Joonas Koponen

Data-analyysi suomalaisesta pimeän verkon huumekaupasta

Tietotekniikan pro gradu -tutkielma

23. elokuuta 2022

Jyväskylän yliopisto

Informaatioteknologian tiedekunta

Tekijä: Joonas Koponen

Yhteystiedot: jolajuko@student.jyu.fi

Ohjaaja: Tapio Frantti

Työn nimi: Data-analyysi suomalaisesta pimeän verkon huumekaupasta

Title in English: Data analysis of the Finnish dark web drug trade

Työ: Pro gradu -tutkielma

Opintosuunta: Sovellettu matematiikka ja laskennalliset tieteet

Sivumäärä: 60+3

Tiivistelmä: Tämä tutkimus tutkii suomalaista huumekauppaa data-analytiikan keinoin pimeässä verkossa. Tutkimuskysymykset käsittelevät huumausaineiden esiintyvyyttä pimeässä verkossa, ja tutkimustulosten perusteella arvoidaan huumausaineiden esiintyvyyttä Suomessa. Saatua tulosta verrataan THL:n tekemään jätevesitutkimukseen. Tutkimuksen aineisto (N $\approx$ 8500) on kerätty tiedonlouhinnan keinoin ja sen pohjalta on laadittu tilastollinen analyysi luokittelija-algoritmia ja sekaannusmatriiseja hyödyntäen. Kyseessä on tapaustutkimus, jonka analyysimenetelmäksi valikoitui luokittelu. Tutkimustuloksen perusteella suomalaisessa huumekaupassa esiintyy eniten kannabista, amfetamiinia ja lääkkeitä. Myös ekstaasia ja subutexia esiintyy verrattain paljon. Hallusinogeenit ja kokaiini muodostavat pienen osuuden jakaumasta ja heroiinia ei esiinny juuri lainkaan. Tulos ei kerro huumausaineiden määrällisestä esiintyvyydestä vaan kyse on enemmän yleisestä esiintyvyydestä tai trendistä. Tutkimustulos ja muut siinä esiin nousseet havainnot ovat linjassa aikaisemman tutkimustiedon kanssa. Tutkimus on yhteiskunnallisesti merkittävä ja antaa tietoa huumausaineiden esiintyvyydestä Suomessa. Lisäksi tutkimus tarjoaa ajankohtaista tietoa jatkuvassa murroksessa olevan pimeän verkon huumekaupan nykytilasta. Tutkimus tarkastelee myös viranomaisten kohdistamia toimia pimeään verkkoon ja tarjoaa heitä hyödyttäviä ideoita. Lisäksi tutkimuksessa esitellään jatkotutkimusaiheita ja pohditaan huumekaupan tulevaisuutta pimeässä verkossa.

Avainsanat: TOR-verkko, huumekauppa, jätevesitutkimus, pimeä verkko, verkkorikollisuus

Abstract: This study examines Finnish drug trafficking in the dark web using data analytics. The research questions deal with the prevalence of drugs in the dark web, and based on the research results, the prevalence of drugs in Finland is estimated. The result obtained is compared with the wastewater study carried out by THL. The data of the study (N $\approx$ 8500) has been collected by means of data mining and a statistical analysis has been prepared using the classifier algorithm and confusion matrices. This is a case study and classification was chosen as the method of analysis. Based on the results of the study, cannabis, amphetamines and medicines are the most abundant in the Finnish drug trade. Ecstasy and subutex are also relatively common. Hallucinogens and cocaine account for a small proportion of the distribution and heroin is almost non-existent. The result is not about the quantitative prevalence of drugs, but more about the overall prevalence or trend. The research result and other findings are in line with previous research data. The study is socially significant and provides information on the prevalence of drugs in Finland. In addition, the drug trade on the dark web is in a constant state of flux, so research provides up-to-date information on its current state. The study also looks at the actions taken by the authorities on the dark web and provides ideas that will benefit them. In addition, the study presents topics for further research and considers the future of drug trafficking in the dark side of the web.

Keywords: TOR network, cybercrime, dark web, drug trafficking, wastewater research

Termiluettelo

.onion	Ylätason verkkotunnus, joka sijaitsee Tor-verkossa. Tarvitsee Tor-selaimen yhteyden muodostamista varten. Pimeän verkko-osoitteen tunnusmerkki.
CAPTCHA	CAPTCHA (engl. Completely Automated Public Turing test to tell Computers and Humans Apart) on *kokonaan automatisoitu julkinen Turingin testi tietokoneiden ja ihmisten erottamiseksi toisistaan*. Turingin koe tarkoittaa mitä tahansa testiä, joka on suunniteltu erottamaan ihminen tietokoneesta. (Tiedepankki 2022)
Dark Web	Pimeä verkko on osa syväverkkoa. Pimeään verkkoon pääsee käytännössä vain Tor-verkon kautta.
Deep Web	Syväverkko on internetin osa, joka käsittää tavallisen pintaverkon ulkopuolelle jäävät verkkosivut, kuten käyttäjätietokannat, kirjautumista vaativat keskustelufoorumit tai maksumuurin takana olevat sivustot.
Dread	Pimeän verkon kauppapaikka
Europol	Euroopan poliisivirasto (Europol) on EU:n lainvalvontavirasto, jonka tehtävänä on parantaa Euroopan turvallisuutta ja avustaa lainvalvontaviranomaisia EU-maissa. (EU 2022)
GDS	Global Drug Survey on maailman suurin huumekysely. (GDS 2022)
IOCTA	Internet Organised Crime Threat Assessment on Europolin vuosittain julkaisema raportti EU:n hallituksiin, yrityksiin ja kansallisiin kohdistuvista kyberuhista- ja rikoksista. (Europol 2022)
KRP	Keskusrikospoliisi on Suomen poliisin valtakunnallinen yksikkö.
Kuvalauta	Kuvalauta eli kuvafoorumi on kuvia ja tekstiä sisältävä keskustelukanava, josta käytetään yleensä ilmausta (kuva)lauta. Tunnettuja kuvalautoja ovat esimerkiksi pintaverkon Ylilauta ja

	Reddit sekä pimeän verkon Silkkitie, Rautatieasema ja Dread.
Lanka	Lanka (engl. thread) on kuvalaudan yksittäinen keskusteluketju.
Rautatieasema	Suomalainen pimeän verkon kauppapaikka (suljettu tutkimuksen aikana)
Silk Road	Pimeän verkon kauppapaikka (suljettu)
Silkkitie	Suomalainen pimeän verkon kauppapaikka (suljettu)
Spurdomarket	Suomalainen pimeän verkon kauppapaikka
Pintaverkko	Pintaverkko on kaikille avoin ja näkyvä internetin osa.
THL	Terveyden ja hyvinvoinninlaitos on kansallinen asiantuntijalaitos, joka tarjoaa luotettavaa tietoa terveys- ja hyvinvointialan päätöksenteon ja toiminnan tueksi.
TOR	The Onion Router eli sipulireititin tai -reititys. Toisinaan viitataan tekniseen verkkototeutukseen, toisinaan Tor Projectin Tor-selaimeen, toisinaan käytetään geneerisenä terminä anonyymille verkkoselaukselle.
Tor Project	Voittoa tavoittelematon järjestö, joka on kehittänyt Tor-selaimen. Taistelevat yksityisyyden suojan puolesta ja internet-sensuuria vastaan.
Unicode	Unicode on tietokonejärjestelmiä varten kehitetty merkistöstandardi, joka määrittää yksilöivän koodiarvon yli 100 000 kirjoitusmerkille. (Wikipedia 2022.)

Kuviot

Kuvio 1. Pintaverkossa käyttäjä on yhteydessä web-palvelimiin internetin välityksellä. Syväverkossa ja pimeässä verkossa käyttäjä on yhteydessä palvelimiin ja tietokantoihin Tor-verkon välityksellä. Wikimedia Commons (2012). Pinta- ja syväverkon vertailu. Haettu osoitteesta https://commons.wikimedia.org/wiki/File: Deep_web_vs_surface_web.svg. CC-BY: https://creativecommons.org/licenses/by/3.0/ 8
Kuvio 2. Vasemmalla on käyttäjän näkökulma (engl. owner's intention) internetissä sijaitsevan informaation suhteen. Näkökulma jaetaan julkiseen ja yksityiseen. Ylhäällä on hakukoneiden näkökulma (engl. search engine technique), joka jakautuu indeksöityihin ja indeksöimättömiin näkökulmiin. Esimerkiksi indeksöity pintaverkko (engl. surface web) on käyttäjän näkökulmasta julkinen. Vastaavasti kaikki näkymättömässä verkossa sijaitseva informaatio on piilossa käyttäjän näkökulmasta. Classification of the information on the web julkaisussa *Measuring the Deep and Dark Web* (Kishi ym. 2002). Saatavilla osoitteesta: https://www.ieice.org/publications/conference-FIT-DVDs/FIT2002/pdf/D/LD_3.PDF ... 9
Kuvio 3. Pinta- ja syväverkon vertailu [mukaillen Wikimedia Commons 2018]. Internet on ikään kuin jäävuori, josta vain pieni osa on nähtävillä. Pinnan yllä on pintaverkko ja pinnan alla näkymätön verkko, joka koostuu syvästä ja pimeästä verkosta. https://commons.wikimedia.org/wiki/File:Deepweb_graphical_representation.svg. CC-BY: https://creativecommons.org/licenses/by/3.0/..... 10
Kuvio 4. Tor-verkon toimintaperiaate visualisoituna [mukaillen Wikimedia Commons 2012]. Värikkäät avaimet esittävät istuntoavaimia. Alice lähettää viestin ("kirjekuoressa") ensimmäiselle vastaanottajalle (solmulle). Kirjekuori avataan punaisella avaimella. Kirjekuoren sisällä on toinen kirjekuori, jossa on seuraava osoite, mutta ei vastaanottajan eikä lähettäjän tietoja. Viesti lähetetään seuraavaan osoitteeseen, jossa kirjekuori avataan keltaisella avaimella. Kuoriminen jatkuu siihen asti, kunnes Bob vastaanottaa sisimmän kirjekuoren, jonka hän avaa vihreällä avaimella. Kirjekuori sisältää Alicen viestin. Haettu osoitteesta: https://upload.wikimedia.org/wikipedia/commons/d/dc/Tor-onion-network.png. CC-BY: https://creativecommons.org/licenses/by/3.0/ 12
Kuvio 5. Dark Webin hakutrendi vuodesta 2004 lähtien. Hakujen määrä lisääntyy selvästi vuoden 2014 jälkeen. Google Trends. Dark Web. Viitattu 14.12.2021. Haettu osoitteesta: https://trends.google.fi/trends/explore?date=all&q=dark%20web..... 13
Kuvio 6. CAPTCHA-kello, joka pitää ratkaista aikarajan puitteissa päästäkseen Dread-sivustolle. Viitattu 7.2.2022. Haettu osoitteesta: https://www.reddit.com/r/onions/comments/no4vre/can_somebody_please_explain_how_to_solve_the/..... 15
Kuvio 7. Domain-tiedolla (engl. domain knowledge) on kuvainnollisesti samoja ominaisuuksia kuin kullalla: louhiminen, säilöminen, myyminen ja jatkojalostaminen. J. Cleland-Huang, "Mining Domain Knowledge [Requirements],"in IEEE Software, vol. 32, no. 3, pp. 16-19, May-June 2015, doi: 10.1109/MS.2015.67. Saatavilla osoitteesta: https://10.1109/MS.2015.67 17

Kuvio 8. KDD-prosessin vaiheet. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), 37. Saatavilla osoitteessa: https://doi.org/10.1609/aimag.v17i3.1230 18
Kuvio 9. KDD-prosessin vaiheet erikseen KD- ja DM-prosesseina. Äyrämö, S. (2006). Knowledge mining using robust clustering (No. 63). University of Jyväskylä. Saatavilla osoitteessa: https://jyx.jyu.fi/handle/123456789/13286 19
Kuvio 10. Myynti-ilmoitus Dread-sivustolta 27
Kuvio 11. Huumausaineisiin luokiteltujen merkkijonojen esiintymien suhteelliset jakaumat Dread-keskustelukanavan viesteissä (N ≈ 5120). N pieneni esiprosessoinnin ja analyysin jälkeen 8500:sta 5120:een. 33
Kuvio 12. Huumeiden käytön kehitys 10 kaupungin alueella (Helsinki, Espoo, Turku, Lahti, Lappeenranta, Tampere, Jyväskylä, Kuopio, Oulu ja Rovaniemi lähialueineen) jätevesitutkimuksen perusteella verrattuna positiivisten löydösten lukumäärään epäillyissä huumerattijuopumuksissa vuosina 2012-2020. Viitattu 7.3.2022. Haettu osoitteesta: https://thl.fi/fi/-/jatevesitutkimus-amfetamiinin-kaytto-jatkaa-kasvuaan-suomessa-kokaiinin-kayttomaarien-kasvu-on-tasaantunut .. 36
Kuvio 13. Huumeiden takaisinlasketut käyttömäärät eri kaupunkien jätevedenpuhdistamojen toiminta-alueilla (milligrammaa käytettyä huumetta/1000 henkilöä/päivä) eri keräysviikkoina vuosina 2012-2021. Viitattu 7.2.2022. Haettu osoitteesta: https://thl.fi/fi/tutkimus-ja-kehittaminen/tutkimukset-ja-hankkeet/jatevesitutkimus .. 37

Taulukot

Taulukko 1. Havainnollistava esimerkki huumausaineiden luokittelusta alaryhmittäin. 23
Taulukko 2. Havainnollistava ja mielivaltainen esimerkki huumausaineiden luokittelusta synonyymeittain. Merkkijonot on luokiteltu huumausaineita vastaaviin alaryhmiin 24
Taulukko 3. Sekaannusmatriisi 30
Taulukko 4. Tutkimuksessa käytetyn luokittelija-algoritmin sekaannusmatriisit. Vasemmalla ylhäällä matriisin matemaattinen esitysmuoto, oikealla ylhäällä sama matriisi binääriluokitteluna ja keskellä alhaalla sama matriisi havaintojen kanssa. .. 30

Sisällys

1 JOHDANTO 1

2 TUTKIMUKSEN TAVOITE 4
2.1 Aihepiirin rajaus ja tutkimuskysymykset 4
2.2 Tutkimuksen toteutus 5

3 KESKEISET KÄSITTEET 6
3.1 Kyberrikollisuus 6
3.2 Pinta- ja syväverkko 7
3.3 Tor-verkko 11
3.4 Pimeä verkko 13
3.5 Pimeän verkon kauppapaikat 14
3.6 Tiedonlouhinta 16
3.7 Luonnollisen kielen käsittely 20

4 AINEISTON KERUU JA ANALYYSI 22
4.1 Käytännön vaiheet 25
4.2 Virheanalyysi 29
4.3 Tunnusluvut 31

5 TULOKSET 33

6 JOHTOPÄÄTÖKSET 35

7 POHDINTA 40

8 YHTEEVETO 42

LÄHTEET 44

LIITTEET 53
A Huumausaineiden suhteelliset jakaumat Dread-keskustelukanavan viesteissä. 53

1 Johdanto

Suurin osa syvän www:n verkkokauppojen markkinoilla tapahtuvasta myynnistä liittyy huumausaineisiin. Euroopan huumausaineiden ja niiden väärinkäytön seurantakeskus EMCDDA:n selvityksessä (EMCDDA 2016), jossa tutkittiin 16 suuren syvän verkon markkinoiden myyntiä vuosina 2011–2015, arvioitiin, että huumeiden myynnin osuus oli yli 90 prosenttia maailmanlaajuisten syvän verkon kauppapaikkojen kokonaistuloista. Lähes puolet (46 %) koko syvän verkon huumemyynnistä oli ilmoitusten mukaan peräisin Euroopassa toimivilta myyjiltä, ja sen arvo oli noin 80 miljoonaa euroa tutkimuksen aikana. Tärkeimmät eurooppalaiset lähdemaat myyntimäärien perusteella olivat Saksa, Alankomaat ja Iso-Britannia, ja suurin osa myyntituloista tuli erilaisten stimulanttien, erityisesti ekstaasin ja kokaiinin, myynnistä.

Verkon kryptomarkkinat tarjoavat huumausaineiden myyjille ja ostajille helpon ja nopean infrastruktuurin rikollisten liiketoimien tekemistä ja palveluiden suorittamista varten. Näillä laittomilla markkinoilla on samankaltaisia piirteitä kuin esimerkiksi eBayn ja Amazonin kaltaisilla laillisilla verkkomarkkinapaikoilla, joissa asiakkaat voivat etsiä ja vertailla tuotteita ja myyjiä keskustelufoorumeilla. Kansainvälisillä kryptomarkkinoilla on havaittu suomalaisten myyjien runsasta toimeliaisuutta: suomalaiset kauppaavat etenkin marihuanaa, jota saadaan ilmeisesti omista kannabiskasviviljelyksistä.

Huolestuttava kehityssuunta liittyy myös huumausaineiden tarjontaan ja huumeiden tai huumekokemusten jakamiseen sosiaalisen median, tätä nykyä erityisesti mobiilisovellusten, kautta. Tätä alaa on vaikea valvoa, ja salattujen verkkojen ja pikaviestinten sekä sosiaalisen median käyttö lisääntyy jatkuvasti. (THL 2020 [EMCDDA 2016; KRP 2020].)

THL:n raportissa puhutaan pimeästä verkosta (engl. dark web), joka on tullut tunnetuksi juuri huumausaineiden ja muiden laittomuuksien kauppapaikkana (Todorof 2019; Hatta 2020). Maailmalla yleisesti tunnettuja kauppapaikkoja ovat Silk Road (2.0), Hydra and Cloud 9. Suomessa yleisesti tunnettuja kauppapaikkoja olivat Sipulikanava ja Silkkitie, jotka suljettiin Suomen tullin toimesta vuosina 2017 ja 2019. (Tulli 2017, 2019) Vuoden 2021 keväällä tilalle tuli Rautatieasema-niminen kuvalauta (Rautatieasema 2020), joka kuitenkin hävisi

verkosta tämän tutkimuksen tekoaikana. Spurdomarket oli ainoa suomalainen kauppapaikka joka jäi verkkoon tutkimuksen tekoajaksi. Sitä ei kuitenkaan sisälletty itse tutkimukseen, koska kyseessä oli enemmän verkkokauppa kuin kuvalauta. Suurin osa kaupankäynnistä siirtyi kansainvälisen Dread-sivuston Suomi-langalle.

Tämän tutkimuksen aiheena on tutustua data-analytiikan keinoin pimeän verkon huumekauppaan Suomessa. Yhteiskunnallisesti aihe on merkittävä ja ajankohtainen. Poliisi kertoi tiedotteessaan syksyllä 2021, että Keskusrikospoliisissa on aloitettu pimeän verkon seurannan projekti. (Poliisi 2021) Lisäksi YLE uutisoi syksyllä 2021 kansainvälisestä huumeiden käyttöä koskevasta tutkimuksesta, jonka mukaan pimeän verkon huumekauppa on verrattain yleistä. (YLE 2021 [Global Drug Survey 2019]) GDS:n (2019) mukaan niistä lähes 2 200 suomalaisesta vastaajasta, jotka olivat käyttäneet tai ostaneet huumeita viimeisen vuoden aikana, 45 prosenttia oli hankkinut huumeita pimeästä verkosta. (Global Drug Survey 2019)

Kansainvälisesti aihetta on tutkittu varsinkin Silkkitien suhteen (Christin 2013; Martin 2014; Van Hout ja Bingham 2014; Broséus ym. 2016; Bakken ja Demant 2019). Suomessa aihetta on tutkittu varsin laajasti Juha Nurmen toimesta (Nurmi ja Kaskela 2015; Nurmi ym. 2017; Nurmi 2019), mutta muuten aihetta on tutkittu varsin vähän suomalaisessa viitekehyksessä. Nurmen väitöskirja (Nurmi 2019) tarkastelee, kuinka Tor-verkossa tapahtuvaa rikollisuutta on mahdollista tutkia ja kuinka rikolliset markkinat toimivat. Pimeän verkon huumekauppaa on myös käsitelty informaatiotutkimuksen (Haasio ja Harviainen 2018; Hakalahti 2019; Haasio, Harviainen ja Savolainen 2020) sekä suoraan Poliisin (Mattila 2020) kuin huumekauppiaiden näkökulmasta (Räisänen 2020). Lisäksi Nurmi ym. (Nurmi ja Kaskela 2015; Nurmi ym. 2017) ovat tutkineet aihetta nimenomaan Silkkitien suhteen.

Tuoreinta tutkimustietoa edustaa syksyllä 2021 ilmestynyt Mikko Ruotsalaisen maisteritutkielma *Rikollinen kaupankäynti Tor-verkon suomenkielisissä piilopalveluissa*, joka käsittelee aihetta samasta viitekehyksestä kuin tämä tutkimus. Ruotsalaisen tutkimuksesta poiketen, tämä tutkimus on toteutettu teknisemmin omaa algoritmia ja koodia hyödyntäen. Lisäksi tutkimus käsittelee aihetta tarkemmin pelkän huumekaupan suhteen. Tutkimus tarjoaa ajankohtaisen katsauksen suomalaisen pimeän verkon nykytilasta ja vahvistaa aikaisempia tutkimustuloksia.

Seuraavassa luvussa esitellään tutkimuksen tavoite, rajataan aihepiiri sekä esitellään tutkimuskysymykset ja tutkimuksen toteutus. Kolmannessa luvussa esitellään keskeisimmät käsitteet tutkimuksen ymmärtämisen kannalta. Keskeisimmät käsitteet ovat kyberrikollisuus, pinta- ja syväverkko, Tor-verkko, pimeä verkko, pimeän verkon kauppapaikat, tiedonlouhinta ja luonnollisen kielen käsittely. Neljännessä luvussa paneudutaan itse asiaan aineistonkeruun ja analyysin muodossa. Luvun alakappaleissa avataan tutkimuksen käytännön vaiheet, esitellään virheanalyysi ja lasketut tunnusluvut. Lopuksi tulokset, johtopäätökset ja pohdinta esitellään erillisinä lukuinaan.

2 Tutkimuksen tavoite

Seuraavassa alakappaleessa käydään läpi tutkimuksen tavoite, rajataan aihepiiri tietyin reunaehdoin ja esitellään tutkimusongelma. Toisessa alakappaleessa käydään lyhyesti läpi tutkimuksen toteutuksen kannalta oleellisimmat asiat.

2.1 Aihepiirin rajaus ja tutkimuskysymykset

Tutkimuksen tavoitteena on arvioida huumekaupan nykytilaa pimeässä verkossa suomalaisessa viitekehyksessä. Aineisto kerätään tiedonlouhinnan keinoin Dread-sivuston Suomilangalta ja sen pohjalta laaditaan tilastollinen analyysi luokittelija-algoritmia ja sekaannusmatriiseja hyödyntäen.

Tutkimuksen aihe on rajattu suomalaiseen viitekehykseen, mikä on perusteltua, koska aihetta on tutkittu Suomessa melko vähän. Tutkimuksen ulkopuolelle jätetään kaikki huumausaineisiin liittymätön kaupankäynti (kuten aseet ja doping) tiedonlouhinnan helpottamiseksi ja analyysin selkeyttämiseksi.

Tarkemmat tutkimuskysymykset ovat:

- Mitä huumausaineita Dread-sivustolla esiintyy?
- Mikä on huumausaineiden jakauma?
- Ketkä niitä myyvät?
- Missä niitä myydään?
- Nouseeko datasta anomalioita?

Tutkimuskysymykset ovat varsin triviaaleja ja ne ovat linjassa aikaisempien tutkimuskysymysten kanssa (Baravalle, Lopez ja Lee 2016; Brooks 2018; Alkhatib ja Basheer 2019). Tutkimus tarjoaa tietoa huumausaineiden esiintyvyydestä, mikä voi auttaa rikosten ennaltaehkäisyssä ja torjunnassa. Toisaalta tutkimus tarjoaa epäsuorasti ideoita huumekaupan edistämiseksi rikkollisesta näkökulmasta: Kuinka peittää jäljet? Kuinka hankaloittaa seurantaa? Kuinka suojata anonymiteetti? Lisäksi tutkimus tarjoaa toimintatapoja ja -periaatteita data-analyysin toteuttamiseen sekä jatkotutkimusaiheita.

2.2 Tutkimuksen toteutus

Tutkimus noudattelee tapaustutkimuksen piirteitä. Tapaustutkimus on tutkimusstrategia, jossa tarkoituksena on tutkia syvällisesti vain yhtä tai muutamaa kohdetta tai ilmiökokonaisuutta. (JYU 2015b)

Tapaustutkimukselle tyypillisiä valintakriteerejä ovat (Oinas 2011):

- Kun kysytään miten, miksi, kuinka
- Kun tutkijalla on vain vähän kontrollia tapahtumiin
- Kun aiheesta on tehty vain vähän empiiristä tutkimusta
- Kun tutkimuksen kohteena on jokin ajassa ja paikassa oleva/tapahtuva ilmiö

Näistä ainakin kolme kriteeriä täyttyy suoraan. Ensimmäiseksi tutkijalla on vain vähän kontrollia tapahtumiin. Rautatieasemalle tuli päivittäin uutta dataa ja sivusto oli välillä alhaalla, mikä vaikeutti aineiston keruuta. Lisäksi aina on olemassa mahdollisuus, että sivusto suljetaan kuten kävi Sipulikanavan, Silkkitien ja nyt Rautatieaseman kanssa. Nämä haasteet liittyvät yleisesti kaikkiin pimeän verkon sivustoihin. Toiseksi aiheesta on tehty vain vähän empiiristä tutkimusta, mikä todettiin myös johdannossa. Kolmanneksi tutkimuskohteena on jokin ajassa ja paikassa oleva/tapahtuma ilmiö, joka on tässä tapauksessa Suomen (paikka) pimeän verkon (ilmiö) huumekauppa (tapahtuma). Tutkimuksessa ei kuitenkaan suoraan haeta vastausta kysymyksiin miten, miksi tai kuinka.

Tapaustutkimuksen näkemys on holistinen (KAMK 2021; UIAH 2021), ja tavallisesti tapaustutkimus on luonteeltaan toteava eli deskriptiivinen (UIAH 2021). Tapaustutkimuksessa halutaan siis kuvailla kohdetta syvällisemmin eikä pelkästään sen ulkoisia ominaisuuksia vaan myös kohteen piilotettuja rakenteita. (UIAH 2021) On myös tyypillistä, että tiedonkeruu ja analyysi eivät rajoitu vain yhteen menetelmään, vaan tutkimuksessa voidaan hyödyntää sekä kvantitatiivisia että kvalitatiivisia menetelmiä. (KvaliMOTV 2021) Tapaustutkimus ei siis ole aineistonkeruutekniikka.

3 Keskeiset käsitteet

Tässä kappaleessa käydään läpi tutkimuksen ymmärtämisen kannalta oleellisimmat käsitteet. Ensimmäiseksi tutustutaan lyhyesti kyberrikollisuuden määritelmään, koska tutkimus käsittelee rikollista toimintaa. Tämän jälkeen siirrytään alakappaleisiin, jotka auttavat ymmärtään paremmin verkkorikollisuuden historiaa ja teknisiä toimintaperiaatteita. Lisäksi tarkastellaan tiedonlouhinnan ja luonnollisen kielen käsittelyn perusperiaatteita, joita varsinkin tutkimuksen aineistonkeruu ja analyysi noudattavat.

3.1 Kyberrikollisuus

Tietoverkosta on tullut huumausainerikollisuuden keskeinen toimintaympäristö, jossa tarjolla olevia välineitä ja mahdollisuuksia hyödynnetään erittäin monipuolisesti ja runsaasti. Tietoverkon hämäräfoorumit muuttavat myös esimerkiksi huumausainemarkkinoiden ilmiöitä yhä laaja-alaisemmiksi ja pirstaleisemmiksi ja kotimaiset ja kansainväliset markkinat liittyvät toisiinsa yhä saumattomammin.

Nettikaupasta on tullut merkittävä huumausaineiden hankinta- ja välityskanava, joka vaikuttaa jo perustavanlaatuisesti markkinoiden rakenteisiin ja valikoimiin. Nuoria aikuisia toimii verkon kautta tapahtuvassa huumausainekaupassa helpon rahan toivossa ilman, että heillä olisi merkittävää rikollista taustaa. Toisaalta verkkovähittäismyyntiä harjoittavat merkittävässä määrin myös ammattimaisesti toimivat järjestäytyneet rikollistahot. (THL 2020 [KRP 2020])

Kyberrikollisuus tarkoittaa verkossa tapahtuvaa tai sitä hyödyntävää rikollisuutta (Kolttola 2020; Rikoksentorjuntaneuvosto 2022; Sisäministeriö 2022), joka on usein myös kansainvälistä (Kolttola 2020; Rikoksentorjuntaneuvosto 2022; Sisäministeriö 2022). Kyberrikollisuus voidaan jakaa kahteen luokkaan (Kolttola 2020; Poliisi 2022):

- Tietoverkkosidonnaisiin ja tietotekniikkaan kohdistuviin rikoksiin
- Tietoverkko- ja tietokoneavusteisiin rikoksiin

Poliisin mukaan: ”Tietoverkkosidonnaiset rikokset kohdistuvat tietoverkkoihin ja tietojär-

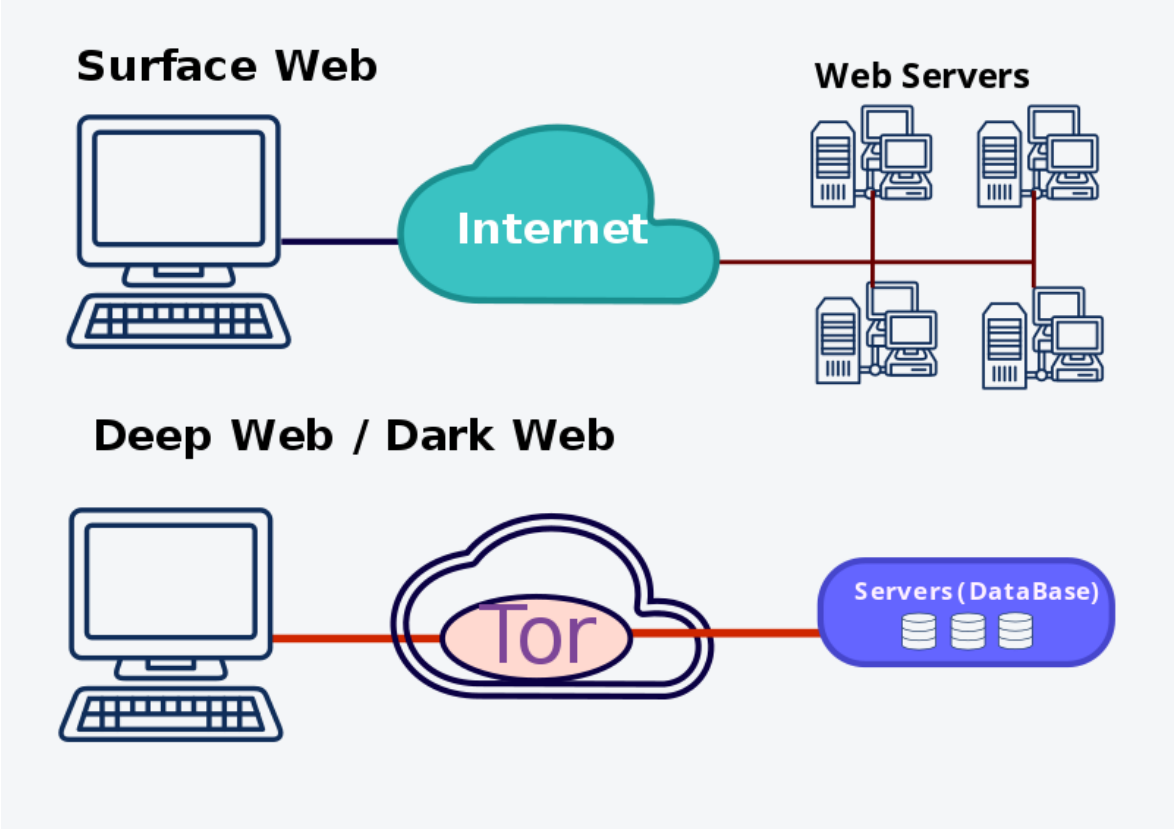
jestelmiin. Rikoksen tekeminen on mahdollista ainoastaan tietokoneita ja tietoverkkoja käyttäen. Tällaisia tekoja ovat esimerkiksi tietojärjestelmän häirintä eli palvelunestohyökkäykset, tietomurrot tai datavahingonteko.” (Poliisi 2022)

Tietoverkkoavusteisissa rikoksissa sen sijaan: ”Hyödynnetään tietoverkkoja tai tietojärjestelmiä osana rikoksen tekemistä. Kyse on - - perinteisestä rikollisuudesta kuten petoksista, huumausainerikollisuudesta, rahanpesusta, mutta joiden toteuttamiseen tietoverkot tuovat uusia tekotapoja. Toisin sanoen tietoverkot mahdollistavat rikoksen tekemisen tietoverkkojen avulla, mutta teko ei kohdistu tietoverkkoon tai tietojärjestelmään.” (Poliisi 2022)

Määritelmän mukaan tämä tutkimus siis tarkastelee tietoverkkoavusteista rikollisuutta. Tässä kontekstissa sillä tarkoitetaan modernia huumekauppaa, joka hyödyntää muunmuassa Torverkkoa, pikaviestimiä sekä kryptovaluuttoja.

3.2 Pinta- ja syväverkko

Internet koostuu pintaverkosta ja syväverkosta. Pintaverkko on internetin käyttäjille avoin ja perinteisten hakukoneiden (Google, Yahoo, Bing ym.) saavutettavissa oleva osa. Pintaverkon alla sijaitsee syväverkko, jota hakukoneet eivät indeksoi. (Bergman 2001; McKiernan 2001; NordVPN 2021b) Toisin sanoen syväverkko on perinteisten hakukoneiden saavuttamattomissa. Verkkojen ero on esitetty kuviossa 1.



Kuvio 1: Pintaverkossa käyttäjä on yhteydessä web-palvelimiin internetin välityksellä. Syväverkossa ja pimeässä verkossa käyttäjä on yhteydessä palvelimiin ja tietokantoihin Tor-verkon välityksellä. Wikimedia Commons (2012). Pinta- ja syväverkon vertailu. Haettu osoitteesta https://commons.wikimedia.org/wiki/File:Deep_web_vs_surface_web.svg. CC-BY: https://creativecommons.org/licenses/by/3.0/

Pimeä verkko on puolestaan osa syvää verkkoa. Kyseessä on joukko piilotettuja verkkosivustoja, joihin pääsee salatulla yhteydellä Tor-verkon avulla. (NordVPN 2021a; Kaspersky 2021; Tor Project 2021) Kishi ym. (2002) määrittelevät pimeän verkon tarkemmin suhteessa muihin verkkoihin (kuvio 2). Syväverkon ja vastaavasti pimeän verkon sivustot jäävät siis indeksöimättöminä perinteisten hakukoneiden ulkopuolelle.

<table>
<tr><td colspan="2"></td><td colspan="3">Search Engine Technique</td></tr>
<tr><td colspan="2"></td><td>Indexable</td><td></td><td>Non-Indexable</td></tr>
<tr><td rowspan="2">Owner's Intention</td><td>Public</td><td>Surface Web</td><td>Deep Web</td><td>Dark Web</td></tr>
<tr><td>Private</td><td colspan="3">Invisible Web</td></tr>
</table>

Kuvio 2: Vasemmalla on käyttäjän näkökulma (engl. owner's intention) internetissä sijaitsevan informaation suhteen. Näkökulma jaetaan julkiseen ja yksityiseen. Ylhäällä on hakukoneiden näkökulma (engl. search engine technique), joka jakautuu indeksöityihin ja indeksöimättömiin näkökulmiin. Esimerkiksi indeksöity pintaverkko (engl. surface web) on käyttäjän näkökulmasta julkinen. Vastaavasti kaikki näkymättömässä verkossa sijaitseva informaatio on piilossa käyttäjän näkökulmasta. Classification of the information on the web julkaisussa *Measuring the Deep and Dark Web* (Kishi ym. 2002). Saatavilla osoitteesta: https://www.ieice.org/publications/conference-FIT-DVDs/FIT2002/pdf/D/LD_3.PDF

Kishin ym. (2002) mukaan internet voidaan siis jakaa kolmeen osaan:

1. Näkymättömään verkkoon, jonka sisältö halutaan piilottaa hakukoneissa sisällön tuottajan puolesta.
2. Syvään verkkoon, jonka sisältö halutaan näyttää hakukoneissa sisällön tuottajan mutta ei hakukoneiden puolesta.
3. Pimeään verkkoon, jonka sisältö halutaan näyttää hakukoneissa sisällön tuottajan ja hakukoneiden puolesta, mutta joka ei ole mahdollista teknisistä syistä.

Kishin ym. (2002) määritelmä on kuitenkin hieman vanhanaikainen, koska syvällä verkol-

la tarkoitetaan nykyään juuri näkymätöntä verkkoa. Syväverkko ja pimeä verkko ovat siis näkymättömän verkon osa (kuvio 3). Siitä huolimatta Kishi ym. (2002) kuvaavat hyvin verkon saavutettavuutta juuri käyttäjän ja sivustojen näkyvyyden suhteen, vaikka arkkitehtuuri hahmotetaankin nykyään hieman eri tavoin.

Kuvio 3: Pinta- ja syväverkon vertailu [mukaillen Wikimedia Commons 2018]. Internet on ikään kuin jäävuori, josta vain pieni osa on nähtävillä. Pinnan yllä on pintaverkko ja pinnan alla näkymätön verkko, joka koostuu syvästä ja pimeästä verkosta. https://commons.wikimedia.org/wiki/File:Deepweb_graphical_representation.svg. CC-BY: https://creativecommons.org/licenses/by/3.0/

Syväverkon sivustot ovat siis implisiittisesti määriteltyjä ja ne yleensä sijaitsevat erilaisten rajoitteiden (maksumuuri, palomuuri, salasanat, lisenssi, hakukone-esto) takana. (Bergman 2001; McKiernan 2001) Syväverkko sisältää myös tietokantoja, sekä esimerkiksi yritysten tai muiden organisaatioiden sisäisiä verkkoja. (NordVPN 2021b)

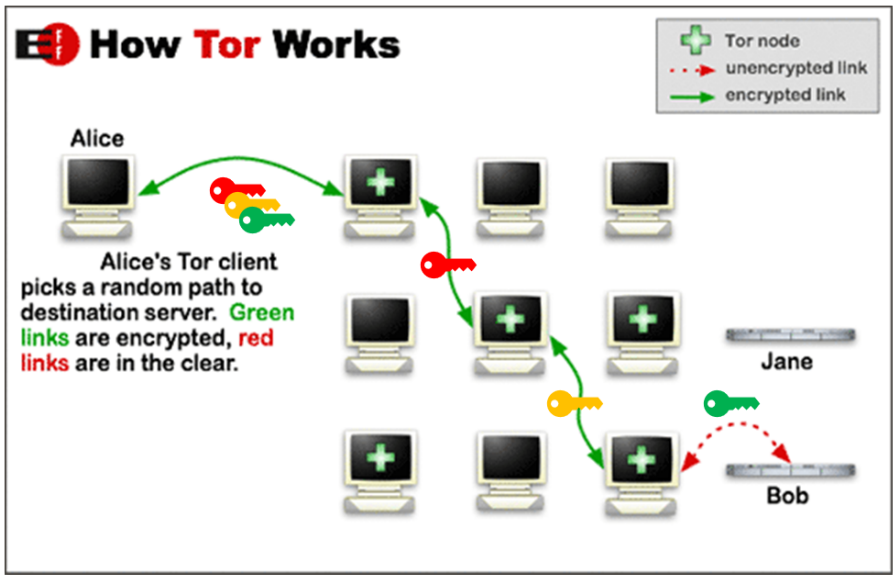
Syväverkko on lähteiden valossa esitelty ensimmäisen kerran vuonna 1996. Bruce Mount ja Matthew B. Koll käyttivät määritelmää näkymätön verkko (engl. Invisible Web) lehdistötiedotteessaan, jossa esiteltiin ensimmäisen sukupolven hakukone AT1. Tuolloin AT1 mahdollisti ”piilotettujen” tietokantojen etsimisen, jotka eivät olleet nähtävissä tietokantaohjelmilla (engl. web spider). (Personal Library Software 1996) Toisin sanoen AT1:n avulla oli mahdollista päästä käsiksi näkymättömän verkon sisältöihin. Tiedeyhteisölle näkymätön verkko esiteltiin syväverkkona ensimmäisen kerran Michael K. Bergmanin julkaisussa *The Deep Web: Surfacing Hidden Value* vuonna 2001.

3.3 Tor-verkko

Tor-reititys eli sipulireititys (engl. The Onion Routing) esiteltiin ensimmäisen kerran tiedeyhteisölle julkaisussa *Tor: The Second-Generation Onion Router*. Sipulireititys määritellään samassa julkaisussa hajautetuksi peiteverkoksi (engl. distributed overlay network), joka on suunniteltu anonymisoimaan TCP-yhteyksiä kuten internet-selausta, salatun liikenteen protokollaa (engl. secure shell, SSH) ja pikaviestejä.

Sipulireitityksessä asiakas valitsee polun verkosta ja muodostaa polun/piirin, jossa jokainen yhteyspiste (engl. node) tietää edeltäjän ja seuraajan, mutta ei muita yhteyspisteitä. Liikenne suuntaa polkua alaspäin vakiokokoisten solmujen kautta, jotka puretaan symmetrisellä avaimella jokaisessa yhteyspisteessä ja lähetetään eteenpäin. Tämä muistuttaa sipulin kuormista. (Syverson, Dingledine ja Mathewson 2004.)

Kuviossa 4 on visualisoitu Tor-verkon toimintaperiaate. Siinä Alicen Tor-asiakas (engl. Tor client) valitsee satunnaisen polun kohdepalvelimelle. Vihreät polut ovat salattuja ja punaiset salaamattomia. Sipulivertaus ilmenee Tor-solmujen välissä, joissa salausavainten määrä vähenee: Ensimmäisessä yhteypisteessä luovutaan punaisesta avaimesta, toisessa keltaisesta ja viimeisessä vihreästä.



Kuvio 4: Tor-verkon toimintaperiaate visualisoituna [mukaillen Wikimedia Commons 2012]. Värikkäät avaimet esittävät istuntoavaimia. Alice lähettää viestin ("kirjekuoressa") ensimmäiselle vastaanottajalle (solmulle). Kirjekuori avataan punaisella avaimella. Kirjekuoren sisällä on toinen kirjekuori, jossa on seuraava osoite, mutta ei vastaanottajan eikä lähettäjän tietoja. Viesti lähetetään seuraavaan osoitteeseen, jossa kirjekuori avataan keltaisella avaimella. Kuoriminen jatkuu siihen asti, kunnes Bob vastaanottaa sisimmän kirjekuoren, jonka hän avaa vihreällä avaimella. Kirjekuori sisältää Alicen viestin. Haettu osoitteesta: https://upload.wikimedia.org/wikipedia/commons/d/dc/Tor-onion-network.png. CC-BY: https://creativecommons.org/licenses/by/3.0/

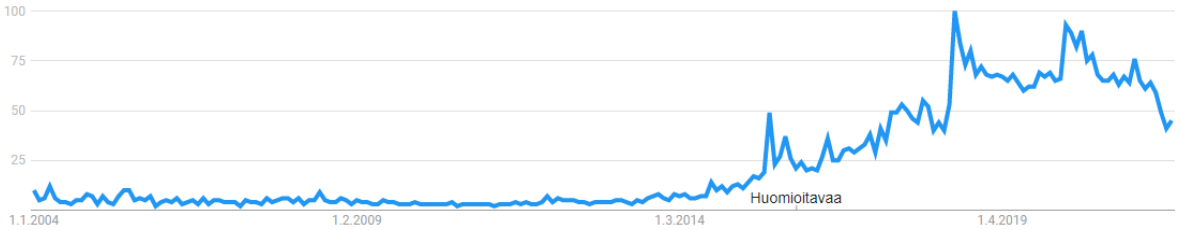
Tor-verkko ei siis ole internetistä irrallaan oleva erillinen verkko (Ruotsalainen M. 2021), vaan muodostuu verkkoon liitetyistä päätelaitteista, jotka toimivat virtuaalisen piirikytketyn verkon solmulaitteina. Tor-verkossa sijaitsevia solmulaitteita kutsutaan sipulireitittimiksi (engl. onion routers), vaikka käytännössä kyse ei ole solmulaitteiden reitityksestä. (Viinikainen 2021; Syverson, Dingledine ja Mathewson 2004.) Alkuperäisen yhteyden paikallistaminen on siis erittäin monimutkaista yhteyspisteiden suuren kappalemäärän vuoksi. Näin Tor-verkko yhdessä Tor-selaimen kanssa mahdollistaa verkkopalveluiden anonyymin käytön.

Tor käsitteenä sekoitetaan joskus piilo/sipulipalveluihin (engl. hidden/onion services). Piilopalvelut ovat kuitenkin verkkopalveluita tai -sivustoja, jotka hyödyntävät Tor-protokollaa. (Tor Project 2015.)

3.4 Pimeä verkko

Pimeä verkon syntyajasta ei ole varmaa tietoa, mutta tiedeyhteisölle käsite pimeä verkko (engl. dark web) on esitelty lähteiden valossa ensimmäisen kerran julkaisussa *Measuring the Deep and Dark Web* (Kishi ym. 2002). Tämä on myös linjassa Tor-projektin historiikkin kanssa, jonka mukaan Tor-verkon lähdekoodi julkaistiin avoimena lähdekoodina vuonna 2002. (Tor Project 2021)

Pimeän verkon käsite on yleistynyt viime vuosina. Vuonna 2011 perustettiin maailman tunnetuin pimeän verkon kauppapaikka Silk Road [ei Suomen Silkkitie] bitcoinin keksimisen vanavedessä (Hatta 2020), joka keksittiin joko vuonna 2008 tai 2009 (Ron ja Shamir 2012; Karishma ja Deshmukh 2017; Harihara Sudhan ym. 2018; Saito ja Iwamura 2019). Tätä ennen pimeän verkon käyttö oli varsin marginaalista. On kuitenkin mielenkiintoista, että yleistyminen näkyy vasta vuodesta 2014 lähtien (kuvio 5). Hakujen lukumäärä on todennäköisesti ollut tuohon asti varsin marginaalinen ja/tai suurin osa niistä on tehty Googlen ulkopuolella.



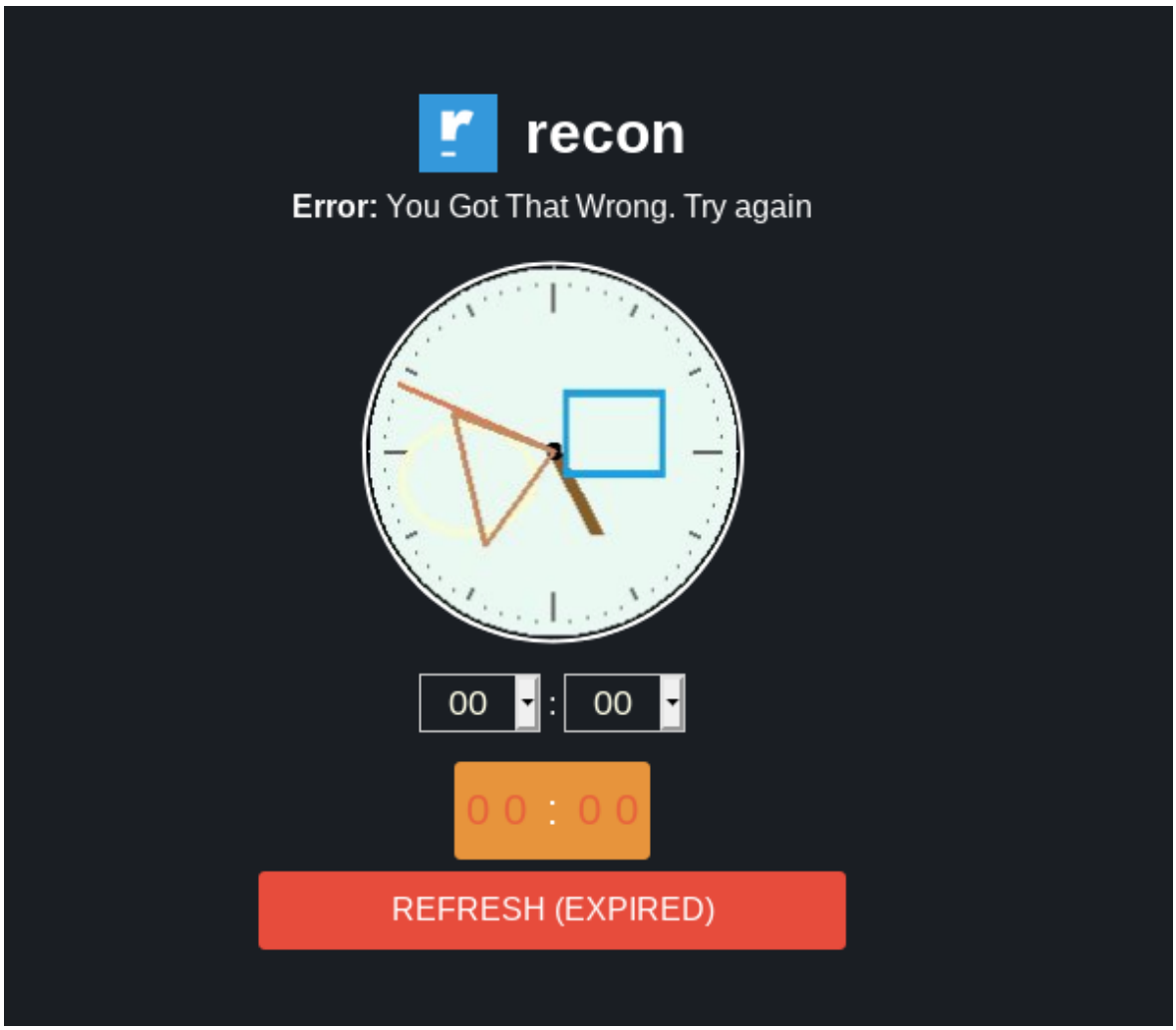
Kuvio 5: Dark Webin hakutrendi vuodesta 2004 lähtien. Hakujen määrä lisääntyy selvästi vuoden 2014 jälkeen. Google Trends. Dark Web. Viitattu 14.12.2021. Haettu osoitteesta: https://trends.google.fi/trends/explore?date=all&q=dark%20web

3.5 Pimeän verkon kauppapaikat

Europolin Internet Organised Crime Threat Assessment eli IOCTA arvioi internetin järjestäytyneen rikollisuuden uhkia vuosittain. IOCTA 2020 (Europol 2020) mukaan suurten kauppapaikkojen lukumäärä on pienentynyt siinä missä pienten lukumäärä on kasvanut. Pienemmistä kauppapaikoista on tullut myös ketterämpiä. (Europol 2020) Esimerkiksi suoraveloituksiin perustuva kaupankäynti (engl. direct deposit / wallet-less market) on yleistynyt. Kaupankäynti tapahtuu yleensä kryptovaluuttojen suoraveloituksella, jossa kaupankäynti tapahtuu myyjän ja asiakkaan välillä ilman välikäsiä tai lompakoita. (Darknetlive 2021.)

Nykyään kauppapaikkoja syntyy melkein sitä mukaan, kun niitä suljetaan. Tämä on osaltaan johtanut Spurdomarketin kaltaisten, suoraveloituksiin perustuvien kauppapaikkojen lisääntymiseen. Oletettavasti kauppapaikkojen lyhyt elinkaari hankaloittaa viranomaisten tekemää työtä. IOCTA 2021 mukaan myös Telegrammin ja Wickrin kaltaiset vahvan salauksen applikaatiot hankaloittavat viranomaistoimintaa, mikä osaltaan pakottaa tutkimaan muitakin alustoja pelkän pimeän verkon sijaan. (Europol 2021)

Globaalit trendit rantautuvat Suomeen tunnetusti hieman viiveellä. Pimeän verkon tapauksessa globaaleiksi trendeiksi voidaan luetella esimerkiksi vahvaan tunnistautumiseen perustuvien kauppapaikkojen määrällinen kasvu sekä Telegrammin ja Wickrin kaltaisten applikaatioiden suosiminen. Tutkimuksen alussa esimerkiksi Spurdomarket oli saavutettavissa ilman sisäänkirjautumista, mutta myöhemmin tutkimuksen aikana sivusto suljettiin rekisteröitymisen taakse. Turingin testejä esiintyy varsinkin suosituimmilla kauppapaikoilla. Esimerkiksi Dread-sivusto oli kelloon perustuvan testin takana (CAPTCHA-kello, kuvio 6).



Kuvio 6: CAPTCHA-kello, joka pitää ratkaista aikarajan puitteissa päästäkseen Dread-sivustolle. Viitattu 7.2.2022. Haettu osoitteesta: https://www.reddit.com/r/onions/comments/no4vre/can_somebody_please_explain_how_to_solve_the/

Keskustelupalstojen perusteella testit tuntuvat yleisesti häiritsevän ja hankaloittavaan käyttäjäkokemusta (Reddit 2022) turvallisuuden lisäämisestä huolimatta. CAPTCHA-testien käyttäjäkokemuksia vertailevassa tutkimuksessa (Gafni ja Nagar 2016) todettiin, että useimmat käyttäjistä eivät pitäneet CAPTCHA-testien tekemisestä. Käyttäjät kyllä ymmärsivät niiden merkityksen, mutta eivät siitä huolimatta kokeneet testien antavan lisäarvoa turvallisuuden kokemukselle. (Gafni ja Nagar 2016)

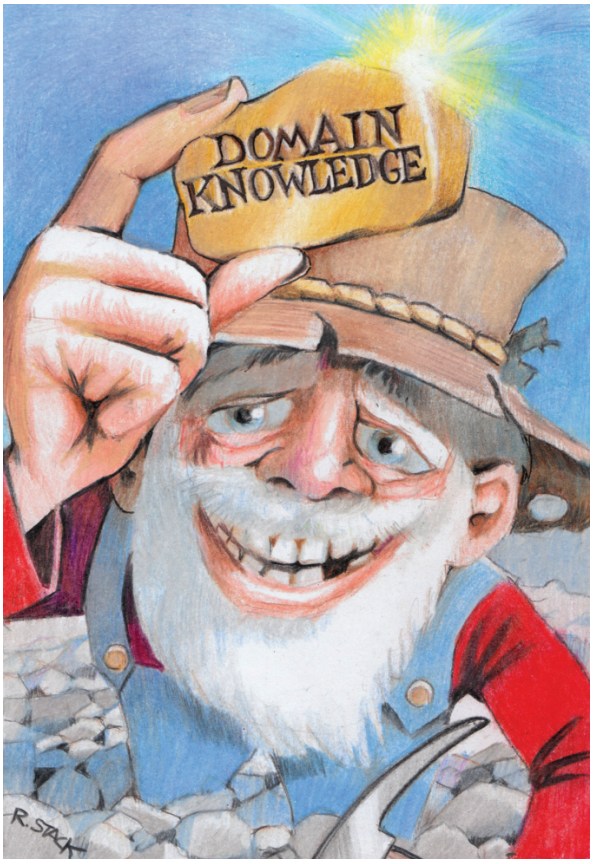
Vaivattomuus sekä nopeus ovat käyttäjäkokemukseen vaikuttava tekijöitä, ja CAPTCHA-testien ratkaiseminen koettiinkin aina jossain määrin negatiivisena (Gafni ja Nagar 2016; Tanthavech ja Nimkoompai 2019). Huumekaupan teknistyminen tuskin vähentää rikollista toimintaa – päinvastoin rikolliset ja uudet tekijät todennäköisesti keksivät vielä luovempia ratkaisuja kaupankäyntiin. THL:n huumeraportissa 2020 todetaankin: ”— *esitutkinnoissa kertyneiden tietojen valossa on arvioitavissa, että ainakin osa Silkkitien huumausaineiden myyjistä on sellaisia, että he tuskin pärjäisivät perinteisessä kovan väkivallan uhkan leimaamassa huumeiden katukaupassa. Näiden ”nörttimyyjien” vahva osaaminen painottuu enemmän tietotekniselle puolelle.*” (THL 2022 [KRP 2020])

3.6 Tiedonlouhinta

Pimeän verkon tiedonlouhintaa on tutkittu jonkin verran ulkomaisessa kontekstissa (Chen ym. 2008; Chen 2011; Baravalle, Lopez ja Lee 2016; Alkhatib ja Basheer 2019). Englanninkielisessä lähdemateriaalissa puhutaan usein aihepiiri-/toimialakohtaisesta tiedosta (engl. domain knowledge, jatkossa domain-tieto) tiedonlouhinnan yhteydessä. Suomeksi käsitteelle ei löydy hyvää määritelmää, mutta englanniksi se voidaan määritellä jonkin tietyn toimialan rajatuksi *ymmärrykseksi* (engl. knowledge). (Alexander ja Judy 1988) Domain-tieto tieto voidaan käsittää selittävänä (mitä), menetelmällisenä (miten) ja ehdollisena (milloin ja miksi) ymmärryksenä, ja se voi ilmetä eksplisiittisenä tai hiljaisena tietona. (Alexander 1992 [Ryle 1949Paris, Lipson ja Wixson 1983Alexander, Schallert ja Hare 1991]) Tiivistetysti voidaan ajatella, että domain-tieto on jonkin alan tai aihepiirin syvällistä ja/tai kokonaisvaltaista ymmärtämistä.

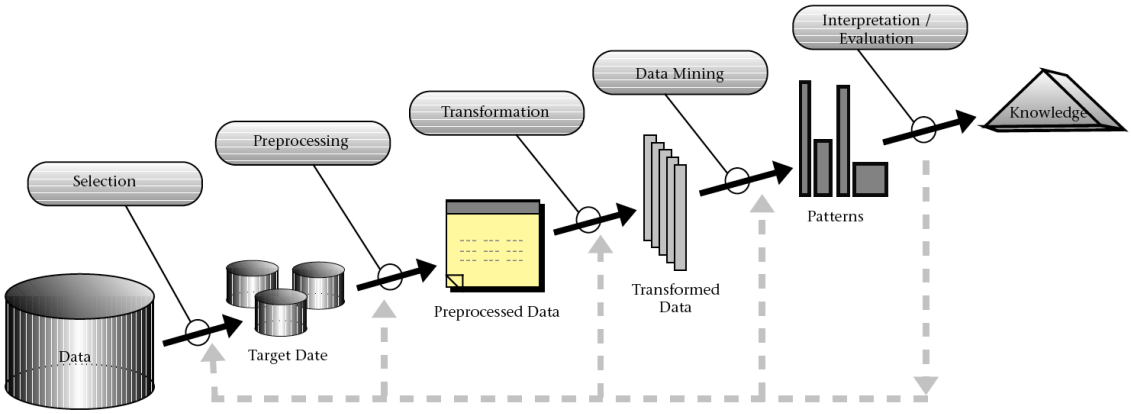
SAS Instituten määritelmän mukaan tiedonlouhinta on ”poikkeamien, mallien ja korrelaa-

tioiden löytämistä massadatasta” (SAS 2022). Käytännössä kyse on usein raakadatan jalostamisesta hyödynnettävään / ennustettavaan muotoon. (TechTarget 2022, Investopedia 2022, SAS 2022) Toisin sanoen tiedonlouhinnan avulla pyritään pääsemään käsiksi domain-tietoon. Kuvio 7 kiteyttää asian varsin hyvin: Domain-tieto on kultaa, jota yritetään louhia. Kuten oikea kulta, domain-tieto voidaan säilöä, myydä sellaisenaan (esimerkiksi henkilötiedot pimeässä verkossa [Affairs 2022]) tai jatkojalostaa johonkin toiseen muotoon (esimerkiksi kohdennetuksi mainonnaksi).



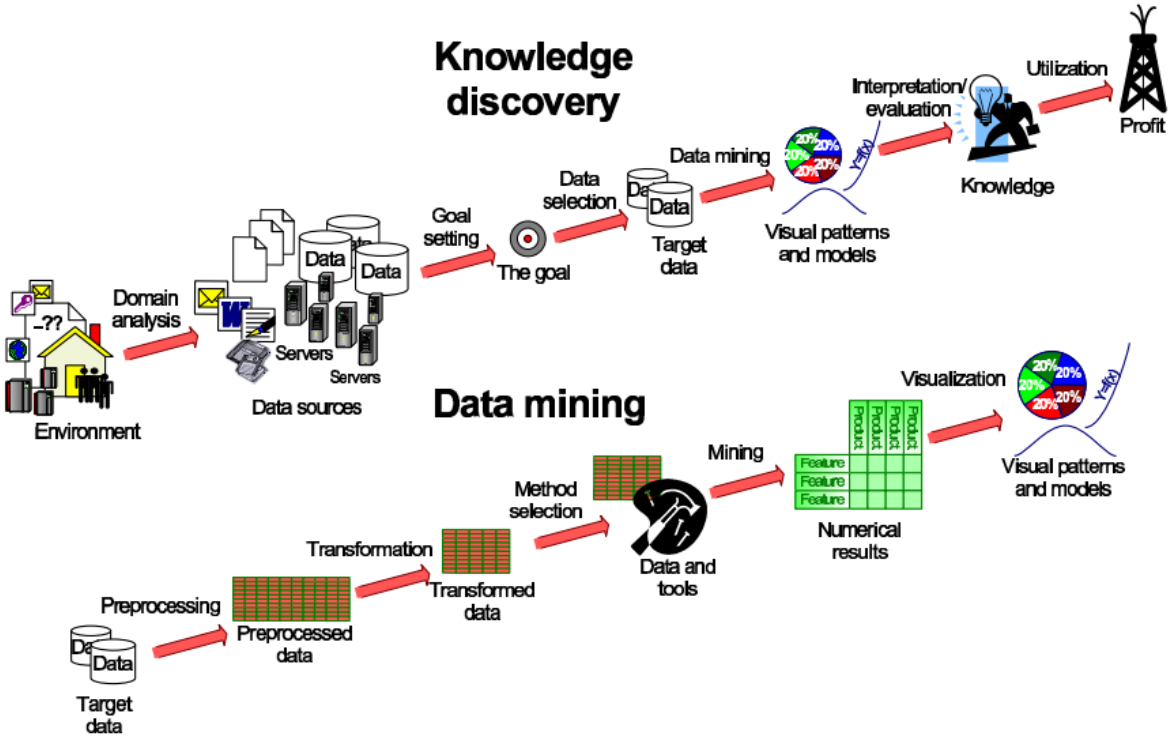
Kuvio 7: Domain-tiedolla (engl. domain knowledge) on kuvainnollisesti samoja ominaisuuksia kuin kullalla: louhiminen, säilöminen, myyminen ja jatkojalostaminen. J. Cleland-Huang, "Mining Domain Knowledge [Requirements],"in IEEE Software, vol. 32, no. 3, pp. 16-19, May-June 2015, doi: 10.1109/MS.2015.67. Saatavilla osoitteesta: https://10.1109/MS.2015.67

Fayyedin ym. (1996) esittelevät KDD-mallin (engl. Knowledge Discovery in Databases) (8), joka kuvaa oppimisprosessia, jossa tiedonlouhinta on yksi prosessin askel. (Fayyad, Piatetsky-Shapiro ja Smyth 1996) Tiedonlouhinta on siis prosessin osa, jossa dataa tai sen osia louhitaan jatkotoimenpiteitä varten.



Kuvio 8: KDD-prosessin vaiheet. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), 37. Saatavilla osoitteessa: https://doi.org/10.1609/aimag.v17i3.1230

Äyrämö puolestaan esittelee väitöskirjassaan *Knowledge Mining Using Robust Clustering* KDD-prosessin, joka on jaettu erikseen KD- ja DM-prosessihin (9). Alkuperäinen KDD-prosessikaavio on laajennettu huomioimaan paremmin domain-kohtaiset kysymykset. Äyrämön mallissa (Äyrämö 2006) kohteen nykytila analysoidaan ensin, jonka jälkeen suoritetaan prosessin muut vaiheet.



Kuvio 9: KDD-prosessin vaiheet erikseen KD- ja DM-prosesseina. Äyrämö, S. (2006). Knowledge mining using robust clustering (No. 63). University of Jyväskylä. Saatavilla osoitteessa: https://jyx.jyu.fi/handle/123456789/13286

KDD-malli (jatkossa KM-malli) jaetaan 1) KD-prosessiin, joka vastaa alkuperäistä KDD-mallia ja 2) DM-prosessiin, joka käsittelee louhittua dataa ja informaatiota. KM-malli on aina konteksisidonnainen suhteessa ympäristöön eli mallin avulla saatu tieto saa merkityksen omassa ympäristössään. (Äyrämö 2006) Tässä tutkimuksessa KM-prosessin lopputulos saa merkityksen suomalaisessa kontekstissa, vaikka saatu tieto seuraakin kansainvälisiä trendejä. Äyrämön mukaan KM:n päätavoitteena on löytää hyödyllisimmät informaatiolähteet ja hyödynnettävin data. Siinä missä KD-prosessi etenee domain-vetoisesti (liiketalous, lääketiede jne.), DM-vaiheet ovat pääasiassa piilossa. DM-prosessi pyrkiikin löytämään ja jatkojalostamaan datan sellaiseen muotoon, jota toimialan asiantuntijat kykenevät käsittelemään. (Äyrämö 2006.)

Tämä tutkimus seuraa KDD-mallia (kuvio 9), siten että kohdetietoa (engl. target data) vastaavat Dread-sivuston keskustelulankojen viestit eli teknisesti merkkijonojoukot. Merkkijonojoukot esiprosessoidaan ja muutetaan sellaiseen muotoon, että niistä voidaan louhia huumausaineisiin liittyvät merkkijonot. Louhitut merkkijonot jatkojalostetaan lopulta taulukoiden ja piirakkakuvioiden muotoon, joita eri alojen asiantuntijat voivat hyödyntää. Prosessi esitellään tarkemmin luvussa 4.

3.7 Luonnollisen kielen käsittely

Yleisesti kieli on viestintäjärjestelmä. (Karlsson 2008) Luonnollisella kielellä tarkoitetaan kieltä, jolla ihmiset kommunikoivat keskenään joko puheen tai merkkien avulla. (Lyons 1991) Luonnollisen kielen ominaisuuksiin voidaan myös liittää historiallis-kulttuurinen tausta ja historia, jossa kieli on kehittynyt ja siirtynyt sukupolvelta toiselle osana normaalia sosiaalista kanssakäymistä. (Karlsson 2008)

Luonnollisen kielen vastakohtana pidetään usein keinotekoista kieltä. (Lehtomäki ja Kukkaniemi 2020 [Stenlund 1990]) Keinotekoisen kielen merkittävin ero luonnolliseen kieleen on se, että keinotekoisen kielen kielioppi ja sanasto on luotu tietoisesti, kun taas luonnollinen kieli on kehittynyt ihmisten vuorovaikutuksen kautta. (Lehtomäki ja Kukkaniemi 2020 [Stenlund 1990; Lyons 1991]) Luonnollisia kieliä eivät siis ole ohjelmointikielet tai esperanton kaltaiset keksityt kielet.

Luonnollisen kielen käsittely (engl. natural language processing, NLP) ja sen tutkiminen pyrkivät ymmärtämään sekä kehittämään ihmisen ja koneiden välistä vuorovaikutusta. (Liddy 2001; Cohen 2014; IBM 2020) Tunnettuja esimerkkejä NLP:stä ovat:

1. Puheentunnistus, jossa ihmisen tuottama puhe (ääniaallot) muutetaan tietoteknisille laittelle ymmärrettävään muotoon (bitit). Tunnetuimpia puheentunnistusta hyödyntäviä avustajia ovat Siri, Google Assistant ja Alexa.
2. Ennakoiva tekstinsyöttö, joka pyrkii tehostamaan ihmisen kirjoitusprosessia. Se pyrkii ennustamaan seuraavan sanan kirjoitettujen kirjainten ja joissain tapauksissa myös kontekstin perusteella. Ennakoiva tekstinsyöttö löytyy nykyään lähes jokaisesta tietoteknisestä laitteesta.
3. Sähköpostisuodattimet, jotka lajittelevat saapuneet sähköpostiviestit roskapostiksi (engl. spam) ja ei-roskapostiksi (engl. ham). Suodattimet perustuvat yksinkertaiseen teknologiaan, jossa lajittelupäätös perustuu todennäköisyyteen, joka lasketaan eri muuttujien – kuten sisällön, lähettäjätietojen ja sähköpostihistorian – suhteen.
4. Kääntäjät, jotka kääntävät puhetta, tekstiä ja kuvia halutulle kielelle. Kääntäjien toiminta perustuu erilaisiin sääntöihin ja todennäköisyyksiin.

Tässä tutkimuksessa NLP liittyy kappaleessa 4.1 esiteltävään skriptiin ja siinä käytettävän lajittelualgoritmin toimintaan. Aihetta käsitellään enemmän kyseisessä luvussa. Tutkimuksen kannalta oleellisinta on ymmärtää, että suomen kielen erityispiirteet aiheuttavat haasteita luonnollisen kielen käsittelyssä.

4 Aineiston keruu ja analyysi

Aineisto (N $\approx$ 8500) koostuu Dread-sivuston Suomi-langan viesteistä. Alunperin aineisto oli tarkoitus kerätä Rautatieasema-sivustolta syksyllä 2021 ja keräämiseen oli tarkoitus käyttää GitHubista löytyvää Flashbackscraper-skriptiä (Kullenberg 2022), joka piti muokata tilanteeseen sopivaksi. Rautatieasema-sivusto hävisi kuitenkin verkosta vain vähän ennen kun skripti saatiin valmiiksi. Tämän jälkeen suuri osa kaupankäynnistä siirtyi Dread-sivuston Suomi-langalle. Langan ensimmäisten viestien ajankohdat ja Ruotsalaisen aiempi havainto (Ruotsalainen M. 2021) siitä, että liikenne siirtyi Dread-sivustolle käyttökatkojen aikana, tukevat tätä päätelmää. Ajan säästämiseksi aineisto kerättiin käsin HTML-tiedostoina, joiden käsittelemiseen luotiin kokonaan uusi ja Flashbackscraperista täysin erillinen Python-skripti. Skripti ja aineisto löytyvät GitHubista: https://github.com/jkpnen/dread-parser

Kaiken kaikkiaan aineiston keräämiseen ja skriptin tekemiseen meni vähemmän aikaa kuin olisi ollut odotettavissa teknisemmällä *scarping/crawling*-toteutuksella. Tämän takia siitä luopuminen oli perusteltua. Aineisto sisältää jokaisen kaupunkilangan käyttäjätunnukset, viestit ja aikaleimat. Aineisto oli tarkoitus analysoida käyttäjätunnusten ja viestien sisällön suhteen, mutta analyysi toteutettiin lopulta pelkkien viestien osalta. Käyttäjätunnusten analysoiminen ei ole tässä tapauksessa perusteltua, koska 1) niiden analysoiminen ei ole tutkimuksen lopputuloksen kannalta oleellista ja 2) käyttäjien anonymiteetti halutaan suojata eettisistä syistä. Saatuja tuloksia verrataan jätevesien huumausainepitoisuuksiin. Tarkoitus on myös tarkastella huumausaineiden esiintyvyyttä viesteissä niin paikkunnittain kuin valtakunnallisesti.

Tutkimuksen analyysimenetelmäksi valikoitui luokittelu, koska siinä voidaan hyödyntää laadullisen ja määrällisen analyysin menetelmiä (JYU 2015a). Lisäksi tutkimuksessa on suurehko, yksittäisistä tapauksista (viesteistä) koostuva aineisto (N $\approx$ 8500), mikä on yksi luokittelun ominaispiirre (UIAH 2022; KAMK 2022). Suuriin, yksittäisistä tapauksista rakentuviin aineistoihin saadaan järjestystä jakamalla tapaukset toisia tapauksia muistuttaviin ryhmiin. Kaikilla tietyn ryhmän tapauksilla tulee siis olla jokin ominaisuus, joka muiden ryhmien tapauksilta puuttuu, tai ainakin tämän ominaisuuden aste on eri ryhmissä erilainen. (UIAH 2022.)

Lisäksi luokittelun päämäärä on aina sama: etsiä aineistossa piilevä systemaattinen rakenne, invarianssi. Ideaalitilanteessa jokaiselle tapaukselle löytyisi vastaava luokka. (UIAH 2022.) Tässä tutkimuksessa viesteissä esiintyvät huumausaineet voidaan luokitella esimerkiksi alaryhmittäin stimulantteihin, opiaatteihin ja hallusinogeeneihin kuten taulukossa 1.

Stimulantit	Opiaatit	Hallusinogeenit
Amfetamiini, ekstaasi	Morfiini, heroiini	LSD, sienet

Taulukko 1: Havainnollistava esimerkki huumausaineiden luokittelusta alaryhmittäin.

Teknisen toteutuksen ja analyysin kannalta mielekkäämpi vaihtoehto on kuitenkin jakaa merkkijonot kyseisiä huumausaineita vastaaviin alaryhmiin taulukon 2 tapaan.

Kannabis	Amfetamiini	Kokaiini
ruoho, heinä, kukka, pilvi, dulla, budi, nuppu, blosse, pajari, paja, ganja, hööni, mari, mauri, höpöheinä, naurutupakka	piri, pöhinä, spiidi, vauhti	koka, koksu, kokkeli
50%	30%	20%

Taulukko 2: Havainnollistava ja mielivaltainen esimerkki huumausaineiden luokittelusta synonyymeittain. Merkkijonot on luokiteltu huumausaineita vastaaviin alaryhmiin

Suomi on taivutuksen ja yhdyssanojen suhteen rikas kieli, mikä tekee siitä monimutkaisen verrattuna indoeurooppalaisiin lauseke-orientoituneisiin kieliin kuten englantiin (Korenius ym. 2004). Kielen yleinen monimutkaisuus tekee siitä myös haastavaa luonnollisen kielen käsittelyn suhteen. Tässä tutkimuksessa haasteita aiheuttaa huumejargonin, paikon varsin luova, käyttö. Suomen kielessä sanojen taivutukset ovat sanan lopussa, kun indoeurooppalaisissa kielissä sanataivutukset tapahtuvat prepositioiden ja postpositioiden kautta. (Korenius ym. 2004) Esimerkiksi virkeessä ”myyn kannabista” sanan ”kannabis” taivutusmuoto on ”kannabista”. Englanniksi sama ilmaisu on ”I sell cannabis” eli sama asia voidaan ilmaista ilman taivutusmuotoja.

Lisäksi aineistoissa esiintyvä kieli on välillä vaikeasti ymmärrettävää jopa natiiville. Haasteita aiheuttavat esimerkiksi huumejargon ja kielioppivirheitä vilisevät ilmaisut. Helsingin yliopiston julkaisemassa tiedotteessa *Tutkijat opettavat tekoälyn käyttämään suomen kielen murteita* todetaankin, että murteiden ymmärtäminen on monitahoinen haaste, eikä mikään malli kykene ymmärtämään ihmisen lailla luonnollista kieltä. (Helsingin yliopisto 2021) Murteet voidaan rinnastaa haasteiden suhteen muihin ammatti- ja erikoiskieliin: Luonnollista kieltä käsittelevät tekniset toteutukset ”ymmärtävät” kieltä parhaiten aina normatiivisessa eli kirjakielisessä muodossa.

4.1 Käytännön vaiheet

Aineisto kerättiin lataamalla viestiketjut käsin Dread-sivuston /d/wickrfinland -langasta. Alkuperäiset viestiketjut ovat HTML-muodossa ja ne ladattiin kaupunkilankojen mukaan seuraavasti:

- Espoo
- Helsinki
- Joensuu
- Jyväskylä
- Kuopio
- Kymenlaakso
- Lahti
- Lappeenranta
- Oulu
- Rovaniemi
- Seinäjoki
- Tampere
- Turku
- Vantaa

HTML-tiedostojen läpikäymiseen ja analysoimiseen luotiin skripti Python 3.8:lla. Skripti toimii pääperiaattein siten, että se poimii jokaisesta HTML-tiedostosta käyttäjänimen, viestin sisällön ja julkaisuajan Beautiful Soup 4:n avulla. Kyseessä on apukirjasto websivujen skreippaamiseen (engl. scraping) eli datan irroittamiseen verkkosivuilta. Tämän jälkeen datasta luodaan xlsx-tiedosto XlsxWriterin avulla, joka on puolestaan apukirjasto excel-tiedostojen luomiseen. Lopullista analyysia varten excel-tiedostot muutetaan csv-tiedostoiksi, jotta niistä voidaan tehdä helpommin käsiteltävä tietokehys (engl. dataframe) pandasin (data-anlyysin apukirjasto) avulla.

Aluksi huumausaineiden esiintyvyyttä yritettiin arvioida sanapilven avulla. Tästä kuitenkin luovuttiin, koska huomattiin nopeasti, että sanapilvi ei toimi riittävällä tarkuudella. Tämä johtui hyvin pitkälti suomen kielen erityispiirteistä ja siitä, että suomen kielen apukirjastot

eivät olleet tarpeeksi kattavia tutkimuksen kontekstissa.

Skriptissä viestiketjut käydään läpi viesti kerrallaan ja jokaisesta viestistä tarkastellaan yksittäisiä sanoja. Ennen sitä yksittäiset viestit esiprosessoidaan luonnollisen kielen käsittelyyn liittyvien haasteiden takia. Keskustelukanavan viestit sisältävät analyysia vaikeuttavia ja ilmentymien laskemista monimutkaistavia asioita, kuten:

- Huumejargonia: amfe, dänkki, sulfa jne.
- Unicode-merkistöä: Tähdet, nuolet, euromerkit ym.
- Emojeja
- Kirjoitusvirheitä
- Ylimääräisiä rivinvaihtoja ja välilyöntejä
- Tyhjiä/poistettuja viestejä

Alla olevasta kuviosta 10 nähdään osa edellä mainituista haasteista.

[SALATTU] holleilla Amfea bentsoja ja vähän priimaa dänkkiä.

Hyvää Amfea tykättyä kuivaa sulfaa laatutakuulla 40-33€/g.

Nopeasti nokkaankin taipuvaa sitä itseään ilman tutkari tai meta kohinoita. Maistuu ja tuoksuu siltä miltä amfen kuuluukin. Iv- safe ja 0.2 toimii veturillakin. Tuu ihmees tsekkaa laatu ja yllättyy. Tuotteen saa huomaamattomasti ja asianmukasta hygieniaa noudattaen tarkastaa ja olla ottamatta jos ei kelpaa. Likanen sormi pussissa on ostopäätös eli järki käteen ja testivedot saat suorittaa omalla ajalla.

Löytyy kans rauhottavia ettei meno mee liian vauhikkaaks eli vähä toimivaksi todettuja serbiapteekin nappeja Galenikan rivoja ja ksaloleita ehjis levyis pitkillä päiväyksillä.
☆☆☆Galenikan 1mg Ksalolit 20-10€/levy☆☆☆
☆☆☆Galenikan Rivat 2mg 30-15€/levy☆☆☆
►Ksalol 20/1 70/5 10/100 ◄
►Rivat 30/1 50/2 100/5 300/20◄

PARHAAT JOULUPAKETIT
2G AMFEA(65) + 2G KUKKAA(35) + 2 LEVYÄ RIVATRIL(35)+ 1 LEVY KSALOL(15)=150€
3G AMFEA(80)+ 1levy KSALOL tai RIVATRIL(20)=100€
1G AMFEA(35) + 1G KUKKAA(15)=50€
1 LEVY RIVATRIL(17.5)+ 1 LEVY KSALOL(15)+1G KUKKAA(17.5)=50€
1.5G AMFEA(50) + 1G KUKKAA(15) + 1LEVY KSALOL(15) + 1LEVY RIVATRIL(20)=80€
↑↑↑↑↑↑↑↑↑↑↑↑↑↑↑↑↑↑↑↑
HO HO HO! JOULUTARJOUS PISTETÄÄ [SALATTU] JOULU
TAATUSTI VALKEEKS! TÄÄLTÄ SITÄ TYKÄTTYÄ JA RAKASTETTUA PUUTERILUNTA ALLE ET PULKKA LUISTAA JA SUKSI KULKEE ???? Ja kylkeen jos tarvii rauhottavaa ryyniä niin lähtee lolit tai rivat kohtuu hintaan samalla

Erittäin rajallinen määrä hyvää suomi KUKKAA! Priimaa rutikuivaa indicaa muiden ostojen kylkeen 1g/20€ akuutteja

Wickr [SALATTU]

Rehti kantasuomalainen skarppi myyjä vastassa.
Kauppa liikkuu metroradan varrella yleensä välillä [SALATTU]

➞Wickr// [SALATTU]

K-18, järkikunnossa ja yksin kaupoille. Rotille löytyy myrkyt ja karkottimet.

Kuvio 10: Myynti-ilmoitus Dread-sivustolta

Esiprosessoinnissa viesteistä poistetaan Unicode-merkistö, emojit sekä ylimääräiset rivinvaihdot ja välilyönnit. Tyhjiä/poistettua viestejä ei kuitenkaan poisteta, koska ne eivät vaikuta lopputulokseen. Lisäksi ne sisältävät tiedon käyttäjätunnuksesta, mikä ei kuitenkaan ole

oleellista tämän tutkimuksen kannalta. Myös merkkijonojen kirjaimet pienennetään. Voidaan olettaa, että viestin pituus lisää sen käsittelyn kompleksisuutta, koska kontekstisidonnaisia muuttujia on enemmän. Esimerkkiviesti on pituudeltaan aineistoin pisimpiä, ja analyysin luotettavuuden kannalta on hyvä, että suurin osa viesteistä on lyhyempiä.

Analyysin tavoitteena on laskea eri merkkijonojen esiintymät ja arvioida niiden perusteella huumausaineiden esiintyvyyttä valtakunnallisesti sekä alueellisesti. Esiintymien laskeminen tapahtuu sanojen alun sekä ennalta määrättyjen avainsanojen ja poikkeuksien perusteella. Kuvion 10 tapauksessa aluksi skripti tarkastelee sanaa ”holleilla”, jonka alku ei kuulu avainsanojen joukkoon. Seuraavaksi skripti ottaa käsittelyyn sanan ”amfea”, joka löytyy avainsanojen joukosta merkkijono-muodossa ”amfe”. Löytynyt esiintymä lisätään Amfetamiini-luokkaa vastaavaan listaan. Lajittelualgoritmi pseudokoodina:

Algorithm 1 Esiintymien lajittelualgoritmi

Require: viesti $\in$ viestit, avainsana $\in$ avainsanat, poikkeus $\in$ poikkeamat

```
for viesti in viestit :
    for merkkijono in viesti :
        if merkkijono is poikkeus
            continue
        else
            if merkkijono starts with avainsana
                Lisää merkkijono listaan
```

Ihminen osaisi helposti päätellä esimerkkiviestin perusteella, että myynnissä on amfetamiinia ja kannabista sekä rauhoittavia. Suurimmat haasteet, jotka vaikuttavat esiintymien laskemiseen ovat:

1. Huumejargon
2. Suomen kielen sijapäätteet ja kielioppivirheet
3. Yksittäisissä viesteissä esiintyvät duplikaatit
4. Englanninkieliset ilmoitukset

Näistä varsinkin 1. ja 3. aiheuttavat virhettä esiintymien tarkkuuteen, koska yhdessä viestissä voidaan mainita useasti sama sana (esimerkiksi kannabis). Tämä oletettavasti lisää kannabis-

sanan esiintymien lukumäärää, minkä takia skriptiä muokattiin siten, että yhdessä viestissä voi olla vain yksi huumausainetta vastaava esiintymä. Vastaavasti osa lääkkeistä jää pakostakin algoritmin ulkopuolelle, koska markkinoilla saattaa olla esimerkiksi uusia muuntohuumeita, joita ei ole osattu listata avainsanojen joukkoon. Avainsanojen lisäksi skriptiin piti lisätä poikkeuksia, joita ei ole mahdollista kiertää muuuten. Esimerkiksi ”kokaiinin” synonyymi ”kola” aiheutti ongelmia, koska viesteissä esiintyi yllättävän useasti sana ”kolahtaa”.

Esiintymä siis merkataan oikeaksi tai vääräksi, mikäli sanan alku löytyy ennaltamääritellystä avainsanojen joukosta. Kola-merkkijonon tapauksessa sana alkaa samoin kuten avainsana ”kola”, mutta skripti ei erottele sanojen loppuja. Tähän ratkaisuun päädyttiin hyvin pitkälti kohdan 2. takia, koska suomen kielessä useat myyntitapahtumiin liittyvät sanat sisältävät sijamuodon. Esimerkiksi virke ”Myyn kannabista” sisältää sijapäätteen ”-ta”. Tätä haastetta ei tietenkään ole englannissa, jossa vastaavat virke olisi ”I sell cannabis”. Myös muut, yleisesti tunnetut ilmaisut kuten ”tarjolla”, ”myynnissä”, ”(pukin)konttiin” sisältävät lähes aina jonkun sijapäätteen. Lisäksi on olemassa tiettyjä sanoja, joita on käytännössä mahdotonta poissulkea ilman konseptuaalista ymmärtämistä. Tästä esimerkkinä käyttäjätunnus, joka voisi olla amfetamiinii kahdella i:llä. Skriptin tapauksessa kyseinen ilmentymä laskettaisiin amfetamiini-esiintymäksi, vaikka kyseessä on käyttäjätunnus.

4.2 Virheanalyysi

Tutkimuksen luokittelija-algoritmin luotettavuutta arvioidaan sekaannusmatriisin (engl. confusion matrix, taulukko 3) avulla. Sekaannus-/virhematriisi on tilastotieteen työkalu, jota käytetään usein juuri luokittelijoiden ennustamiskyvyn arvioimiseen. (Nieminen 2010; Suomen avoimien tietojärjestelmien keskus 2017; Scikit-learn 2022b, 2022a)

Määritelmän mukaan sekaannusmatriisi C ja tarkemmin C_{ij} on lukumäärä havainnoista, jotka kuuluvat luokkaan i ja jotka on luokiteltu luokkaan j. (Nieminen 2010; Suomen avoimien tietojärjestelmien keskus 2017; Scikit-learn 2022a, 2022b) Esimerkiksi binääriluokittelijalla C_{11} on lukumäärä havainnoista, jotka kuuluvat luokkaan **False**, mutta jotka on luokiteltu luokkaan **Positive**. Tässä tutkimuksessa käytetään binääriluokittelijaa, joka antaa vastauksen kyllä (huumausaine) tai ei (muu merkkijono). Luokittelija yrittää arvata kuuluuko käsiteltävä

merkkijono huumausaineiden merkkijonojoukkoon vai ei. Luokittelijan arvaus on siis oikein tai väärin verrattuna testiaineiston tunnettuun totuuteen.

Taulukko 3: Sekaannusmatriisi

C_{01}	C_{11}
C_{00}	C_{10}
$C_{00} + C_{01}$	$C_{10} + C_{11}$

False Positive = FP	True Positive = TP
True Negative = TN	False Negative = FN
FP + TN	TP + FN

FP = 12	TP = 303
TN = 8751	FN = 25
FP + TN = 8763	TP + FN = 328

Taulukko 4: Tutkimuksessa käytetyn luokittelija-algoritmin sekaannusmatriisit. Vasemmalla ylhäällä matriisin matemaattinen esitysmuoto, oikealla ylhäällä sama matriisi binääriluokitteluna ja keskellä alhaalla sama matriisi havaintojen kanssa.

Binääriluokittelijassa oikeiden negatiivisten luokittelujen lukumäärä on paikassa C_{00}, väärien negatiivisten C_{01}, oikeiden positiivisten C_{11} ja väärien positiivisten C_{10}. Sekaannusmatriisi kertoo, kuinka hyvin luokittelija suoriutuu tehtävästään. Diagonaalialkiot vasemmalta oikealle (C_{01} ja C_{11}) vastaavat oikeita ja muut vääriä luokitteluja. Mitä suurempi diagonaalialkioiden lukumäärä on suhteessa muihin alkioihin, sitä parempi luokittelija on kyseessä. (Nieminen 2010; Suomen avoimien tietojärjestelmien keskus 2017; Scikit-learn 2022b, 2022a) Sekaannusmatriisin tapauksessa paremmuudella tarkoitetaan oikeita arvauksia.

Sekaannusmatriisia varten viestit käytiin manuaalisesti läpi siten, että tarkasteluun otettiin 20 viestiä jokaisesta kaupunkilangasta – Rovaniemen kaupunkilankaa lukuunottamatta, sillä siinä oli yhteensä alle 20 viestiä. Jokaisessa viestissä on $\mathrm{N}_{\mathrm{merkkijonot}}$ määrä merkkijonoja, jotka jaetaan huumausaine-merkkijonoihin ja muihin merkkijonoihin. Tarkastelun ulkopuolelle jätettiin esimerkiksi tyhjät ja pelkän WICKR-tunnuksen sisältävät viestit. Näin sekaannusmatriisi arvioi viestejä, jotka sisältävät aidosti vähintään yhden huumasuaine-merkkijonon.

4.3 Tunnusluvut

Tarkkuus (engl. accuracy) on yksinkertaisesti oikeat luokittelut jaettuna kaikilla vastauksilla. Tarkkuus kertoo miten luokittelija onnistuu luokitellussaan:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} = \frac{303 + 8751}{303 + 8751 + 12 + 25} \approx 0,99.$$

Positiivinen ennustearvo (engl. positive predictive value) kertoo, millä todennäköisyydellä luokittelija luokittelee huumausaineen oikein huumausaineeksi:

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} = \frac{303}{303 + 12} \approx 0,96.$$

Negatiivinen ennustearvo (engl. negative predictive value) kertoo, millä todennäköisyydellä luokittelija luokittelee ei-huumausaineen ei-huumausaineeksi:

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{FN} + \mathrm{TN}} = \frac{8751}{25 + 8751} \approx 0,99.$$

Sensitiivisyys (engl. sensitivity) kertoo huumausaineiksi luokiteltujen merkkijonojen osuu-

den huumausaineista:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{303}{303 + 25} \approx 0,92.$$

Spesifisyys (engl. specificity) kertoo merkkijonoiksi luokiteltujen merkkijonojen osuuden merkkijonoista:

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} = \frac{8751}{12 + 8751} \approx 0,99.$$

F-arvo on posiitivisen ennustearvon ja sensitiivisyyden harmoninen keskiarvo ja kertoo näiden kahden välisestä tasapainosta:

$$\text{F}_1 = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} = \frac{2{\cdot}303}{2{\cdot}303 + 12 + 25} \approx 0,94.$$

Tarkkuudeksi saadaan 99%, mikä on epäilyttävän hyvä tulos. Taulukon 4 perusteella voimme todeta, että aineistossa on muita merkkijonoja määrällisesti enemmän kuin huumausaine-merkkijonoja. Tarkemmin sanottuna huumausaineiden osuus kaikista merkkijonoista on:

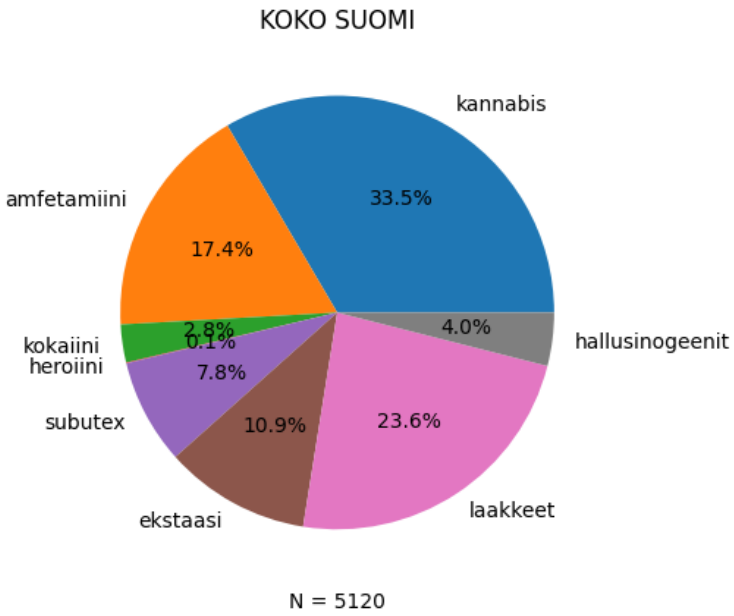
$$\frac{\text{TP+FN}}{\text{TP+FN+FP+TN}} = \frac{303 + 25}{303 + 25 + 12 + 8751} \approx 0,03.$$

Tällaisissa tapauksissa voidaan laskea F-arvo, joka huomioi myös matriisin alkioiden jakaumat. F-arvoksi saadaan 94%, mikä on jo uskottavampi tulos.

Kaiken kaikkiaan ylläolevien tunnuslukujen perusteella voidaan todeta, että algoritmi suoriutuu tehtävästään merkittävän hyvin. Varsinkin väärien merkkijonojen luokittelu menee käytännössä aina oikein ja huumausaineden luokittelu lähes aina. Toisaalta voidaan arvioida, että FN-esiintymien osuus on todellisuudessa suurempi, koska esimerkiksi lääkkeitä esiintyy toiseksi eniten (taulukko 11) koko aineistossa ja niissä on myös eniten hajontaa kirjoitusasun sekä uniikkien esiintymien suhteen. Virhemarginaali voidaan arvioida kaikissa skenaarioissa sen verran pieneksi, että tarkempaa analyysia aiheesta ei ole mielekästä tehdä. Toisin sanoen algoritmissa esiintyy virhettä, joka ei huonoimmassakaan tapauksessa muuta lopullista tulosta merkittävästi.

5 Tulokset

Huumausaineisiin luokiteltujen merkkijonojen esiintymien suhteelliset jakaumat on esitelty kuviossa 11. Kuviosta nähdään, että Dread-sivuston Suomi-langan myynti-ilmoituksissa esiintyy eniten kannabista, amfetamiinia ja lääkkeitä. Myös ekstaasia ja subutexia esiintyy verrattain paljon. Hallusinogeenit ja kokaiini muodostavat pienen osuuden ja heroiinia ei esiinny juuri lainkaan. Tulos ei tietenkään kerro kuinka paljon huumausaineita ostetaan ja myydään määrällisesti vaan kyse on enemmän yleisestä esiintyvyydestä tai trendistä.



Kuvio 11: Huumausaineisiin luokiteltujen merkkijonojen esiintymien suhteelliset jakaumat Dread-keskustelukanavan viesteissä (N $\approx$ 5120). N pieneni esiprosessoinnin ja analyysin jälkeen 8500:sta 5120:een.

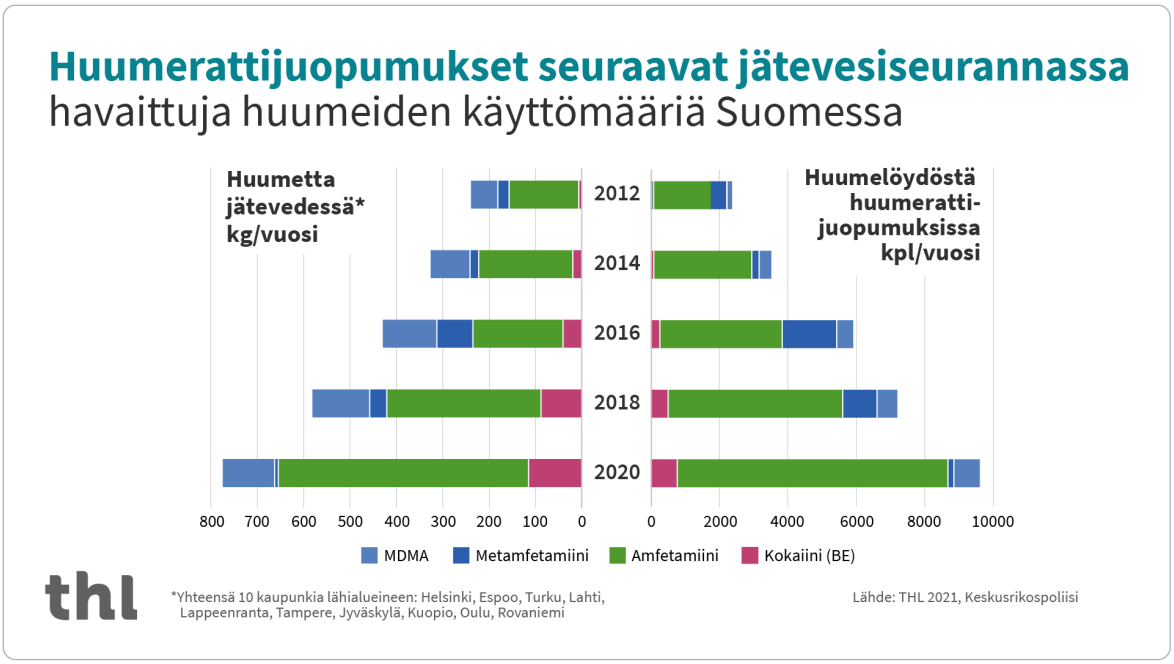
Liitteessä A on esitelty merkkijono-esiintymien suhteelliset jakaumat paikkakunnittain. Suurin osa paikkakunnista on linjassa kuvion 11 kanssa. Alle 50 esiintymän paikkakuntia tulee tulkita pienellä varauksella. Esimerkiksi Rovaniemen tapauksessa esiintymiä oli vain 13, minkä takia suhteelliset osuudet antavat hieman vääristyneen kuvan jakaumasta. 13:sta viestistä yksi viesti sisälsi yhden esiintymän jokaisesta huumausaine-luokasta (heroiinia lukuu-

nottamatta), minkä takia kuvaajasta saa käsityksen, että Rovaniemellä esiintyy yllättävän paljon kaikkia huumausaineita. Lisäksi Helsinki, Espoo ja Vantaa muodostavat yhdessä yli puolet aineistosta:

$$\frac{\mathrm{N}_{\text{pääkaupunkiseutu}}}{\mathrm{N}} = \frac{2973}{5120} \approx 58\ \%.$$

6 Johtopäätökset

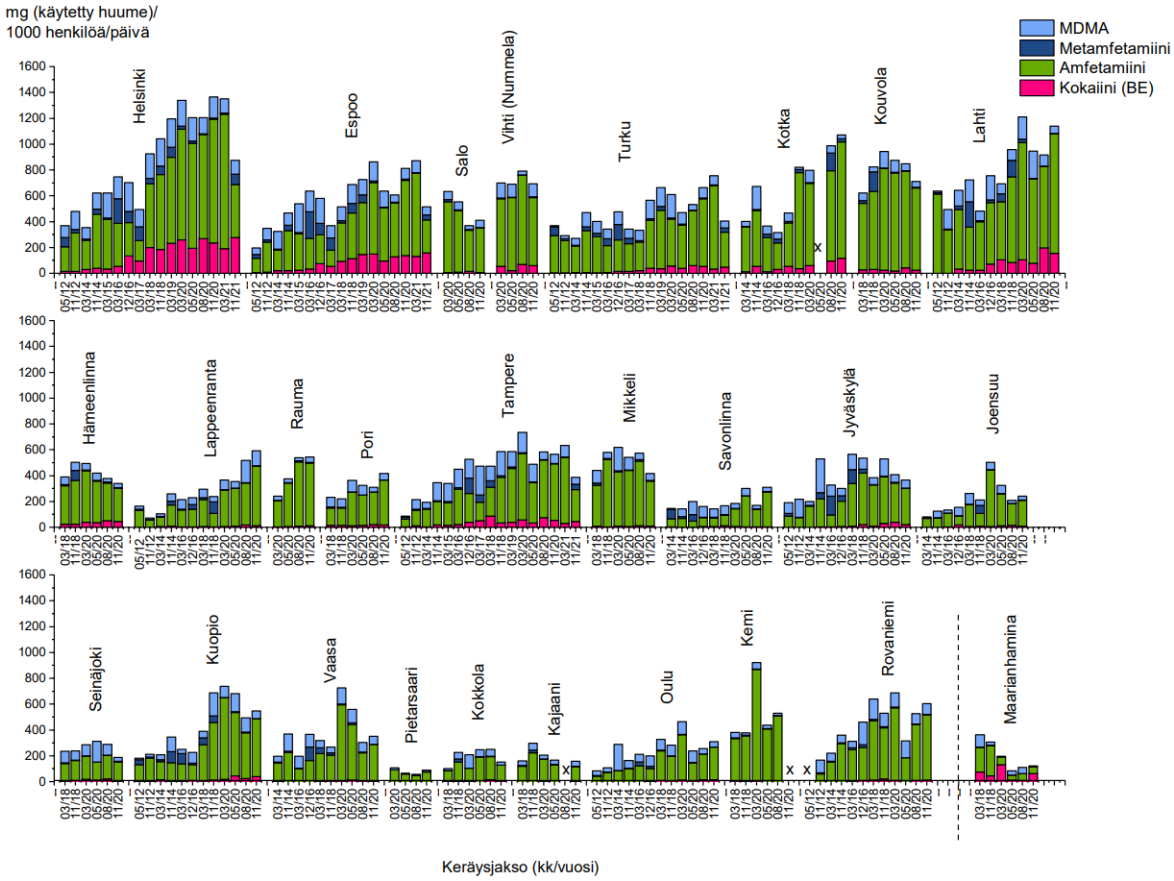
Tämän tutkimuksen tulos (kuvio 5) on linjassa THL:n huumeraportin (THL 2020) kanssa. Suomen huumausainemarkkinoilla esiintyy ennen kaikkea kannabista – sekä amfetamiinia, ekstaasia ja muita synteettisiä huumausaineita, korvaushoitovalmiste Subutexia sekä runsaasti huumausaineeksi luokiteltuja lääkevalmisteita, etenkin bentsodiatsepiinilääkevalmisteita. THL:n mukaan jätevesitutkimuksilla saadaan riippumatonta tietoa jätevesiverkoston alueella tapahtuvasta huumeiden käytöstä lähes reaaliaikaisesti. Menetelmää voidaan käyttää myös huumetilanteen pitkäaikaisseurantaan. (THL 2022) Kuviosta 12 nähdään, että koko Suomen tulokset ovat olleet pääsääntöisesti hyvin yhteneviä huumerattijuopumustilastojen kanssa. Kappaleen 5 kuvio 11 on linjassa amfetamiinien ja ekstaasin suhteen. Toisaalta ekstaasin ja kokaiinin välinen suhde on kuitenkin ristiriitainen tutkimustulosten perusteella. THL:n kehittämispäällikkö Aino Kankaanpään mukaan: ”Kokaiinin käyttö on tasaantunut ja ekstaasin käyttö jopa vähentynyt koronarajoitusten aikana useissa tutkimuskaupungeissa. Tähän ovat mahdollisesti vaikuttaneet paitsi huumeiden saatavuus, myös kokoontumisrajoitukset sekä iltaelämää koskevat rajoitukset”. (THL 2021)



Kuvio 12: Huumeiden käytön kehitys 10 kaupungin alueella (Helsinki, Espoo, Turku, Lahti, Lappeenranta, Tampere, Jyväskylä, Kuopio, Oulu ja Rovaniemi lähialueineen) jätevesitutkimuksen perusteella verrattuna positiivisten löydösten lukumäärään epäillyissä huumerattijuopumuksissa vuosina 2012-2020. Viitattu 7.3.2022. Haettu osoitteesta: https://thl.fi/fi/-/jatevesitutkimus-amfetamiinin-kaytto-jatkaa-kasvuaan-suomessa-kokaiinin-kayttomaarien-kasvu-on-tasaantunut

Jätevesitutkimuksiin nojaten voidaan ihmetellä, minkä takia Dread-sivustolla kokaiinia esiintyy varsin marginaalisesti verrattuna ekstaasiin. Yksi selittävä tekijä voi olla kokaiinin kallis katukauppahinta. Voidaan siis epäillä, että jakelijoita on vähemmän ja että kokaiini on pienemmän ja/tai varakkaamman piirin huumausaine.

THL:n mukaan eri huumausaineiden käytössä ja niiden muutoksissa todetaan merkittäviä alueellisia eroja siten, että huumeita käytetään keskimääräisesti enemmän eteläisen Suomen suurissa kaupungeissa (kuvio 13). (THL 2022) Liite A tukee tätä tulosta – toisaalta etelästä ja muista suurista kaupungeista oli myös enemmän dataa.



Kuvio 13: Huumeiden takaisinlasketut käyttömäärät eri kaupunkien jätevedenpuhdistamojen toiminta-alueilla (milligrammaa käytettyä huumetta/1000 henkilöä/päivä) eri keräysviikkoina vuosina 2012-2021. Viitattu 7.2.2022. Haettu osoitteesta: https://thl.fi/fi/tutkimus-ja-kehittaminen/tutkimukset-ja-hankkeet/jatevesitutkimus

THL:n mukaan valtakunnalliset jätevesitutkimukset osoittavat, että väestötasolla yhteenlas-

ketun amfetamiinin, metamfetamiinin, metyleenidioksimetamfetamiinin (MDMA, ”ekstaasi”) ja kokaiinin käyttö on lisääntynyt Suomessa tutkimusaikana. Viime vuosina merkittävin muutos on ollut jätevesitutkimuksen perusteella jo ennestään suositun amfetamiinin käytön kasvu Suomessa. (THL 2022.) Tässä tutkimuksessa amfetamiini ja metamfetamiini on kuitenkin sijoitettu samaan luokkaan stimulanttisten ominaisuuksiensa vuoksi. Kappaleen 5 kuviosta 11 nähdään, että amfetamiineja esiintyy kannabiksen jälkeen eniten. Tämä on linjassa jätevesitutkimusten kanssa. Lisäksi ekstaasi muodostaa varsin suuren osan viestien sisällöstä. Molemmat löydökset ovat linjassa aikaisemman tutkimustiedon (THL 2020) kanssa.

Toinen mielenkiintoinen havainto liittyy kokaiinin käyttöön. THL:n mukaan se on vajaassa vuosikymmenessä voimakkaasti kasvanut eurooppalaisittain erittäin matalalta tasolta, painottuen yhä voimakkaasti eteläisen Suomen suuriin kaupunkeihin. Kokaiinin käyttömäärät ovat kuitenkin edelleen Helsingin alueella eurooppalaista keskiarvoa alhaisemmat, mutta toisaalta kokaiinin osuus huumeiden käytössä Suomessa ei ole enää aivan marginaalinen, kuten ennen. (THL 2022.) Kappaleen 5 kuvion 11 ja liite A kuvioiden perusteella kokaiinia esiintyy marginaalisesti koko maassa ja suhteellisesti hieman enemmän etelässä. Toisaalta pohjoisesta ja idästä on vähemmän dataa kuin etelästä, minkä takia kokaiinin esiintyvyyttä on hieman hankala arvioida maantieteellisesti. THL:n (2020) selvityksen mukaan kokaiini on yleistynyt ja ainetta on takavarikoitu aikaisempaa enemmän Suomessa viimeisen 10 vuoden aikana, mutta se on edelleen markkinoilla melko marginaalinen huumausaine. (THL 2020) Samassa selvityksessä todetaan, että heroiinia on Suomen huumausainemarkkinoilla esiintynyt edelleen hyvin vähän. Sen sijaan Subutexin käyttö huumausaineena on Suomessa edelleen hyvin runsasta. (THL 2020) Molemmat väitteet ovat linjassa tämän tutkimuksen tulosten (kuvio 5) kanssa. Subutex ym. opioidijohdannaiset ovat todennäköisesti korvanneet heroiniin käytön, minkä takia sitä ei esiinny juuri lainkaan tutkimusaineistossa.

Kolmas mielenkiintoinen havainto liittyy ekstaasin ja amfetamiinin suhteeseen verrattuna niiden väliseen suhteeseen tässä tutkimustuloksessa. Jätevesissä amfetamiineja esiintyy noin viisi kertaa enemmän kuin ekstaasia, mutta tutkimustuloksen (kuvio 5) perusteella amfetamiinia esiintyy (vain) puolitoista kertaa enemmän. Yksi selittävä tekijä voi olla sama kuin kokaiinin tapauksessa – eli tutut ja turvalliset myyntiringit, jolloin kauppaa ei tarvitse käydä pimeän verkon sivustoilla.

Neljäs havainto liittyy viime vuosien suosituimman muuntohuumeen alfa-PVP:n käyttöön, joka oli vuonna 2018 hyvin alhaisella tasolla, mutta joka on kasvattanut hieman suosiotaan painottuen pääkaupunkiseudulle ja sen läheisyyteen. (THL 2022) Tutkimustuloksen perusteella (kuvio 5) lääkkeitä esiintyy varsin paljon koko Suomessa. Tutkimus ei kuitenkaan erittele muuntohuumeita, minkä takia tutkimus ei ota kantaa erikseen alfa-PVP:n esiintyvyyteen.

Huumausainemarkkinat näyttäytyvät organisoituina ja ammattimaisina. Toisaalta huumemarkkinat koostuvat erilaisista ”tasoista”, joissa ylemmän tason maahantuojat ja tukkurit, keskitason välittäjät ja alemman tason myyjät toimivat eri tavoin (THL 2022.) Pelkästään keskustelukanavien viestien perusteella on vaikea vetää johtopäätöksiä, mille tasolle pimeä verkon kaupankäynti sijottuu Suomessa. Myös toiminnan ammattimaisuutta on vaikea arvioida pelkästään tämän tutkimuksen tulosten perusteella. Tietyt käyttäjätunnukset saattavat kuitenkin käydä tutuiksi aineistoa läpikäydessä ja niitä saattaa esiintyä muissakin kaupunkilangoissa. Keskustelukanavilla esiinyy myös näitä edellä mainittuja maahantuojia ja tukkureita, joista voidaan käyttää nimitystä *bulkkimyyjä*. Tämä oli havaittavissa myyntiilmoitusten perusteella, joissa esiintyi määrällisesti yllättävän paljon myyntieriä. Esimerkiksi Perälä (2011) toteaa Helsingin huumekaupasta, että kaupanteko ylemmästä toimijasta alempaan toimijaan asti ei aina näyttäydy kovinkaan rationaalisena toimintana. Toiminnassa kyse on ennemmin päihderiippuvuudesta tai -riippuvuuksista ja muista ongelmista kuin järjestelmällisyydestä. (Perälä 2011) Tämäkin oli havaittavissa keskustelukanavien viestien peruusteella. Jos toisessa ääripäässä ovat bulkkimyyjien tukkutarjoukset, niin toisessa ääripäässä ovat kirjoitusvirheitä vilisevät ostotarjoukset, joissa etsitään nopeita ja yleensä pieniä kauppoja. Osa myyjistä olikin varautunut tähän asettamalla minimitilausmäärän. KRP:n (2020) mukaan keskitason huumekauppiailla on nykyään aikaisempaa enemmän huumausaineita hallussaan, minkä seurauksena huumausaineita liikkuu Suomessa yhä suurempia määriä. (THL 2020 [KRP 2020]) Tämän perusteella bulkkimyyjien viesteistä on loppujen lopuksi vaikea vetää johtopäätöksiä esimerkiksi järjestäytyneen rikollisuuden toiminnasta.

7 Pohdinta

Rikoksentorjunnan ja kansanterveyden kannalta voidaan kysyä, mitä hyötyjä pimeän verkon markkinoiden seuraamisella ja niiden alasajamisella loppujen lopuksi saavutetaan? Historian perusteella vaikuttaa siltä, että mitä paremmin poliisi ja tulli onnistuvat tehtävissään, sitä luovempia ratkaisuja kauppiaat ja ostajat keksivät. Johtaako sivustojen ja palvelimien sulkeminen sekä ylläpitäjien rankaiseminen lopulta siihen, että kaupankäynti muuttuu rikoksentorjunnan kannalta yhä vain monimutkaisemmaksi?

THL:n selvityksen mukaan vakavaa huolta aiheuttaa erityisesti nuorten henkilöiden kasvava osuus laittomien toimijoiden määrästä. Nettikaupan vaivattomat, ajasta ja maantieteellisestä sijainnista riippumattomat laittomat palvelut tavoittavat helposti uusia nuoria ja usein vailla aikaisempaa rikostaustaa olevia käyttäjiä sekä syrjäisiä haja-asutusalueita, joilla ei muuten vielä olisi huumausaineiden tarjontaa. Monet huumausaineiden käyttäjät ovat lainvalvontaviranomaisten esitutkinnoissa kertoneet, että huumausaineita ei ilman nettimarkkinoita olisi heidän saatavillaan. Nettikaupasta tilatut huumeet toimitetaan ostajalle yleensä posti- tai kuriirilähetyksinä tai maastokätköjen kautta. (THL 2022 [KRP 2020].) Tämän perusteella nettimarkkinat tarjoavat matalan kynnyksen rikollisen uran aloittamiselle ja kaupat aiheuttavat onnistuessaan vakavia lisähaittoja varsinkin nuorille. Lisäksi nettimarkkinoiden sulkeminen on varsin lyhytnäköinen ratkaisu ongelmaan, joka palaa aina kehittyneempänä takaisin. Toisaalta internet, sosiaalinen media ja pikaviestimet ovat tulleet jäädäkseen, minkä takia viranomaisten tulisi jatkaa ja kehittää ennaltaehkäisevää työtä. Myöskään vanhempien vastuuta nuorten suojelemisen ja digialustojen käytön valvomisen suhteen ei voi peräänkuuluttaa liikaa. Poliisin ja Tullin tulisi kehittää täsmällisempiä menetelmiä verkkorikollisuuden hillitsemiseksi sekä kohdentaa pakkokeinoja maahantuojiin ja tukkumyyjiin. Toisaalta voidaan myös pohtia onko olemassa joku kaupankäyntiin liittyvä (tekninen) raja, joka ennaltaehkäisee juuri nuorten ja uusien tekijöiden kynnystä ajautua rikolliseksi?

Jo pelkästään tämän tutkimuksen aineistonkeruun aikana kaupankäynti siirtyi enenevissä määrin pikaviestimien puolelle. Saman havainnon on tehnyt aiemmin KRP (2020): ”Huolestuttava kehityssuunta liittyy myös huumausaineiden tarjontaan ja huumeiden tai huumekokemusten jakamiseen sosiaalisen median, tätä nykyä erityisesti mobiilisovellusten, kautta.

Tätä alaa on vaikea valvoa, ja salattujen verkkojen ja pikaviestinten sekä sosiaalisen median käyttö lisääntyy jatkuvasti.” (THL 2022 [KRP 2020]) Tämä hankaloittaa ja on jo hankaloittanut pimeän verkon seuraamista ja yleisesti tutkimuksen tekemistä, koska tarkemmat kaupankäyntitiedot ym. yksityiskohdat jaetaan salaisesti ostajan ja myyjän välillä. Vertailun vuoksi kaupankäyntiin liittyvää informaatiota (määrät, kuvat, hintatiedot, sijainnit) jaettiin paljon avoimemmin esimerkiksi internetistä poistuneen Rautatieasema-kuvalaudan langoilla.

Tutkimusta ei ole tietenkään mahdotonta toteuttaa pikaviestimien puolella, mutta siihen liittyy jo enemmän eettisiä ja teknisiä kysymyksiä – Kuinka laillista on seurata ja tallentaa systemaattisesti kahden (tai useamman) henkilön välistä rikolliseen toimintaa liittyvää kommunikointia? Miten seuranta on teknisesti laillista ja mahdollista toteutaa? Entä kuinka parantaa ja jatkokehittää tutkimuksessa esitettyä skriptiä sekä lajittelu-algoritmia pikaviestimiin sopivaksi? Näihin kysymyksiin ei vastata tässä tutkimuksessa, mutta ne tarjoavat aiheita jatkotutkimuksiin.

Jatkossa voidaan myös pohtia kaupankäyntiin liittyviä transaktioita ja niiden jäljitystä. Tämä on teknisistä syistä todennäköisesti vaikeaa ellei jopa mahdotonta. Kvanttitietokoneet voivat ratkaista siihen liittyviä haasteita, mutta toisaalta olla itsessään uusi uhka, kun esimerkiksi salausten murtaminen muuttuu tehokkaammaksi. Lisäksi olisi hyvä miettiä bottien roolia nyt ja lähitulevaisuudessa. Tullaanko pimeän verkon kauppapaikolla näkemään samanlaista kalastelua joko bottien tai huijareiden toimesta kuten Facebook-kirpputoreilla ja Tori.fi:n kaltaisilla sivustoilla?

Viranomaisten puolestaan tulisi ennakoivasti siirtää resursseja pikaviestimien seurantaan ja pohtia niihin liittyviä eettisiä ja juridisia haasteita. Näennäisesti voi näyttää hyvältä, kun jokin iso kauppapaikka saadaan alas viranomaisyhteistyöllä. Tämä kuitenkin näyttää siirtävän kaupankäynnin juuri pikaviestimien puolelle, mikä oletettavasti hankaloittaa kaupankäynnin seurantaa ja tutkimista. Kaupankäynti ikään kuin lakaistaan maton alle, jolloin viranomaistoiminta näyttäytyy tehokkaana. On toki tavoiteltavaa, että rikollisuutta ja huumekaupasta aiheutuvia haittoja vähennetään viranomaistoiminnalla, mutta tutkimustiedon valossa pelkkä kauppapaikkojen sulkeminen ei välttämättä ole tähän pitkällä tähtäimellä tehokkain keino.

8 Yhteeveto

Tutkimuksessa tutkittiin suomalaista huumekauppaa pimeässä verkossa data-analytiikan keinoin. Tutkimuksen tavoitteena oli selvittää:

1. Mitä huumausaineita Dread-sivustolla esiintyy?
2. Mikä on huumausaineiden jakauma?
3. Ketkä niitä myyvät?
4. Missä niitä myydään?
5. Nouseeko datasta anomalioita?

Tutkimuskysymykset käsittelevät huumausaineiden esiintyvyyttä pimeässä verkossa, ja tutkimustulosten perusteella arvoidaan huumausaineiden esiintyvyyttä Suomessa. Tutkimustulosta verrattiin THL:n tekemään jätevesitutkimukseen. Tutkimuksen aineisto ($N \approx 8500$) kerättiin tiedonlouhinnan keinoin ja sen pohjalta on laadittiin tilastollinen analyysi luokittelijaalgoritmia ja sekaannusmatriiseja hyödyntäen. Kyseessä on tapaustutkimus, jonka analyysimenetelmäksi valikoitui luokittelu. Tutkimustuloksia verrattiin pääasiassa THL:n tekemään jätevesitutkimukseen sekä muihin THL:n sekä KRP:n tietoihin.

Tutkimuskysymyksistä vastaukset saatiin kysymyksiin 1, 2, 4 ja 5. Kysymykseen 3 ei voitu vastata tämän tutkimuksen pohjalta, koska myyjistä ja ostajista ei ole käyttäjänimeä tarkempia tietoja. Tutkimustuloksen perusteella suomalaisessa huumekaupassa esiintyy eniten kannabista, amfetamiinia ja lääkkeitä. Myös ekstaasia ja subutexia esiintyy verrattain paljon. Hallusinogeenit ja kokaiini muodostavat pienen osuuden jakaumasta ja heroiinia ei esiinny juuri lainkaan. Tutkimustuloksen perusteella amfetamiinilla ja kokaiinilla on todennäköisesti tutut ja turvalliset myyntiringit, jolloin kauppaa ei tarvitse käydä niin paljoa pimeän verkon puolella. Subutex ym. opioidijohdannaiset ovat todennäköisesti korvanneet heroiniin käytön, minkä takia sitä ei esiinny juuri lainkaan tutkimusaineistossa. Yleisesti näyttää siltä, että huumekauppaa käydään koko Suomessa vaikka kaupankäynti näyttäisi vakiintuneen suurimpiin kaupunkeihin. Tämä ilmenee viestien määrän ja käyttäjien aktiivisuuden perusteella paikkakunnittain. Tutkimustulos ja muut tutkimuksessa esiin nousseet havainnot ovat linjassa aikaisemman tutkimustiedon kanssa. Tutkimustulosta voidaan siis kaiken kaikkiaan

pitää luotettavana.

Lähteet

Affairs, Privacy. 2022. “Web Scrapers Claim to Possess and Sell Personal Data on 1.5 Billion Facebook Users on a Hacker Forum”. https://www.privacyaffairs.com/facebook-data-sold-on-hacker-forum/.

Alexander, Patricia A. 1992. “Domain knowledge: Evolving themes and emerging concerns”. *Educational psychologist* 27 (1): 33–51.

Alexander, Patricia A, ja Judith E Judy. 1988. “The interaction of domain-specific and strategic knowledge in academic performance”. *Review of Educational research* 58 (4): 375–404.

Alexander, Patricia A, Diane L Schallert ja Victoria C Hare. 1991. “Coming to terms: How researchers in learning and literacy talk about knowledge”. *Review of educational research* 61 (3): 315–343.

Alkhatib, Bassel, ja Randa S. Basheer. 2019. “Mining the Dark Web: A Novel Approach for Placing a Dark Website under Investigation”. *International Journal of Modern Education and Computer Science.*

Bakken, Silje Anderdal, ja Jakob Johan Demant. 2019. “Sellers’ risk perceptions in public and private social media drug markets”. *International Journal of Drug Policy* 73:255–262. ISSN: 0955-3959. https://doi.org/https://doi.org/10.1016/j.drugpo.2019.03.009. https://www.sciencedirect.com/science/article/pii/S0955395919300659.

Baravalle, Andres, Mauro Sanchez Lopez ja Sin Wee Lee. 2016. “Mining the Dark Web: Drugs and Fake Ids”. Teoksessa *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW),* 350–356. https://doi.org/10.1109/ICDMW.2016.0056.

Bergman, Michael K. 2001. “White Paper: The Deep Web: Surfacing Hidden Value”. *Journal of Electronic Publishing* 7 (1). https://doi.org/https://doi.org/10.3998/3336451.0007.104.

Brooks, Nick. 2018. “DARK NET: Light in the Dark”. https://www.kaggle.com/nicapotato/dark-net-light-in-the-dark.

Broséus, J., D. Rhumorbarbe, C. Mireault, V. Ouellette, F. Crispino ja D. Décary-Hétu. 2016. "Studying illicit drug trafficking on Darknet markets: Structure and organisation from a Canadian perspective". Special Issue on the 7th European Academy of Forensic Science Conference, *Forensic Science International* 264:7–14. ISSN: 0379-0738. https://doi.org/https://doi.org/10.1016/j.forsciint.2016.02.045. https://www.sciencedirect.com/science/article/pii/S0379073816300676.

Chen, Hsinchun. 2011. "Dark Web: Exploring and Mining the Dark Side of the Web". Teoksessa *2011 European Intelligence and Security Informatics Conference,* 1–2. https://doi.org/10.1109/EISIC.2011.78.

Chen, Hsinchun, Wingyan Chung, Jialun Qin, Edna Reid, Marc Sageman ja Gabriel Weimann. 2008. "Uncovering the dark Web: A case study of Jihad on the Web". *Journal of the American Society for Information Science and Technology* 59 (8): 1347–1359. https://doi.org/https://doi.org/10.1002/asi.20838. eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.20838. https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.20838.

Christin, Nicolas. 2013. "Traveling the Silk Road: A Measurement Analysis of a Large Anonymous Online Marketplace", 213–224. WWW '13. Rio de Janeiro, Brazil: Association for Computing Machinery. ISBN: 9781450320351. https://doi.org/10.1145/2488388.2488408. https://doi.org/10.1145/2488388.2488408.

Cohen, Kevin Bretonnel. 2014. "Chapter 6 - Biomedical Natural Language Processing and Text Mining". Teoksessa *Methods in Biomedical Informatics,* toimittanut Indra Neil Sarkar, 141–177. Oxford: Academic Press. ISBN: 978-0-12-401678-1. https://doi.org/https://doi.org/10.1016/B978-0-12-401678-1.00006-3. https://www.sciencedirect.com/science/article/pii/B9780124016781000063.

Darknetlive. 2021. "Cryptonia Market". https://darknetlive.com/markets/cryptonia/.

EMCDDA. 2016. "EU Drug Markets Report: In-depth Analysis", https://www.emcdda.europa.eu/publications/joint-publications/eu-drug-markets-2016-in-depth-analysis_en.

EU. 2022. "Europol". https://european-union.europa.eu/institutions-law-budget/institutions-and-bodies/institutions-and-bodies-profiles/europol_fi.

Europol. 2020. “Europol: Internet Organised Crime Threat Assessment (IOCTA) 2020”, https://www.europol.europa.eu/publications-events/main-reports/internet-organised-crime-threat-assessment-iocta-2020.

———. 2021. “Europol: Internet Organised Crime Threat Assessment (IOCTA) 2021”, https://www.europol.europa.eu/publications-events/main-reports/internet-organised-crime-threat-assessment-iocta-2021.

———. 2022. “IOCTA”. https://www.europol.europa.eu/publications-events/main-reports/iocta-report.

Fayyad, Usama, Gregory Piatetsky-Shapiro ja Padhraic Smyth. 1996. “From Data Mining to Knowledge Discovery in Databases”. *AI Magazine* 17, numero 3 (maaliskuu): 37. https://doi.org/10.1609/aimag.v17i3.1230. https://ojs.aaai.org/index.php/aimagazine/article/view/1230.

Gafni, Ruti, ja Idan Nagar. 2016. “CAPTCHA – Security affecting User Experience”, 905. Tammikuu. https://doi.org/10.28945/3469.

GDS. 2022. “Global Drug Survey”. https://www.globaldrugsurvey.com/.

Global Drug Survey. 2019. “Global Drug Survey”. http://www.globaldrugsurvey.com/gds-2019/.

Haasio, Ari, ja J. Tuomas Harviainen. 2018. “"Laudan parasta lakkaa"– Tor-verkon huumekaupan informaatiokäytännöt”. *Informaatiotutkimus* 37, numero 3 (lokakuu): 29–32. https://doi.org/10.23978/inf.76070. https://journal.fi/inf/article/view/76070.

Haasio, Ari, J. Tuomas Harviainen ja Reijo Savolainen. 2020. “Information needs of drug users on a local dark Web marketplace”. *Information Processing Management* 57 (2): 102080. ISSN: 0306-4573. https://doi.org/https://doi.org/10.1016/j.ipm.2019.102080. https://www.sciencedirect.com/science/article/pii/S0306457319300469.

Hakalahti, Maria. 2019. “Tor-verkon Sipulitorilla tapahtuva identiteettiin liittyvien tietojen jakaminen sekä niihin liittyviä tiedontarpeet”, https://doi.org/https://urn.fi/URN:NBN:fi:tuni-201911075812.

Harihara Sudhan, R Iyer, ym. 2018. “How Good is the Bitcoin as an Alernative Asset for Hedging”. *International Journal of Pure and Applied Mathematics* 119 (17): 497–508.

Hatta, M. 2020. “Deep Web, Dark Web, Dark Net: A Taxonomy of “Hidden” Internet”, https://doi.org/https://doi.org/10.7880/abas.0200908a.

Helsingin yliopisto. 2021. “Tutkijat opettavat tekoälyn käyttämään suomen kielen murteita”. https://www.helsinki.fi/fi/uutiset/tekoaly/tutkijat-opettavat-tekoalyn-kayttamaan-suomen-kielen-murteita.

IBM. 2020. “Natural Language Processing (NLP)”. https://www.ibm.com/cloud/learn/natural-language-processing.

Investopedia. 2022. “Data Mining”. https://www.investopedia.com/terms/d/datamining.asp.

JYU. 2015a. “Luokittelu”. https://koppa.jyu.fi/avoimet/hum/menetelmapolkuja/menetelmapolku/aineiston-analyysimenetelmat/luokittelu.

———. 2015b. “Tapaustutkimus”. https://koppa.jyu.fi/avoimet/hum/menetelmapolkuja/menetelmapolku/tutkimusstrategiat/tapaustutkimus.

KAMK. 2021. “Tapaustutkimus”. https://www.kamk.fi/fi/opari/Opinnaytetyopakki/Teoreettinen-materiaali/Tukimateriaali/Tutkimustyypit/Kuvaileva/Tapaus.

———. 2022. “Luokittelu”. https://www.kamk.fi/fi/opari/Opinnaytetyopakki/Teoreettinen-materiaali/Tukimateriaali/Laadullisen-analyysi-ja-tulkinta/Luokittelu.

Karishma, Miss, ja Pratibha Deshmukh. 2017. “Bitcoin : A secure Medium” (kesäkuu): 103–107.

Karlsson, 1946-, F. 2008. *Yleinen kielitiede.* Uud. laitoksen 3. p. Toimittanut 1946- Karlsson F. Helsinki: Gaudeamus Helsinki University Press.

Kaspersky. 2021. “Mitä syvä ja pimeä verkko ovat?” https://www.kaspersky.fi/resource-center/threats/deep-web.

Kishi, Nobuko, Brian Lavoie, Richard Bennett ja Edward O’Neill. 2002. “Measuring the Deep and Dark Web”.

Kolttola, Ilari. 2020. “Rikollisuustilanne 2020 : Rikollisuuskehitys tilastojen ja tutkimusten valossa”, http://hdl.handle.net/10138/337259.

Korenius, Tuomo, Jorma Laurikkala, Kalervo Järvelin ja Martti Juhola. 2004. “Stemming and Lemmatization in the Clustering of Finnish Text Documents”. Teoksessa *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management,* 625–633. CIKM ’04. Washington, D.C., USA: Association for Computing Machinery. ISBN: 1581138741. https://doi.org/10.1145/1031171.1031285. https://doi.org/10.1145/1031171.1031285.

KRP. 2020. “Keskusrikospoliisin huumausainerikollisuutta koskevat julkaisemattomat tilannekatsaukset sekä Keskusrikospoliisin antamat lisätiedot”.

Kullenberg, Christopher. 2022. “Flashbackscraper”. https://github.com/christopherkullenberg/flashbackscraper.

KvaliMOTV. 2021. “Tapaustutkimus”. https://www.fsd.tuni.fi/menetelmaopetus/kvali/L5_5.html.

Lehtomäki, E., ja R. Kukkaniemi. 2020. “Luonnollisen suomen kielen ymmärtäminen koneellisesti”, https://doi.org/http://urn.fi/URN:NBN:fi:jyu-202005183289. https://jyx.jyu.fi/handle/123456789/69035.

Liddy, Elizabeth D. 2001. “Natural language processing”.

Lyons, John. 1991. *Natural Language and Universal Grammar: Essays in Linguistic Theory.* Nide 1. Cambridge University Press. https://doi.org/10.1017/CBO9781139165877.

Martin, James. 2014. “Lost on the Silk Road: Online drug distribution and the ‘cryptomarket’”. *Criminology & Criminal Justice* 14 (3): 351–367. https://doi.org/10.1177/1748895813505234. https://doi.org/10.1177/1748895813505234.

Mattila, Miisa. 2020. “Tor-verkon käyttö Porin poliisiasemalla”, https://doi.org/https://urn.fi/URN:NBN:fi:amk-2020111022503.

McKiernan, Gerry. 2001. “The Hidden Web”. *Science & Technology Libraries* 20 (4): 65–74. https://doi.org/10.1300/J122v20n04_08.

Nieminen, Paavo. 2010. “Jyväskylä yliopiston IT-tiedekunnan konenäkö-kurssi”. http://users.jyu.fi/~nieminen/konenako2010/konenako_luokittelu_ja_mlp_2010h.pdf.

NordVPN. 2021a. “Mikä on dark web eli pimeä verkko, ja kannattaako sitä käyttää?” https://nordvpn.com/fi/blog/pimea-verkko/.

———. 2021b. “Syväverkko: mikä on deep web ja minkälaisia vaaroja siihen liittyy?” https://nordvpn.com/fi/blog/mika-on-deep-web/.

Nurmi, Juha. 2019. “Understanding the Usage of Anonymous Onion Services: Empirical Experiments to Study Criminal Activities in the Tor Network”, https://urn.fi/URN:ISBN:978-952-03-1091-2.

Nurmi, Juha, ja Teemu Kaskela. 2015. “Silkkitie. Päihteiden suomalaista nappikauppaa”, https://urn.fi/URN:NBN:fi-fe2015120722877.

Nurmi, Juha, Teemu Kaskela, Jussi Perälä ja Atte Oksanen. 2017. “Seller’s reputation and capacity on the illicit drug markets: 11-month study on the Finnish version of the Silk Road”. *Drug and Alcohol Dependence* 178:201–207. ISSN: 0376-8716. https://doi.org/https://doi.org/10.1016/j.drugalcdep.2017.05.018. https://www.sciencedirect.com/science/article/pii/S0376871617302910.

Oinas, Pirjo. 2011. “YMEN1332 Laadullisen tutkimuksen suuntaukset”. https://view.officeapps.live.com/op/view.aspx?src=https%5C%3A%5C%2F%5C%2Fwww.ulapland.fi%5C%2Floader.aspx%5C%3Fid%5C%3D70f29511-8a76-4b33-a5dd-50214af6c2c9&wdOrigin=BROWSELINK.

Paris, Scott G, Marjorie Y Lipson ja Karen K Wixson. 1983. “Becoming a strategic reader”. *Contemporary educational psychology* 8 (3): 293–316.

Personal Library Software. 1996. “PLS introduces AT1, the first ’second generation’ Internet search service”. https://web.archive.org/web/19971021232057/http://www.pls.com/news/pr961212_at1.html.

Perälä, Jussi. 2011. “"Miksi lehmät pitää tappaa?": Etnografinen tutkimus 2000-luvun alun huumemarkkinoista Helsingissä”.

Poliisi. 2021. “Keskusrikospoliisissa luodaan tilannekuvaa pimeän verkon huumekaupasta”. https://poliisi.fi/-/keskusrikospoliisissa-luodaan-tilannekuvaa-pimean-verkon-huumekaupasta.

———. 2022. “Kyberrikokset”. https://poliisi.fi/kyberrikokset.

Rautatieasema. 2020. “Portti Rautatieasemalle”. https://torilauta.net/rautatieasema.

Reddit. 2022. “Can somebody please explain how to solve the Clock-styled captcha?” https://www.reddit.com/r/onions/comments/no4vre/can_somebody_please_explain_how_to_solve_the/.

Rikoksentorjuntaneuvosto. 2022. “Kyberrikokset”. https://rikoksentorjunta.fi/kyberrikokset.

Ron, Dorit, ja Adi Shamir. 2012. “Quantitative Analysis of the Full Bitcoin Transaction Graph”, 6–24. Marraskuu. ISBN: 978-3-642-39883-4. https://doi.org/10.1007/978-3-642-39884-1_2.

Ruotsalainen M. 2021. “Rikollinen kaupankäynti Tor-verkon suomenkielisissä piilopalveluissa”, https://doi.org/http://urn.fi/URN:NBN:fi:jyu-202112105917.

Ryle, Gilbert. 1949. *The concept of mind.* London: Hutchinson.

Räisänen, Tuomas. 2020. “Sniikit diilit, vai hiipiikö hiitit? : Sosiologinen katsaus huumekauppiaan työhön ja internetvälitteiseen huumekauppaan”, https://doi.org/http://urn.fi/URN:NBN:fi-fe2020061242913.

Saito, Kenji, ja Mitsuru Iwamura. 2019. “How to make a digital currency on a blockchain stable”. *Future Generation Computer Systems* 100:58–69. ISSN: 0167-739X. https://doi.org/https://doi.org/10.1016/j.future.2019.05.019. https://www.sciencedirect.com/science/article/pii/S0167739X18300475.

SAS. 2022. “Data Mining”. https://www.sas.com/en_us/insights/analytics/data-mining.html.

Scikit-learn. 2022a. “Confusion matrix”. https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html.

Scikit-learn. 2022b. “sklearn.metrics.confusion$_matrix$”. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html.

Sisäministeriö. 2022. “Kyberrikollisuus”. https://intermin.fi/poliisiasiat/kyberrikollisuus.

Stenlund, S. 1990. *Language and Philosophical Problems.* Cambridge University Press.

Suomen avoimien tietojärjestelmien keskus. 2017. “Sekaannusmatriisi”. https://coss.fi/wp-content/uploads/2017/12/4-Koneoppiminen.pdf.

Syverson, Paul, Roger Dingledine ja Nick Mathewson. 2004. “Tor: The secondgeneration onion router”. Teoksessa *Usenix Security,* 303–320.

Tanthavech, Nitirat, ja Apichaya Nimkoompai. 2019. “CAPTCHA: Impact of Website Security on User Experience”. Teoksessa *Proceedings of the 2019 4th International Conference on Intelligent Information Technology,* 37–41. ICIIT ’19. Da, Nang, Viet Nam: Association for Computing Machinery. ISBN: 9781450366335. https://doi.org/10.1145/3321454.3321459. https://doi.org/10.1145/3321454.3321459.

TechTarget. 2022. “Data Mining”. https://www.techtarget.com/searchbusinessanalytics/definition/data-mining.

THL. 2020. “Huumetilanne Suomessa 2020”, https://urn.fi/URN:ISBN:978-952-343-576-6.

———. 2021. “Jätevesitutkimus: Amfetamiinin käyttö jatkaa kasvuaan Suomessa, kokaiinin käyttömäärien kasvu on tasaantunut”. https://thl.fi/fi/-/jatevesitutkimus-amfetamiinin-kaytto-jatkaa-kasvuaan-suomessa-kokaiinin-kayttomaarien-kasvu-on-tasaantunut?redirect=%2Ffi%2F.

———. 2022. “Jätevesitutkimus: väestötason huumeiden käyttö”. https://thl.fi/fi/tutkimus-ja-kehittaminen/tutkimukset-ja-hankkeet/jatevesitutkimus.

Tiedepankki. 2022. “CAPTCHA”. https://termipankki.fi/tepa/fi/haku/captcha.

Todorof, M. 2019. “FinTech on the Dark Web: the rise of cryptos”. *ERA Forum* 20 (1): 1–20. https://doi.org/https://doi.org/10.1007/s12027-019-00556-y.

Tor Project. 2015. “Nine Questions about Hidden Services”. https://blog.torproject.org/nine-questions-about-hidden-services/.

Tor Project. 2021. “History”. https://www.torproject.org/about/history/.

Tulli. 2017. “Tulli sulki Sipulikanavan yhteistyössä poliisin kanssa”. https://tulli.fi/-/tulli-sulki-sipulikanavan-yhteistyossa-poliisin-kanssa.

———. 2019. “Tuhansien rikosepäilyjen Silkkitie”. https://tulli.fi/web/tullinvuosi/2019/valvonta/huumausainerikokset/tuhansien-rikosepailyjen-silkkitie.

UIAH. 2021. “Tapaustutkimus”. http://www.uiah.fi/virtu/materiaalit/tuotetiede/html_files/14111_totea.html.

———. 2022. “Luokittelu”. http://www.uiah.fi/virtu/materiaalit/tuotetiede/html_files/14113_totea.html.

Van Hout, Marie Claire, ja Tim Bingham. 2014. “Responsible vendors, intelligent consumers: Silk Road, the online revolution in drug trading”. *International Journal of Drug Policy* 25 (2): 183–189. ISSN: 0955-3959. https://doi.org/https://doi.org/10.1016/j.drugpo.2013.10.009. https://www.sciencedirect.com/science/article/pii/S0955395913001722.

Viinikainen, Ari. 2021. “Tor (The Onion Router - anonyymi verkko)”. https://tim.jyu.fi/view/kurssit/tie/itkp104/2020-2021/materiaali#tor-the-onion-router---anonyymi-verkko.

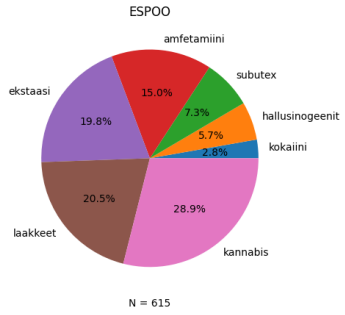
Wikipedia. 2022. “Unicode”. https://fi.wikipedia.org/wiki/Unicode.

YLE. 2021. “Suomi on pimeän verkon huumekaupan kärkimaa, ja se sai KRP:n lisäämään rikollisuuden valvontaa verkossa”. https://yle.fi/uutiset/3-12087690.

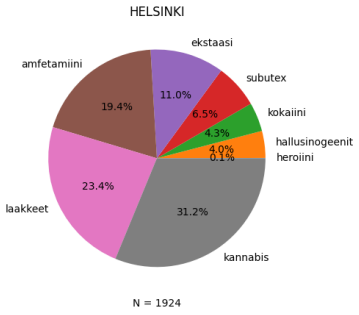
Äyrämö, Sami. 2006. *Knowledge mining using robust clustering*. 63. University of Jyväskylä.

Liitteet

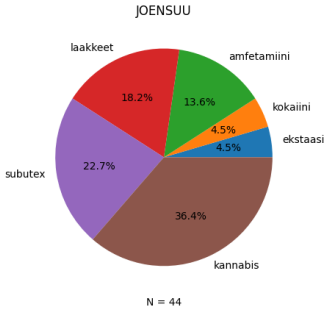
A Huumausaineiden suhteelliset jakaumat Dread-keskustelukanavan viesteissä.



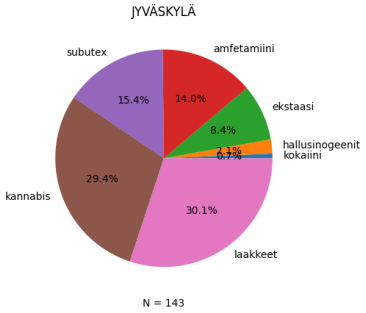
(a) Espoo (n = 568)



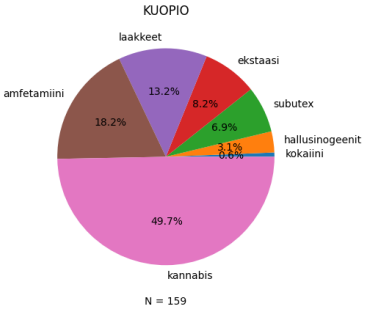
(b) Helsinki (n = 1752)



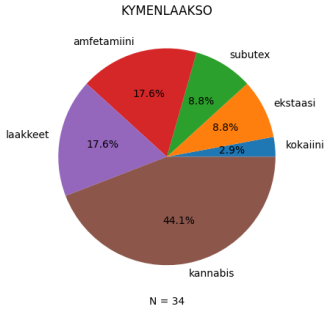
(c) Joensuu (n = 44)



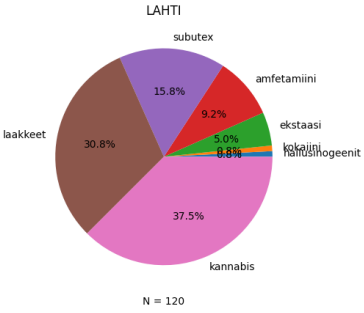
(d) Jyväskylä (n = 132)



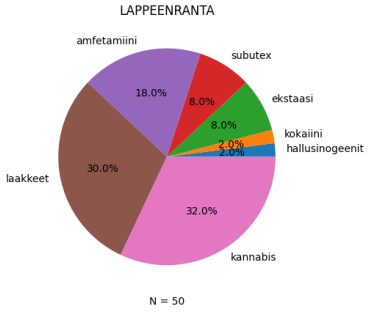
(e) Kuopio (n = 151)



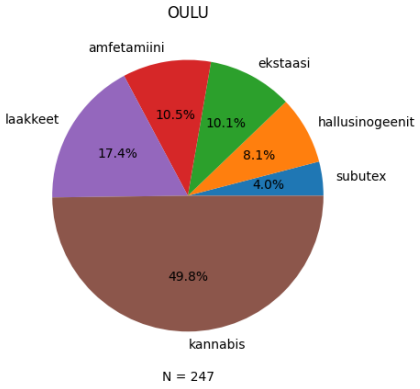
(f) Kymenlaakso (n = 33)



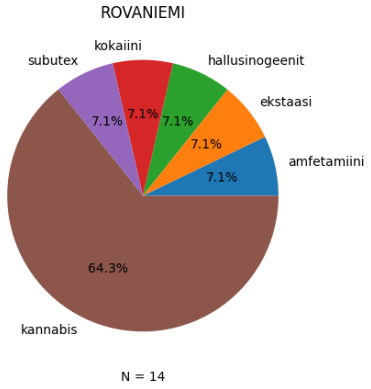
(g) Lahti (n = 105)



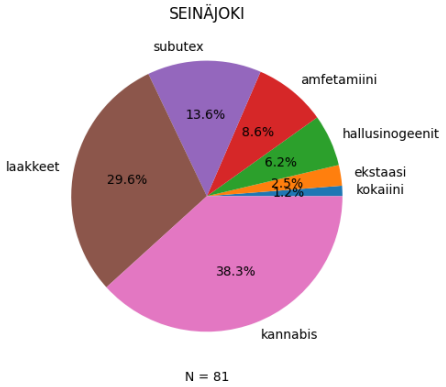
(h) Lappeenranta (n = 45)



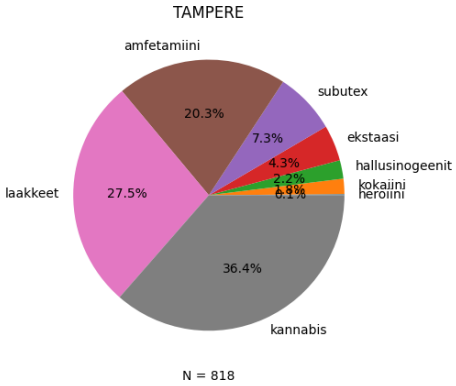
(a) Oulu (n = 230)



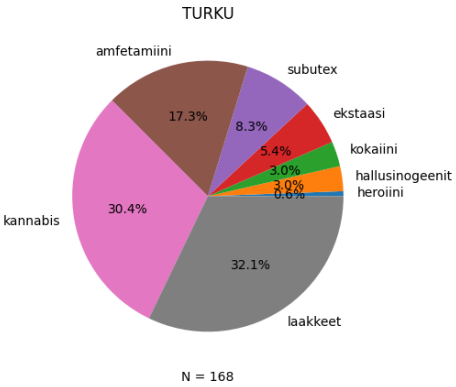
(b) Rovaniemi (n = 13)



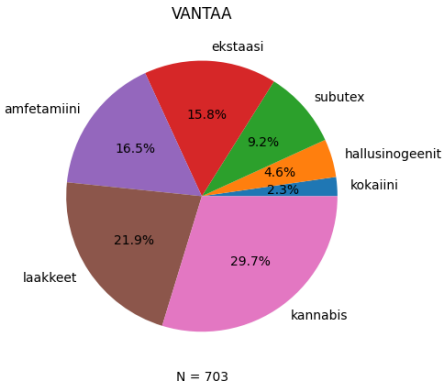
(c) Seinäjoki (n = 76)



(d) Tampere (n = 742)



(e) Turku (n = 154)



(f) Vantaa (n = 653)