Domača naloga 2 - "Big data"

Viri podatkov

Alen Kahteran

18. 03. 2021

Opis problema

Google je opazil, da CDC objavlja podatke o prevalenci gripe v populaciji z nekaj tedensko zamudo. Se pravi, podatki ki jih objavi CDC v nekem trenutku so pravzaprav stanje pred 1 do 2 tedni. Cilj je bil, da bi z iskalnimi nizi, ki jih oseba vtipka v Googlov iskalnik, poiskali takšne nize, ki so visoko korelirani z obiskom pri osebnem zdravniku (kar CDC objavlja z zamudo) in s tem predčasno zaznali porast širjenja gripe. Iskalnih nizov je bilo ogromno (v članku omenijo okrog 50 miljonov iskalnih nizov) zato je bilo potrebno za računanje uporabiti več 100 računalnikov (distribuirani sistemi), ki so dlje časa iskali ustrezne iskalne nize. To je ena izmed tipičnih primerov "Big data" - da ni mogoče v nekem razumnem času z osebnim računalnikom doseči rezultata.

Zasnova modela

Podatke pridobijajo na tak način, da vsakič ko neka oseba vtipka iskalni niz v njihov iskalnik, si ga shranijo. Poleg tega, če se iskalni niz ponavlja večkrat shranijo pogostost. Za ta primer so shranili še lokacijo osebe, zato da so lahko kasneje preverjali uporabo modela tudi lokalno.

Podatke so pripravili tako da so iskalne nize najprej normalizirali s številom vseh iskalnih nizov na določeni regiji v tednu. Poskušali so pripraviti enostaven model sledeče oblike

$$logit(I(t)) = \alpha logit(Q(t)) + \epsilon\,,$$

kjer $I(t)$ predstavlja delež obiskov pri osebnem zdravni oseb s simptomi podobnimi gripi, $Q(t)$ predstavlja iskalne nize ki so povezani s simptomi podobnimi gripi v času t, α predstavlja multiplikator, ϵ pa napako.

Da jim je uspelo najti ustrezne iskalne nize, ki so dovolj dobro napovedovali delež obiskov pri osebnem zdravni oseb s simptomi podobnimi gripi, so uporabili vsak iskalni niz kot edino spremenljivko. Za mero "kvalitete" iskalnega niza so uporabili korelacijo. Te iskalne nize so razvrstili po korelaciji in nato izbrali 45 najboljših, saj se je pokazalo da več iskalnih nizov ne izboljša modela.

Iskalne nize so nato utežili glede na pogostost niza v določenem okraju, saj so model pripravili na dveh ravneh (lokalno in skupno).

Večje težave

Kmalu se je pokazalo, da model ki so ga pripravili, ni več vračal dobrih rezultatov in sicer je precenjeval št. ljudi, ki bo obiskalo osebnega zdravnika (tudi do 2x v nekaterih obdobjih). Težav je bilo več. Ena večjih je bila, da niso redno kalibrirali modela. To pomeni da bi morali redno in ustrezno nastavljati uteži v modelu, da bi odražali "stanje" iskalnih nizov. Dodaten problem je v temu, da so algoritem njihovega iskalnika spreminjali, ker so želeli izboljšati kako deluje iskalnik. Na drugi strani to povzroči, da se pogostost

določenih iskalnih nizov zelo hitro spremeni, saj algoritem lahko predlaga druge iskalne nize. Torej iskalni nizi, ki so se uporabili za razvoj modela niso več tako pogosti zaradi samih predlogov algoritma iskalnika. Posledično bi bilo potrebno to tudi inkorpirirati v model saj se je kvaliteta podatka spreminjala v času (ali pa upoštevati na nek način). Podobno so pozabili pomisliti na to, da se iskalni nizi spreminjajo v času, saj se uporabniki iskalnikov starajo oz. prihajajo nove generacije. Druga večja težava je najverjetneje bila, da so pravzaprav našli iskalne nize ki so sezonsko korelirani z gripo (bolj popularno v zimskem obdobju). To so tudi sami omenili ko so v članku omenili da so odstranili niz "high school basketball", ki je očitno nepovezan z gripo, medtem ko je bil visoko koreliran saj se srednješolska košarka dogaja ravna v obdobju gripe (jesen-zima-pomlad). Poleg tega nikoli niso razkrili, kateri iskalni nizi so bili tisti, ki so jih izbrali za končni model. To je bil velik problem saj se je najbrž zgodilo to, da so nekateri nizi bili po naključju korelirani z obiski pri osebnem zdravniku. Verjetnost, da je 1 primer niza (oz. nekaj primerov) koreliran z obiski pri osebnem zdravniku, je visoka. Ignorirali so eno večjih logičnih zmot in to je da korelacija še ne pomeni vzročne zveze. To da teh 45 nizov nikoli niso objavili bi lahko bila ravno tako rdeča luč za druge raziskovalce, ker nihče od vpletenih ni bil epidemiolog/infektolog, ki bi lahko preveril te nize. Ker so nizi najverjetneje bili v stilu "kašelj", "vročina", ipd. bi tu zagotovo "peer review" proces opravil svoje. To je klasičen problem "Big data", kjer se velikokrat zgodi da podatki, ki se zajemajo, ne sledijo nekim standardom, ki bi zadovoljili znanstvene raziskave.

Dobre plati

Ideja sama ni bila nujno napačna, saj je bil cilj dobronameren. Tak tip podatkov se lahko zagotovo uporabi v dobre namene, saj se lahko zelo dobro spremlja vedenje posameznikov. Če na same podatke ne bi bilo mogoče vplivati z več različnih kotov, bi lahko dosegli boljšo kvaliteto podatkov in bi model lahko bolje deloval. Prepričan sem da bi lahko s takšnimi podatki (iz iskalnikov) zagotovo zelo dobro postavljali mrežne modele družb, ter kategorizacije posameznikov (v katere "skupine" sodi posameznik). Podobne podatke oz. modele lahko najdemo pri Facebooku, Twitterju, Netflixu, itd., ki najverjetneje že imajo zelo dobre vedenjske modele. Sicer jih ne uporabljajo za takšne namene, vendar glede na to da sem mnenja da bi se dalo na tak način (ne nujno samo s podatki iz iskalnikov) zelo dobro analizirati širjenja bolezni, sem pravzaprav nekoliko razočaran, da do tega še ni prišlo.

Sam primer je zagotovo dvignil nekaj prahu okoli transparentnosti in ponovljivosti. To je pomemben proces pri vseh modelih, ker če ni mogoče ponoviti raziskave se hitro pojavi vprašanje koliko so rezultati takšne raziskave koristni.

Zaključek

Kot se je pokazalo v več študijah, je kombinacija več stvari boljša kot samo Googlov model. Primer tega je bil ko so kombinirali Googlov model s standardnim CDC monitoringom in so že samo s tem precej izboljšali rezultate. Po temu ko je Google objavil podatke so nekateri drugi znanstveniki z drugačnimi modeli, ravno tako dosegli izboljšane rezultate. Podobno idejo peljejo dalje na inštitutu za kognitivno znanost v sklopu Univerze v Osnabrücku, s tem da uporabljajo še podatke od Twitterja.

Največji "aha" moment, ki sem ga doživel je to, da Google kljub svoji velikosti, ni uspel korektno postaviti relativno enostavnega modela in ga nato vzdrževati. To da se tudi "velikim" dogajajo napake je le odraz tega da smo ljudje.

"Big data" ni nujno slaba stvar, če se jo ustrezno uporabi. Tolikšne količine podatkov kot jih imajo nekatera podjetja je lahko že nekoliko zaskrbljujoče, saj v primeru da ti podatki "uidejo" bomo priča velikim posegom v zasebnost posameznikov. Hkrati pa vidimo, da se tak tip podatkov lahko uporabi v dobre namene, sicer je vprašljivo če bodo podjetja to storila samoiniciativno. V najboljšem primeru lahko mogoče pričakujemo le podatke, pa še to je po mojem mnenju malo verjetno. To, ko količina podatkov že dosega absurdne velikosti, je del "industrije 4.0", ko digitalizacija poteka povsod. Posledično lahko najdemo senzorje oz. nek objekt ki zajema podatke na vsakem koraku. Zato je potrebno še posebej paziti, da se podatke korektno zajema, obdeluje in nato tudi uporablja.