelmassa käyttäjä voi opettaa järjestelmän hakemaan haluamansa tiedot nettisivulta, jonka jälkeen ohjelmisto luo algoritmin, joka hakee samantyyliset tiedot eri nettisivuilta. (Karatas 2024.)

Muita verkkoharavointiin liittyviä ongelmia on verkkosivujen haravoinnin estävät tekniikat ja mekanismit. Verkkosivujen estäessä hakurobottien toiminnan ei hakuohjelmien käyttö ole enää eettisesti oikein. Eettisestä verkkoharavoinnista kerrotaan lisää seuraavassa kappaleessa.

2.3.4 Verkkoharavoinnin lailliset ja eettiset ongelmat

Verkkoharavointi itsessään ei ole laitonta, eikä sen käyttämiseen ole omaa lainsäädäntöä. Verkkoharavointia ohjaa enemmänkin useat toisiinsa liittyvät oikeudelliset säädökset ja kysymykset, kuten tekijänoikeudet, tietosuoja ja nettisivujen käyttöehdot. (Szwed 2021.) Kerättyä dataa ei saa käyttää laittomiin toimiin, esimerkiksi myymällä maksumuurin takana olevaa dataa eteenpäin omalla alustalla (Krotov & Silva 2018).

Verkkosivujen ja tietokantojen sisältö saattaa olla tekijänoikeussuojan alaisena, jolloin niitä ei saa kopioida tai käyttää ilman lupaa. Tekijänoikeussuojan alaisena voi olla yksittäisiä kuvia, logoja tai tekstin pätkiä. Nettisivujen ohjelmointikoodi tai sen graafinen ilme taas on harvemmin suojattuja. Vaikka verkkosivujen yksittäiset sivut eivät yltäisi teostason suojaan, voi sivusto kokonaisuudessaan muodostaa tietokannan ja näin olla tekijänsuojan alaisena. (Korpela 2013.) Verkkoharavoinnin työkalua rakentaessa tuleekin olla tarkkana, että ei kerätä ja jaeta tekijänoikeussuojan alaista materiaalia.

Euroopan Unionin yleinen tietosuoja-asetus (GDPR) suojaa EU:n kansalaisten henkilötietoja ja asettaa rajoituksia sille, mitä tietoja yritykset saavat kerätä, säilyttää ja hallinnoida. Henkilötietoja ovat tiedot, joiden perusteella henkilö voidaan tunnistaa suoraan tai välillisesti yhdistämällä yksittäisiä tietoja. Näitä on muun muassa nimi, osoite, IP-osoite, terveystiedot tai sukupuolinen suuntautuminen. Yritys ei voi vapaasti kerätä tietoa, joita käyttäjät jakavat esimerkiksi LinkedIn- tai Facebook sivustoilla, vaan sen tulisi kysyä käyttäjiltä erikseen lupa tietojen keräämiseen ja rekisterin muodostamiseen. (EU 2022.)

Nettisivujen käyttöehdot saattavat kieltää koneellisen keräämisen, jolloin sääntöjen rikkominen johtaa sopimusrikkomukseen. Tämä kiellon laillisuus on kuitenkin epäselvää sillä, verkkosivujen käyttäjän on ensin hyväksyttävä sivuston käyttöehdot esimerkiksi klikkaamalla valintaruutua verkkosivuilla. Tietoa koneellisesti haravoitaessa ei ole mahdollista tutustua ja hyväksyä kaikkien verkkosivujen käyttöehtoja, lisäksi on epäselvää voiko kone hyväksyä sääntöjä ihmisten puolesta. Verkkosivujen omistajan tulisikin huolehtia siitä, että haravointikielto on kommunikoitu niin, että tietokone ohjelmisto pystyy sen lukemaan. (Szwed 2021.)

Haravointikielto voidaan toteuttaa verkkosivujen robots.txt tiedostolla, jossa määritellään, mitkä sivujen tiedot saa indeksoida tai käydä läpi. Robos.txt tiedosto sijaitsee verkkotunnuksen päähakemistossa ja löytyy lisäämällä verkkosivun loppuun robots.txt. Siinä käytetään kahta eri muuttujaa määrittelemään asetukset: User-agent ja Disallow tai Allow. Ensimmäisessä User-agent kohdassa määritellään mitä hakukoneita tai robotteja ohje koskee ja toisessa kohdassa määritellään, mitkä verkkosivujen polut ovat sallittuja tai kiellettyjä haravoida. (Google 2024.)

Lainsäädäntö ei ole pysynyt mukana teknologian ja varsinkin tekoälyn nopeassa kehityksessä ja verkkoharavoinnin laillisuus onkin vaikeasti määriteltävissä. Tämän vuoksi on tärkeää pohtia verkkoharavoinnin eettisiä ongelmia. Eettisestä näkökulmasta verkkoharavoinnissa tulisi ainakin pohtia aiheutuuko haravoinnista haittaa sen kohteena oleville nettisivuille. Verkkoharavoinnin tarkoitus on kerätä ajantasaista dataa, jolloin liian tiheä datan päivitys saattaa aiheuttaa ongelmia sivuille, josta tietoja pyydetään. Päivityspyyntöjen tiheyttä ja pyyntöjen määrää olisikin hyvä rajoittaa, jotta sivustojen toiminta ei häiriinny. Monet nettisivut tarjoavat pääsyä dataan API-rajapinnan kautta, jota olisi hyvä suosia haravoinnin sijaan. API-kuormittaa verkkosivuja vähemmän, kuin koneellinen haravointi, mutta saattaa sisältää vanhaa dataa. (Jurgenstojku 2020.)

Kilpailijoiden hintojen koneellinen hakeminen on yksi yleisimpiä nettiharavoinnin kohteita. Tähän liittyy kuitenkin omat eettiset ongelmat reilusta kilpailusta. Kerättyä dataa käytetään dynaamisessa hinnoittelussa, joka voi johtaa kilpailun vastaisiin käytäntöihin tai epäreilujen etujen saavuttamiseen. (Bhagyashree 2023.)

3. Analysointityökalun toteutus

Tutkimuksen tavoitteena on luoda kuluttajalle hinta-analysointityökalu verkkoharavoinnin pohjalta, käyttäen Harvard Business Schoolin tiedon elinkaari viitekehystä. Työssä käydään läpi kaikki elinkaaren vaiheet datan luomisesta johtopäätösten tekoon. Datan kerääminen ja osa käsittelystä toteutetaan verkkoharavoinnin ohjelmistoprojektilla ja datan analysointi, sekä visualisointi tehdään Microsoft Power BI ohjelmalla. Työn lopputulokset eli Jupyterissa Pythonilla luotu verkkoharavoinnin robotti, koodin hakema CSV-tiedosto, sekä Power BI:llä luotu analysoinnin ja visualisoinnin näkymät jaetaan GitHubissa https://github.com/johranta/oppari_analysointi/tree/main.

3.1 Tutkimuksen taustat

Hinta-analysoinnin työkalu tehdään sattumanvaraisesti valitun suomalaisen kosmetiikka-alan yrityksen Lumenen näkökulmasta. Lumenen valikoimaan kuuluu ihon- ja hiustenhoidon tuotteita, sekä kosmetiikkaa. Tuotteista suurin osa valmistetaan Suomessa. Lumene on Pohjoismaiden johtava kauneuden kiertotalouden toimija, joka on sitoutunut ilmastotyöhön ja pyrkii olemaan hiilineutraali vuoteen 2050 mennessä. (Lumene 2024.) Yrityksen tuotteita myydään kaikissa Suomen suurimmissa tavarataloissa ja sillä on oma verkkokauppa. Työkalussa verrataan Lumenen tuotteiden hintaa eri verkkokaupoissa, jotta kuluttajan on helpointa löytää edullisin sivusto tuotteiden tilaamiseen. Edullisuutta mitataan tuoteryhmän keskihinnoilla ja halvimpien tuotteiden lukumäärällä, tarkastelussa on myös valikoiman laajuus.

Lumenella on satoja eri tuotteita, joten tuotemäärien rajaamiseksi valittiin yksi tuoteryhmä eli Lumene Nordic Bloom [Lumo] sarja, jossa on 14 erillistä tuotenimeä. Tiedot haetaan kolmen eri vähittäiskaupan nettisivuilta, jotka ovat sallineet verkkoharavoinnin sivuiltaan.

3.2 Datan luominen

Analysoinnissa käytetään Lumene Nordic Bloom [Lumo] sarjan 14 erillisen tuotteen tuotenimeä ja hintatietoja. Data haetaan kolmelta eri verkkosivulta, jotka ovat luoneet tiedot lisäämällä ne verkkosivuilleen.

Elinkaaren ensimmäisessä vaiheessa tarkistetaan, mitkä sivut sallivat verkkoharavoinnin ja valitaan yritykset ja niiden luomat tiedot sen perusteella eettisen verkkoharavoinnin takaamiseksi. Tarkastaminen tapahtuu hakemalla verkkosivujen robots.txt tiedostot. Projektin ensimmäisessä vaiheessa käydään läpi sivustoja, joissa myydään Lumenen tuotteita. Robots.txt tiedostot saattavat olla pitkiä ja verkkosivut ovat voineet kieltäneet verkkoharavoinnin haku funktion kautta, jota joudutaan käyttämään haettaessa tiettyjä tuotteita. Ensimmäiseksi luodaan koodi, joka etsii, esiintyykö

verkkosivujen robots.txt Disallow kohdassa ”search” tai ”haku” ja luodaan tiedoista taulukko (Kuva 3).

```
Search Allowance Table:
{'Site': 'https://www.stockmann.com', 'Search Allowed': False}
{'Site': 'https://www.k-citymarket.fi', 'Search Allowed': False}
{'Site': 'https://www.sokos.fi', 'Search Allowed': True}
{'Site': 'https://www.lumene.com', 'Search Allowed': False}
{'Site': 'https://www.lyko.com', 'Search Allowed': True}
{'Site': 'https://www.tokmanni.fi', 'Search Allowed': False}
{'Site': 'https://www.prisma.fi', 'Search Allowed': True}
{'Site': 'https://www.kicks.fi', 'Search Allowed': False}
{'Site': 'https://www.cocopanda.fi', 'Search Allowed': False}
{'Site': 'https://www.nordicfeel.fi', 'Search Allowed': False}
```

Kuva 3. Robots.txt taulukko verkkosivuista, jotka kieltävät tai sallivat verkkoharavoinnin

Suurin osa yrityksistä, joita opinnäytetyössä oli tarkoituksena käyttää, on kieltänyt verkkoharavoinnin tuotehaun kautta. Ainoastaan Sokos, Prisma ja Lyko ovat taulukon mukaan sallineet verkkoharavoinnin. Ennen hakurobotin tekoa tulee vielä varmistaa, että verkkoharavointi näiltä sivuilta on tosiaan sallittua ja sitä ei ole kielletty muilla ehdoilla. Tämä onnistuu tulostamalla kyseisten yritysten verkkosivujen robots.txt tiedot (Kuva 4).

```
Robots.txt content for https://www.sokos.fi:
User-agent: *
Allow: /


Robots.txt content for https://www.lyko.com:
User-agent: *
Sitemap: https://lyko.com/ext/sitemapindex.xml
Disallow: /*.htm$

Robots.txt content for https://www.prisma.fi:
# *
User-agent: *
Allow: /
```

Kuva 4. Verkkoharavoinnin sallivien verkkosivujen robots.txt tiedostot

Sokos ja Prisma ovat sallineet kaikki hakurobotit ja verkkosivujen sisällöt indeksoinnille. Lyko on kieltänyt htm loppuisten verkkosivujen verkkoharavoinnin, jota haravoinnissa käytetty hakusivusto ei sisällä. Analyysityökalun rakentamisen näkökulmasta kolmelta eri verkkosivulta saadaan riittävästi tietoja työkalun rakentamiseksi. Kuluttajalle työkalusta olisi enemmän hyötyä, jos useampi kauppa olisi sallinut verkkoharavoinnin käyttöehdoissaan, mutta eettisen haravoinnin toteuttamiseksi tätä ei voida tehdä. Robots.txt tiedostot voivat myös muuttua, joten ennen haravoinnin aloittamista ne tulee tarkistaa uudestaan.

3.3 Esivalmistelut verkkoharavointiohjelman luomiseksi

Opinnäytetyön verkkoharavoinnin työkalu laaditaan Python ohjelmointikielellä, sen yksinkertaisuuden, monipuolisuuden ja runsaiden verkkoharavointiin tarkoitettujen kirjastojen vuoksi. Ennen verkkoharavointiohjelman luomista tulee varmistaa, että tietokoneelle on ladattu viimeisin Python versio virallisilta nettisivuilta. Tämän jälkeen tarvitaan ohjelmointiympäristö koodin luomista varten. Tässä työssä käytetään JupyterLab kehitysympäristöä, joka on ilmainen ohjelmoinnin ja laskennan työkalu. JupyterLab ladataan terminal komennon <pip3 install jupyterlab> komennolla.

Ohjelmistojen lataamisen jälkeen tarvitaan vielä ylimääräiset Python kirjastot Beautiful Soup, requests, lxml ja pandas, jotka ladataan terminal komennolla <pip3 install requests pandas beautifulsoup4 lxml>.

3.4 Datan kerääminen verkkoharavoinnin ohjelmalla

Datan kerääminen verkkoharavoinnilla on melko yksinkertaista. Jupyteriin ladataan ensin ylimääräiset asennetut Python kirjastot <import> komennolla (Kuva 5).

```
import pandas as pd
from bs4 import BeautifulSoup
import requests
import lxml
```

Kuva 5. Python import komennot

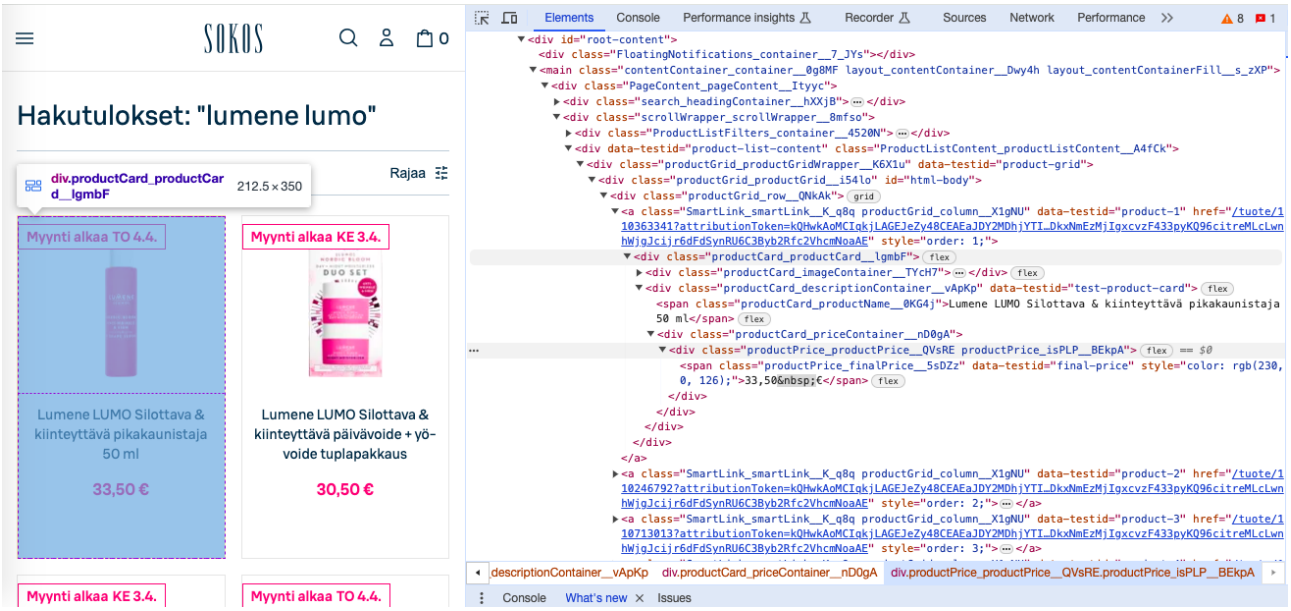
Tämän jälkeen määritellään verkkosivu, jolta tiedot halutaan hakea, lähetetään HTTP-pyyntö ja palautetaan verkkosivun HTML-tiedot. Ennen tietojen käsittelyä tarkistetaan, että pyyntö on onnistunut ja verkkosivu on palauttanut halutut tiedot. Tarkistuksen jälkeen jäsennetään haettu HTML-dokumentti. Oheisessa kuvassa esiteltynä verkkoharavoinnin ensimmäiset vaiheet (Kuva 6).

```
#Määritellään verkkosivut, joista tieto halutaan hakea
url1 = "https://www.sokos.fi/haku?q=lumene%20lumo"
#Lähetetään HTTP-pyyntö verkkosivuille ja palautetaan HTML-tiedot
response1 = requests.get(url1)
# Tarkistetaan, että pyyntö onnistui
if response1.status_code == 200:
    #Jos pyyntö onnistuu parsitaan haetut HTML-tiedot BeautifulSoup-kirjastolla ja lxml-jäsentimellä
    soup1 = BeautifulSoup(response1.content, 'lxml')
```

Kuva 6. Python HTTP-pyynnön laatiminen ja HTML-tietojen jäsentäminen

HTML-tietojen hakemisen ja jäsentelyn jälkeen valitaan elementit, jotka sivulta tarvitaan. Tässä työssä elementit haetaan tarkastelemalla sivustojen HTML-rakenteita selaimen kehitystyökalulla, sillä tarvittavien elementtien luokat tai nimet eivät ole tiedossa. Klikkaamalla eri elementtejä

kehitystyökalussa, korostuvat ne samalla nettisivuilla. Näin voidaan helposti etsiä, missä tarvittavat tiedot sijaitsevat, sekä tunnisteet, joilla niitä voidaan hakea. Oheisessa kuvassa on esitelty selaimen kehitystyökalun toiminnallisuutta (Kuva 7).



Kuva 7. Elementtien tunnistaminen HTML-rakenteista selaimen kehitystyökalulla

Työkalua varten tarvitaan tuotteen nimi ja hinta, sekä ylätason elementti, jonka sisällä nämä tiedot ovat. Koska sivustot ovat dynaamisia, saattavat elementit muuttua sivustoa päivittäessä. Näin kävi muun muassa Sokoksen verkkosivuille 3+1 tarjouspäivien aikana, jolloin tuotteiden normaalihinnat ja 30pv alimmat hinnat hävisivät ja tilalle tuli s-etukortilla saatavat tarjoushinnat. Jos sivustoilla ei olisi vierailtu uudestaan, tuotteen halvin s-etukortilla saatava hinta olisi jäänyt huomaamatta.

Tunnisteiden hakemisen jälkeen voidaan luoda koodi, jonka avulla käydään läpi kaikki sivuston tuotekortit ja haetaan tuotteen nimi ja hinnat tuotekorteista (Kuva 8).

```
# Haetaan kaikki tuotekortit, jotka sisältävät tarvittavia teitoja
product_cards = soup1.find_all('div', class_='productCard_productCard__lgmbF')
#Käydään läpi kaikki tuotekortit ja haetaan tarvittavat tiedot niistä
for card in product_cards:
    # Haetaan tuotteen nimi
    product_name = card.find('span', class_='productCard_productName__0KG4j').text.strip()
    # Tuotteen tämän hetkinen hinta
    final_price = card.find('span', class_='productPrice_finalPrice__5sDZz').text.strip()
     # S-etukortti alin hinta, jos löytyy
    s_card_price_tag = card.find('span', class_='productPrice_coopPrice__QzJ2h')
    s_card_price = s_card_price_tag.text.strip() if s_card_price_tag else "N/A"
    #Tulostetaan haetut tiedot ja lisätään tyhjä rivi loppuun
    print("Tuote:", product_name)
    print("Hinta:", final_price)
    print("S-etu hinta:", s_card_price)
    print()
```

Kuva 8. Tuotetietojen haku verkkosivuilta Python ohjelmointikoodilla

Vastaukseksi saamme listan sen hetkisistä tuotteista ja hinnoista (Kuva 9). Tuotetiedot ja hinnat saattavat vaihdella päivän mukaan.

```
Tuote: Lumene LUMO Silottava & kiinteyttävä pikakaunistaja 50 ml
Hinta: 33,50 €
S-etu hinta: N/A

Tuote: Lumene LUMO Pikakaunistaja + kollageeniseerumi tuplapakkaus
Hinta: 33,50 €
S-etu hinta: N/A

Tuote: Lumene LUMO Silottava & kiinteyttävä päivävoide + yövoide tuplapakkaus
Hinta: 30,50 €
S-etu hinta: N/A

Tuote: Lumene LUMO VITALITY Silottava & elvyttävä päivävoide + yövoide tuplapakkaus
Hinta: 34,50 €
S-etu hinta: N/A

Tuote: Lumene LUMO Kimmoisuutta lisäävä silmänympärysseerumi 10ml
Hinta: 30,50 €
S-etu hinta: N/A

Tuote: Lumene LUMO VITALITY Silottava & elvyttävä yövoide 50 ml
Hinta: 34,50 €
S-etu hinta: 25,80 €
```

Kuva 9. Verkkoharavoinnilla haetut tuotetiedot listattuna

Tiedot kerätään samalla tavalla kaikilta kolmelta verkkosivulta, koodia tulee kuitenkin muuttaa hieman verkkosivujen rakenteen ja hintatietojen mukaan. Lykon ja Prisman verkkosivuilla on vain yksi nykyinen hinta, joten koodia muutetaan sen osalta. Prisman sivuilla ”Lumene Lumo” tuoteryhmään tulee haussa yksi ylimääräinen tuote, joka ei kuulu Lumo tuotesarjaan. Tuotteen nimi sisältää Lumene ja Lumo tekstit, mutta ei peräkkäin, kuten Lumene Lumo -tuotesarjassa. Koodiin tuleekin lisätä ylimääräinen if lause, joka rajaa tämän ylimääräisen tuotteen verkkohaun ulkopuolelle (Kuva 10).

```python
# Haetaan kaikki tuotekortit, jotka sisältävät tarvittavia teitoja
product_cards2 = soup2.find_all('div', class_='ProductCard_card__lf3Ri')
#Käydään läpi kaikki tuotekortit ja haetaan tarvittavat tiedot niistä
for card2 in product_cards2:
    # Haetaan tuotteen nimi
    product_name2 = card2.find('a', class_='ProductCard_heading__eOM2y').text.strip()
    # Rajataan if lauseella tuotteet, joissa ei lue "lumene lumo" peräkkäin isoilla tai pienillä kirjaimilla verkkohaun ulkopuolelle
    if "lumene lumo" in product_name2.lower():
```

Kuva 10. Prisman Python lähdekoodiin lisätty if lause

3.5 Datan käsittely

Datan keräämisen jälkeen, se käsitellään strukturoituun muotoon analysointia varten. Hakurobotti tulostaa tuotetiedot listattuna allekkain, joten tietoja ei voida sellaisenaan käyttää analysoinnissa. Datan käsittelyssä käytetään Pythonin Pandas kirjastoa, sillä sen avulla voidaan luoda haetuista tiedoista helposti taulukot ja yhdistää ne yhdeksi isoksi taulukoksi. Ensimmäisessä vaiheessa verkkosivujen hakurobotteihin lisätään koodi tyhjän taulukon luomiseksi tietojen tallennusta varten (Kuva 11).

```
 # Luodaan tyhjä DataFrame tuotteiden tietojen tallentamista varten
dfSokos = pd.DataFrame(columns=['Tuote', 'Sokos_Hinta','S-etu_Sokos'])
```

Kuva 11. Python Pandas tyhjän taulukon luonti

Taulukon luomisen jälkeen lisätään for loopin sisälle koodi, joka lisää tuotteiden tiedot yksitellen taulukkoon (Kuva 12).

```
 # Lisätään tuotteen tiedot DataFrameen
dfSokos = pd.concat([dfSokos, pd.DataFrame({'Tuote': [product_name1], 'Sokos_Hinta': [final_price1], 'S-etu_Sokos':[s_card_price1]})],
ignore_index=True)
```

Kuva 12. Python Pandas tietojen lisääminen taulukkoon

Kun kaikille verkkosivuille on luotu omat taulukot, yhdistetään tiedot yhdeksi isoksi taulukoksi yhteisen tuote sarakkeen perusteella ja tallennetaan CSV-tiedostoksi (Kuva 13).

```
# Yhdistetään DataFramejen df_prisma, df_sokos ja df_lyko 'Tuote'-sarakkeen perusteella
df_combined = dfPrisma.merge(dfSokos, on='Tuote', how='outer').merge(dfLyko, on='Tuote', how='outer')

# Tulostetaan yhdistetty DataFrame
print(df_combined)
# Tallennetaan DataFrame CSV-tiedostoksi
df_combined.to_csv('tuotetiedot.csv', index=False)
```

Kuva 13. Python Pandas taulukkojen yhdistäminen yhdeksi isoksi taulukoksi

Yhdistelyn jälkeen yhdellä tuotteella tulisi olla neljä tai kolme eri hintaa, jos tuotetta myydään kaikilla verkkosivuilla. Yhdistelyssä kuitenkin huomattiin, että Lyko käyttää tuotteiden englanninkielisiä nimiä, kun taas Sokos ja Prisma käyttävät suomenkielisiä nimiä. Tämän vuoksi Lykon tuotenimet ja hinnat näkyvät taulukossa omina riveinään oheisessa yhdistetyssä taulukossa (Kuva 14).

	Sokos_Hinta	S-etu_Sokos	Tuote_yhd	Hinta_Prisma	Hinta_Lyko
1	30,50 €	22,80 €	Lumene LUMO Kiinteyttävä hajusteeton päivävoide SPF30 50 ml		
2	30,50 €	N/A	Lumene LUMO Kimmoisuutta lisäävä silmänympärysseerumi 10ml	26,50 €	
3	33,50 €	N/A	Lumene LUMO Pikakaunistaja + kollageeniseerumi tuplapakkaus		
4	33,50 €	N/A	Lumene LUMO Pre-retinoli kasvoöljy 30 ml	29,95 €	
5	33,50 €	N/A	Lumene LUMO Silottava & kiinteyttävä pikakaunistaja 30 ml	28,50 €	
6	33,50 €	N/A	Lumene LUMO Silottava & kiinteyttävä pikakaunistaja 50 ml		
7	30,50 €	N/A	ımene LUMO Silottava & kiinteyttävä päivävoide + yövoide tuplapakkaus		
8	30,50 €	22,80 €	Lumene LUMO Silottava & kiinteyttävä päivävoide 50 ml	26,50 €	
9	30,50 €	22,80 €	Lumene LUMO Silottava & kiinteyttävä päivävoide mineraali SK30 50ml		
10	31,50 €	N/A	Lumene LUMO Silottava & kiinteyttävä silmänympärysvoide 15 ml	26,50 €	
11	30,50 €	22,80 €	Lumene LUMO Silottava & kiinteyttävä yövoide 50 ml	26,50 €	
12	34,50 €	N/A	UMO VITALITY Silottava & elvyttävä päivävoide + yövoide tuplapakkaus		
13	34,50 €	25,80 €	Lumene LUMO VITALITY Silottava & elvyttävä päivävoide 50 ml	28,50 €	
14	34,50 €	25,80 €	Lumene LUMO VITALITY Silottava & elvyttävä yövoide 50 ml	28,50 €	
15	36,90 €	N/A	Lumene LUMO VITALITY Silottava & elvyttävä öljyseerumi 30 ml	32,50 €	
16	33,50 €	N/A	Lumene Lumo Kimmoisuutta lisäävä kollageeniseerumi 30ml	28,50 €	
17			ɔrdic BloomAnti-Wrinkle & Firm Day Cream SPF30 Fragrance Free50 ml		29,50 €
18			Nordic BloomAnti-wrinkle & Firm Day Fluid Mineral SPF 3050 ml		29,50 €
19			Nordic BloomAnti-wrinkle & Firm Day Moisturizer50 ml		29,50 €
20			Nordic BloomAnti-wrinkle & Firm Moisturizing Eye Cream15 ml		29,50 €
21			Nordic BloomAnti-wrinkle & Firm Moisturizing V-Shape Serum30 ml		32,50 €
22			Nordic BloomAnti-wrinkle & Firm Night Moisturizer50 ml		29,50 €
23			Nordic BloomBerry Pre-Retinol Facial Oil30 ml		29,50 €
24			Nordic BloomVegan Collagen Essence30 ml		22 €
25			Nordic BloomVegan Collagen Eye Serum10 ml		20 €
26			Nordic BloomVitality Anti-Wrinkle & Revitalize Oil Serum30 ml		35,90 €
27			Nordic BloomVitality Anti-Wrinkle & Revitalize Overnight Balm50 ml		31,50 €
28			Nordic BloomVitality Anti-Wrinkle & Revitalize Rich Day Cream50 ml		31,50 €

Kuva 14. Python Pandas yhdistetty taulukko

Raakadatan käsittelyä jatketaan Power BI-ohjelmassa, jossa on valmiita helppokäyttöisiä ominaisuuksia datankäsittelyyn. Järjestelmä on valittu, sillä siellä tehdään myös datan analysointi ja visualisointi. Ensimmäiseksi Power BI:ssa yhdistetään englannin- ja suomenkieliset nimet järjestelmän valmiilla ryhmittely ominaisuudella, jossa valitaan yksitellen tuotteet, jotka halutaan yhdistää ja luodaan uusi ”Tuote (groups)” mittari. Näin saadaan kaikki verkkokauppojen hinnat samalle tuotteelle.

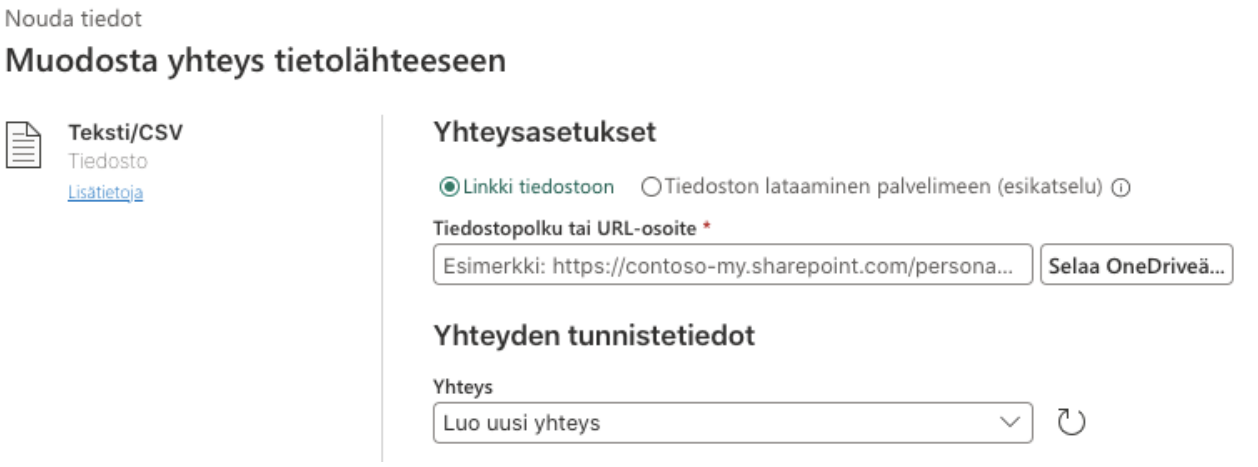
Seuraavaksi poistetaan Sokoksen ja Prisman hinnoista ylimääräinen euromerkki Power BI:n DAX-ohjelmointikielellä ja muutetaan tekstimuotoiset mittari numeraaliseksi. Luodaan myös uusi ”halvin hinta” mittari Sokokselle, joka etsii tuotteelle halvimman hinnan Sokoksen normaaleista ja s-etu-hinnoista. Näiden muutosten jälkeen data on analysoitavassa muodossa (Kuva 15).

Tuote	Average of HintaPrisma	Average of Sokos_halvin_hinta	Average of Hinta_Lyko
Lumene LUMO Kiinteyttävä hajusteeton päivävoide SPF30 50 ml & ...		22,80	29,50
Lumene Lumo Kimmoisuutta lisäävä kollageeniseerumi 30ml & ...	28,50	33,50	22,00
Lumene LUMO Kimmoisuutta lisäävä silmänympärysseerumi 10ml & ...	26,50	30,50	20,00
Lumene LUMO Pikakaunistaja + kollageeniseerumi tuplapakkaus		33,50	
Lumene LUMO Pre-retinoli kasvoöljy 30 ml & ...	29,95	33,50	29,50
Lumene LUMO Silottava & kiinteyttävä pikakaunistaja 30 ml & ...	28,50	33,50	32,50
Lumene LUMO Silottava & kiinteyttävä pikakaunistaja 50 ml		33,50	
Lumene LUMO Silottava & kiinteyttävä päivävoide + yövoide tuplapakkaus		30,50	
Lumene LUMO Silottava & kiinteyttävä päivävoide 50 ml & ...	26,50	22,80	29,50
Lumene LUMO Silottava & kiinteyttävä päivävoide mineraali SK30 50ml & ...		22,80	29,50
Lumene LUMO Silottava & kiinteyttävä silmänympärysvoide 15 ml & ...	26,50	31,50	29,50
Lumene LUMO Silottava & kiinteyttävä yövoide 50 ml & ...	26,50	22,80	29,50
Lumene LUMO VITALITY Silottava & elvyttävä päivävoide + yövoide tuplapakkaus		34,50	
Lumene LUMO VITALITY Silottava & elvyttävä päivävoide 50 ml & ...	28,50	25,80	31,50
Lumene LUMO VITALITY Silottava & elvyttävä yövoide 50 ml & ...	28,50	25,80	31,50
Lumene LUMO VITALITY Silottava & elvyttävä öljyseerumi 30 ml & ...	32,50	36,90	35,90
Total	**28,25**	**29,64**	**29,20**

Kuva 15. Käsitelty data

3.6 Datan varastointi

Verkkoharavoinnilla haetaan ajantasaista dataa, joten datan varastoinnin tulisi olla mahdollisimman automaattista. Tallennettava CSV-tiedosto on yksinkertainen strukturoitu taulukko, joten se voidaan varastoida Microsoftin OneDrive-pilvitallennustilaan, jonne on mahdollista luoda suora yhteys Power BI -ohjelmalla (Kuva 16). Kun tiedosto päivitetään OneDrive ohjelmassa, päivittyy se myös Power BI -järjestelmässä. Tässä työssä tiedosto viedään manuaalisesti OneDrive tallennustilaan. Jupyterista olisi mahdollista viedä tiedosto automaattisesti OneDrive työtilaan ohjelmoimalla, jolloin data päivittyisi automaattisesti Power BI -ohjelmassa, mutta tämä on rajattu opinnäytetyön ulkopuolelle puuttuvien oikeuksien vuoksi.



Kuva 16. Power BI yhteyden luominen OneDrive pilvitallennustilaan

3.7 Datan hallinta

Datan hallinta kattaa datan organisoinnin, säilytyksen ja palauttamisen koko analysointiprosessin aikana. Siihen kuuluu muun muassa käyttöoikeuksien hallintaa, dataan tehtyjen muutosten seurantaa ja datan suojausta. Opinnäytetyöprosessin aikana oikeudet tehdä muutoksia ohjelmiin on vain työn kirjoittajalla. Oikeudet käyttää Power BI ja Jupyter-ohjelmaa sekä OneDrivea tulevat koulun puolesta. Työn lopputulokset eli Jupyterissa Pythonilla luotu verkkoharavoinnin robotti, koodin hakema CSV-tiedosto, sekä Power BI -ohjelmalla luotu analysoinnin ja visualisoinnin näkymät jaetaan GitHubissa https://github.com/johranta/oppari_analysointi/tree/main niin, että kuka tahansa voi niitä käyttää.

3.8 Datan analysointi

Analysoinnin tarkoituksena on vastata kysymykseen, mistä kuluttajan olisi tällä hetkellä edullisinta ostaa Lumenen Lumo-sarjan tuotteita. Kyseessä on yksinkertainen kuvailevan analyysin datan kokoaminen, jonka tarkoituksena on luoda yleiskatsaus tietyn hetken hinnoista. Datasta ei voida tehdä johtopäätöksiä siitä, miksi jotain on tapahtunut tai todennäköisesti tapahtuu tulevaisuudessa, sillä historiatietoja ei ole kerätty.

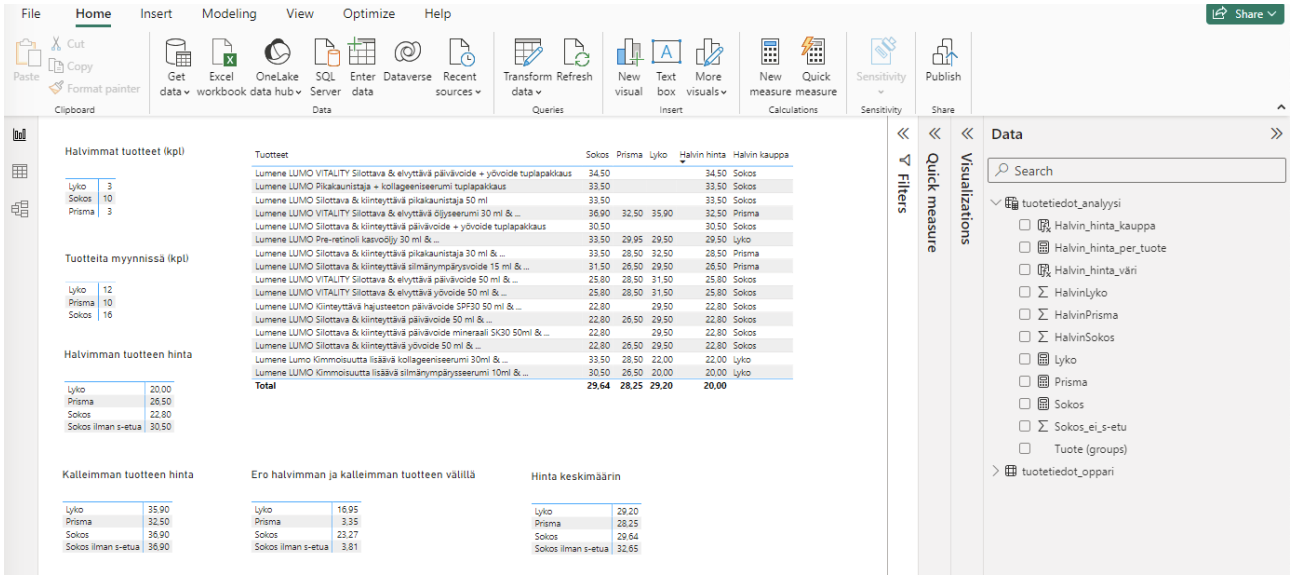
Datan analysointi tapahtuu Power BI -ohjelmassa, jossa luodaan oma analysointitaulukko ”tuotetiedot_analyysi”, jonne haetaan edellisessä vaiheessa luotu yhdistetty tuoteryhmä ”Tuote (groups)”, sekä Lykon, Prisman ja Sokoksen hinnat. Luodaan lisäksi ”Halvin_hinta_per_tuote” mittari, joka hakee halvimman hinnan tuotteelle kolmesta vaihtoehdosta, sekä mittari, joka kertoo, missä kaupassa tämä halvin hinta on. Taulukon avulla voidaan etsiä nopeasti, mistä tuote kannattaa käydä ostamassa ja kuinka paljon se maksaa. Analyysiä varten luodaan myös mittarit Sokokselle, Prismalle ja Lykolle, jotka laskevat, kuinka monta halvinta tuotetta kyseisessä kaupassa on.

Mittareiden mukaan Sokoksessa on eniten edullisimpia tuotteita 10 kpl, kun taas Prismassa ja Lykossa on vain 3 kpl halvimpia tuotteita. Tuotetietoja tutkimalla selviää, että Sokoksessa on myynnissä enemmän tuotteita (16 kpl), kuin Prismassa (10 kpl) ja Lykossa (12 kpl). Luodusta analysointitaulukosta nähdään, että Sokoksen neljän tuotteen halvin hinta johtuu siitä, että tuotetta ei ole saatavilla muissa kaupoissa. Vaikka näitä tuotteita ei huomioitaisi, olisi Sokoksessa silti kappalemääräisesti eniten halvimpia tuotteita (6 kpl).

Luodaan yksinkertaiset taulukot halvimpien ja kalliimpien tuotteiden hintojen tarkasteluun, josta huomataan, että Lykossa on yksittäinen halvin hinta, joka on 20 euroa. Sokoksen halvin hinta on 22,80 euroa ja Prisman 26,50 euroa. Sokoksessa on s-etukortilla ja ilman s-etukorttia kallein tuotehinta 36,90 euroa, Lykossa 35,90 euroa ja Prismassa 32,50 euroa. Sokoksen kalleimman tuotteen saa Prismasta siis 4,4 euroa halvemmalla. Prismassa on halvin keskihinta 28,25 euroa ja pienin

ero kalleimman ja halvimman tuotteen välillä 3,35 euroa. Lykossa on toiseksi halvin keskihinta 29,20 euroa ja toiseksi pienin ero kalleimman ja halvimman tuotteen välillä 16,95 euroa. Sokoksessa on siis kallein keskihinta 29,64 euroa ja suurin ero halvimman ja kalleimman tuotteen välillä 23,27 euroa. Ilman s-etukorttia Sokoksen keskihinta on 32,65 euroa eli keskimäärin 3,01 euroa kalliimpi, kuin s-etukortin kanssa.

Kuten kuvan 17 Power BI -analysointinäkymästä voidaan päätellä, Prismassa on keskimäärin halvimmat hinnat. Hintojen hajonnan ollessa pientä asiakkaalle ei tule suurempia yllätyksiä tuotteita ostaessa. Sokoksessa on eniten halvimpia tuotteita, mutta hintojen hajonta on suurta. Sokoksessa asioidessa asiakkaan kannattaa käyttää s-etukorttia, jolloin hän säästää keskimäärin 3,01 euroa per tuote. Lykosta taas löytyy parin tuotteen osalta yksittäiset halvimmat tuotehinnat. Yhteenvetona asiakkaan kannattaa tarkistaa taulukosta haluamiensa tuotteiden hinnat ja valita kauppa sen perusteella. Prismassa on keskimäärin halvimmat hinnat, jos tuotehintoja ei päästä tarkistamaan. Prismasta ei kuitenkaan löydy kaikkia Lumene Lumo sarjan tuotteita.



Kuva 17. Power BI analysointi näkymä

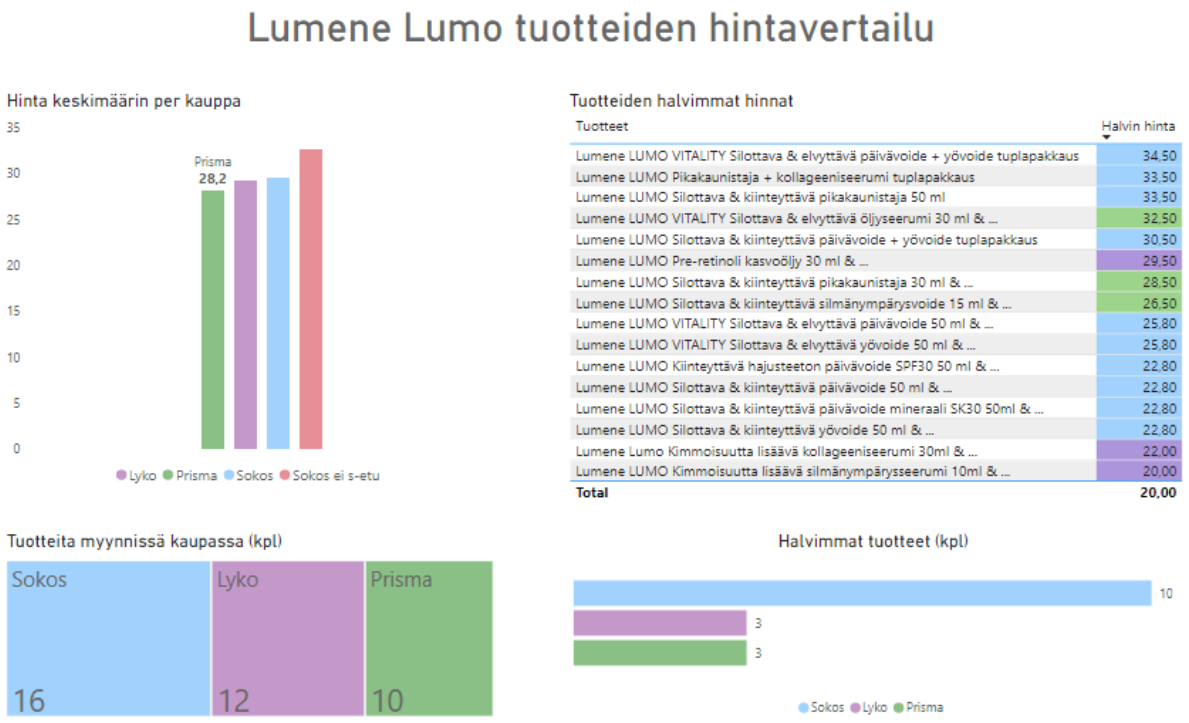
3.9 Datan visualisointi

Datan visualisoinnilla pyritään esittämään johtopäätökset selkeästi ja ymmärrettävästi. Visualisointiin käytetään Power BI -ohjelmaa, johon luodaan näkymä Lumenen Lumo tuotteiden hintavertailulle.

Hintavertailu näkymässä käytetään BI-järjestelmän valmiita interaktiivisia visualisointeja, joista ensimmäiseksi valitaan pylväskaavio. Se kertoo Lumo tuotteiden keskihinnan kaupoittain. Pylväsdiagrammi osoittaa selkeästi, että Prisman keskihinta on halvin. Tämä on viesti, jota halutaan

painottaa visualisoinnin käyttäjälle ja tämän vuoksi vain Prisman keskihinta on merkitty kuvaajaan. Visualisoinnista näkyy näin heti, että asiakkaan kannattaa suunnata Prismaan ostoksille, sillä siellä on halvin keskihinta tuotteille. Pelkästään tämä yksi kaavio voisi riittää asiakkaan toivoman johtopäätöksen osoittamiseen. Näkymään on kuitenkin tuotu lisäksi tuotetietoja, jos asiakas haluaa tarkistaa tiettyjen tuotteiden hinnat tai saatavuudet.

Keskihintojen alle on luotu puukartta, jossa kunkin suorakulmion koko riippuu mitattavan arvon suuruudesta. Puukartta osoittaa eri kauppojen valikoimien koon kauppakohtaisella värikoodauksella. Kaaviosta on helppo tarkistaa, että Sokoksella on laajin tuotevalikoima, joten tietty tuote löytyy todennäköisimmin sieltä. Puukartan vieressä on palkkikaavio, joka osoittaa, että Sokoksessa on kappalemäärällisesti eniten halvimpia tuotteita, joten tietyn tuotteen hinnan tarkastaminen kannattaa aloittaa Sokoksesta. Viimeisenä näkymässä on taulukko, jossa on kaikki Lumo sarjan tuotteet ja niiden halvin hinta. Hinnat on värikoodattu sen mukaan, mistä kaupasta se löytyy. Taulukosta asiakas voi tarkistaa tietyn tuotteen hinnan ja mistä kaupasta se kannattaa ostaa. Kaikki taulukot ovat interaktiivisia, joten napauttamalla tuotetta tuotetaulukossa asiakas voi samalla tarkistaa pylväsdiagrammista, kuinka paljon halvempi tuote on verrattuna muihin kauppoihin. Puukartasta voidaan tarkistaa, löytyykö tuote kaupan valikoimasta. Kuvassa 18 valmis Power BI -visualisointi.



Kuva 18. Power BI visualisointi näkymä

3.10 Datan tulkinta

Datan tulkinnan tarkoituksena on analysoinnin ja visualisoinnin perusteella ymmärtää, mitä data kertoo, mutta myös mitä vaikutuksia sillä on. Kuten datan analysointi osiossa pääteltiin, data kertoo meille, että Prisma on halvin kauppa tehdä ostoksia ja asiakkaan kannattaa suunnata sinne ostoksille.

4. Johtopäätökset

Opinnäytetyön tavoitteena oli luoda työkalu tuotehintojen vertailuun datan elinkaarimallin avulla, sekä selvittää onnistuuko työkalun luominen verkkoharavoinnilla haetun datan perusteella. Teoriaosuudessa käytiin ensin läpi tiedon elinkaarimalli, jonka perusteella tutkimus laadittiin. Tiedon elinkaari koostuu kahdeksasta eri vaiheesta datan luomisesta, keräämisestä, käsittelystä, varastoinnista, hallinnasta, analysoinnista, visualisoinnista ja tulkinnasta. Näistä jokainen vaihe käytiin läpi empiriaosuudessa. Tämän jälkeen teoriassa käsiteltiin verkkoharavoinnin teknistä toteuttamista, jota tarvittiin elinkaarimallin kolmessa ensimmäisessä vaiheessa. Teorian perusteella verkkoharavointi päätettiin toteuttaa Python ohjelmointikielellä, sen yksinkertaisuuden ja monipuolisten verkkoharavointiin tarkoitettujen kirjastojen vuoksi. Verkkoharavoinnin teoriassa käsiteltiin myös haravoinnin teknisiä ja eettisiä ongelmia.

Kuten verkkoharavoinnin teoriassa kerrottiin, liittyy hakurobotin tekoon useita rakenteellisia ja eettisiä ongelmia. Ensimmäinen ongelma havaittiin jo elinkaarimallin datan luomisvaiheessa, jossa useampi verkkosivusto oli kieltänyt hakurobotit robots.txt tiedostossa. Vaikka hinnat pystyisi tiedostosta huolimatta hakemaan, olisi hakurobotin käyttö eettisesti väärin. Tämän vuoksi hinta-analysointityökalun keräämä datamäärä pieneni alkuperäisestä suunnitelmasta kolmeen verkkokauppaan, jotka eivät olleet kieltäneet hakurobotteja.

Seuraavaan rakenteelliseen ongelmaan törmättiin elinkaarimallin toisessa, datan keräämisen vaiheessa, jossa Sokoksen verkkosivujen hintaelementit olivat muuttuneet, kun sivustoilla vierailtiin uudestaan. Sivustolle oli tullut uusi hintaelementti s-etukortilla saatavista alennuksista. Jos sivustoilla ei olisi vierailtu uudestaan, olisi uudet halvemmat hinnat jääneet huomaamatta ja työkalu antanut virheellistä dataa. Tämä oli erinomainen esimerkki siitä, miksi hakurobotin säännöllinen päivittäminen on ensiarvoisen tärkeää tietoja kerätessä.

Datan kerääminen verkkosivuilta osoittautui teknisesti luultua yksinkertaisemmaksi. Keräämiseen riitti melko yksinkertainen koodi, joka oli helppo kopioida eri verkkosivuille. Ainoastaan oikeiden hinta- ja tuotetieto elementtien etsiminen vei hieman aikaa, kun ne etsittiin hakukoneen kehittäjä moduulilla. Pythonin käyttäminen hakurobotin luomisessa osoittautui oikeaksi valinnaksi koodin intuitiivisuuden ansiosta.

Kolmannessa elinkaaren vaiheessa huomattiin, että vaikka dataa oli melko yksinkertaista haravoida verkkosivuilta, tuli sitä käsitellä melko paljon ennen analysointia. Ensin verkkosivujen tiedot yhdistettiin yhdeksi strukturoiduksi taulukoksi Pythonin Pandas kirjaston avulla. Strukturoidun taulukon luominen Pandas kirjaston avulla oli helppoa, mutta datan käsittely muuten osoittautui vaikeaksi, jonka takia luotu taulukko ladattiin Power BI järjestelmään. Power BI järjestelmä valittiin datan

analysointiin ja visualisointiin sillä se on helppokäyttöinen BI-järjestelmä, jossa on valmiita visualisointeja. Sillä oli myös helpompi muokata dataa analysoitavaan muotoon valmiiden ominaisuuksien ja DAX-koodin ansiosta. Tietojen lataamisen jälkeen huomattiin, että Prisman ja Sokoksen hinnat olivat tekstimuodossa euromerkin vuoksi ja ne tuli muuntaa numeraaliseksi. Lykolla oli eri tuotenimet ja tuotekoodit, kuin Prismalla ja Sokoksella, jonka vuoksi saman tuotteen hinnat menivät eri tuotteelle. Tuotenimet tuli yhdistää manuaalisesti Power BI:n ryhmittelyllä, joka vähensi työkalun automaation tasoa. Jos Lumene lanseeraisi uusia tuotteita Lumo sarjaan, ei työkalu enää toimisi kunnolla vaan sitä tulisi päivittää manuaalisesti, jotta hinnat menisivät oikeille tuotteille. Samoin, jos työkalua haluttaisiin jatkossa laajentaa hakemalla uusien verkkokauppojen hintoja, tulisi ensin manuaalisesti varmistaa, että tuotetiedot menisivät oikeille tuotteille, sillä verkkokaupat saavat nimetä ja numeroida tuotteet haluamallaan tavalla ja pienikin ero tuotenimessä tai -koodissa aiheuttaa virheen tuotteen hintavertailussa ja näin tuottaa virheellistä dataa.

Datan varastoinnin osalta tiedot vietiin manuaalisesti OneDrive pilvitallennustilaan. OneDrive valittiin sen vuoksi, että Power BI -järjestelmästä on mahdollista luoda suora yhteys pilvitallennustilaan ja näin dataa on helppo päivittää. Työkalun automaation tasoa kuitenkin heikensi CSV-tiedoston manuaalinen tallentaminen OneDrive ohjelmaan. Jupyter ohjelmointiympäristöstä on mahdollista luoda automaattinen tallennus OneDriveen, jolloin työkalua olisi helpompi päivittää automaattisesti. Tämä on kuitenkin rajattu opinnäytetyön ulkopuolelle, sillä opiskelijoiden oikeuksia on rajattu oikeuksien jakamisen osalta. Manuaalinen ratkaisu on kuitenkin riittävä tämän opinnäytetyön osalta, sillä haettavia tietoja on vain vähän ja datan käsittely tapahtuu automaattisesti Pandas kirjaston ja Power BI:n avulla. Datan varastoinnin automatisoinnin ratkaisut verkkoharavoinnilla haetun tiedon osalta olisi kuitenkin mielenkiintoinen jatkotutkimuksen aihe.

Datan hallintaa opinnäytetyössä sivuttiin vain hieman analyysityökalun ollessa yksityisessä käytössä. Jos työkalu tulisi yrityksen käyttöön, oikeuksien ja versioiden hallintaan tulisi kiinnittää enemmän huomiota. Työkalun näkymään olisi tällöin hyvä lisätä päivä, jolta tiedot on haettu. Mielenkiintoinen jatkotutkimuksen aihe olisi rakentaa tietokanta, johon hintatiedot haettaisiin päivittäin ja näin voitaisiin tarkastella, miten hinnat ovat eri verkkokaupoissa kehittyneet.

Datan analysoinnin vaiheessa tarkasteltiin verkkoharavoinnilla haettuja hintoja tarkemmin. Haetusta datasta oli mahdollista tehdä vain kuvailevan analyysin datan kokoaminen. Siinä data esitettiin tiivistetyssä muodossa yleiskatsauksen luomiseksi, sillä verkkoharavoinnilla haettiin vain tietyn hetken hinnat. Analysoinnin vaiheessa luotiin uusia mittareita selvittämään tuotteen halvin hinta, missä kaupassa halvin hinta on sekä, missä kaupassa on eniten halpoja tuotteita. Jos työkaluun haluttaisiin jatkossa lisätä uusia verkkokauppoja, tulisi mittareiden koodia muuttaa sillä se huomioi vain nykyiset verkkokaupat. Näin ollen analysoinnin työkalu ei ole automaattisesti skaalautuva.

Datan visualisoinnissa luotiin analysoinnin tueksi Power BI -näkymä, jossa esitetään johtopäätökset selkeästi ja ymmärrettävästi. Opinnäytetyössä onnistuttiin luomaan näkymä, josta näkee ensimmäisellä vilkaisulla, että Prismassa on halvimmat hinnat. Näkymä on visuaalisesti yksinkertainen ja selkeä. Siitä on helppo löytää yksittäisen tuotteen halvin hinta ja kauppa, jossa tuotetta myydään. Visualisointi on interaktiivinen eli tuotetaulukon tuotetta klikatessa myös muut visualisoinnit muuttuvat. Tämä aiheuttaa haasteen keskihinta pylväskuvaajan kanssa, sillä siihen on selkeyden vuoksi jätetty vain Prisman datapisteelle kaupan nimi ja keskihinta tekstinä. Vaikka yksittäinen tuote olisi jossakin muussa kaupassa edullisempi, näyttää kuvaaja vain Prisman hinnan tekstinä. Edullisimman hinnan voi kuitenkin tarkistaa tuotetaulukosta, joten tämä visualisointi tehtiin tarkoituksella halutun johtopäätöksen viestimiseksi.

Datan tulkinta oli helppoa, sillä sen pohjalla käytetty data on melko suppea katsaus tietyn hetken hinnoista. Tulkintaa tuki myös selkeät analysoinnit ja visualisoinnit.

Kaiken kaikkiaan opinnäytetyö onnistui tavoitteessaan selvittää, miten Python ohjelmointikielellä voidaan luoda verkkoharavoinnin ohjelma. Lopputuloksena oli toimiva hakurobotti, joka haki tarvittavat tiedot verkkosivuilta. Opinnäytetyö onnistui vastaamaan myös, toiseen tavoitteeseen eli voidaanko verkkoharavoinnilla haetun datan perusteella luoda analysoinnin työkalu datan elinkaarimallin avulla. Työssä onnistuttiin luomaan analysoinnin työkalu käymällä kaikki datan elinkaaren vaiheet läpi. Kuten tässä luvussa on kuvailtu, haasteita oli melko paljon ja lopputuloksena oli yksinkertainen työkalu, jossa on manuaalisia vaiheita. Työkalua tulisi jatkokehittää, jotta siitä saisi täysin automaattisen ja siihen olisi mahdollista lisätä uusia verkkosivuja ilman manuaalisia työvaiheita. Mielenkiintoinen jatkotutkimuksen aihe olisi jalostaa hakurobotista päivittäiseen käyttöön sopivan automaattisen työkalun, joka tallentaisi myös historiatietoja.

Opinnäytetyöprojekti kasvatti merkittävästi ymmärrystäni verkkoharavoinnista, tiedon elinkaaresta ja Pythonista. En ollut aiemmin käyttänyt verkkoharavoinnin työkaluja ja yllätyin siitä, kuinka helppoa ohjelma oli Pythonilla laatia, sillä ennen opinnäytetyötä Python osaamiseni oli aloittelijan tasolla. Tutkimuksen tekeminen opetti minulle Python ohjelmointikielen ja sen kirjastojen käyttöä, jatkossa uskallan rohkeammin lähteä ohjelmoimaan, vaikka en alkuun täysin tietäisi, miten ohjelma pitäisi tehdä. Tutkimuksen tekeminen syvensi Power BI taitojani analysoinnin ja visualisoinnin osalta.

Opinnäytetyöprojekti oli itselleni haastava perheenlisäyksen takia ja valitettavasti aikataulu venyi puolella vuodella. Työtä kirjoittelin suurimmaksi osaksi iltaisin vauvan mentyä nukkumaan. Tämän vuoksi työtä piti pilkkoa pienempiin osiin niin, että pystyin parissa tunnissa paneutumaan käsiteltävään asiaan. Aluksi kokonaisuus vaikutti liian suurelta, mutta alun hankaluuksien jälkeen

kirjoittaminen alkoi sujua. Projektin myötä opin tehostamaan ajankäyttöäni ja jakamaan suuria kokonaisuuksia pienempiin osiin. Uskon, että näistä taidoista on huomattavasti hyötyä myös työelämässä.

5. Lähteet

Amos, D. 2022. A Practical Introduction to Web Scraping in Python. Luettavissa: https://realpython.com/python-web-scraping-practical-introduction/. Luettu 4.6.2023.

Besinsky, A. 2024 What Is Web Scraping? Definitive Guide 2024. Luettavissa: https://brightdata.com/blog/how-tos/what-is-web-scraping. Luettu: 20.2.2024.

Bhagyashree 2023. Ensuring Ethical Price Scraping: Best Practices and Guidelines. Luettavissa: https://www.promptcloud.com/blog/ensuring-ethical-price-scraping-best-practices-and-guidelines/. Luettu: 14.2.2024.

Big data framework 2019. Three different data structures. Luettavissa: https://www.bigdataframework.org/data-types-structured-vs-unstructured-data/. Luettu: 27.2.2024.

Czaban, T. 2023. How To Use AI for Data Visualizations and Dashboards. Luettavissa: https://www.gooddata.com/blog/how-to-use-ai-for-data-visualizations-and-dashboards/. Luettu: 15.3.2024.

Daivi 2024. 7 Python Libraries For Web Scraping To Master Data Extraction. Luettavissa: https://www.projectpro.io/article/python-libraries-for-web-scraping/625. Luettu: 19.4.2024.

Dhanashree 2023. Web Scraping with Python Tutorial. Luettavissa: https://nanonets.com/blog/web-scraping-with-python-tutorial/. Luettu: 25.3.2024.

Dilmegani, C. 2024. 3 Benefits of Using Web Scraping as a Service in 2024. Luettavissa: https://research.aimultiple.com/web-scraping-as-a-service/. Luettu: 1.3.2024.

Educative 2024. What is Beautiful Soup? Luettavissa: https://www.educative.io/answers/what-is-beautiful-soup. Luettu 12.2.2024.

Ermakovich, G. 2023. Web Scraping: What It Is and How to Use It. Luettavissa: https://scrape-it.cloud/blog/web-scraping-what-it-is-and-how-to-use-it. Luettu: 8.2.2024.

EU 2022. Yleinen tietosuoja-asetus. Luettavissa: https://europa.eu/youreurope/business/dealing-with-customers/data-protection/data-protection-gdpr/index_fi.htm. Luettu: 5.2.2024.

Evolytics 2024. Data Driven Storytelling Tip #1: Know Your Audience. Luettavissa: https://evolytics.com/blog/1-know-your-audience/. Luettu: 15.3.2024.

Gallagher, J., R. & Beveridge, A. 2021. Project-Oriented Web Scraping in Technical Communication Research. Journal of Business and Technical CommunicationVolume 36, Issue 2, April 2022, Pages 231-250.

Google 2024. How to write and submit a robots.txt file. Luettavissa: https://developers.google.com/search/docs/crawling-indexing/robots/create-robots-txt. Luettu: 14.4.2024.

Harvard Business School Online (HBO). A Beginner´s Guide to Data & Analytics. Luettavissa: https://online.hbs.edu/Documents/a-beginners-guide-to-data-and-analytics.pdf. Luettu 4.6.2023.

IBM. What is data visualization? Luettavissa: https://www.ibm.com/topics/data-visualization. Luettu: 10.3.2024.

Insightsoftware 2023. How can AI help with data visualization?. Luettavissa: https://insightsoftware.com/blog/how-can-ai-help-with-data-visualization/. Luettu: 19.3.2024.

Javaid, S. 2024. 6-Step AI Data Collection Process & Roadmap in 2024. Luettavissa: https://research.aimultiple.com/data-collection-process/. Luettu: 20.2.2024.

Jurgenstojku 2020. Is web scraping legal in 2020?. Luettavissa: https://medium.com/@jurgenstojku_62417/is-web-scraping-legal-in-2020-63cbcf0d5ec. Luettu: 14.2.2024.

Kamaly, T. 2022. The Importance of Data Lifecycle Management (DLM) and Best Practices. Luettavissa: https://www.computer.org/publications/tech-news/trends/the-importance-of-data-lifecycle-management. Luettu: 13.3.2024.

Karatas, G. 2024. AI-Powered Web Scraping in 2024: Best Practices & Use Cases. Luettavissa: https://research.aimultiple.com/ai-web-scraping/. Luettu: 12.2.2024.

Korpela, J. 2013. Tekijänoikeus: vastauksia usein esitettyihin kysymyksiin, luku 5 Tekijänoikeus ja netti. Luettavissa: https://jkorpela.fi/tekoik/5.6.html. Luettu 20.2.2024

Krotov, V. & Silva, L. 2018. Legality and Ethics of Web Scraping. Emerhent Research Forum (ERF).

Ksn-developer 2023. Webscraping using pandas. Luettavissa: https://dev.to/ksndeveloper/web-scraping-using-pandas-fph. Luettu: 1.3.2024.

Lumene 2024. Vastuullisuus. Luettavissa: https://fi.lumene.com/pages/sustainability. Luettu: 25.4.2024.

LXML 2024. Parsing XML and HTML with lxml. Luettavissa: https://lxml.de/parsing.html#the-feed-parser-interface. Luettu: 30.2.2024.

Miller, K. 2021. CREATING DATA VISUALIZATIONS IN EXCEL: WHAT TO KEEP IN MIND. Luettavissa: https://online.hbs.edu/blog/post/data-visualizations-in-excel. Luettu: 20.3.2024.

MDN 2024a. CSS: Cascading Style Sheets. Luettavissa: https://developer.mozilla.org/en-US/docs/Web/CSS. Luettu: 26.4.2024.

MDN 2024b. Domain. Luettavissa: https://developer.mozilla.org/en-US/docs/Glossary/Domain. Luettu. 27.4.2024.

MDN 2024c. HTML: HyperText Markup Language. Luettavissa: https://developer.mozilla.org/en-US/docs/Web/HTML. Luettu: 26.4.2024.

MDN 2024d. URL. Luettavissa: https://developer.mozilla.org/en-US/docs/Glossary/URL. Luettu: 27.4.2024.

MDN 2024e. HTTP. Luettavissa: https://developer.mozilla.org/en-US/docs/Glossary/HTTP. Luettu: 27.4.2024.

MDN 2024f. SQL. Luettavissa: https://developer.mozilla.org/en-US/docs/Glossary/SQL. Luettu: 27.4.2024.

Microsoft 2023. DAXin yleiskatsaus. Luettavissa: https://learn.microsoft.com/fi-fi/dax/dax-overview Luettu: 27.4.2024.

Nigam, H. & Biswas, P. 2021. Web Scraping: From Tools to Related Legislation and Implementation Using Python. Innovative Data Communication Technologies and Application. Lecture Notes on Data Engineering and Communications Technologies, Springer, 59, s.149-164.

Pandas 2024. About Pandas. Luettavissa: https://pandas.pydata.org/about/index.html. Luettu: 1.3.2024.

Penn LPS 2022. 5 key reasons why data analytics is important to business. Luettavissa: https://lpsonline.sas.upenn.edu/features/5-key-reasons-why-data-analytics-important-business. Luettu: 15.4.2024.

ProWebScraper 2024.What Is Web Scraping: The Comprehensive Guide For 2024 Luettavissa: https://prowebscraper.com/blog/what-is-web-scraping/, Luettu: 15.4.2024.

Ramirez, F., M. 2023. Choosing the Right HTML and XML Parsers: “html”, “html.parser”, “html5lib” and “xml”. Luettavissa: https://medium.com/@juanfernandomoyanoramirez/choosing-the-right-html-and-xml-parsers-html-html-parser-html5lib-and-xml-f9c117b3e6bf. Luettu: 30.2.2024.

Ronquillo, A. 2024. Python's Requests Library (Guide). Luettavissa: https://realpython.com/python-requests/. Luettu: 6.3.2024.

Segment 2023. Data lifecycles in 2024: phases, use cases, & tips. Luettavissa: https://segment.com/blog/data-life-cycle/. Luettu: 27.2.2024.

Sharma, R. 2024. Web Scraping Using Jupyter. Luettavissa: https://medium.com/@ronak.d.sharma111/web-scraping-using-jupyter-b7aace016d38. Luettu: 26.4.2024.

Skakun, V. 2024a. A Comprehensive Guide to Web Scraping with Selenium WebDriver in Python. Luettavissa: https://hasdata.com/blog/web-scraping-using-selenium-python. Luettu: 20.2.2024.

Skakun, V. 2024b. 8 Best Python Libraries and Tools for Web Scraping in 2024. Luettavissa: https://hasdata.com/blog/best-python-libraries-for-web-scraping. Luettu: 3.3.2024.

Stevens, E. 2023. The 4 Types of Data Analysis [Ultimate Guide]. Luettavissa: https://careerfoundry.com/en/blog/data-analytics/different-types-of-data-analysis/. Luettu: 3.3.2024.

Stobierski, T. 2021a. 8 Steps In The Data Life Cycle. Luettavissa: https://online.hbs.edu/blog/post/data-life-cycle. Luettu: 20.2.2024.

Stobierski, T. 2021b. TOP DATA VISUALIZATION TOOLS FOR BUSINESS PROFESSIONAL. S. Luettavissa: https://online.hbs.edu/blog/post/data-visualization-tools Luettu: 13.3.2024.

Suramaparna 2024. Excel vs Tableau vs Power BI: Choosing the Right Tool for Your Data Analysis Needs. Luettavissa: https://medium.com/@suramaparna/excel-vs-tableau-vs-power-bi-choosing-the-right-tool-for-your-data-analysis-needs-67739b3c782a. Luettu: 20.3.2024.

Szwed, P. 2021. Is web scraping legal? A short guide on scraping under EU law. Luettavissa: https://discoverdigitallaw.com/is-web-scraping-legal-short-guide-on-scraping-under-the-eu-jurisdiction/. Luettu: 20.2.2024.

Urazaliev, F. 2023. Data Processing, Storage, and Organization. Luettavissa: https://medium.com/@urazaliev_f/data-processing-storage-and-organization-b47464f6f18a. Luettu: 15.3.2024.