Google flu trends – »Big data«

Sezonska gripa je velik zdravstveni problem, ki letno povzroči na milijone dihalnih obolenj in 250.000-500.000 smrti vsako leto. Gripa se začne pojavljati večinoma oktobra, vrh doseže decembra in februarja ter izzveni do maja. Znano je, da je gripa virus, ki vsako leto mutira in s tem pride do pojava novih sevov. Težko je napovedati kakšen sev bo krožil v populaciji naslednjo sezono (zimo), zato se samo cepivo prilagaja glede na obstoječe seve gripe, ki krožijo v populaciji. Ob pojavu novega seva, s katerim človek še ni prišel v stik, lahko ta povzroči epidemijo in posledično tudi več smrtnih žrtev, zato je potrebno, da se lokacije širjenja virusa hitro odkrije in ustrezno ukrepa. Namen Google flu trends-a je bil razviti avtomatizirano metodo, ki bi na podlagi iskalnih poizvedb napovedovali tedensko aktivnost gripe v posamezni zvezni državi v Združenih državah Amerike. S tem pripomogli k odkrivanju novih žarišč izbruha gripe. Iz modela so izključili napovedovanje prisotnosti gripe izven sezone (od maja od oktobra).

Center za preprečevanje in obvladovanje bolezni (CDC) pridobiva podatke od 8 ključnih ustanov (ponudniki zdravstvenih storitev, bolnišnice, klinični in javni laboratoriji itd.) nato združijo podatke v 5 kategorij, iz katerih poskušajo ugotoviti: kdaj in kje se je gripa pojavila, določiti kateri sevi virusa gripe krožijo v populaciji, spremembe v virusu (ali je prišlo do mutacije) in izmeriti kakšen vpliv bo imela gripa na ostale bolezni, hospitalizacijo in smrti. Podatke o deležu pozitivnih primerov in obiskov zdravnika v povezavi z gripo in njej podobnimi simptomi objavljajo z enotedenskim zamikom, podatke o smrtnosti pa objavljajo z zamikom dveh tednov.

Google je na začetku vzel zgodovino iskalnih poizvedb, ki so jih uporabniki iskali med leti 2003 in 2008. Iz teh so izračunali časovno serijo, kjer so za vsak teden pridobili 50 milijonov najpogostejših iskalnih poizvedb v ZDA. To so naredili za vsako zvezo državo posebej. Nato so oblikovali avtomatizirano metodo, kjer so kot pojasnjevalno spremenljivko v linearnem modelu vzeli eno poizvedbo, ter pogledali kako se model prilega podatkom CDC in izračunali korelacijo. Ta samodejni postopek je ustvaril seznam najpogostejših iskalnih poizvedb, ki so bile razvrščene glede na vrednost korelacije. Nato so seznam razdelili še na n število najpogostejših iskalnih poizvedb. Na koncu so izbrali 45 poizvedb, ki so imele najvišjo vrednost korelacije in so se najbolje prilegale CDC podatkom. Teh 45 poizvedb naj bi vsebovalo besedne zveze, ki so povezane z gripo. Ostale poizvedbe, ki so dajale dobre rezultate so odstranili, saj so bile besedne zveze povezane s košarko. Sprva so hoteli že v

naprej predpostaviti, katere poizvedbe so povezane z gripo, vendar jih je skrbelo, da bi s predhodnim filtriranjem podatkov izgubili del pomembne informacije. Končni model je uspel doseči zelo dobro ujemanje s CDC podatki, s povprečno korelacijo 0.9.

Spodletelo jim je leta 2012/13, ko so napovedali dvakrat več aktivnih primerov gripe, kot so poročali podatki CDC, kar si ne bi smeli dovoliti. Eden izmed razlogov zakaj je prišlo do tega je, da je Google takrat uvedel avtomatizacijo predlogov poizvedb v brskalniku. Predlagane besede se prikažejo glede na to, kaj so drugi uporabniki iskali oziroma izbrali. Posledično pride do povečanja števila teh iskalnih poizvedb in do povečanja napovedi. Na pogostost iskalnih poizvedb lahko vplivajo tudi zunanji dogodki (npr. novice v medijih), odpoklic zdravila za prehlad in/ali gripo, radovednost ljudi,… Presenetilo me je, da z uvedbo te novosti Google ni posodobil in prilagodil algoritma modela novemu načinu iskanja.

Dobri vidiki Google flu trends-a so, da bi s trenutnimi ocenami in napovedmi, omogočili hitro odkrivanje novih žarišč, boljši odziv zdravstva, aktivacija dodatnih resursov znotraj zvezne države/regije, nakupili bi več cepiv, po potrebi bi ljudi ozaveščali preko medijev, itd. Google flu trends bi lahko bil tudi dober dodatni vir k tradicionalnemu zbiranju podatkov, tako da bi prispeval dodatne informacije, ki jih CDC ne more zbirati na lokalni ravni. Dober vidik je tudi, da se podatki zbirajo hitro in avtomatizirano, vendar tu lahko prihaja do raznih napak.

Problem, ki ga vidim je, da ti podatki niso javno dostopni tako, da ni možno preveriti točnosti podatkov. Nekdo bi lahko za svoje potrebe in cilje tudi prirejal podatke. Ravno tako je vprašljiva tudi kakovost podatkov, saj so med najpogostejšimi poizvedbami pojavili tudi tisti, ki niso bili povezani z gripo, a vendar so njihovi modeli dajali zelo dobre rezultate. Ker podatki niso javni, ni možno pridobiti informacije katerih 45 poizvedb so uporabili za napovedovanje. Podatke ravno tako ni možno pridobiti v znanstvene in raziskovalne namene, da bi raziskovalci ponovili analizo in ugotovili kje je prišlo do napake, da so napovedovali dvakrat več primerov kot jih je v resnici bilo. Pojavlja se tudi vprašanje ali je identiteta vsakega uporabnika zaščitena in ni vključena v bazo podatkov kot to zatrjuje Google. To bi mogoče bil tudi odgovor na vprašanje zakaj ti podatki niso dostopni.

Do neke mere gre za množične podatke, kjer imamo ogromno količino podatkov, ki se ustvarjajo v času, se generirajo z veliko hitrostjo, se hitro spreminjajo, ter so hitro na voljo.

Sama ideja Googla se mi je zdela zanimiva. Mislim, da so imeli dober namen, vendar sama ideja ni bila v celoti dovolj premišljena, saj niso imeli nekega strokovnega znanja iz področja epidemiologije, infektologije. Preveč so se osredotočali na korelacijo, ki pa ne predstavlja

vzročnosti. Bolje bi bilo, da bi naredili napovedi z modelom, v katerega bi vključili podatke CDC od zadnjih dveh tednov in model bi se posodabljal na tedenski bazi.

Viri:

Ginsberg, J. et al. (2009). Detecting influenza epidemics using search engine query data. Nature, 457, 1012-2015. Najdeno na http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html (19. 3. 2021)

Arthur, C. (27.3.2014). Google Flu Trends is no longer good at predicting flu, scientists find. The Guardian.Najdeno na https://www.theguardian.com/technology/2014/mar/27/google-flu-trends-predicting-flu (19. 3. 2021)

Lazer, D., Kennedy, R., King, G., Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis.Science, 343(6176), 1203-1205. Najdeno na http://science.sciencemag.org/content/343/6176/1203.full (19. 3. 2021)

Center for Disease Control and Prevention (19.10.2018). Overview of Influenza Surveillance in the United States. Najdeno na https://www.cdc.gov/flu/weekly/overview.htm (19. 3. 2021)