ESTONIAN NEWS

Magnus Paal, Tõnis Hendrik Hlebnikov, Alvar Antson

Introduction

The idea of our project was to scrape data from Postimees Group news websites and to analyse the data and perform machine learning to find patterns that will bring the most views/reads on an article. Because Postimees group profits primarily from advertising, the read count on an article is the primary indicator on how well the article has done.

Data and preprocessing

We ended up gathering data from Postimees Group API, because we could get the read count from there. Our plan was to gather as many columns as we could find. In the end we got the article's newspaper, datetime, id, title, content, read count, and author, section. Since section is not really helpful, we excluded it. Also newspaper only had one value, so we excluded that as well. Columns that might need explaining are id, which is a unique identifier of the article, read count, which is the number of people who have seen or read the article. From approximately 250,000 articles we got around 214,000 after dropping duplicates.

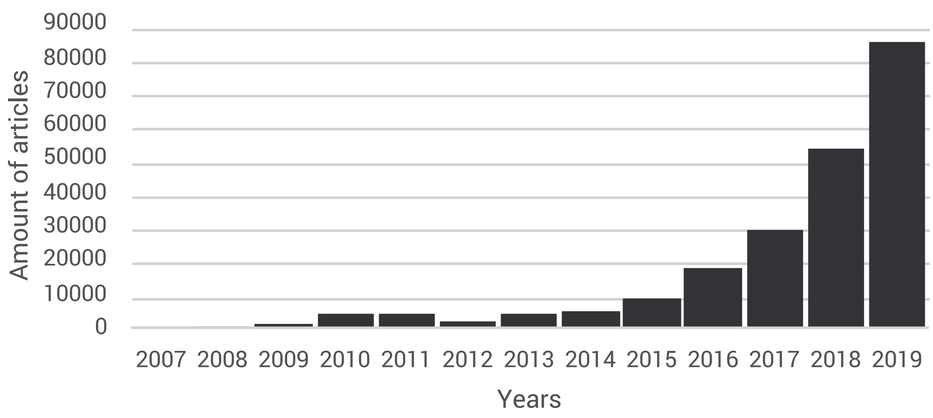

Figure 1. Articles from each year.

Data mining

The main objective of data mining was to find what influenced the articles read count. The best way to do that, was to find out average read count and compare it to the article's feartures. Most interesting was title's length compared to the average read count (Figure 2).

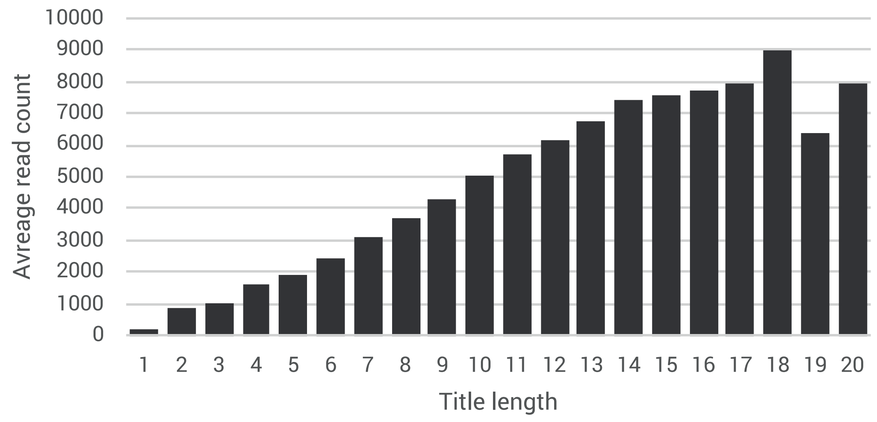

Figure 2. Average read count of articles with the given read count.

We also wanted to know what words and names in the article resulted in the highest read count. For this we used only words in the title, because the article is mainly read because of the title or the image. We converted each of the words in the title to its lemma form using estoninan natural language processing tool estnltk. If a word occured twice in the title then we left only one instance of it. Then we summed the read count of the articles the word was in the title of and divided it by the number of articles it was in the title of. The same was found for bigrams. The 10 words, names and bigrams (names not included) with the highest average read count can be seen on the next page

word	amount	freq	avg_read_count
Kingo	1022718	69	14822
Sardiinia	831500	58	14336,2069
Nolani	957272	70	13675,31429
Nolan	990296	75	13203,94667
ELU24	1261853	98	12876,05102
Simsoni	718970	59	12185,9322
15aastane	686132	58	11829,86207
Wales	1101638	94	11719,55319
Pullerits	783965	68	11528,89706
Kuzitškin	690164	60	11502,73333

Figure 3. Average read count of articles with given names in title (50 to 100 occurences).

word	amount	freq	avg_read_count
Tänak	9908107	508	19504,14764
Neuville	1746996	107	16327,06542
Järveoja	1434702	114	12585,10526
Toyota	1966349	160	12289,68125
Kaja	2184303	186	11743,56452
Ogier	1298714	111	11700,12613
Helme	4683796	403	11622,32258
EKRE	5535438	494	11205,34008
Ott	3510210	333	10541,17117
Tänaku	2691976	259	10393,72973

Figure 4. Average read count of articles with given names in title (100 to 1000 occurences)

word	amount	freq	avg_read_count
maksimum	1514856	41	36947,70732
doping	1004853	40	25121,325
bikiinipilt	888147	43	20654,5814
blogi	1257681	73	17228,50685
sotsmeedia	778237	47	16558,23404
koobas	771851	52	14843,28846
seksima	783002	56	13982,17857
vägistamine	590604	43	13734,97674
postitama	749950	55	13635,45455
ränk	1258835	94	13391,8617

Figure 5. Average read count of articles with given words in title (40 to 100 occurences).

word	amount	freq	avg_read_count
ralli	5842185	468	12483,30128
sõitnud	1771145	154	11500,94156
kahtlustatav	1060365	104	10195,81731
modell	1735445	173	10031,47399
eestlanna	1219443	124	9834,217742
seksikas	1097160	113	9709,380531
vaatepilt	1059400	111	9544,144144
neiu	1810381	196	9236,637755
kleit	951635	104	9150,336538
kaader	2078249	236	8806,139831

Figure 6. Average read count of articles with given words title (40 to 100 occurences).

bigram	reads	freq	avg_reads
('lahku', 'läinud')	318421	10	31842,1
('mees', 'hukkus')	297304	10	29730,4
('võitis', 'kindlalt')	248208	10	24820,8
('noor', 'naine')	446749	18	24819,389
('ei', 'varja')	286739	12	23894,917
('teisel', 'kohal')	447395	19	23547,105
('kahe', 'auto')	233457	10	23345,7
('kukkus', 'kokku')	254574	11	23143,091
('kahe', 'lapse')	261476	12	21789,667
('levib', 'uus')	243083	12	20256,917

Figure 7. Average read count of articles with given bigrams in title (10 to 20 occurences).

bigram	reads	freq	avg_reads
('hukkus', 'liiklusõnnetuses')	364882	21	17375,33
('viimsele', 'teekonnale')	274470	21	13070
('välja', 'sõitnud')	339601	26	13061,58
('löögi', 'saanud')	271572	25	10862,88
('tunned', 'ära')	270932	25	10837,28
('juhtub', 'kehaga')	259802	25	10392,08
('peidab', 'end')	299866	29	10340,21
('uue', 'omaniku')	224143	22	10188,32
('väga', 'halb')	201091	20	10054,55
('miks', 'tekib')	212239	23	9227,783

Figure 8. Average read count of articles with given bigrams in title (20 to 30 occurences).

Machine Learning

The plan was to predict read count based on the title. Best results were achieved when articles were divided to two categories by read count. To avoid class imbalance, the classes needed to be equal in size. So the articles were divided by read count under 1000 (0) and read count over 1000 (1). Word feature vectors with dimensions 1 to 3 were generated from titles and assigned frequencies. Training and test sets were made by 70/30 division. Best results at predicting the category were achieved by using Linear Support Vector Classification with accuracy of around 0.73 on the test set.

	Predicted: 0	Predicted: 1
Actual: 0	25560	7847
Actual: 1	9218	21581

Figure 9. Confusion matrix of predicted and actual labels.

Conclusion

- The data was too biased by the year 2019, which can be seen on figures 4 and 6 where relevant things of that year prevail, mostly the World Rally Championship.
- Processing large amounts of text takes long, so we didn't end up using the article's content much, and it was decided that it is actually not what influences read count.
- Initially we hoped to achieve better accuracies with the machine learning model, but we are content with the accuracy we got, considering how difficult it was to achieve.

REPOSITORY:

TARTU ÜLIKOOL
1632
arvutiteaduse instituut