Domača naloga 2: Big data – Cene

**Billion Prices project**

Prvi vir je the Billion Prices project. Projekt sta ustvarila Alberto Cavallo in Roberto Rigobon, ki sta to razvila iz njunega prejšnjega projekta Inflacion Verdadera. Ta stran združuje podatke iz več različnih projektov teh ustvarjalcev. Na strani je 6 različnih baz podatkov. Prva se nanaša na podatke o cenah za vse artikle v dveh ameriških veletrgovcih. Podatki so pridobljeni s strani ameriškega urada za delovno statistiko za 94511 artiklov od leta 2017 do 2020. Drugi vir so podatki o indeksih cen za Argentino in ZDA, ter pariteti kupne moči za Argentino in Avstralijo, poleg tega je še podatek o inflaciji za nekaj največjih držav. Gre za zbirko starejših podatkov o teh ekonomskih kazalnikih za leta od 2007 do 2015. Tretji vir so podatki iz raziskave za Inflacion Verdadera, ki jih boš podrobneje opisal kasneje. Četrti vir so dnevne cene za vse artikle prodane s strani štirih globalnih trgovcev prisotnih v skupno 85 državah za leta od 2008 do 2013. Peti vir je primerjava cen na spletu in v fizični trgovici za 10 velikih držav sveta od leta 2014 do 2016. Šesti vir pa so podatki o pariteti kupne moči za 11 velikih držav po svetu od leta 2010 do leta 2017. Te baze so bile primarno zbrane za več različnih raziskav na Harvardu in so tu dostopne za ostale. . Glavne prednosti da gre za podatke, ki so ponavadi težje dostopni sploh za različne države, tu pa so vsi na enem mestu, zelo primerno za sekundarni vir pri raziskavah, večinoma uradni viri in zanesljivi viri. Glavne slabosti pa so, da gre za relativno stare podatke, ni nobenih aktualnih podatkov, ki bi se trenutno zbirali, večina podatkov je o ZDA in Argentini, manj o ostalih državah, podatki so že zbrani in omejeni. Gre torej za knjižnico več različnih že zbranih baz, ki so prosto dostopne in na voljo za nadaljnje raziskave.

Prvi vir zame ni big data. Gre za knjižnice že prej zbranih podatkov, ki večinoma niti niso zelo veliki viri (samo ena datoteka večja od 1 GB, pa tudi ta samo 6 GB) v primerjavi z drugimi primeri big data podatkov, kjer govorimo o več terabajtih podatkov. Podatki so tudi precej omejeni običajno za nekaj trgovcev ali nekaj držav. Poleg tega so tudi vse to stari podatki, nobeden se ne zbira aktualno in podatki so večinoma ena ali dve informacije zbrane dnevno za nekaj let, kar je vseeno relativno malo informacij, če primerjamo z večjimi big data bazami podatkov. Ni tudi eksponentne rasti podatkov, ki je običajna pri big data, tu niti ni rasti podatkov.

So pa podatki v prvem viru zelo točni glede na to, da gre za podatke s strani uradnih virov, ki so ponavadi najbolj kvalitetni, preostali podatki pa so iz uradnih spletnih strani trgovin, kjer tudi gre za prave vire teh podatkov. Poleg tega so ti podatki tudi preverjeni, uporabljeni v raziskavah v preteklosti in tudi kot sekundarni vir v več raziskavah. Zato bi ocenil, da gre za zelo točne podatke sploh ker tudi gre za raziskovalce iz priznanih univerz.

**Inflacion Verdadera**

Drugi vir je Inflacion Verdadera, ki je prejšnji projekt kreatorjev Billion Prices project. Na spletni strani sta dva projekta prvi je bil o merjenju prave inflacije v Argentini v letih od 2007 do 2012. Na ta projekt se nanaša tudi tretji vir, kjer je na voljo članek s strani ustvarjalcev te strani. Inflacion Verdadera je nastal kot odgovor na hiperinflacijo v Argentini in na sume, da inflacija, ki jo je dajala za uradno država ni resnična. Podatke o inflaciji so preverili tako, da so ugotovili rast cen največjih trgovcev Latinske Amerike. Kljub temu, da je pojem inflacije precej kompleksnejši od cene veletrgovcev, se je v raziskavi izkazalo, da sta vrednosti indeksov pogosto precej podobni. Podatki so zbrani tudi v več državah in v vseh so bile razlike med indeksom spletnih cen in inflacijo zelo majhne, manj kot ena indeksna točka z izjemo Argentine, kjer je bila razlika kar približno 12 indeksnih točk. Rezultate so tudi preverili z dodatnim trgovcem, cenami v fizični trgovini in vse se je ujemalo z njihovimi ugotovitvami, zato sta raziskovalca prišla do sklepanja, da uradni podatek za inflacijo v Argentini ni realen. Iz tega projekta je nastal tudi PriceStats, ki meri inflacijo glede na spremembe cen. Podatki so bili zbrani v Inflacion Verdadera na enak način kot na PriceStats in sicer s tako imenovanim web scrapingom. To je tehnologija, ki iz html podatkov spletne strani razbere ceno produktov in jo shrani v strukturno bazo. PriceStats pa ima poleg tega še določene dodatne nadgraditve, kjer oceni kateri so najboljši trgovci za kasnejše merjenje inflacije in paritete kupne moči, torej izbira trgovce z velikim tržnim deležom prisotne na spletu in v fizičnih trgovinah. Dnevno tudi preverjajo podatke, da preverjajo potencialne težave in jih sestavijo v dnevne statistike. Prednosti PriceStats, ki je nekako tehnološki razvoj iz Inflacion Verdadera so da gre za aktualne podatke, ki se zbirajo in objavljajo takoj, vključujejo večino največjih držav, podatki so tudi redno preverjani, precej dostopni in metodologija je razložena vsaj okvirno. Slabosti ki bi se lahko pojavile so potencialne razlike med ceno na spletu in v trgovinah, relativno težko jih je še naprej širiti in dodajati nove spletne strani, zbiranje podatkov odvisno od tretje osebe torej od trgovcev. Gre torej za globalen vir podatkov o spremembah cen proizvodovna podlagi katerega lahko na relativno enostaven način opazujemo realno dnevno inflacijo, ki je načeloma zelo težko merljiv pojem. Na ta način pa relativno enostavno in učinkovito izmerijo ta pojem. Poelg tega so podatki dostopni za veliko držav sveta in omogoča določeno mero primerjave med državami.

Pri PriceStats gre po mojem za primer big data, podatki o cenah so na voljo za 23 držav in se širijo. Na voljo so tudi za vedno več trgovcev, zato kljub temu da ni podatka o velikosti, sklepam da gre za ogromno bazo cen, obenem je baza še rastoča, sklepam da raste tudi število produktov. Tako da gre za veliko količino aktualnih in še naprej rastočih podatkov, tako da sklepam da gre za primer big data.

Podatki so glede na metodologijo zelo točni. Redno so preverjani tako s strani avtomatskih sistemov, kot tudi ročno. Poleg tega se podatki tudi redno objavljajo in so iz javno dostopnih virov, tako da so izredno enostavno preverljivi. Avtor je tudi zelo izkušen na tem področju in na podoben primer zbira podatke že več kot 10 let, zato tudi on sam daje podatkov nekoliko večjo veljavo.

Drugi projekt na spletni strani pa je Inflacion Verdadera Venezuela, ki pa meri pravo inflacijo v Venezueli. Uporablja pa povsem drugačno metodo kot PriceStats, saj podatke zbirajo fizične osebe, ki slikajo bar kode v trgovinah. Podatke preverjajo algoritem in tudi fizične osebe iz drugih držav, ki ocenijo če gre za pravi podatek. Sistem pa potem te podatke zbere, da v skupno bazo in izračuna spremembo cen. Začeli so z zbiranjem v maju 2017 in še vedno poteka projekt. Prednosti tega projekta so da tu uradni in spletni viri podatkov niso relevantni, zato se zdi da je to edin učinkovit način ugotavljanja spremembe cen, podatki so tudi povsem aktualni in realni. Glavna slabost pa je da je precej odvisno od odziva ljudi, če podatke zbira veliko ljudi so ti precej bolj natančni, ena oseba tudi ne zbere veliko podatkov (samo za do 10 artiklov) in ker so ljudje prostovljci, lahko zbirajo podatke na slab način (neredno, iz različnih trgovin, različni artikli in podobne napake). Sam projekt je primer kako bi lahko pravi podatki pomagali pri reševanju enega najhujših modernih problemov hiperinflacije.

Projekt vsekakor poiskuša postati big data, kjer bi podatke zbiralo čim več prostovoljcev, ampak glede na to, da je vse skupaj še v beta razvojnem obdobju, sklepam da teh podatkov še ni tako veliko. Sicer se število podatkov povečuje, glede na to, da jih še vedno aktualno zbirajo, a ta odvisnost od prostovoljcev jih precej omeji, sploh ker je to prostovoljstvo omejeno na Venezuelo. Zaradi tega sklepam, da samih podatkov ni tako ogromno, tudi ker vsak prostovolec slika približno 10 produktov, vsakič ko gre v trgovino, torej ne zbira na dnevni ravni.

Podatki sami po sebi predvidevam da tudi niso precej točni. Sicer jih preverjajo tudi razvijalci projekta, a podatke zbirajo nestrokovnjaki in tudi preverjajo jih nestrokovnjaki, zato tu nekakšna visoka kvaliteta podatkov ni mogoča, a tudi se mi zdi, da ni v interesu, bolj je namen pomoč Venezueli, ki se bori s hiperinflacijo.

**Numbeo**

Zadnji vir pa je spletna stran Numbeo, ki je baza življenskih stroškov. Za države in mesta sveta ima stran zbrane informacije o cenah v trgovinah, restavracijah, transportu, položnicah, oblačilah, nepremičninah in ostalih stroških. Uporabno je predvsem pri primerjanju življenja med različnimi mesti po svetu. Dodali so še druge indikatorje kot so kvaliteta življenja in cena nepremičnin. Podatki so zbrani s strani uporabnikov in s spleta (spletne strani trgovin, članki in ankete). Iz spletnih virov vnašajo podatke dvakrat letno. Podatki s strani uporabnikov so preverjeni s strani obstoječih podatkov in tako izloči lažne in napačne podatke. Prednosti Numbea so, da gre za mešani način vzorčenja, torej niso odvisni samo od enega vira informacij, rezultati so javno dostopni, zato jih lahko preverimo, podatki so tudi aktualni, dostopni za ogromno držav in mest in imajo veliko število uporabnikov in ljudi ki dajejo te informacije. Glavne slabosti pa so, da po pričanjih uporabnikov precej izloči potencialne spremembe ali pa prepogosto izloči določene ekstreme, dogajajo se tudi primeri prirejanja podatkov, kjer mesta želijo dati občutek, da so cenejša, poleg tega gre za projekt enega človeka v osnovi, ki vseeno ne prikaže celotne metodologije javno, zato se pojavijo vprašanja o vplivih na podatke. Gre torej za projekt, ki ljudem prikaže cene dobrin po svetu, a gre za projekt s strani ene osebe in projekt je doživel veliko rast glede na svoj začetek.

V tem primeru po mojem mnenju je v ozadju projekta big data. Glede na to, da so zbrani podatki za večino mest tudi v Sloveniji sklepam, da gre za ogromno bazo podatkov v ozadju, baza tudi rasti z rastjo števila uporabnikov, podatki so tudi vedno boljši. Poleg tega imajo tudi bazo podatkov iz interneta, ki sklepam, da je zelo velika. Podatki se zbirajo tudi celoten čas, vsak obiskovalec strani, lahko vnese nove podatke.

Kljub temu, da sem opozarjal na pomisleke o metodologiji in temu, da je samo en ustvarjalec se mi zdi, da vseeno gre za precej točne podatke. Podatke sem na hitro preveril in vsaj za Slovenijo je moj občutek o cenah podoben, obenem je stran zelo popularna in nima veliko kritikov na internetu. Ti napadajo sicer občasno to, da so cene večinoma nekoliko prenizke v razvitem svetu, kar bi lahko pripisali manjšim manipulacijam, kjer poskušajo ljudje olepšati svoje kraje. Niso pa uporabniki zaznali neke večje manipulacije, katera bi verjetno tudi izredno znižala popularnost strani in jo s tem uničila.

Moji aha momenti:

* Najbolj šokanten so mi tu primeri, kako lahko zbiranje kvalitetnih podatkov pomaga državam in predvsem državljanov. Izpostavil bi primer Inflacion Verdadera, kjer so najprej razkrili goljufije države, sedaj pa pomagajo ugotoviti pravo inflacijo in razkriti cene v Venezueli. Nisem si predstavljal, da lahko s samo zbiranjem podatkov imaš tako velik vpliv na državo.
* Obenem pa sem bil presenečen, da so ekipe ljudi ki delajo na teh res velikih projektih zelo majhne, v primeru Numbeo ima ustvarjalec in direktor res veliko moč nad veliko količino podatkov manipulacija katerih bi lahko imela resne posledice, podobno tudi pri Inflacion Verdadera.