Predikcia trendov vo vedeckých článkoch

Tím Normy

(Lea Gogová, Hana Hladíková, Jana Maľová, Natália Kňažeková, Kristína Sásiková)

Predmet: Princípy dátovej vedy

Študijný odbor: Dátová veda (DAV)

Ročník: tretí

Školský rok : 2022/2023 (Zimný semester)

Garant predmetu: Vladimír Boža, Radoslav Harman, Tomáš Vinař

Obsah:

1 Úvod

V súčasnej dobe má široká verejnosť vďaka Internetu a globálnej spolupráci vedcov voľný prístup k veľkému množstvu dát. Jedným z príkladov je databáza biomedicínskych článkov PubMed, ktorá sa využíva na vzájomnú komunikáciu a spoluprácu v biomedicíne. My budeme prostredníctvom nej skúmať, ktoré témy sú v tejto oblasti najdiskutovanejšie a naopak, ktoré z nich sa stávajú zabudnutými.

Špecifikácia otázok

V našom projekte nám prišlo zaujímavé urobiť nasledovné analýzy:

- Ako máme dobre a efektívne čistiť textové dáta?
- Ako sa vysporiadať z rôznymi tvarmi jedného slova v texte?
- Ako získať kľučové slová z textu?
- Aké slová, trendy boli najpopulárnejšie v priebehu rokov?
- Aké rôzne predikčné metódy by sme mohli použiť v našej analýze?
- Vieme predpovedať úspešnosť slova či dokonca názvu článku na základe minulých dát?
- Čo zaujímavé sa nachádza v dátach pre konkrétne lieky či ochorenia?

2 Prehľad dátových zdrojov

2.1 Zdroj dát

Naším hlavným a jediným zdrojom dát bola databáza PubMed, ktorá momentálne obsahuje viac ako 35 miliónov záznamov biomedicínskej literatúry.

2.2 Charakteristika dát

Pracovali sme s dvoma sadami dát zodpovedajúcim dvom časovým obdobiam, 1930-1990 a 2007-2022, kde za druhé obdobie to neboli všetky dáta, ale iba ich vzorka.

Každý vedecký článok obsahoval o sebe veľa informácií ako napríklad autora, rok vydania, rok úpravy, MeSH headings (medical subject headings), abstrakt, názov. V tomto projekte sme sa rozhodli využívať nasledovné atribúty:

- **id** - id článku v databáze (PMID)
- **year** - rok vydania článku
- **month** - mesiac vydania článku
- **title** - názov článku

2.3 Sťahovanie dát

Pre efektívnejšie (opakované) spracovanie dát sme si najprv stiahli lokálne celú PubMed databázu. NCBI ponúka na prehľadávanie databáz okrem API aj command-linovú utilitu EDirect, ktorou vieme prehľadávať a prefiltrovať aj lokálne stiahnuté dáta. Dáta o článkoch sú poskytované vo forme xml súborov. Preto sme ich prostredníctvom EDirectu preparsovali do tabuľkového csv formátu s iba potrebnými údajmi.

Množstvo dát za posledných 30 rokov v PubMed databáze rapídne rástlo. Preto sme nemohli pracovať so všetkými článkami za jednotlivé roky 2007-2022. Pracovali sme s náhodnou vzorkou: na každý rok sme vybrali 10% náhodne zvolených článkov.

3 Použité metódy

3.1 Príprava dát

3.1.1 Čistenie dát

Čistenie dát prebiehalo v jazyku python pomocou knižníc pandas a re. Najskôr sme zmenili pomocou funkcie astype z knižnice stĺpec title na string pretože po načítaný csv súboru bol načítaný ako object. To isté sme spravili s atribútmi year a month, ktoré sme zmenili na int64 (tohto typu je aj atribút id).

Ďalej sme stĺpec title zmenili na malé písmená a pomocou knižnice re sme odstránili všetky znaky okrem písmen, číslic a medzier. Číslice sme sa rozhodli v článkoch nechať, keďže sú to hlavne medicínske články, teda máme názvy ako napríklad ako covid-19. Vyčistené dáta boli uložené do nového stĺpcu title_cleaned.

3.1.2 Lematizácia

Kedže z jednou z našich úloh je porovnávanie frekvencií jednotlivých slov, tak bolo pre nás dôležité, aby sme nedostávali rôzne formy slov ako napríklad effect a effects alebo study a studying. Preto sme použili techniku lematizácie.

Lematizácia je technika používaná v spracovaní prirodzeného jazyku, v ktorom sa prevedie dané slovo do základného tvaru. Teda z minulého času ho prevedie do prítomného, odstráni -ing formu, prevedie plurál na singulár - teda nám zostane slovo v základnom tvare. Využíva kontext slova v danom jazyku, teda sa nestane, že zo slova caring by spravila slovo car (ako napríklad metóda stemming, ktorú sme sa rozhodli nevyužiť napriek tomu, že je rýchlejšia).

Na lematizáciu sme využili knižnicu spacy a jeho model en_core_web_sm, ktorý je vyvinutý pre anglický jazyk. Stiahli sme jeho menšiu verziu, ktorá má veľkosť 12 MB namiesto väčšej lg ktorá má 560 MB. Ten väčší má síce o niečo vyššiu presnosť ale oveľa pomalšie sa načítava a pre naše účely postačoval aj menší.

Názvy článkov, na ktorých bola použitá lematizácia boli uložené do nového stĺpca lemma.

3.1.3 Stopwords

Stopwords sú slová, ktoré zámerne vynechávame z textu. Nižšie spomínaná metóda Countvectorizer ponúka argument stopwords = "english". Nám to ale nestačilo. Preto sme k štandardným anglickým stopwords pridali tie nami vybrané (v dokumente stop_words.txt) a aplikovali ich do Countvectorizer, ktorá ich vynechá. Do csv súboru sme pridali stĺpec s názvom lemma_without_stopwords.

3.1.4 Výber kľúčových slov

CountVectorizer

Počítače nerozumejú vetám, slovám, ani znakom. Preto pri práci s dátami v tomto formáte potrebujeme text reprezentovať pomocou čísel. Jednou z metód ako previesť text do čísel je metóda Countvectorizer. Je to nástroj poskytovaný knižnicou scikit-learn v Pythone. Slúži na transformáciu textu do číselného vektora na základe frekvencie každého slova, ktoré sa v texte nachádza. Toto je najmä užitočné pri súbore viacerých textov, kde chceme každé slovo v každom texte previesť na vektory. Toto je veľmi prospešné pre rôzne textové analýzy.

TfidfTransformer

TF-IDF, term frequency - inverse document frequency, je štatistika, ktorej cieľom je priradiť každému slovu hodnotu na základe významnosti slova pre daný dokument, ktorý súčasťou súboru viacerých dokumentov. TF-IDF závisí na tom, že sa pre každé slovo vypočíta počet výskytov v rámci jeho dokumentu a počet výskytov vo všetkých dokumentoch dokopy. Každému slovu sa tak priradí jeho skóre vypočítané z týchto dvoch hodnôt. Čím je slovo jedinečnejšie pre svoj dokument, tým má väčšie skóre a naopak.

Ďalšou funkciou knižnice scikit-learn je Tfidftransformer. Využitím Countvectorizer priradí každému slovu v texte jeho TF-IDF skóre tiež vzhľadom na všetky dokumenty. My si v našej práci zvolíme 5 slov s najlepším TF-IDF skóre a prehlásime ich za kľúčové slová nášho dokumentu. Do csv súboru sme pridali finálny stĺpec s názvom tfidf_keywords.

3.1.5 Charakteristika finálneho .csv súboru

Máme 2 súbory csv súbory s upravenými dátami a to 1970_90_keywords.csv a 2007_22_keywords.csv. Rôznymi technikami popísanými vyššie sme vyčistili dáta a k pôvodným stĺpcom id, year a title sme pridali nové stĺpce:

- **Title_cleaned** - vyčistené názvy článkov
- **Lemma** - zlematizované názvy článkov
- **Lemma_without_words** - zlematizované názvy článkov, z ktorých boli odstránené stopwords
- **Tfid_keywords** - klúčové slová vybrané zo stĺpca lemma_withour_words pomocou techniky TfidfTransformer

3.2 Predikcia

Existuje veľa rôznych spôsobov, ako predpovedať budúcnosť na základe minulých dát. No na druhú stranu sú tu problémy, ako napríklad, aký model použiť na dané dáta? Nie je daný model príliš zložitý? Preto sme sa rozhodli na predikciu využiť viacero modelov.

3.2.1 Predpoveď na základe priemeru pár predošlých hodnôt

Ako prvý model sme si vybrali jednoduchú predikciu na základe priemeru zopár predošlých hodnôt. Postup výpočtu vyzeral nasledovne. Najskôr sme si pre každú za sebou idú dvojicu rokov vypočítali ich rozdiel v percentách. Nasledovne sme z týchto hodnôt vyrobili Simple Moving Average z predošlých 2 hodnôt. Tento posledný priemer sme následne pripočítali k poslednej hodnote a zistili sme novú hodnotu pre nepoznaný rok.

3.2.2 Polynomiálna regresia

Ďalším spôsobom, akým sme sa rozhodli predpovedať, ako úspešná bude daná vedecká téma v ďalšom roku, je pomocou polynomiálnej regresie. Všeobecne, polynomiálna regresia sa dá definovať ako popis polynomiálneho vzťahu medzi závislou premennou Y, ktorú sa snažíme

predpovedať a nezávislou premennou X. Jeho všeobecný tvar pri n stupňoch vypadá nasledovne:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_n X^n + \epsilon$$

Na túto predikciu je potrebné čo najväčšie množstvo dát, aby sa určil správny tvar regresie a správny odhad rádu. V našom výskume sme pre rád regresie používali hodnoty 2-5, no najlepšie odhadoval pri 3-tom stupni.

3.2.3 Model ARIMA

ARIMA (Auto Regressive Integrated Moving Average) je model na predpovedanie budúcich hodnôt na základe minulých dát a chýb predpovede (forecast errors). Vznikol spojením dvoch jednoduchších modelov, Auto Regresívny a Moving Average, ku ktorým je ešte kvôli stacionarite pridáva diferencovanie, teda rozdiel súčasnej a minulej hodnoty. Preto je tento model charakterizovaný 3 premennými: p,d,q.

"p" zodpovedá rádu Auto Regresívnemu modelu. Všeobecný predpis tohto modelu má tvar:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + ... + \beta_p Y_{t-p} + \epsilon_t$$

,kde Y sú zistené hodnoty, α je konštanta modelu, β_1 až β_p sú koeficienty modelu a ϵ je chyba.

"q" zodpovedá rádu Moving Average modelu. Všeobecný predpis tohto modelu má tvar:

$$Y_t = \alpha + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + ... + \phi_q \epsilon_{t-q} + \epsilon_t$$

,kde Y je odhadovaná hodnota, α je konštanta modelu, ϕ_0 až ϕ_q sú koeficienty modelu a ϵ_t až ϵ_{t-q} sú chyby. Tie sa dajú vypočítať z nasledujúcich rovníc:

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + ... + \beta_0 Y_0 + \epsilon_t$$

$$Y_{t-1} = \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + ... + \beta_0 Y_0 + \epsilon_{t-1}$$

"d" zodpovedá číslu, koľko minulých hodnôt odčítame od súčasnej.

$$\nabla y_t = y_t - y_{t-1}$$

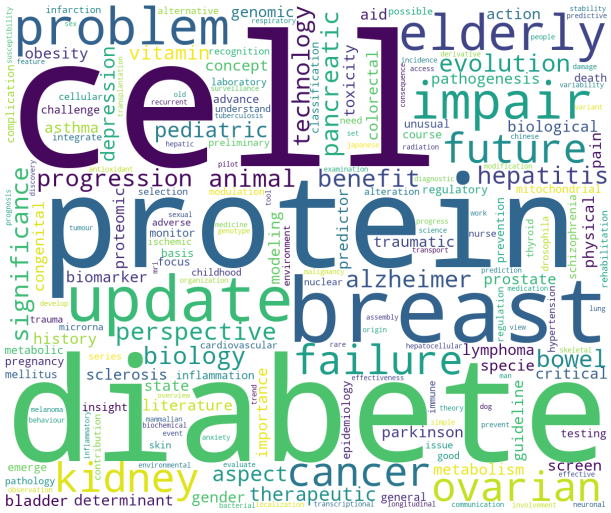
Spolu výsledný predpis vyzerá nasledovne:

$$\nabla Y_t = \alpha + \beta_1 \nabla Y_{t-1} + \beta_2 \nabla Y_{t-2} + ... + \beta_p \nabla Y_{t-p} + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + ... + \phi_q \epsilon_{t-q}$$

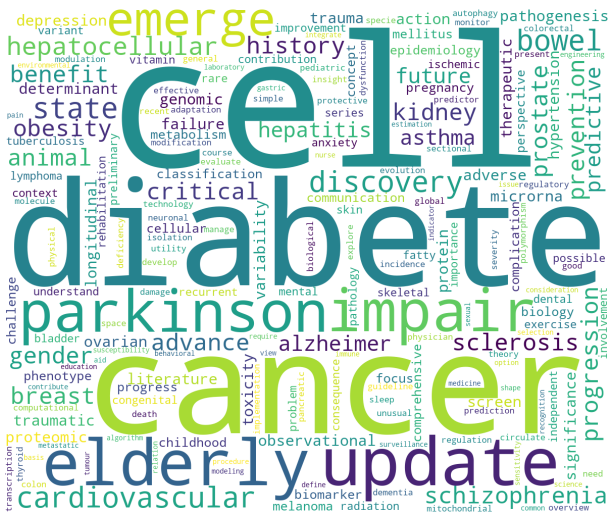
4 Výsledky analýzy

4.1 Popularita jednotlivých slov v časovom období

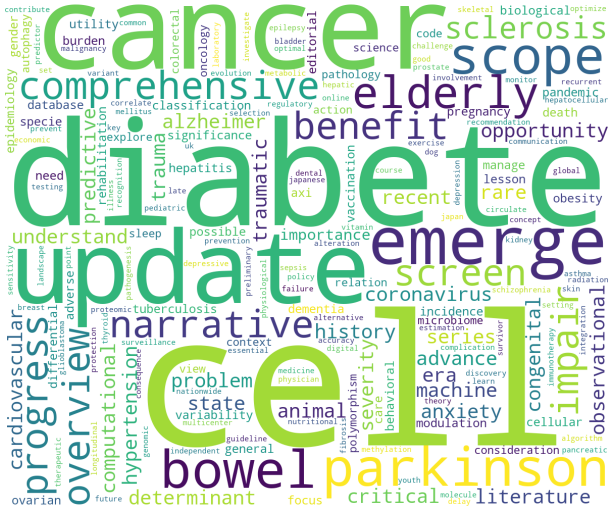
Rozhodli sme sa analyzovať dáta aj pomocou word cloudov. Novšie dáta, dáta zo súboru 2007_22_keywords.csv, sme si chronologicky rozdelili do 3 skupín a vytvorili word cloudy pre každú z nich. Dostali sme nasledovné výsledky.



Obr. 1: Wordcloud pre roky 2007 až 2012



Obr. 2: Wordcloud pre roky 2013 až 2017



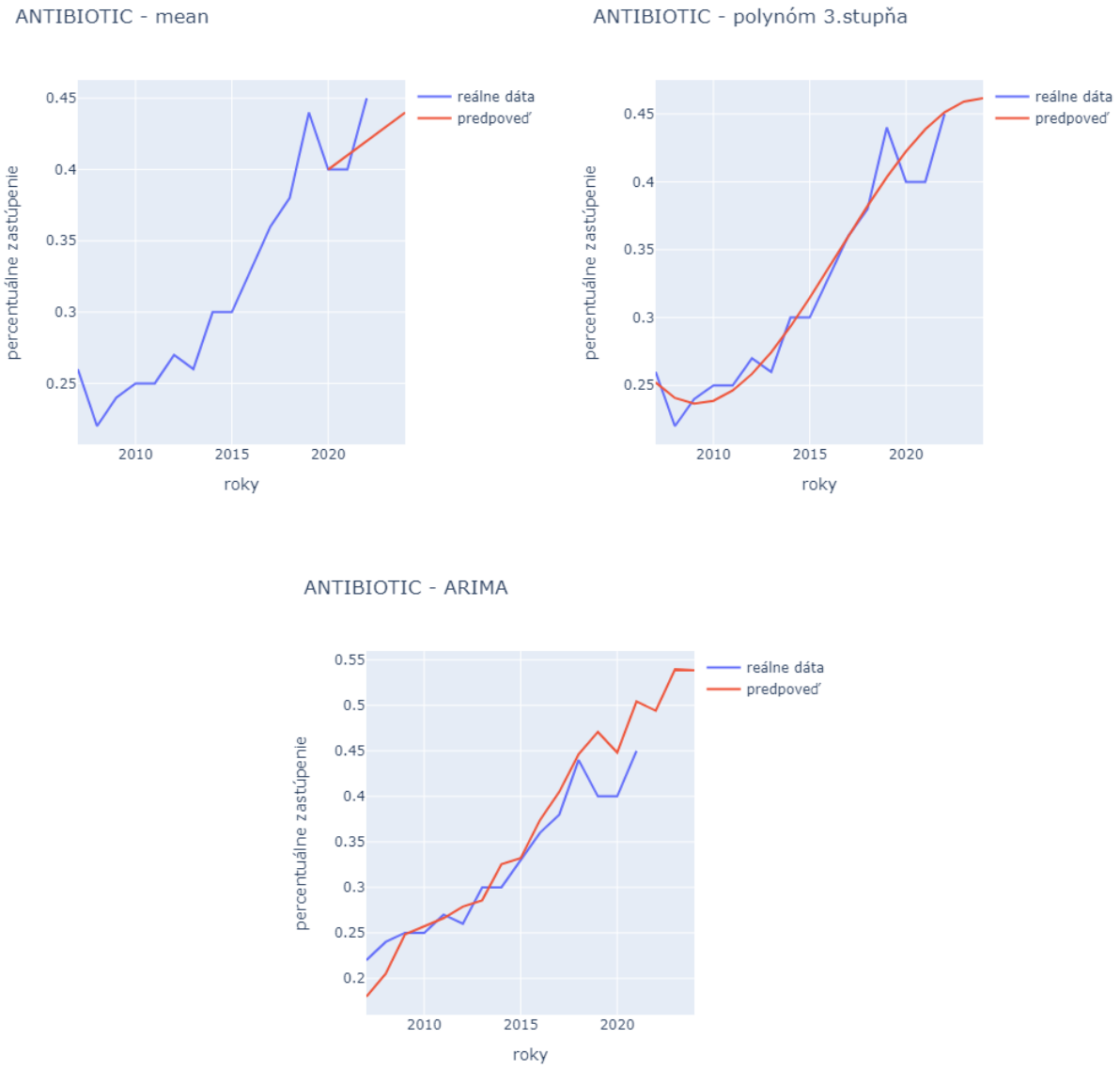
Obr. 3: Wordcloud pre roky 2018 až 2022

Môžeme pozorovať, že v každom období sa nám opakujú niektoré slová. Tie najčastejšie sú napríklad update, cell, diabete, cancer, problem, bowel, obesity. Kvôli nášmu predspracovaniu textu, kvôli lemmatization, sú kľúčové slová vo word cloudoch vo svojich základných tvaroch. V biomedicínskych článkoch sa ale mohli používať aj v iných tvaroch. Napríklad diabete ako diabetes, history ako historical, significant ako significance, depression v tvare depressed a tak ďalej. Vďaka tomuto sa v článkoch mohli vyskytovať tiež ako slovné spojenia ako napríklad breast cancer, cancer prevention alebo obesity screening.

V rokoch 2007 až 2012 boli celkom populárne slová kidney a ovarian, ktoré neboli v ďalších rokoch. V rokoch 2013 až 2017 bolo jedinečne populárne slovo napríklad hepatocellular. Slovo schizophrenia bolo spomedzi všetkých období používané v týchto rokoch výrazne najviac. V období 2018 až 2022 sa zrazu objavilo populárne slovo coronavirus, čo sa dalo čakať. Spolu s ním sa spopularizovali tiež slová ako depression, obesity alebo anxiety, ktoré ako iné boli tiež následky pandémie covid-19.
Priebehom období rokov sa niektoré slová stávali postupne trendami. Neskôr ich ale iné trendy vystriedali. Môžeme teda povedať, že najstabilnejšími trendami za obdobie 2007 - 2022 boli už spomínané slová ako cell, cancer a problem. Mohli by sme teda očakávať, že najbližšie budúce roky sa tiež objavia medzi tými najpoužívanejšími.

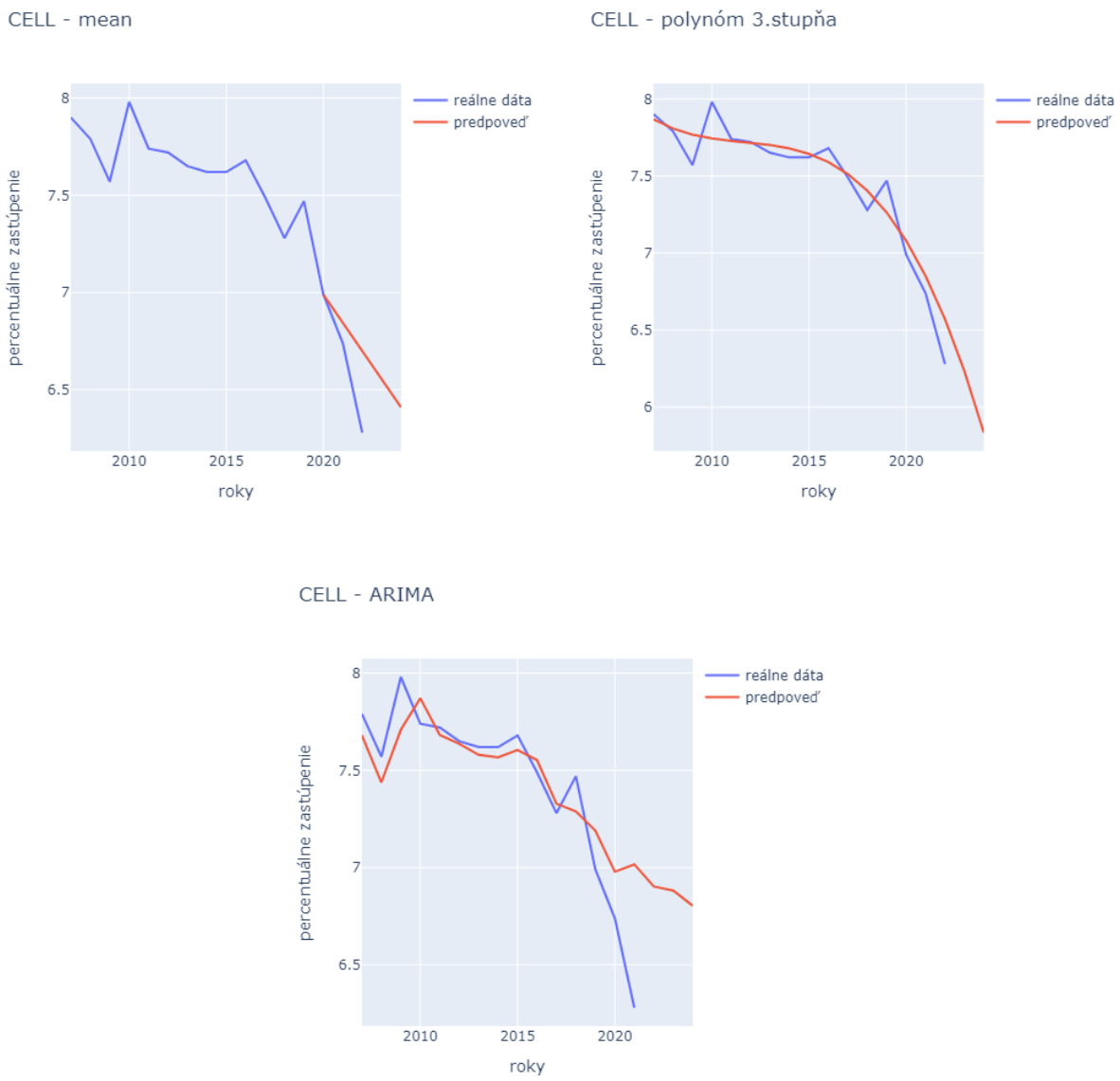
4.2 Porovnanie predikčných metód

Jedným z najväčších bodov našej analýzy bolo predpovedanie trendovosti v budúcich rokoch. Preto v tejto časti porovnáme nami vybrané metódy, teda priemer rozdielov posledných rokov, polynomiálnu regresiu a ARIMA model.
Na porovnanie sme si vybrali pár slov z novšieho datasetu, teda z rokov 2007-2022. Vybrali sme tie, ktoré sa pri pohľade na dáta najviac vyskytovali alebo boli pre nás zaujímavé. Sú to slová antibiotic, covid, cell, hiv a mouse. Výsledné grafy vyzerajú nasledovne:



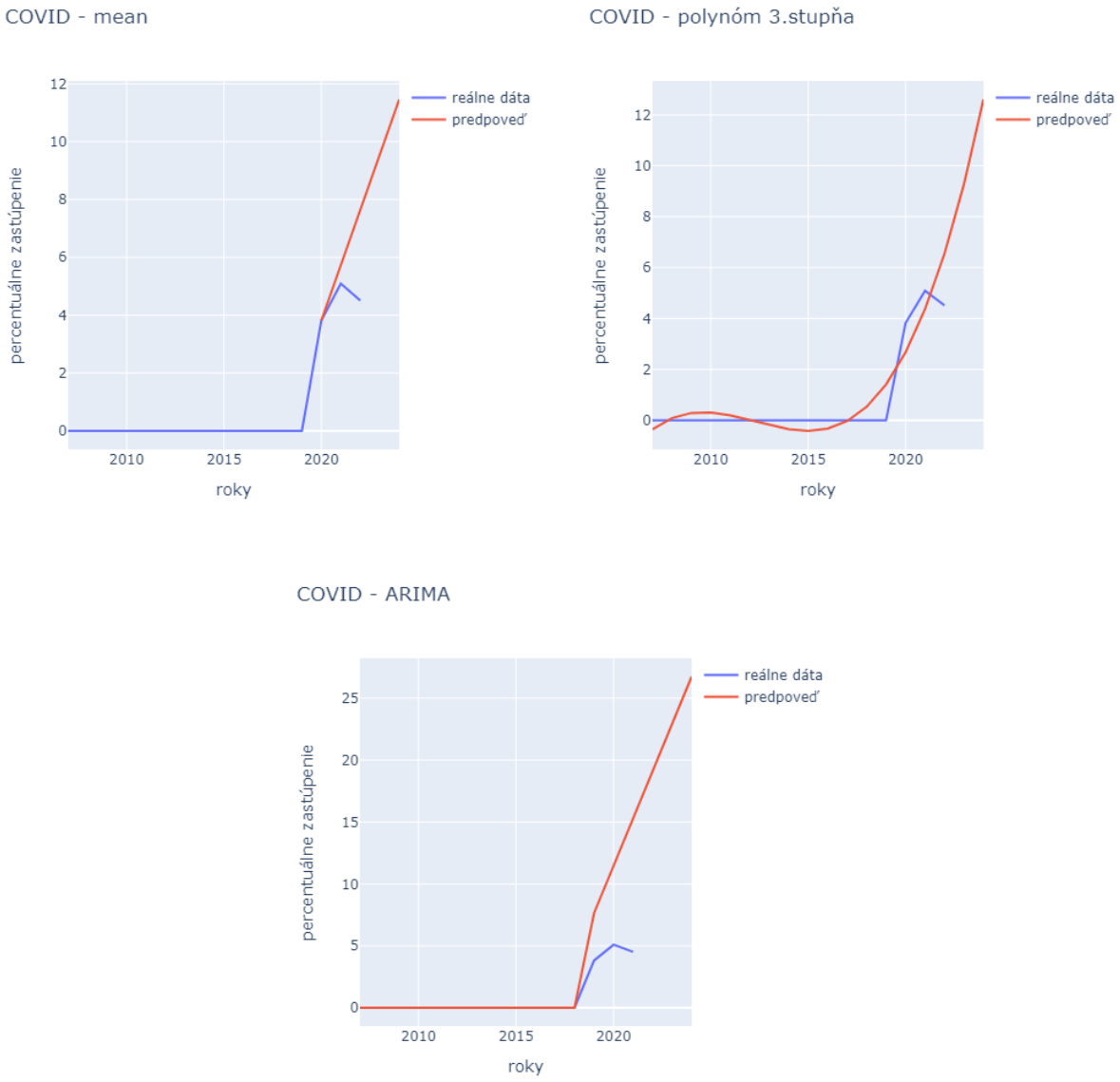
Obr. 4: Predpoveď pre slovo "antibiotic"

Percentuálne zastúpenie slova antibiotic malo v posledných rokoch nasledovný priebeh. Najskôr bol výrazný nárast, potom nasledoval prepad a ďalšie obdobie znova nastal výrazný nárast. Vidíme 3 rôzne prístupy. Všetky 3 prístupy predikujú pokračujúci rast trendu. Metóda priemerov sa na testovacích 2 rokoch žiaľ výraznejšie odlišuje Najoptimistickejšie predikuje práve model ARIMA.



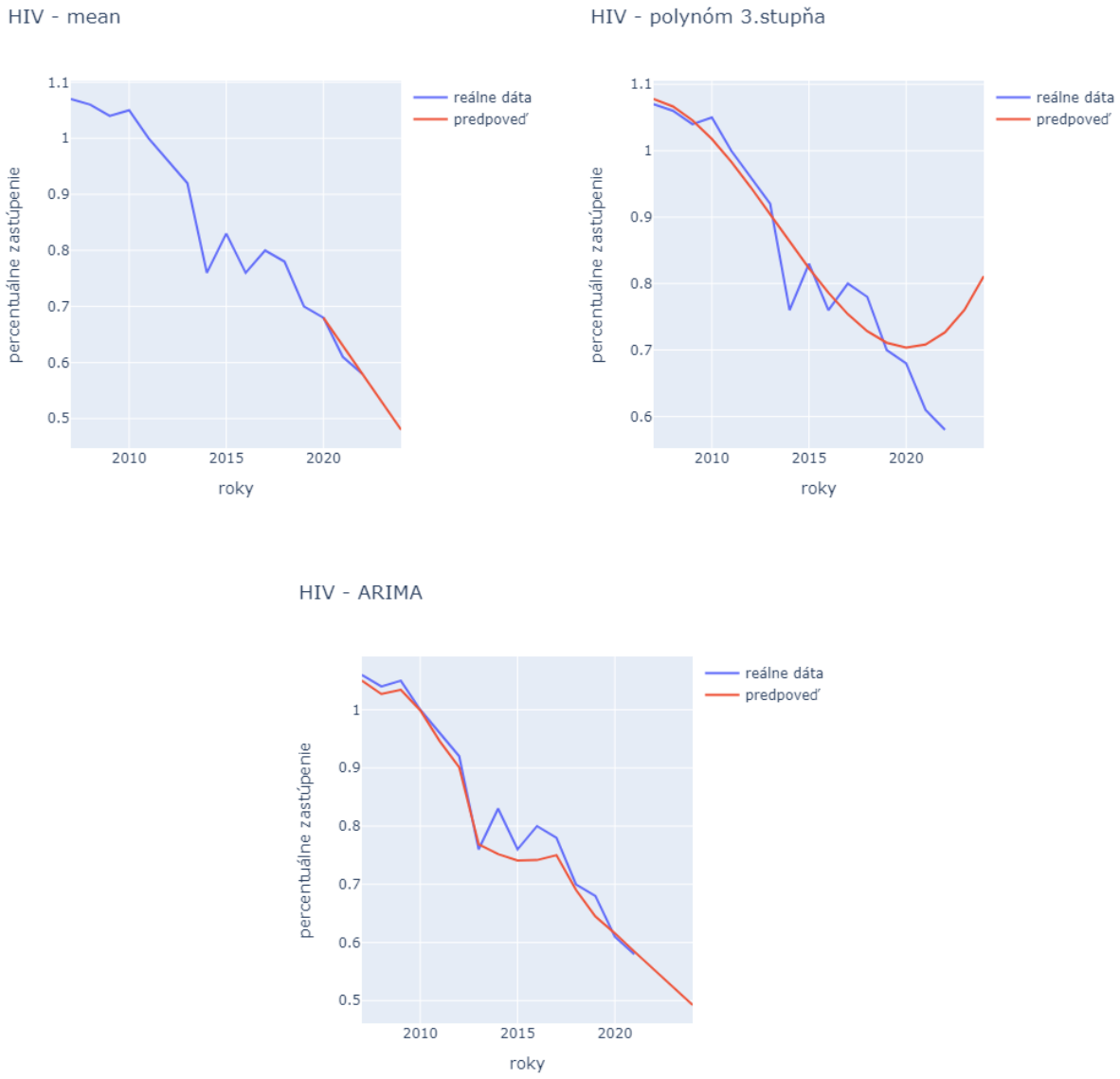
Obr. 5: Predpoveď pre slovo "cell"

Pri tomto trende slova cell vidíme, že metóda polynómom sa ukazuje ako najlepšia. ARIMA model nenaznačuje taký pokles ako je v skutočnosti, klesá oveľa miernejšie. Metóda priemerov zjemňuje pokles.



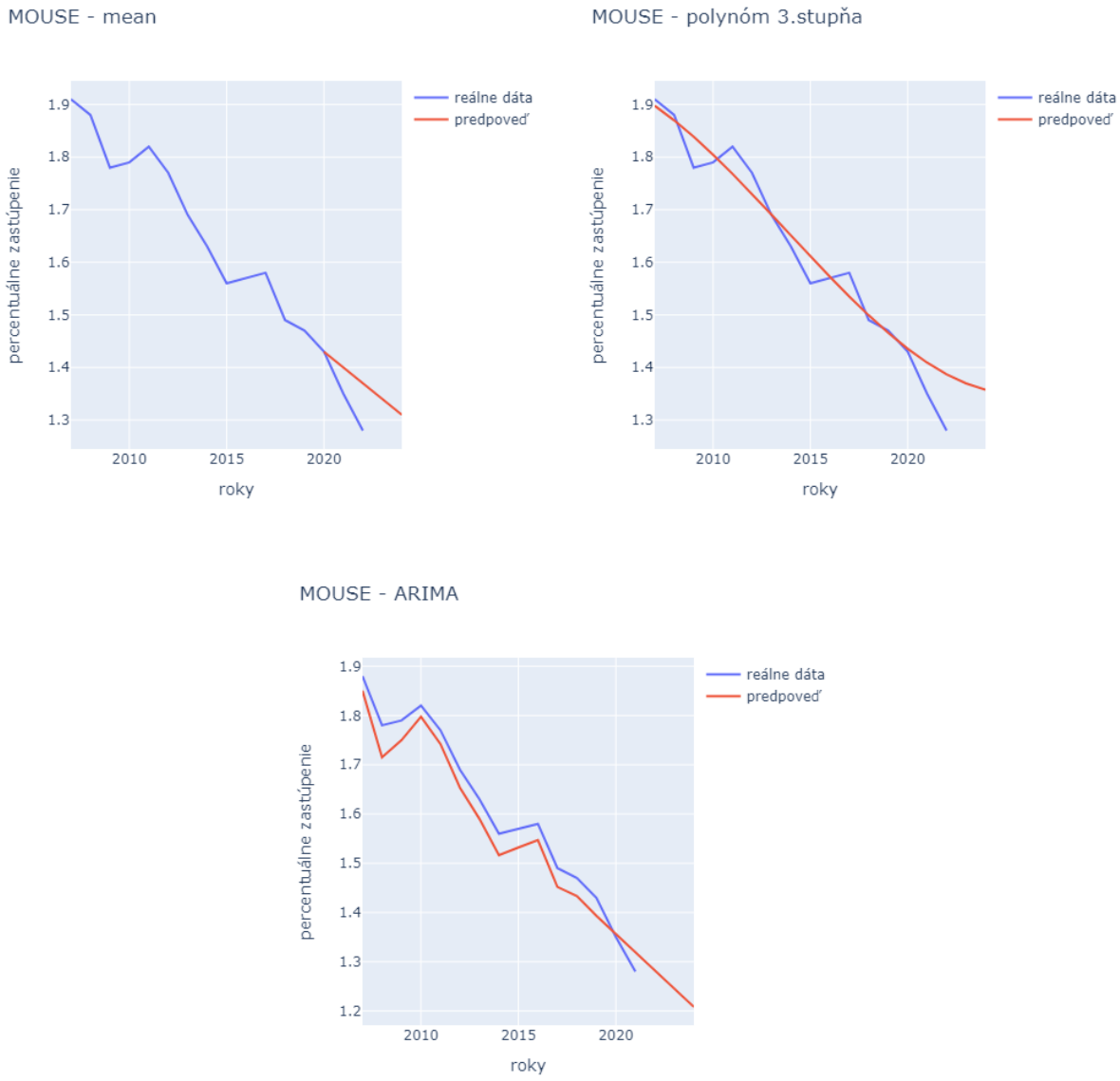
Obr. 6: Predpoveď pre slovo "covid"

Covid téma je veľmi špecifická, kedže sa počas obdobia 2007-2019 veľmi málo až vôbec nevyskytovala v článkoch, ale za posledné 3 roky nastal veľký boom a trend vyskočil vysoko, ktorý ale už aj za krátky čas začal klesať. Všetky metódy predikujú výrazný rast, avšak medzi rokmi 2021 a 2022 nastal pokles, čo by malo v nalsedujúcich rokoch aj pokračovať.



Obr. 7: Predpoveď pre slovo "hiv"

Z pozorovania skutočných dát môžeme vidieť, že trend klesá. ARIMA a metóda priemerov tento trend zaznamenali a podobne predikujú aj na ďalšie roky. A však polynómom 3. stupňa predikuje opak, teda nárast.



Obr. 8: Predpoveď pre slovo "mouse"

Ani jeden z modelov nepredikuje taký výrazný pokles, ktorý by sme predpokladali "voľným okom" na základe pozerania sa na posledné roky. Dokonca polynóm naznačuje miernu zmenu trendu, z klesajúceho na stáli.

4.3 Analýza datasetov

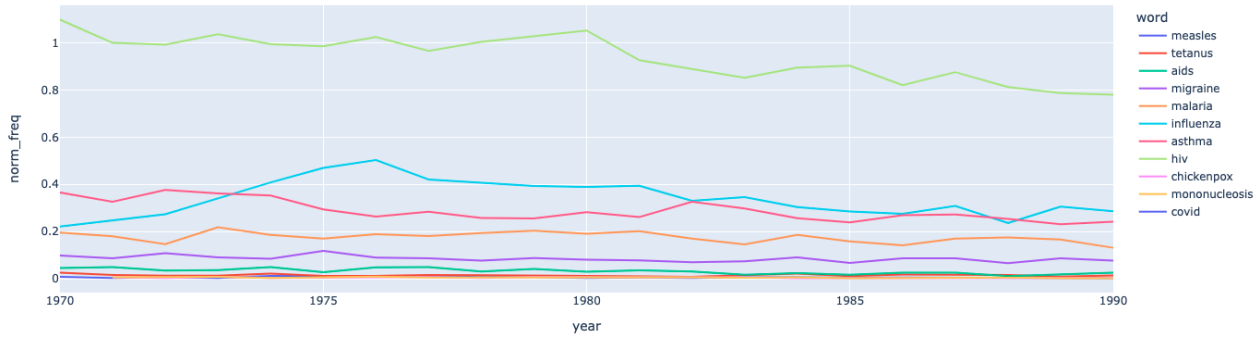
Po predikcii sme sa rozhodli pustiť do hlbšej analýzy jednotlivých liekov či ochorení. Ako prvú analýzu sme porovnávali novšie (2007 - 2022) a staršie (1970 - 1990) dáta. Najskôr sme skúšali vyhladať najčastejšie slová vo vyčistených názvoch článkov. Najčastejšie používané slovo bolo cell a nasledovali cancer a base. V kľúčových slovách bolo tiež najpoužívanejšie slovo cell.

4.3.1 Početnosť výskytu rôznych chorôb

Pozreli sme sa vývoj rôznych chorôb ako často sa vyskytovali v názvoch vedeckých článkov za obdobie 1970-90 a 2007-2022. Vybrali sme nasledovné: chickenpox (kiahne), hiv, influenza (chrípka), mononucleosis (mononukleóza), aids, asthma (astma), malaria (malária, measles (osýpky), migraine (migréna), tetanus (tetánia), rabies (besnota) a covid.

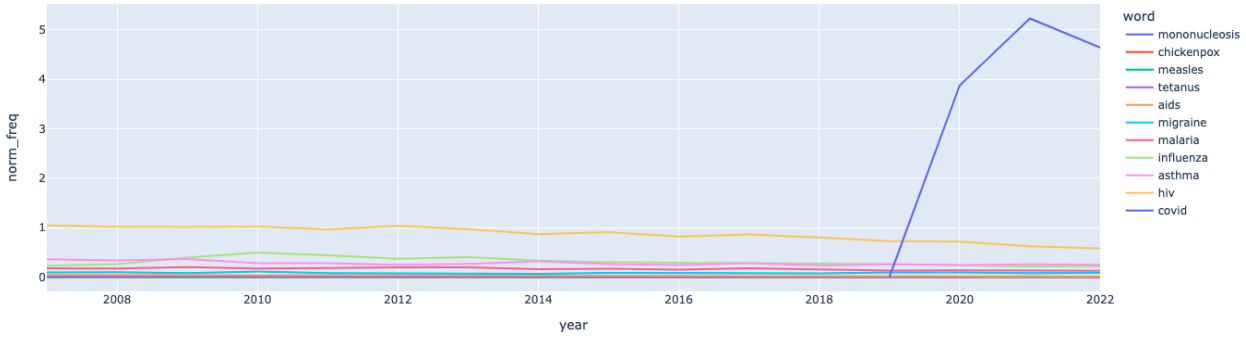
Za obdobie 1970-90 je najviac spomínanou chorobu HIV s počtom spomenutím vyše 10 000. Na druhom mieste sa nachádza chrípka so 4060 výskytmi a na treťom mieste astma s 3550. Nie je prekvapivé, že na prvom mieste je HIV, keďže okolo 1970 sa začala vyskytovať v USA v meste New York, aj keď do prvého širšieho povedomia prišla až o desaťročie neskôr, kedy sa vyskytovala už aj na opačnej strane USA a to v San Franciscu.

Pozreli sme sa ešte bližšie na chorobu HIV za obdobie 1970-1990 a 2007-2022. V období 2007-2022 sa vyskytuje v 13060 článkoch a 4459 krát ako kľúčové slovo. Za obdobie 1970-90 to bolo 9900 krát a 3241 krát ako kľúčové slovo. Po vytriedení článkov obsahujúce slovo HIV sme zistili, že najčastejšie slová sa v obidvoch obdobiach prelínajú a sú to 1, infection, man, antiretroviral, therapy. Toto naznačuje, že obsahom daných článkov je nájdenie spôsobu ako vyliečiť alebo aspoň napomôcť človeku, ktorý má daný vírus, keďže do dnešného dňa je táto choroba nevyliečiteľná. Číslovka 1 hovorí o type vírusu HIV-1, ktorý je bežnejší ako typ 2. Avšak, počet článkov o HIV za posledné roky klesá.



Obr. 9: Ochorenia v rokoch 1970-1990

Za obdobie 2007-2022 bol na prvom mieste covid-19 s 20446 výskytmi, väčšina z nich však bola za roky 2019-2022. V období 1970-90 sa covid spomínal iba štyrikrát. Na druhom mieste je HIV a na treťom je astma. Na danom grafe je vidieť, ako na epidémiu covidu promptne zareagovali vedci a študovali jej výskyt a možnú liečbu.



Obr. 10: Ochorenia v rokoch 2007-2022

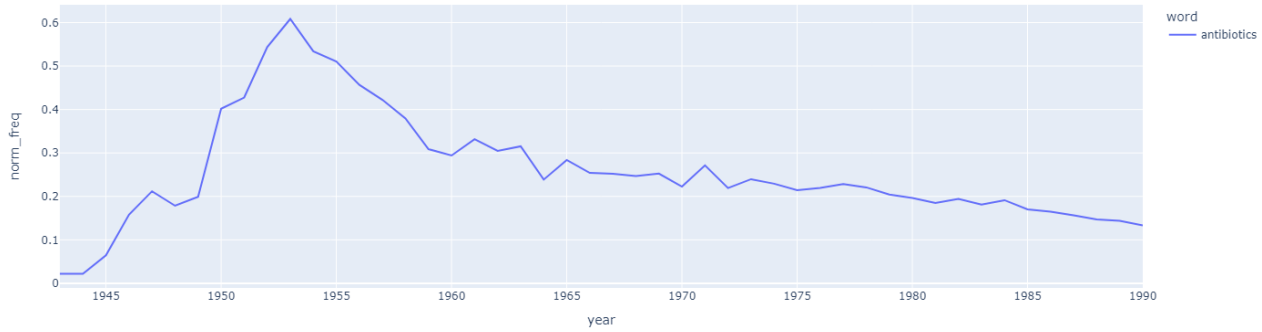
Za tieto dve obdobia sa nezmenilo poradie rôznych chorôb ako často sa vyskytujú v jednotlivých článkoch, jediný rozdiel bol covid 19, ktorý sa presunul z predposledného miesta na prvé.

Počet výskytov chorôb v názvoch článkov za obdobie 1970-1990 = hiv: 10778, influenza: 4060, asthma: 3550, malaria: 2371, migraine: 976, aids: 345, tetanus: 157, measles: 48, mononucleosis: 39, chickenpox: 23, covid: 4, rabies: 2

Počet výskytov chorôb v názvoch článkov za obdobie 2007-2022 = covid: 20446, hiv: 14232, influenza: 5286, asthma: 4966, malaria: 3176, migraine: 1428, aids: 403, tetanus: 203, measles: 73, mononucleosis: 55, chickenpox: 29, rabies: 6

4.3.2 História antibiotík

Ako ďalšiu analýzu sme sa rozhodli pozrieť ešte na viac historické dáta, a to až od roku 1930 do 1990. Ako je z histórie známe, 20 storočie je významné svojim množstvom objavov. Preto sme sa rozhodli bližšie pozrieť na jednu oblasť a to antibiotiká. Na nasledujúcom grafe je zobrazený "vývoj" antibiotík vo vedeckých článkoch, presnejšie, jeho percentuálne zastúpenie v každom roku.



Obr. 11: Percentuálne zastúpenie článkov, ktoré obsahujú slovo "antibiotic"

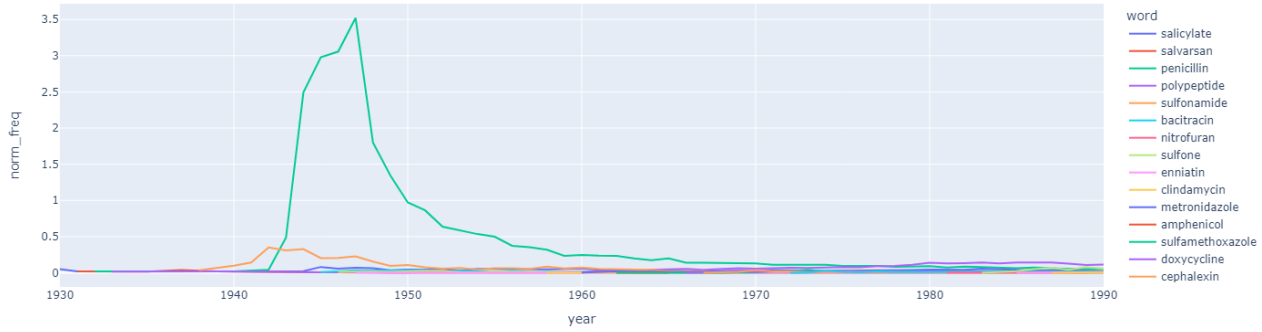
Ako vidíme na grafe, jeho vrchol nastal začiatkom 50. rokov 20. storočia, hoci objav prvého antibiotika, penicilínu, nastal už v roku 1929. Približne od roku 1955 začal postupne klesať, no ako vieme, nikdy už z našich životov antibiotiká nezmiznú.

No poďme sa na ne pozrieť bližšie, na jednotlivé druhy. Nasledujúci graf zobrazuje vývoj rôznych druhov antibiotík, ako napríklad spomínaný penicilín, sulfonamíd, polypeptíd a podobne v historických dátach.

Z grafu vidíme jasné vedenie penicilínu, ktorý svojim objavom prekopal celú medicínu v 40. rokoch. Jeho vrchol je posunutý od jeho objavu o viac ako 10 rokov, čo je spôsobené jeho výskumom. Vďaka nemu sa na trh prvýkrát objavil už v roku 1945 a odvtedy patrí medzi najznámejšie lieky na svete.

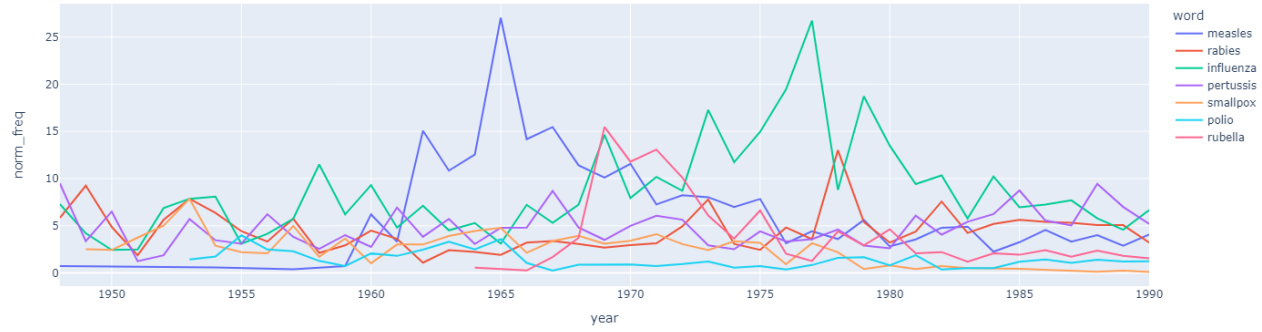
4.3.3 Pandémie a vakcíny

V nedávnej covidovej dobe sme o vakcínach počuli toho dosť veľa. No ako to však bolo v minulosti? Na aké choroby a kedy vznikli prvé vakcíny? Na to sme sa pozreli v ďalšej etape.



Obr. 12: Percentuálne zastúpenie článkov obsahujúcich dané antibiotikum

Na nasledujúcich dvoch grafoch sú znázornené rôzne choroby a ich vakcíny. Na jednotlivé pandémie sme sa pozreli jednotlivo.

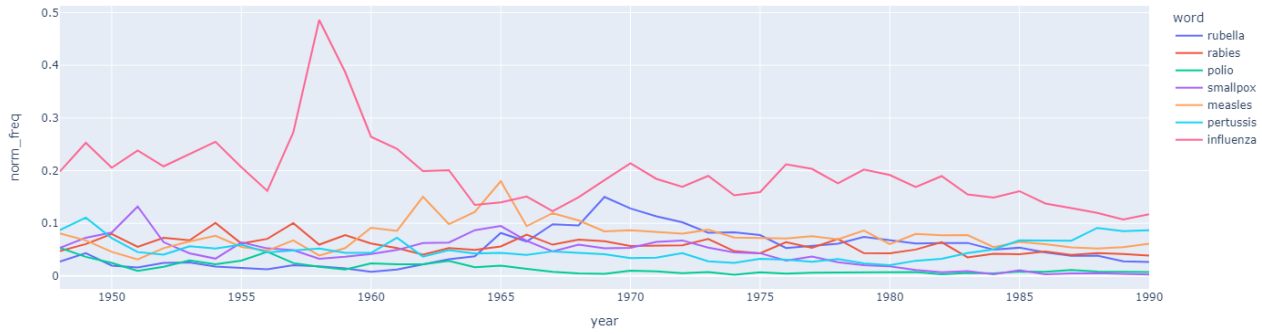


Obr. 13: Percentuálne zastúpenie článkov obsahujúcich danú chorobu a slovo "vaccine"

Ako prvú sme si vybrali measles, čo sú osýpky. Choroba sama o sebe ma staršiu históriu ako sú naše dáta. Preto na oboch grafoch je ich vrchol v rovnakom časovom období, teda keď bola vyvynutá vakcína proti nej. Pri pohľade na graf vidíme, že to bolo okolo roku 1965, čo potvrdzujú aj zdroje.

Ďalej sme sa pozreli na "influeanzu", čoho slovenský preklad je chrípka. Časové obdobia na grafoch sa nezhodujú. V celkovom precentuálnom zastúpení je jej vrchol v roku 1958, čo súvisí s tzv. Ázijskou chrípkou, ktorá prepukla v rokoch 1957-1958. V druhom grafe je jej najväčší vrchol v 70. rokoch. To súvisí s rozšírením prasacej chrípky a vakcinácií proti nej. Podľa zdrojov, jedna z najúčinnejších vakcín proti nej vznikla v roku 1978.

Tretiu a zároveň poslednú sme sa pozreli na "rubellu", alebo po slovensky ružienku. Jej



Obr. 14: Percentuálne zastúpenie článkov obsahujúcich danú chorobu

priebeh je podobný osýpkam. Teda na rovnakom časovom úseku majú vrchol, presnejšie v roku 1969. Práve toto súvisí s tým, že aj prvá vakcína na toto ochorenie vznikla v roku 1969.

5 Diskusia

Počas práce sme sa stretli s nasledovnými problémami:

Získanie dát

Jednou výzvou bolo ako dáta z databázy získať a spracovať do vhodného formátu. Najprv sme skúšali dáta ťahať prostredníctvom API, ktoré PubMed ponúka. Avšak pri sťahovaní väčšieho množstva sa celý proces extrémne spomalil, práve kvôli obmedzeniu 3 requests za sekundu.
Neskôr sme skúšali lokálne stiahnuť pár xml súborov a tie spracovať cez Python, ani to však nebolo dostatočne rýchle a navyše xml súbory zaberali pomerne veľa miesta. Práca s dátami za väčší počet rokov by tým pádom neprichádzala do úvahy. Nakoniec sme pristúpili k možnosti vytvoriť si lokálny archív PubMedu a dáta spracovať pomocou utility EDirect, čo síce trvalo dlhšie, avšak z dlhodobého hľadiska sa oplatilo, keďže sme ešte nevedeli s akými dátami budeme nakoniec pracovať. Po stiahnutí bolo získanie dát oveľa rýchlejšie ako všetky predchádzajúce možnosti.

Ďalšou výzvou bolo ako získať dáta iba za určité časové obdobie. Parsovanie dát do csv prostredníctvom EDirectu bolo možné iba po poskytnutí PMIDs. A síce sú záznamy článkov v PubMed databáze usporiadané, no iba podľa dátumu úpravy, nie dátumu vydania. Preto sme jedenkrát prešli celú databázu a vytvorili sme si pomocný csv súbor kde sme k danému PMID článku priradili jeho dátum vydania. Následne sme jednoducho vyfiltrovali PMIDs s požadovanými rokmi a k nim získali dáta.

Analýza príliš veľkých dát

Pri niektorých metódach spracovania textu trval celý proces príliš dlho. Ak sme chceli analyzovať dáta za viacero rokov, museli sme pracovať iba s výberovou vzorkou.

Vybrali sme celkovo 20% dát z rokov 1970-90, a aj 2007-2022 (a 10% zastúpenie z každého roku), čo malo okolo 1 130 651 záznamov pre roky 1970-90 a 1 638 187 záznamov z rokov 2007-2022. Rozdelilo sa to na 3 časti po 400 000 slovách (pri 2007-2022 550 000 slov) a lematizácia spolu s keywords zabrala vyše hodinu pre jednu tretinu dát, takže dokopy to trvalo vyše 3 hodiny. Naša výpočtová technika nepostačovala na to, aby sme skúšali viac záznamov.

Boli skúšané aj iné knižnice na lematizáciu, ako napríklad NLTK, ale spacy sa ukázala ako

lepšia a rýchlejšia, aj napriek tomu, že to trvalo vyše hodinu. Dlho trval proces tokenizácie slov, kde si slová prevádza na objekt typu doc.
Bol vyskúšaný aj iný postup, kde sme nebrali vždy iba jeden riadok tabuľky a ten lematizovali, ale zobrali sme všetky slová v stĺpci title_cleaned, spravili proces tokenizácie a následne lematizovali. Použili sme oddelovač " | ", aby sme vedeli, ktorý oddeľoval jednotlivé vety. Avšak, pri tomto spôsobe veľmi často padal jupyter, keďže si musel pamätať veľa znakov.

Na začiatku bola skúšaná aj technika stemming, ktorá na rozdiel od lematizácia nerozoznáva kontext slova (a preto je aj rýchlejšia), len ho prevádza do základného tvaru na základe pravidiel, pre anglický jazyk to je napríklad odstránenie prípony -ing na konci, -ity, -ful a iné.
Skúšali sme PorterStemmer z knižnice NLTK, avšak po zanalyzovaní výsledkov, ktoré neboli dostačujúce sme sa rozhodli ponechať účinnejšiu avšak pomalšiu techniku lematizácie.

Predikcia

Ďalším problémom, s ktorým sme sa stretli bola predikcia, resp. výber modelu, s ktorým by sme predpovedali.

Najskôr sme vyskúšali jednoduché predikcie na základe pár minulých dát, ako napríklad rozdiely medzi jednotlivými rokmi, percentuálne rozdiely či ich priemery. Potom sme chceli vyskúšať niektoré zabudované modely pomocou knižníc. Možností bolo veľa, no modely boli ťažšie na pochopenie a niektoré aj na implementáciu. Veľký problém bol aj ten, že tým, že sme pracovali s menším počtom rokov, tak počet časových období nebol dostatočný na predpoveď pomocou daných zložitejších metód a výsledky neboli správne.

6 Záver

Táto téma má veľa ďalších smerov, ktorými sa môže vyvíjať a pokračovať, v budúcnosti by sme sa mohli venovať daným témam:

- Predikovať, do akých kategórií patria kľúčové slová z článkov (napríklad či dané slovo je choroba alebo liek)
- Mali sme predpripravené zlematizované všetky unikátne slová, ktoré sa vyskytovali v článkoch - ktoré sme nakoniec nevyužili
- Ak by sme mali lepšiu výpočtovú techniku, mohli by sme zanalyzovať aj abstrakty nielen názvy článkov