// ENGLISH BELOW

Ebben az elemzésben arra voltam kíváncsi, hogy a “Putyin” kifejezést tartalmazó cikkeknek megváltozik-e a szentiment pontszáma a háború kitörésével.

Másrészt pedig arra voltam kíváncsi, hogy a cikkek szentiment pontszámára volt-e hatással a szerkesztőség cseréje.

Az elemzés nem jut szignifikáns eredményekre, mivel a kutatást nem előzte meg megfelelően körültekintő tervezés.

Ugyanakkor a projekt segíthet bemutatni az adattisztítási, és scrapelési készségeket.

==

Sikerült lescrapelni azokat a cikkeket az index.hu-ról, amelyek tartalmaznak a “Putyin” kifejezést és a “háború” címke alatt találhatóak.

A “Putyin” kifejezést kb. 1500 cikk tartalmazza. - ezeket scrapeltem le.

Felmondás előtt [2019-2021]

Felmondás után [2021-2023]

**Fájlok:**

Nyersen scrapelt adatok a scraper alapján

putin\_text\_raw.csv

df\_clean-ben tisztított adatok

cleaned\_putin\_text.csv

dfm a big\_code-ból

putin\_dfm.csv

**Scrapelés: Index\_putyin\_scraper.ipynb** -

**Scrapelés után marad kb. 1260 cikk**

**Nyers fileok tisztítása: df\_cleaner.ipynb** – ezt a változók átirogatásával futtattam 2x

Itt:

NaN törlés

HTML tagek törlése

/n törlés

whitespace törlés

duplicate törlés

“Kövessen minket facebookon is” törlés

Ezt követően marad kb. 1160 cikk

**A szentiment-elemzést a putin\_big\_code.R file végzi**.

A fájlok alapvető szűrésen, és dátumkezelésen estek át.

Itt különleges karakterek, whitespacek, és a dátumok string formátumból date formátumba való konvertálása történt.

Előkészítés:

- létrehoztam egy corpust

- tokenizáltam

- lemmatizáltam a tokeneket.

- kivettem a stopszavakat a tokenek közül.

- végül egy documentum-feature mátrixot hoztam létre. Ezeket kimentettem.

**Elemzés:**

- Leellenőriztem, hogy a cikkek hosszúsága változik-e a háború kitörésével. Ez beigazolódott. Ugyanakkor a szerkesztőség leváltása nem hozott trendszerű különbséget a cikkek tekintetében.

- Két részre bontottam az elemzést. A szentimentek ábrázolásást nem csak cikk szinten aggregáltam dátumok alapján. Létrehoztam egy külön sort a putyin kifejezés környékén található 300-300 karakterláncra. Így nem volt olyan nagy a szórás szentiment terén, és egyfajta csökkenés is megfigyelhető.

- Napokra aggregélt szentiment pontszámokat ábrázoltam a ggplot, és a ggplotly segítségével.

- Leggyakoribb szavak ábrázolása.

// ENGLISH

I have been doing sentiment analysis on index.hu articles which contain the term “Putin”

I was curious how the sentiment scores change as the war rages. Also during this time majority of the journalists of the site resigned, so perhaps the change in sentiment scores of articles containing the Putin term may be explained by the date of resignation.

I was able to scrape the articles which contain the “Putin” term, while it also had the “háború” (“war” in hungarian) tag as well.

==

On the website I have found roughly 2100 articles, while only 1500 have been scraped.

I also tried to take into account the fact that the majority (roughly 80 percent of the writers) of the site resigned. I tried to check if it had any effect so I split the data to 2 groups:

Before resignation [2019-2021]

After resignation [2021-2023]

As the resignation happened at 2021.07.24. I set the difference at 2021.08.24. as resignation takes 1 month in Hungary. If there had been sentiment score difference explained by the resignation itself I would have had anticipated it only after 08.24. because that is the time when newly recruited people arrived to the site.

**Files:**

Raw Scraped data:

putin\_text\_raw.csv

Data cleaned in df\_clean:

cleaned\_putin\_text.csv

document feature matrix from big\_code (after cleaning)

putin\_dfm.csv

**Scraping: Index\_putyin\_scraper.ipynb** -

**After scraping is complete roughly 1260 articles remain**

**Cleaning raw files: df\_cleaner.ipynb**

Itt:

deleting NaN

deleting HTML tags

deleting /n

deleting whitespace

deleting duplicates

deleting “Kövessen minket facebookon is” boilerplate text

After this 1160 articles remain

**Sentiment analysis performed by putin\_big\_code.R file**

Files have been further filtered and the dates have also been formatted from string to date.

how:

- made a corpus

- tokenized the text by words

- lemmatized tokens

- deleted stopwords from tokens

- Eventually made a document feature matrix, and saved it.

**Analysis:**

-I used the HunMineR dictionary to determine a token’s sentiment score. I summed the token’s scores which can also be negative. Scores had been grouped by articles, and summed by days. Eventually plotted the results.

- However I tried to reduce the deviation. So I reduced the scope of the analysis and I only checked for every 300 characters before, and after the “Putin” word in any given article. I also summed the same way, and plotted the results as mentioned before. The standard deviation had been reduced, but I could not find any conclusive result based on this relatively outdated method (compared to neural network models).

- Checked weather the length of the articles has changed as the war starts. It is confirmed, however the number of articles before was extremely low compared to before the war. At the same time the resignation did not bring any change in the length of the articles.

- Sentiment scores have been plotted out with ggplot and ggplotly.