UPORABA INOVATIVNIH METOD IN PODATKOV V URADNI STATISTIKI

Črt Grahonja[1]
[1]Statistični urad RS, Litostrojska 54, Ljubljana
crt.grahonja@gov.si

Dandanes uporabniki zahtevajo vedno bolj pogoste, natančnejše in hitreje dostopne statistike, zaradi česar je inovacija postopkov v uradni statistiki vedno pomembnejša. Statistični urad Republike Slovenije v okviru Evropskega statističnega sistema (ESS) zato že leta sodeluje v projektih Eurostata in Evropske komisije, ki se ukvarjajo z zbiranjem in obdelavo podatkov iz digitalnih virov, razvojem in uporabo novih tehnologij in ustvarjanjem okolja, ki olajša prestop v prakso na teh področjih.

V prispevku je prikazan primer projekta, kjer se besedilni podatki strgajo z interneta in obdelujejo z uporabo strojnega učenja in metod besedilnega rudarjenja. Projekt se letos končuje in s prispevkom bi radi opisali, kakšne so splošne izkušnje ob takem sodelovanju.

Cilj projekta je bil razvoj procesnega okolja, ki bi omogočalo vsem vključenim državam članicam pripravo lastnih internetnih strgalnikov, obdelavo postrganih podatkov, izmenjavo izkušenj in delovnih procesov ter standardizacijo skupnih postopkov. Od samega začetka so bila izpostavljena tri najbolj problematična področja: standardizacija procesov z upoštevanjem velikega števila zelo različnih jezikov, pridobitev dostopov do širokega nabora virov podatkov z nejasno populacijo in priprava modelov za obdelavo podatkov z ustrezno kakovostjo.

Za namen pridobivanja podatkov je bilo vzpostavljeno okolje za zbiranje podatkov *Web Intelligence Hub* (WIH). V Hubu lahko vsak uporabnik naloži vhodne internetne naslove, pripravi in testira strgalnike in sproži proces strganja, ki podatke shranjuje v skupno bazo podatkov. Do podatkov se lahko potem dostopa v obdelovalnem okolju *Datalab*, kjer se izvajajo tudi analiza, čiščenje in obdelava podatkov. Vsi procesi so pripravljeni modularno, kar omogoča izbiro ustreznih postopkov in proste roke pri delu s podatki, obenem pa tudi ohranja standardnost vsakega koraka.

Strgalnike je možno nastaviti na številne načine, ki omogočajo zbiranje urejenih podatkov, preko besedilnih pravil, na podlagi prisotnosti HTML značk, z uporabo dinamičnih elementov itd. Uporabniki lahko sami razvijamo rešitve za analizo in obdelavo podatkov in jih delimo preko repozitorija v *GitHubu.* Predvsem se spodbuja razvoj prosto-kodnih rešitev, po navadi v Pythonu, R-u, JavaScriptu in podobno.

Ena večjih dodatnih vrednosti je bila vzpostavitev procesa za zbiranje podatkov o prostih delovnih mestih v evropskem okolju, ki deluje ne glede na jezik in zbira podatke za 9 lastnosti, kot so poklic, jezik, zahtevana izobrazba, država in kraj dela itd.

V okviru iskanja visoko kakovostnih rešitev se je organiziralo več manjših podenot, ki so se specializirale za določene naloge, kot je zbiranje lastnosti podjetij, priprava učnih množic in meril uspešnosti za metode strojnega učenja. Vzpostavila se je posebna delovna skupina, ki se je ukvarjala izključno s preverjanjem in razvojem ustrezne metodologije za računanje statistik na področju prostih delovnih mest.

Ključne besede: internetno strganje, besedilno rudarjenje, NLP, strojno učenje, evropsko sodelovanje

USE OF INNOVATIVE METHODS AND DATA SOURCES IN OFFICIAL STATISTICS

The Statistical Office of Slovenia often cooperates in Eurostat and European Commission projects to answer user demand for new, timelier and more detailed statistics. One such project, dealing with internet scraping and machine learning, is ending this year.

Its goal is to create a processing environment that enables all member states to prepare custom scraping, analysing and processing pipelines using modular open-code solutions. Three main fields of interest were recognised: process standardisation regardless of language, accessing data with an unknown population, and quality assurance.

The *Web Intelligence Hub* was developed to extract internet data. In the Hub each user can customise their scrapers, test them, and initiate scraping processes in a standardised fashion. Users can then access the data and manipulate them in the *Datalab* using Python, R or other open-code programming languages.

Participants are organised in smaller groups focusing on specific tasks like collecting online job postings, collecting enterprise characteristics, methodology and quality in machine learning, etc.

Keywords: internet scraping, text mining, NLP, machine learning, European cooperation