Kaj so podatki?

Vloga podatkov v sistemu umetne inteligence

V splošnem digitalnem smislu podatki ustrezajo informacijam, ki jih uporablja, obdeluje in ustvarja programska oprema v računalniškem sistemu.

Brez podatkov ni umetne inteligence. Podatki imajo osrednjo vlogo v vseh procesih strojnega učenja, saj se uporabljajo tako za usposabljanje kot za testiranje. Prihajajo tudi v obliki parametrov, ki se uporabljajo za upravljanje procesov usposabljanja. Končno je sistem umetne inteligence kombinacija določene programske arhitekture z vsemi učenimi parametri, tako imenovanim modelom, ki je prav tako podatek.

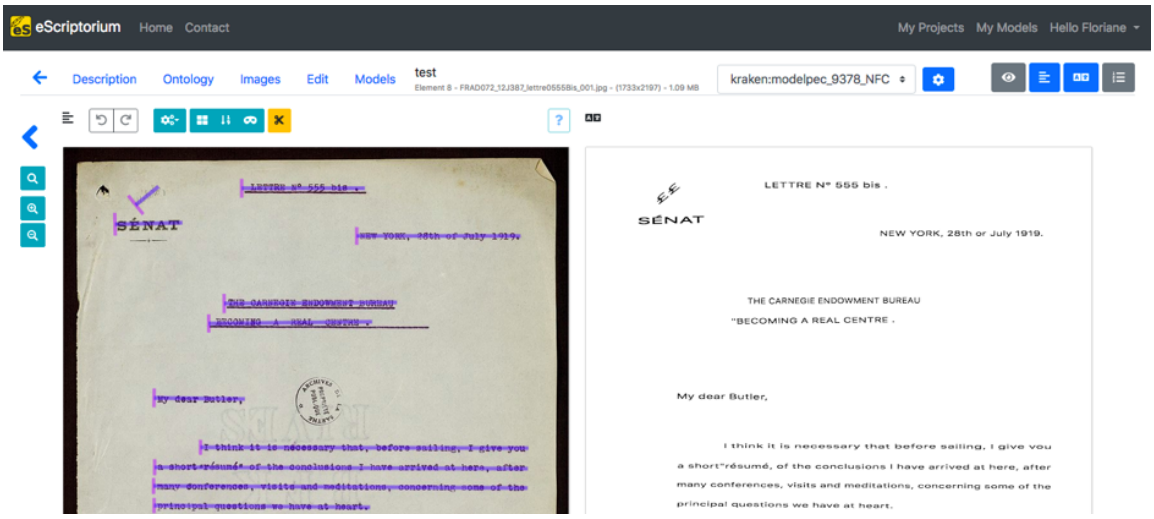
Razumevanje vloge podatkov v sistemih umetne inteligence skupaj z načinom njihovega izbiranja, dokumentiranja in razširjanja je bistvenega pomena za oceno obnašanja sistema umetne inteligence. To je pomembno z vidika ponovljivosti ali za primerjavo dveh različnih sistemov umetne inteligence.

V svetu obdelave naravnega jezika je na primer razpoložljivost velikih količin govorjenih in pisanih podatkov ključnega pomena za delovanje programov za preverjanje črkovanja, napovedovanje v iskalnikih ali seveda za strojno prevajanje. Uporabljajo se za izgradnjo tako imenovanih jezikovnih modelov, ki nadaljnjim postopkom zagotavljajo statistične predstavitve kombinacij besed ali stavkov.

Trajnost sistemov umetne inteligence je torej tesno odvisna od metod upravljanja podatkov, ki se uporabljajo pri njihovi zasnovi.

Podatki za nadzorovane in nenadzorovane sisteme umetne inteligence

Kot smo že videli, so sistemi umetne inteligence na voljo v dveh različicah, odvisno od načina uporabe podatkov za njihovo usposabljanje. Nadzorovani sistemi se zanašajo na zagotavljanje vhodnih podatkov skupaj z ustreznimi predvidenimi izhodnimi podatki. Usposabljanje je torej sestavljeno iz učenja sistema, da iz neznanih vhodov ustvari najverjetnejši izhod. Takšne podatke je mogoče pridobiti na različne načine. Na primer slikovna zbirka podatkov, kjer je vsaka slika povezana s ključnimi besedami, ali zbirka digitaliziranih dokumentov, ki so jih prepisali anotatorji (glej spodnjo sliko).



Slika: Slika 1: Samodejni prepis pisma Paula D'Estournellesa (z dovoljenjem F. Chiffoleauja, Coll. Archives de la Sarthe)

Sistemi umetne inteligence, ki temeljijo na nenadzorovanem usposabljanju, ne bodo zasnovani za določeno vedenje, temveč bodo oblikovani tako, da bodo upoštevali statistične lastnosti podatkov za usposabljanje. To velja na primer za jezikovne modele, kot je BERT, ki besedam z enakim skladenjskim ali pomenskim vedenjem običajno pripišejo podobne položaje v matematičnem prostoru, kar se opazi na podlagi zagotavljanja velikega števila vzorčnih stavkov za vsako besedo. Takšni modeli so na primer zelo dobri pri napovedovanju sinonimov ali nadaljnjih besed danega zaporedja.

Viri - izbor, dokumentacija, priprava, anotacija

Zasnova sistema umetne inteligence v bistvu temelji na ustrezni zasnovi nabora podatkov, ki se uporablja za njegovo usposabljanje. Med različnimi dejavniki, ki pridejo v poštev, lahko navedemo ustreznost podatkov za obravnavano nalogo, velikost podatkov, ki mora ustrezati zapletenosti programske arhitekture umetne inteligence --- več kot je matematičnih parametrov za usposabljanje, več podatkov potrebujete --- in raznolikost vzorcev, ki mora odražati zapletenost naloge.

Glede na vire podatkov je treba podatke izbrati in pogosto očistiti, preden se vključijo v postopek usposabljanja. Če vzamemo na primer usposabljanje jezikovnega modela na podlagi spletne vsebine, je treba različne vzorce razvrstiti glede na dejanski jezik, jih očistiti spremljajoče spletne kode (HTML, Javascript itd.) in po možnosti premešati, da se preprečijo kršitve avtorskih pravic. Dober primer takšne priprave podatkov je oblikovanje korpusa OSCAR[1].

Oblikovanje anotiranih podatkov za nadzorovane sisteme umetne inteligence je bolj zapleteno, saj vključuje oblikovanje sheme anotacije, organizacijo kampanj anotacije in nadzor kakovosti anotiranih podatkov, na primer z ocenjevanjem soglasja med anotatorji istih podatkov.

Na splošno je bistveno, da je postopek načrtovanja dobro dokumentiran, da bi lahko izsledili vir morebitnega neuspešnega obnašanja v rezultatnem usposobljenem sistemu.

pristranski sistemi umetne inteligence

Kot smo že omenili, obnašanje sistema umetne inteligence tesno odraža naravo podatkov, ki so bili uporabljeni za njegovo usposabljanje. To pa lahko povzroči celo vrsto možnih pristranskosti, ki so lahko posledica osnovne izbire podatkovnih nizov. Tako bo na primer jezikovni model, ki je usposobljen samo na časopisnih člankih, zajemal popolnoma drugačne vrste izrazov in tem kot model, za katerega je bila izbrana literatura ali vsebina družabnih omrežij. Na enak način bodo sistemi za generiranje slik odražali velikost in raznolikost podatkovnih zbirk izvornih slik (npr. umetniških del), ki so bile upoštevane.

Pri nadzorovanih sistemih lahko posebna pristranskost izhaja iz načina oblikovanja oznak za anotacijo in tudi iz načina, kako bodo anotatorji s svojim kulturnim ozadjem razlagali podatke. Če želite na primer prepoznati sovražni govor v družabnih omrežjih, se lahko način, kako si anotatorji razlagajo čustva, razlikuje glede na starost, kulturo in osebne občutke anotatorjev v zvezi z gradivom, ki ga je treba anotirati.

Na splošno je treba vedno upoštevati, da so sistemi umetne inteligence že po svoji zasnovi zelo konservativni glede svojih podatkov za učenje in s tem glede obstoječih opazovanih podatkov. Od sistema umetne inteligence ne moremo pričakovati prave ustvarjalnosti.

Gostovanje, združevanje in distribucija podatkov

Zaradi velikosti in morebitne zapletenosti podatkov za učenje v sistemih umetne inteligence ter iz njih izhajajočih modelov so bile vzpostavljene različne pobude, ki omogočajo njihovo gostovanje in distribucijo.

Odprte podatkovne nize in modele je mogoče gostiti v specializiranih skladiščih (npr. vir slikovnih podatkov[2]) ali v splošnih nacionalnih ali mednarodnih (npr. Zenodo[3]) skladiščih. Takšni repozitoriji običajno zagotavljajo potrebno infrastrukturo za upravljanje avtorstva, licenciranja, verzioniranja in arhiviranja svoje vsebine.

Pri kompleksnih nalogah, kjer različne skupine vzporedno delajo na anotiranju različnih podatkovnih vzorcev, nekatere pobude delujejo kot katalogi za ustrezne podatkovne vire. Tako je na primer v primeru pobude HTR United[4], ki zbira metapodatke anotiranih dokumentov za prepoznavanje (rokopisnega) besedila.

---

1. OSCAR corpus: https://oscar-corpus.com/ ↩
2. Image data resource: https://idr.openmicroscopy.org/ ↩
3. Zenodo: https://zenodo.org/ ↩

4. HTR United: https://htr-united.github.io ↵