Analiza podatkov Food And Drug Administration z orodjem Bokeh

Neža Belej (63120340), Matej Dolenc (63120178)

January 28, 2017

1 Bokeh

Bokeh je interaktivna knjižnica programskega jezika Python. Omogoča elegantno in interaktivno vizualizacijo nad veliko množico podatkov. Arhitektura Bokeha sestoji iz dveh delov: izdelava grafov s programskim jezikom Python in pa izris v brskalniku s knjižnico BokehJS. Grafi v Pythonu se pretvorijo v JSON format, saj to zahteva BokehJS. Takšen dizajn je zelo fleksibilen, saj omogoča, da delo tudi z drugimi programskimi jeziki (R, Scala, Lua,...) lahko privede do enakih Bokeh grafov v brskalniku.

Če želimo sinhronizacijo Bokeh grafov in interaktivne vizualizacije v brskalniku, moramo uporabiti tudi strežnik Bokeh. Tako imamo omogočeno avtomatično posodabljanje uporabniškega vmesnika v brskalniku glede na naše klike in vnose.

1.1 Težave

Med spoznavanjem orodja Bokeh smo opazili, da ima orodje še kar nekaj hroščev. V okviru izdelave smo velikokrat naleteli na repozitorij na Githubu, kjer je trenutno kar 764 odprtih nalog ("issues": https://github.com/bokeh/bokeh/issues). Orodje je preprosto za uporabo, ima dobro dokumentacijo, vendar ima še veliko lukenj.

Ob prikazu grafa najpogostejših reakcij smo želeli na interaktiven način izvesti prikaz reakcij v različnih časovnih obdobjih. Zato smo najprej želeli uporabiti element DatePicker, kjer bi lahko izbrali začetni in končni datum. Opazili smo, da ima element v trenutni fazi zelo slab izgled, kar je opisano v sledeči povezavi (issue: https://github.com/bokeh/bokeh/issues/4503). Zato smo poizkusili z uporabo elementa DateRangeSlider, ki naj bi imel na različnih straneh drsnika začetni in končni datum. Element se ni prikazal. To je opisano v povezavi (issue: https://github.com/bokeh/bokeh/issues/2268). Zato smo bili primorani uporabiti dva navadna drsnika, enega za začetno, drugo za končno leto.

2 Pyton

Programska koda je spisana v jeziku Python. Uporabljen je bil za poizvedovanje, obdelavo podatkov in na koncu še za gradnjo grafov, na katerih so predstavljeni podatki, ki so bili pridobljeni s poizvedbami.

Vsa programska koda se nahaja v priloženih *.py* datotekah in sicer:

- *fda.py* - ta datoteka je namenjena pisanju poizvedb. V njej se nahajo vse poizvedbe, ki so bile uporabljene v tej seminarski nalogi.
- *main.py* - v tej datoteki se nahajajo funkcije, ki prejmejo podatke pridobljene iz predhodno opravljeni poizvedb. Ko so podatki obdelani gredo v izris.
- *plotters.py* - v tej datoteki smo definirali funkcije, ki nam omogočajo izris grafov, kot so vidni na rezultatih analize podatkov. To vključuje lomljenke, histograme in kombinacijske grafe.

3 OpenFDA

OpenFDA je projekt, ki ga je ustanovil Taha Kass-Hout z namenom ozaveščanja prebivalstva, odprte dostopnosti do podatkov in, potencialno, reševanja življenj.

OpenFDA (Food and Drug Administration) nam na svojih straneh omogoča dostop do 100 GB velike množice podatkov, kjer lahko poizvedujemo o medicinskih poročilih o zdravilih in hrani; na primer stranski učinki zdravil ali odpoklic prehrambenih produktov. OpenFDA je namenjena predvsem za poizvedbe preko njihovega zmogljivega API-ja, ki ima v ozadju implementiran učinkovit Elastic Search. Ta nam omogoča hitro in preprosto poizvedovanje po podatkih. Podatki, namenjeni prenosu, so razbiti na veliko število datotek v JSON formatu. Če želimo prenesti podatke, moramo paziti, da ob vsaki posodobitvi podatkov znova prenesemo celotno zbirko podatkov. Podatki so v dokumentni, nenormalizirani obliki, kar omogoča hitro iskanje.

4 Navodila za izvajanje

Za zagon projekta je potrebno imeti nameščeno orodje Anaconda. Nato iz konzole Bash ali pa Windows-ovega CMD-ja zaženemo ukaz:

- *conda install bokeh*

Nato se premaknemo v direktorij, ki vsebuje main.py našega projekta in natipkamo ukaz:

- *bokeh serve .*

V brskalniku se pomaknemo na *localhost:5006*. 5006 tukaj predstavlja številko vrat (port), ki se nam ob zagonu strežnika izpiše v konzoli.

5 Analiza podatkov

Analizo podatkov sva izvedla na dveh različnih bazah in sicer na bazi Food in na bazi Drugs. Baza Food vsebuje podatke o hrani, prehrambenih dopolnilih in presenetiljivo tudi o kozmetiki. Razloga za to žal ne poznava. Na drugi strani baza Drugs vsebuje podatke o zdravilih in njihovi uporabi.

Pri bazi Food naju je zanimala proizvodnja hrane v ameriških zveznih državah, stranski učinki in njihova frekvenca ter odpoklic hrane iz trga. Pri bazi Drugs naju je zanimalo še posebej kateri stranski učinki so najbolj pogosti pri ljudeh (pri moških, ženskah in skupaj), kako pogosto se stranski učinki pojavljajo pri kombinaciji dveh različih zdravil.

Pri analizi sva pazila, da sva se držala zahtevanega števila obdelanih vrstic (vsaj 1 milijon). To je tudi razvidno iz grafov. Primer: Na sliki o najpogostejših stranskih učinkih (Figure 2) je vsota vseh poročanj zagotovo večja kot 1 milijon.

Na koncu bi omenila še, da vsi podatki veljajo samo za Združene države Amerike.

5.1 Število poročil o stranskih učinkih glede na neko kombinacijo zdravil

Graf (Figure 1) nam ponazarja število poročanih stranskih učinkov glede na kombinacijo dveh zdravil. Siva barva pomeni, da ni bilo poročanj o reakcijah; bolj intenzivna barva pomeni večje število poročanih reakcij. Na spodnji sliki tako vidimo, da se največje število stranskih učinkov pojavi pri zdravilih Methotrexate in Humira ter Methotrexate in Enbrel. Ob premiku na želen kvadratek se nam prikažejo imena zdravil in število poročanj. Za vizualizacijo grafa smo opravili dvojno poizvedovanje:

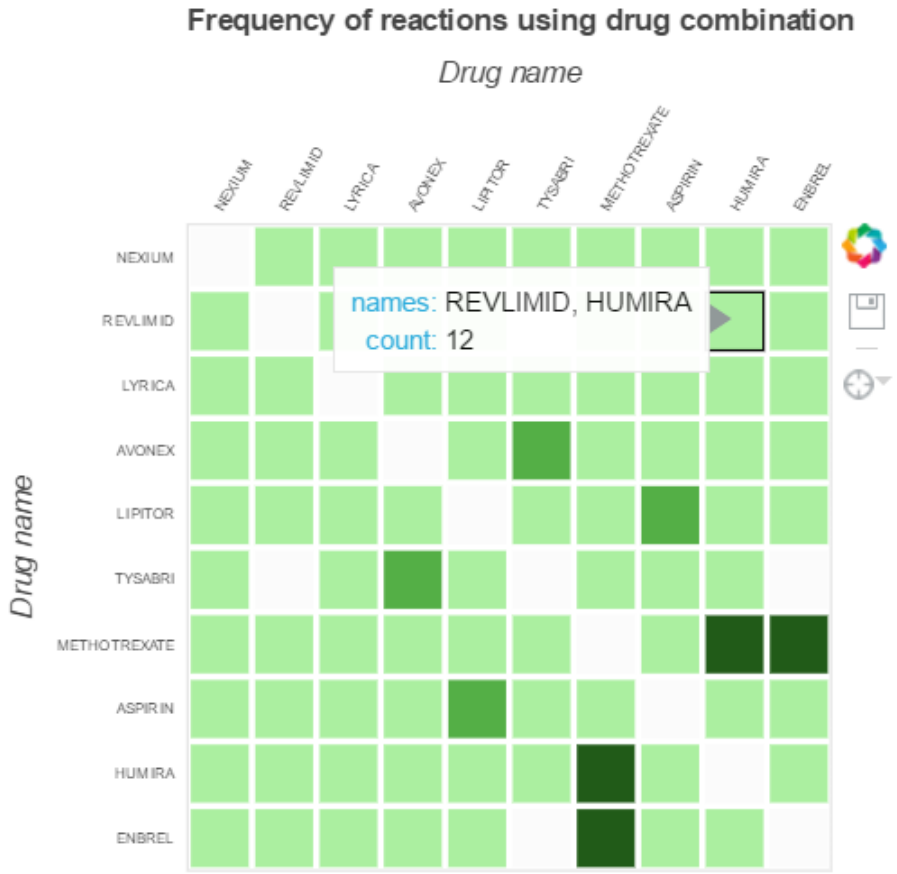
- Pridobitev 10 najpogostejših zdravil:

  ```
  https://api.fda.gov/drug/event.json?
  search=receivedate:[20040101+TO+20161230]
  &count=patient.drug.medicinalproduct.exact&limit=10
  ```

- Štetje poročanj pri tej kombinaciji zdravil:

  ```
  https://api.fda.gov/drug/event.json?
  search=receivedate:[20040101+TO+20170106]
  +AND+patient.drug.medicinalproduct:REVLIMID
  +AND+patient.drug.medicinalproduct:HUMIRA
  ```

Figure 1: Število poročil o stranskih učinkih glede na kombinacijo dveh zdravil



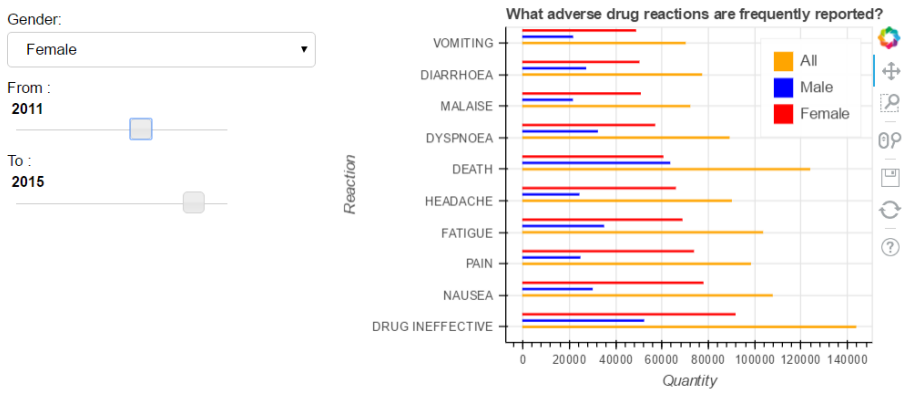
5.2 Najpogostejši stranski učinki zdravil

V spodnjem grafu (Figure 2) prikazujemo najpogostejše stranske učinke zdravil. Na abscisni osi imamo tako ponazorjeno število poročanj za neko reakcijo. Graf nam omogoča sortiranje glede na spol, čeprav je prikaz obeh spolov ves čas prikazan na grafu. To nam omogoča transparentno pregledovanje podatkov. Filtriranje je možno tudi glede na leto: izberemo lahko začetno in končno leto, torej obdobje, za katero nas zanima število poročanih stranskih učinkov.

Poizvedba:

```
https://api.fda.gov/drug/event.json?
search=receivedate:[20040101+TO+20170106]+AND+
patient.patientsex:2&
count=patient.reaction.reactionmeddrapt.exact
```

Figure 2: Najpogostejši stranski učinki



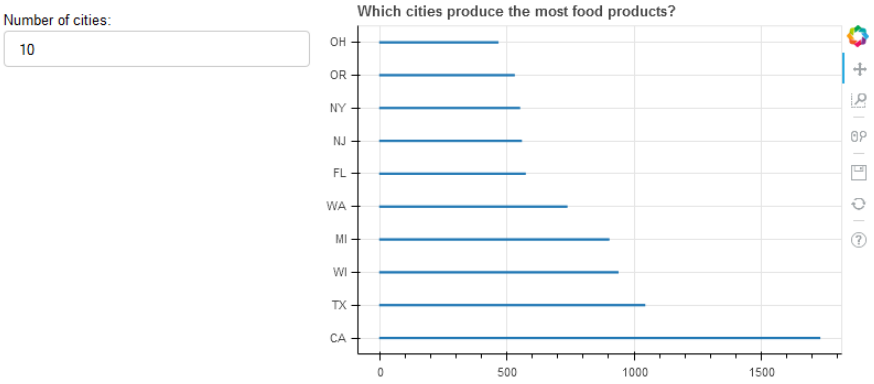
5.3 Proizvodnja hrane v ameriških zveznih državah

Na sledečem grafu (Figure 3) je predstavljena proizvodnja hrane v ameriških zveznih državah. S pomočjo tekstovnega polja lahko poljubno izbiramo število mest, ki jih želimo prikazati na grafu. Mesta so urejena od najmanjše proizvodnje do največje proizvodnje. Največ hrane proizvedejo v zvezni državi California (skoraj dvakrat več kot v Teksasu, ki je po proizvodnji hrane na drugem mestu).

Do teh podatkov pridemo s pomočjo sledeče poizvedbe:

```
https://api.fda.gov/food/enforcement.json?count=state
```

Figure 3: Proizvodnja hrane v ameriških zveznih državah



5.4 Odpoklic hrane s prodajnih polic po letih

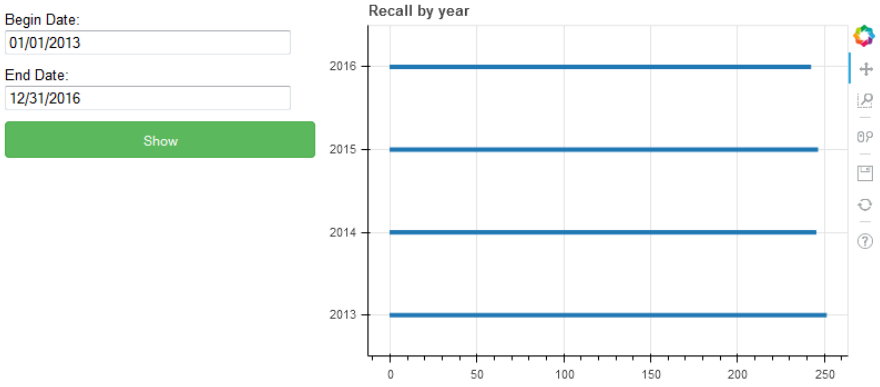
S pomočjo elementa DatePicker, ki je implementiran v orodju Bokeh, lahko izbiramo tudi datume. To smo uporabili v sledečem primeru, kjer nas je zanimal

odpoklic hrane s prodajnih polic po letih. Preden izvedemo poizvedbo, moramo izbrati začetni in končni datum. Kot smo že omenili, je v orodju Bokeh prisotnih kar nekaj hroščev, zato sam izgled elementa DatePicker ni najlepši. Nato poizvedba poišče vse odpoklice, ki padejo v ta datumski interval. Rezultat poizvedbe smo nato obdelali v Pythonu. Končni rezultat je prikazan na grafu (Figure 4).

Do teh podatkov pridemo s pomočjo sledeče poizvedbe:

```
https://api.fda.gov/food/enforcement.json?
search=recall_initiation_date:[20150101+TO+
20171231]&count=recall_initiation_date
```

Figure 4: Odpoklic hrane s prodajnih polic po letih



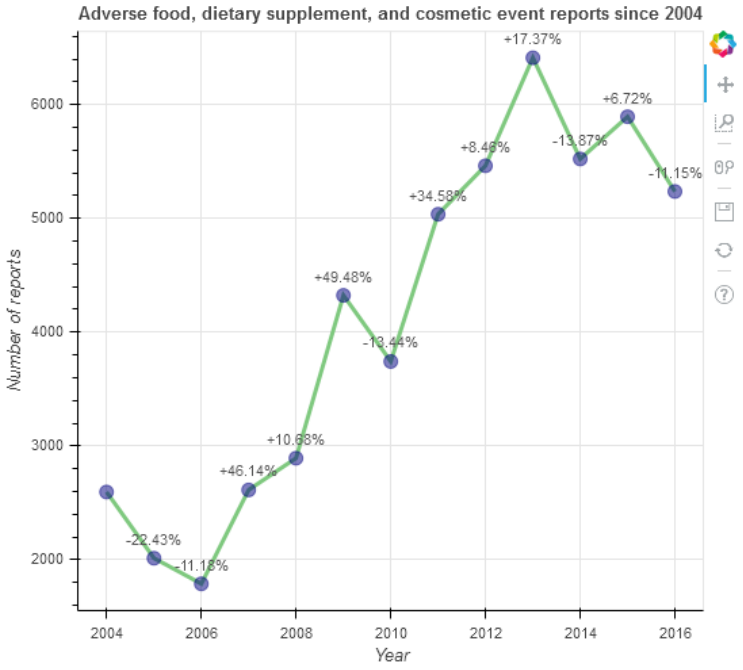
5.5 Stranski učinki hrane, prehranskih dopolnil in kozmetike po letih

Na sledečem grafu (Figure 5) sva želela prikazati število negativnih dogodkov, ki so bili posledica hrane, prehranskih dopolnil ali kozmetike. Za vsako leto je tudi prikazano kakšen je bil skok ali padec glede na prejšnje leto (v procentih). Zanimivo je videti, da so dogodki od leta 2006 do leta 2013 skoraj izključno samo naraščali (izjema je leto 2010). Iz tega bi lahko sklepali, da se v naši hrani, prehrambenih dodatkih in kozmetiki nahaja vedno več umetnih snovi, na katere človek negativno reagira.

Do teh podatkov pridemo s pomočjo sledeče poizvedbe:

```
https://api.fda.gov/food/event.json?count=date_created
```

Figure 5: Stranski učinki hrane, prehranskih dopolnil in kozmetičnih dodatkov po letih



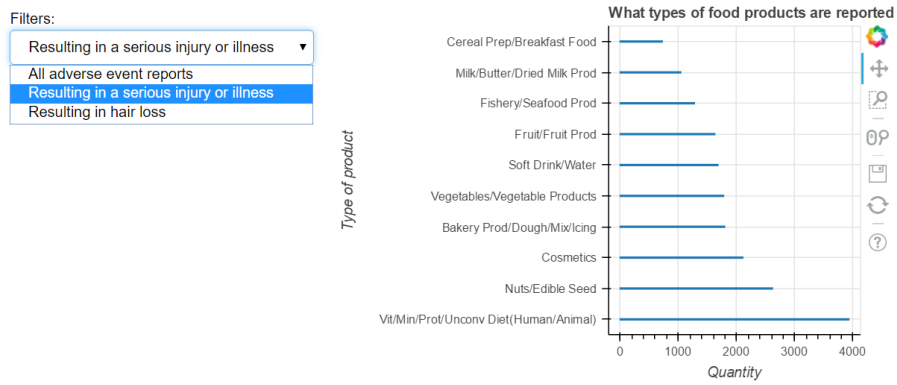
5.6 Najbolj pogoste skupine hrane, prehranskih dopolnil ali kozmetike, ki povzročajo stranske učinke

Pri tem grafu (Figure 6) nas je zanimalo katere skupine hrane ali prehranskih dopolnil povzorčajo največ stranskih učinkov. Stranske učinke je mogoče tudi prosto izbirati s pomočjo filtra. Tako lahko pogledamo za specifičen stranski učinek ali pa za vse stranske učinke skupaj. Tukaj sva dobila zelo zanimiv rezultat, saj največ stranskih učinkov povzročajo dodatki kot so vitamini in minerali.

Primer poizvedbe:

```
https://api.fda.gov/food/event.json search=outcomes:"serious
+injuries"&count=products.industry_name.exact
```

Figure 6: Najbolj pogosta hrana ali prehranska dopolnila, ki povzročajo stranske učinke



6 Reference

- http://bokeh.pydata.org/en/latest/
- https://github.com/bokeh/bokeh/
- https://open.fda.gov/