Suurandmed ja Google Trends

Sotsiaalse analüüsi alused, Kadri Rootalu

Üha enam räägitakse andmeanalüüsi ja otsuste langetamise puhul suurandmetest ja isetekkelistest andmetest. Käesolevas õppetunnis teeme sissejuhatuse suurandmetesse ja vaatame võimalusi nende analüüsimiseks Google Trends näitel. Ülesannete eduka lahendamise eest võib saada kursuse arvestusse ka 2 boonuspunkti.

Mis on suurandmed?

Suurandmete iseloomustamisel on räägitud eelkõige 3V mudelist. Need kolm V tähte tähendaks inglise keeles:

Volume (maht)

Suurandmete puhul ei tehta (tavaliselt) valimeid. Arvesse võetakse kõik huvipakkuva teema kohta käivad vaatlused või mõõtmised. Sellest tulenevalt on suurandmete maht sõna otseses mõttes suur. Just suurandmete suuruse argument on olnud peamine esimestes suurandmete kohta tehtud töödes ja definitsioonides. Siin võib muidugi tekkida küsimus, et kui suur on suur? Üks tüüpiline vastus oleks, et suurandmed hakkavad sellest suurusest, mis ei mahu enam Excelisse (http://www.cmswire.com/cms/information-management/what-is-big-data-anything-that-wont-fit-in-excel-emetrics-020502.php). Antud kriteerium ei ole siiski absoluutne. Andmete maht on vaid üks argumente, mida arvesse võtta suurandmete kasulikkuse hindamisel.

Velocity (liikuvus, kiirus)

Suurandmed tekivad ja on kättesaadavad reaalajas. See omakorda on seotud andmete mahuga. Näiteks kus hoida andmeid, mis mõõdavad näiteks iga Eesti majapidamise energiatarbimist minuti täpsusega (s.t iga majapidamise kohta on 1440 mõõtmistulemust päevas) ja kuidas saadud ajasõltuvat infot analüüsimisel kõige paremini ära kasutada? Kuidas seda infot reaalajas otsuste langetamisel kasutada?

Variety (mitmekesisus)

Suurandmed esinevad väga mitmekesises vormis: arvuliselt, sõnaliselt, pildiliselt, heliliselt jne. Lisaks tüüpilistele ja meile rohkem tuttavatele arvulisel kujul andmetele analüüsitakse järjest suureneva intensiivsusega teksti ja keelekasutust (näiteks sotsiaalmeedia postitused), pilte ja videoid (näiteks satelliidipiltide klassifitseerimine, meditsiiniliste uuringute käigus tekkinud piltide analüüs, soovi korral loe juurde näiteks: http://analytics-magazine.org/images-a-videos-really-big-data/), geograafilised andmed (näiteks telefonikõnede tegemisel määratud asukoht) jms.

Andmete mitmekesisusega käib kaasas ka andmete puudulikkuse teema. Väga tihti suurandmed ei ole täielikud. Näiteks: on inimesi, kes näiteks iga päev ei kasuta sotsiaalmeediat. Sel päeval on nende kohta käiv info puudu. Vahetevahel on võimalik puuduolevat infot muude allikate põhjal asendada või täiendada. Sellisel juhul kasutatakse koos mitmeid erinevaid andmetüüpe, mis omakorda esitab väljakutseid analüüsile.

Lisaks kasutatakse suurandmetest rääkimisel veel kahte V-d, mis viitavad pigem andmete analüüsimisel tekkivatele probleemidele ja raskustele. Nendeks on:

Veracity (tõepärasus): andmete kvaliteet võib olla väga kõikuv, palju on müra. Tüüpiliseks näiteks on sotsiaalmeediakasutuse analüüsimisel saadavad andmed. Kui vaatame näiteks Twitteris päeva jooksul tehtud säutse, siis vaid osa nendest on originaalsed. Silma hakkavad aga ka robotid ja teiste inimeste säutsude edastamine (retweet).

Variability (muutuvus): andmete tähendus on muutumises. Näiteks sama märksõna aastal 2010 või omada hoopis teist tähendust kui aastal 2017.

Võimaluste poolelt on uute V-dena toodud välja:

Visualization (visualiseerimine): suurte andmehulkade paremaks hoomamiseks on väga kasulik neid visualiseerida. Samas (olete ehk isegi praktikumide käigus märganud) õige ja infot kõige paremini iseloomustava joonise valik ja ettevalmistamine võib olla vägagi töömahukas protsess. Teiselt poolt aga on hea visualiseerimisega võimalik ka info selguses ja järelduste arusaadavuses väga palju võita.

Value (väärtus): olemasolevatel andmetel (kui neid analüüsida osata) on organisatsioonide jaoks väga suur väärtus (ka rahaline). Andmetest räägitakse isegi kui uuest maavarast.

Soovijad võivad lisaks lugeda ka diskussiooni V-de üle (inglise keeles, pole kohtustuslik)
https://datafloq.com/read/3vs-sufficient-describe-big-data/166

Suurandmete kasutamise võimaluste ja probleemide kohta lugege läbi järgmised artiklid:

Anu Masso „Suurandmed – «maavara» või «radioaktiivsed jäätmed»?“

http://tehnika.postimees.ee/3980975/suurandmed-maavara-voi-radioaktiivsed-jaatmed

E.-M. Tiit „Suurandmed statistikas“

https://statistikaamet.wordpress.com/2016/08/25/suurandmed-statistikas/

Eelnevate materjalide põhjal on ka üks osaülesanne praktikumi ülesande juures.

Vaadata võite ka videot suurandmete kohta sotsiaalteadustes (inglise keeles)

https://www.youtube.com/watch?v=XRVIh1h47sAjaindex=51jalist=PLtjBSCvWCU3rNm46D3R85efM0hrzjuAIg

Praktilised ülesanded

Üks lihtsamaid viise suurandmete visualiseerimiseks on Google otsingute analüüsimiseks välja töötatud Google Trends, mis asub lehel https://trends.google.com/trends/ minge sellele lehele.

Seejärel lugege läbi järgmised materjalid Inglise keeles ning teise materjali lõpus tehke test (quiz).

What is Google Trends data — and what does it mean?

https://medium.com/google-news-lab/what-is-google-trends-data-and-what-does-it-mean-b48f07342ee8

Google Trends: See what's trending across Google Search, Google News and …
https://newslab.withgoogle.com/lesson/4781275034419200

Proovige Google Trends lehel läbi erinevaid (ja erinevate piirangutega) otsinguid nii eesti kui inglise keeles. Proovige ka sõnu, millel on erinevates keeltes erinev tähendus. Seejärel vastake küsimustele:

1) Millised on erinevused esiletõstetud lugudes USA ja Suurbritannia vahel (võite valida lisaks ka mõne muu riigi, mille keelt Te valdate)?
2) Tehke otsing kogu maailmas sõnaga „Trump" (jutumärke ärge pange). Millist trendi märkate?
3) Lisage võrdlusesse ka sõna „Clinton" (+ märgi järel kõrvallahtrisse „Võrdle" juurde). Kirjutage, mida näete?
4) Kitsendage eelmine otsing selliselt, et näidataks vaid viimast aastat.
5) Kitsendage otsing Eestile (kogu maailm asemel valige Eesti)
6) Vaadake allpool olevaid Eesti kaarte ja selgitage välja: millises maakonnas otsiti kõige enam Trumpi kohta ja millises maakonnas kõige enam Clintoni kohta

7) Tehke joonis selle kohta, kui palju otsiti Eestis alates 2016a algusest kuni tänaseni sõna „Reformierakond" ja „Keskerakond" (selleks tuleb ajaperioodi juures valida „Kohandatud"

   Saadud joonist saaks jagada või alusandmeid salvestada. Joonise paremal üleval nurgas on noole märk, mis võimaldaks:

   a. joonist kohta käivat linki salvestada/postitada
   b. joonist sotsiaalmeedias jagada
   c. joonise aluseks olevad andmed salvestada csv formaadis (avaneks Exceliga). Selleks peaksite olema aga Google kasutajaga sisse loginud.

8) Tehke omal soovil üks võrdlev otsing Teile huvi pakkuvate otsingusõnadega. Postitage saadud pilt Moodle foorumisse (selleks valige noolenupu alt „<> Manustamine" ning kopeerige ja kleepige saadud link soovitud kohta). Kirjutage juurde, milliseid piiranguid või probleeme antud otsingu juures tuleks arvesse võtta (vt loenguosa materjale).