1 CZ check

We first run a check for cz language. For this purpose we used fasttext model. It can only work on lines so we had to split to lines and run it. At first we also applied the filtering to brief. But we found it that is had a lot a false negatives. We decided to drop it and oniy use the filterin on article This produces a percentage of czech lines. We tried to use ratio o 1.0 and checked the results. We got: TODO It eliminiated a lot of articles so we decided to inspect it more closely. We generated histogram graphs based on ratio to see. We decided to inspect the result with ratio of 0.0. As expected we found a lot of ukraine article in seznamzpravy.cz Deník had a lot of articles with just one line with 'VIDEOSOUHRN SPUSTíTE ZDE' etc... Idnes had problems with tables being parsed cell per line. So for mobile phones it had no chance to be czech. Other problems was english that was prevalent. We also found few slovak articles.

0.3 outliers. Stil the same problems but more of false negatives. Often due to table or sport matches

Matches problem: CELKOVÉ POŘADÍ - JEDNOTLIVCI 1. a 2. místo TC0029 Milan Matoušek, Tachov 13 84 TC0059 Radek Mazanec, Planá 13 84 3. až 12. místo

Tables problem

We the proceed to NULL author that weren't real people.

2 Article inspection

After that we inspect articles per server basic

3 Irozhlas

We first inspect the oldest articles. It turns out there is one from 1992. Idk why it's there but i check url and it is really from that year. Should I erase all to 2017 because on github I said i will use only 2017-2019? Articles are correclty parsed because irozhlas has only one version feeelsgood.

3.1 headline

First we checked both ends of headlin we didn't find that the end of headline would be forceully cut off. Howerver there we problems at low number of characters. We found few articles with just two characters or one words. When preprocessing we would also split headline by [-] an then take a first part. However it turns out that it might not have been good idea as we would then get this headline. Fotbalisté Boleslavi deklasovali v play or Proč krachují start. We will have to run the extractor again I guess without this rule. After that we will set threshold to 25/30 characters.

We also seen holes in the graph. However we listen the parts where there are supposed to be 0 headline length but there we no holes. Idk why it shows it like this.....

We also checked why the mean line is so perfect around 2021 for mean by date. 1 reason is simply because of graph density. Second is that the write are force to write healines of length 110 characters. We isnpected headlines and they were not cut off. We also checked the outliers with high values but they were perfeclty fine.

3.2 article length

No problem at big lengths. Those are podcast transcripts. I don't see reason to cut them off. As artciles over 4k words are jsut 5

We found that articles with length ¡ 100 were either bad parsed or just ome weird articles. This was also seen ¡ 300. We decided to cut off articles with length ¡ 300.

We also inspected the big outlier with huge aritcle length that happend in 2022. It's this url and it is truly feaking long https://www.irozhlas.cz/zpravy-svet/polsko-evropska-unie-soudni-dvur-eu-europoslanec-marek-belka$_2112151341_e rn$.

3.3 avg word length

We found out that low avg word length were the sport results. We decided to cut off articles with avg word length ¡ 4.3. We haven't found anything wrong with high values. Otherwise we can see from graphs that there is high correlation between avg word length and article length.

3.4 word num

Since we could easly see huge correlation between word num and article length we decided to use ratio. We found the sport results articles to have ratio ¿ 0.22.

3.5 brief

We found out that we could filter weather articles based on brief. noc -6 až -15 °C — den -9 až -5 °C. Also from what I have seen author article usually ahve longer brief than 50. If not it's usually sport report or aut oreport. No problems with long brief. We also check strange same brief at the 2022 but it is just bad visualization.

3.6 non-alpha

We found the non-alpha to be quite useless as it again corellated with article length. We decided to use ratio of non-alpha/article lengt. We found out that articles with 0.045 are the ones with sport results. low values not particularry interesting. We also observed the outlier in 2004 and those were again sport resuts.

3.7 num of words per line

We haven't found any interesting with high value. But we found that with value lower than 14 there we related articles list at the end. Lower were also weather prediction.

4 Idnes

We first inspect the oldest articles. First of all there is a article from 1900 which is href[https://www.idnes.cz/revue/zajimavosti/v-britanii-se-podarilo-vypestovat-dvojbarevne-jablko.A090928$_1$05911$_z$$ajimavosti_nhthisone.Probablyabugonidnesside.Iremoveditfromanalysis$

4.1 headline

We first checked ¿ 0.999 outliers. We didnt' find any probelms there. For ¡ 20 it again showed the problem with [-]. We inspected 2014+ articles and didn't find any cutting

We also checked drop after 74 but didn't find any cutting.

4.2 article length

No problem at big lengths. Those were some astornout lnog articles. I don't see reason to cut them off. As artciles over 6k words are jsut 5There was also one with sport results

We found that articles with length ¡ 100 were either bad parsed or just ome weird articles. ¡200 typicall short news. This was also seen ¡ 300. We decided to cut off articles with length ¡ 300.

example of short one: https://www.idnes.cz/hry/kratke-zpravy/nadherny-pohled-na-theed-ve-star-wars-battlefront-ii.A170608$_1$10306$_b$$w-kratke-zpravy_oz$

example of short one at 199x https://www.idnes.cz/hry/recenze/quest-for-glory-v-dragon-fire.A000405$_q$$uestforglory5_bw$

We also inspected strange outliers around 2000 but they were correctl parsed.

4.3 avg word length

We haven't found anything strange on both sides. We found some sport result ¡4 but it wasn't prevalent.

4.4 word num

Again we found 0.22 ¿ ratio to be sport results Smaller.

4.5 brief

WE found that value ¡ 20 were typically not articles but rather manuals or file pages eg čeština do hry x or it was weather

We also check strange outliers around 2000 no idea if they are fine ???

4.6 non-alpha

Again 0.045 ¿ ration was sport results. We have also checked peaks at ¡ 2005 but they were correctly parsed sport articles.

4.7 num of words per line

Clealry there is problem with newlining at first extractor which ends at 2011. It doesn't correctly add . Will check it but I am not sure if it's even needed to fix it as we will just tokenize whole article and this should make a difference.

WE also check the low outliers but those were probbaly wrongly parsed as some of them had like 10 words per line. Tho not problem if tokenized.

4.8 Key findings

Article length peaks are ok as per section. Same for brief length. Num words per line wrongly parsed line but should metter. non-alpha ratio is also fine

5 novinky

Again we found old article from 1900. I removed it from analysis as it fucked up the graphs.

5.1 headline

Nothing wrong with long ones Nothing interesint at == 20. But ¡ 20 were strange. Nothing strange with them date ¿ 2021.

5.2 article length

BIG PROBLEM !!! a lot of articles with ¡ 200 length. Must cut it. Like this one https://www.seznamzpravy.cz/clanek/fond-kinematografie-chce-vic-na-pobidky-zahranicnim-stabum-119863

Long articles not problematic as they are interviews.

which only parsed the headings. Strange thing is that we have two extractors for novinky but the proportion of articles with ¡ 200 length is the same. So it's not due to extractor.

5.3 avg word length

Outliers over 0.99 shows the problematinc articles with only heading.

5.4 word num ratio

Both high and low values shows problematic articles

5.5 brief

Overall brief looks kinda scatchy as they don't look like briefs. Low ¡ 20 briefs show that. Will have to recheck extracting.

5.6 non-alpha

Again we can see clealry which articles are problematic. Because they have very low non-alpha ratio as those are only headings. No common denominator of high values.

5.7 num of words per line

Same problem low values show headings.

5.8 Key findings

BIG PROBLEM with content extraction oftentimes extracted only headings. No idea why there are holes in headline but shouldn't be there basedo on data. No way to inspect articles peak at 2018-07-19 as the articles are not there anymore.

6 SeznamZpravy

6.1 headline

Classic - problem. No probelm otherwise. No idea why there is steap drop around 2019. INspected it and found nothing strange. No idea how to debug the spikes.

6.2 word length

Nothing interesting

num words ratio Same problem as with novinky ¿ 0.22 shows only headings.

6.3 article length

The long articles at 2022 are due to covid articles or interviews. IIMPORTANT again with articles ¡ 200 we spotteed problem with headings. We also check

6.4 brief length

The step drop at 230 was inspected and is probably restricted by senzamzpravy as no cut was seen.

6.5 non-alpha

¿ 0.06 shows heading probllem.

6.6 num of words per line

No problem with parsing now. High values looked fine, just long paragraphs. Low values typically contained table.

6.7 Key findings

Again problem with heading. Brief is not probellmatic with drop. No idea how to debug steep drop at 2022-02 of num of articles.

7 aktualne

7.1 headline

Same problem with [-]. Check the drops and haven't found any cutting. Check change by date

I have also checked the dramatic drops around 2008 and 2017. No idea what happend at 2008. I checked the headline before and after and they were fine.

2017 was when layout of page change and we use extractor 2.0 for that. However I again checked both ends and they are correct articles thus it's likely that writers were told to write longer articles.

Same applies to 2019.

7.2 article length

The long articles are transcripts of documents. like this https://zpravy.aktualne.cz/domaci/dokument-vladni-navrh-na-zruseni-delnicke-stany/r i:article:649724/

I think we should cut it by setting max length to something similiar to outher sources.

¡ 200 are typically photogalleris or videos. We should cut it by setting min length to like 300. Maybe even higher

7.3 avg word length

No problem with high values. ¡ 3.8 usually soprt results.

7.4 word num ratio

¿ 0.22 usually sport results. ¡ 0.11 Usually short pages like photogallery or videos but not as prevalent as with aritcle length.

7.5 brief

Low values typically sport results live Slava - Spatta 1:2, but there not exclusseive.

We found high values like 300 to be valid briefs and correctly parsed. For some reason these long ones are from second extractor period. No idea why.

FOUND IT. The difference is that there are basically two briefs for the article. One that is short is in metadata and one is at the start of the article. NEEDS TO BEEEE FIXED.

7.6 non-alpha

¿ 0.06 usually sport results

7.7 num of words per line

Nothing interesting

7.8 Key findings

Problem with brief. Must be extracted from article. Otherwise okay. Must cut short articles to elimate gallerys and videos.

denik

7.9 headline

Reason for peaks is simply that bin size is too big and thus some bins conainta only partial data. I checked 80 drop and haven't found any cutting. ¿ 80 headlines were correctly pardsed. Same probles as other [-]

7.10 article length

No problems found at that department. Even ¡ 200 were okay.

7.11 avg word length

¿ 10 problem as it was just urls. ¡ 3.8 just sport results.

7.12 word num ratio

¿ 0.22 just sport results. ¡ 0.11 just ulrs.

7.13 brief

Briefs correctly parsed. Even ones with 1000 words. ¡ 20 were fine too.
non-alpha ¿ 0.06 sport results/electoin results. ¡ 0.01 not interesting.

7.14 num of words per line

¡ 10 table like articles with listings etc. Long one probably has wrong splitting, but no problem with tokenization.

7.15 Key findings

Overall it was very well parsed. The only problem are those super short articles with long word length.