Results of similarity analysis of online news in Czech republic

Ing. Ondřej Veselý

Motivation and dataset

- the most common situation where students are confronted with text reuse is online journalism
- curiosity to quantify; no available datasets
- six-month long press monitoring of major online news publishers in the Czech Republic
- 60,000 articles; both regional and national

Motivation and hypothesis

Thank to IPPHEAE we have a lots of data about text reuse in academia but there is no precise data about the situation in news journalism. We start with the fact, that “**it is known that more than 10--20% of articles collected by portal sites are nearly identical or quite similar**” (Chang-Keon R., 2009)

Process

1. dataset creation
2. plain text data extraction
3. similarity analysis
4. data visualisation

Dataset creation

14	http://www.brnenskadrbna.cz/rss/	Brněnská drbna
15	http://www.zitbrno.cz/feed	Žít Brno
16	http://www.super.cz/rss2	Super.cz
17	http://www.denik.cz/rss/z_domova.html	Deník.cz
18	http://zpravy.aktualne.cz/rss/	Aktuálně.cz
19	http://www.ceskatelevize.cz/ct24/rss/vsechny-zprav...	ČT24.cz
20	http://idnes.cz.feedsportal.com/c/34387/f/625936/i...	iDnes.cz
21	http://hn.ihned.cz/?p=500000_rss	Hospodářské noviny
23	http://servis.metro.cz/rss.aspx	Metro
24	http://echo24.cz/rss/s	Echo24
25	http://www.parlamentnilisty.cz/export/rss.aspx	Parlamentní Listy

CHCU Žít Brno » Languages » Čeština

Stanovisko německých sociálních demokratů k Deklaraci smíření
4.6.2015 14:31

Níže si můžete přečíst překlad vyjádření předsedkyně frakce SPD v německém chápat jako oficiální postoj německých sociálních demokratů. Berlín, 29.05.201 poškodily a... Read more »

Zápisky z cest – Holdingová divadelní společnost v Grazu
2.6.2015 13:02

Vícezdrojové (kooperativní) financování, to je klíčová agenda v oblasti kultury. živá umění (divadlo, opera, balet, filharmonie) na obce. A to včetně finanční zá

Asi mi to dojde až později
1.6.2015 20:18

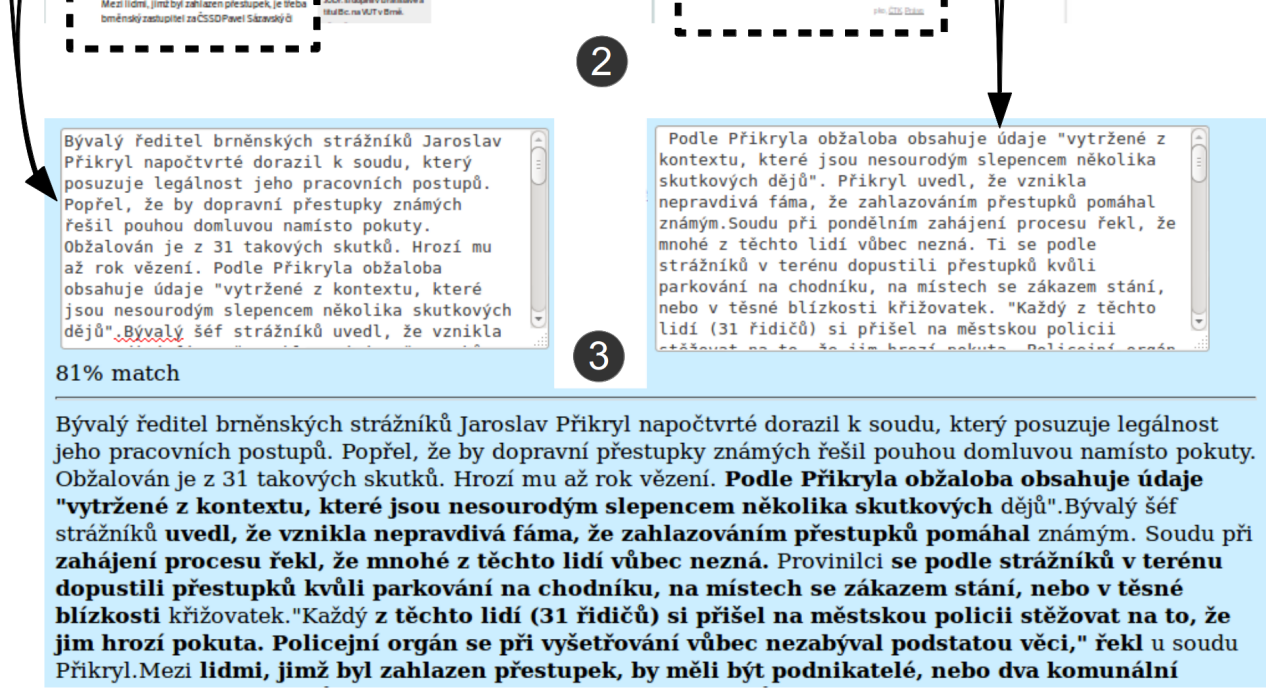
„Asi mi to dojde až později." Nemyslel jsem si, že kdy budu nucen použít tuto Stalo se něco divného. Stalo se to, že stovky Brňáků a mnoho německy mluvíc

url	perex	date	address	lat	lng	title	source	updated	hash	fulltext
http://www.plzenskenovinky.cz/sid=nis8q1cenjg779s8...	Na Hrách VII. letní olympiády dětí a mládeže se ml...	10.06.2015		0	0	Na dětských Hrách VII. letní olympiády uvidíte i ...	Plzenske novinky	2015-06-10 10:45:58	4f3eb6fe	<html><body> <p>Zapálením...

relevant text
identification
and
plain text
extraction

similarity processing

blog.orwen.org/textmatcher



Every similarity analysis of a pair of the articles A and B produces two values

1. how A is similar to B <=> s_{AB} = sim(A,B)
2. how B is similar to A <=> s_{BA} = sim(B,A)

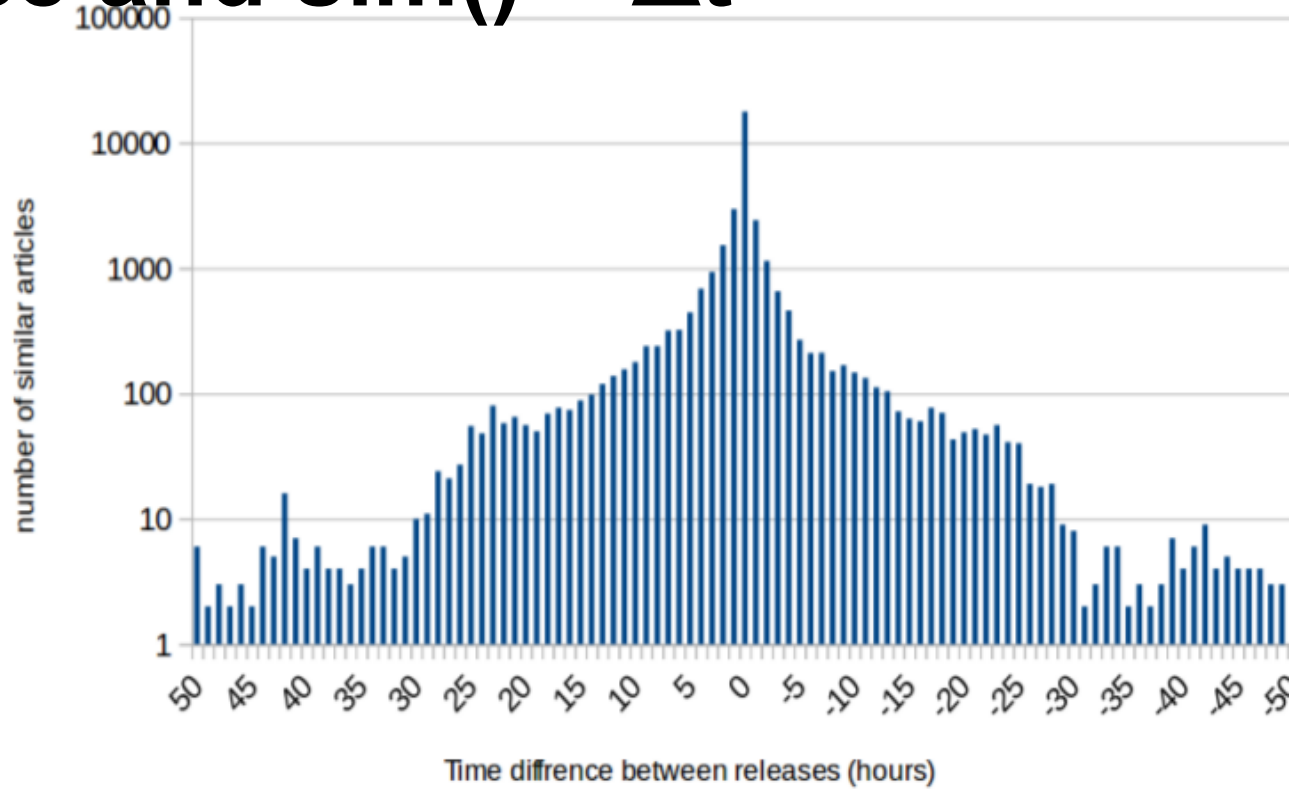
For example for texts A = “aaabbb” and B = “aaa”, the results are

s_{AB} = sim(A,B) = 50%

s_{AB} = sim(A,B) = 100%

Performance and sim() ~ Δt

- hundreds new articles per day
- not matching every article just those which were published in the same two-week window
- explain ČTK

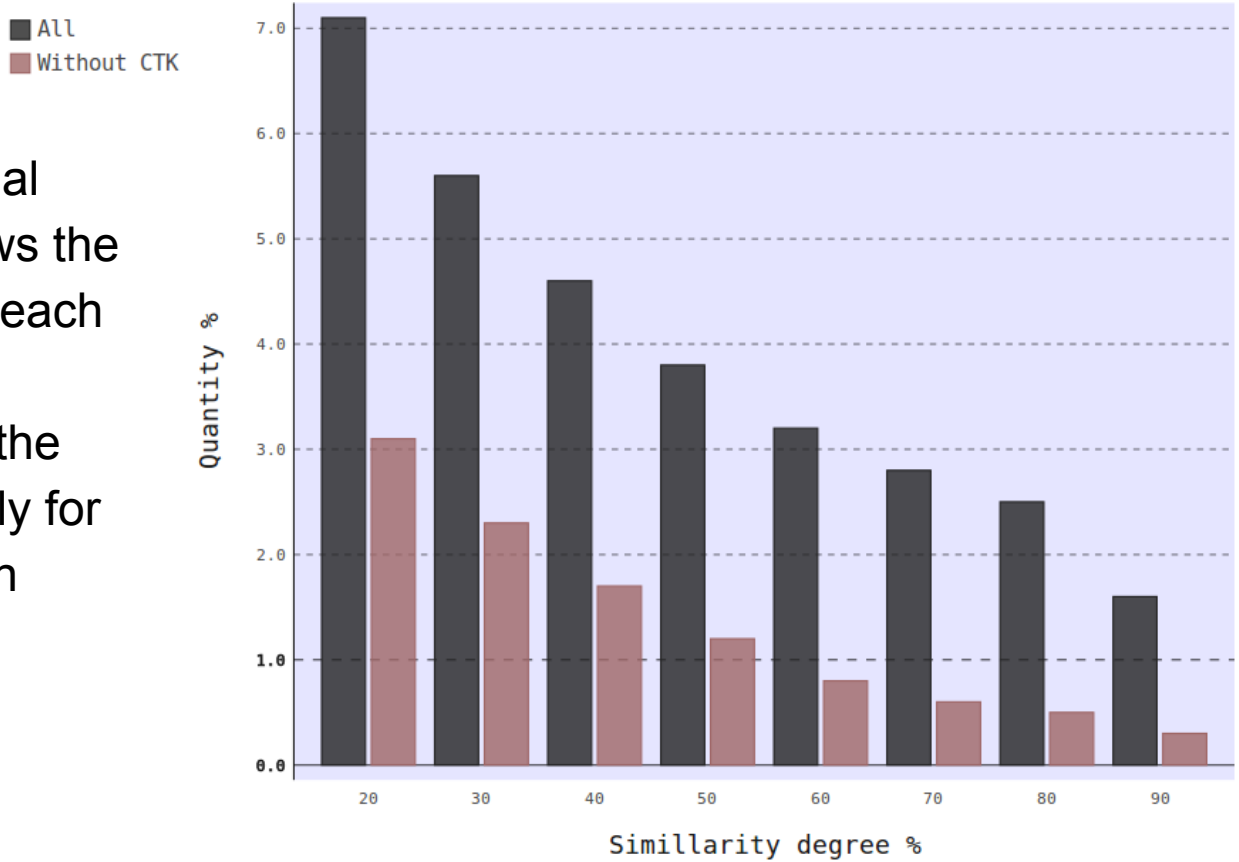


Relation of number of similar articles based on their published time difference (please note the logarithmic scale).

Straight to results (regional media)

Overall similarity on regional news servers dataset shows the number similar articles for each similarity class.
The lighter columns show the overall similarity values only for articles which had not been produced by ČTK

~ 2% of similar articles

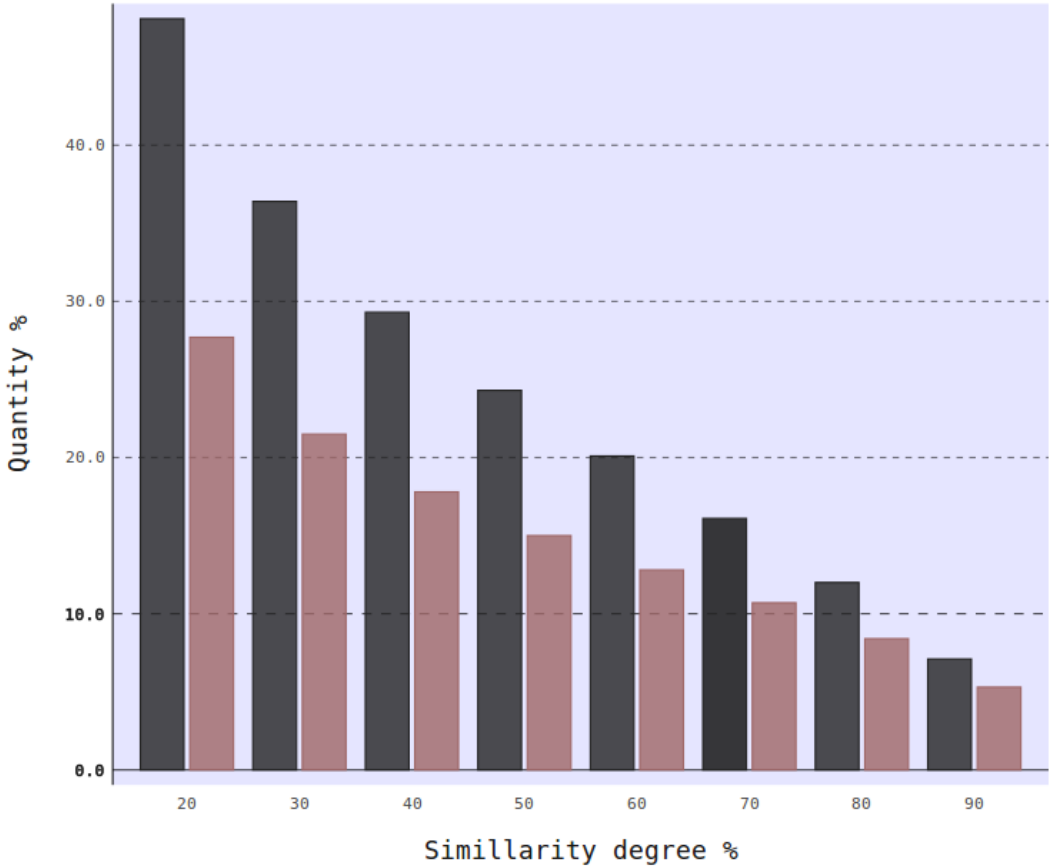


Straight to results (national media)

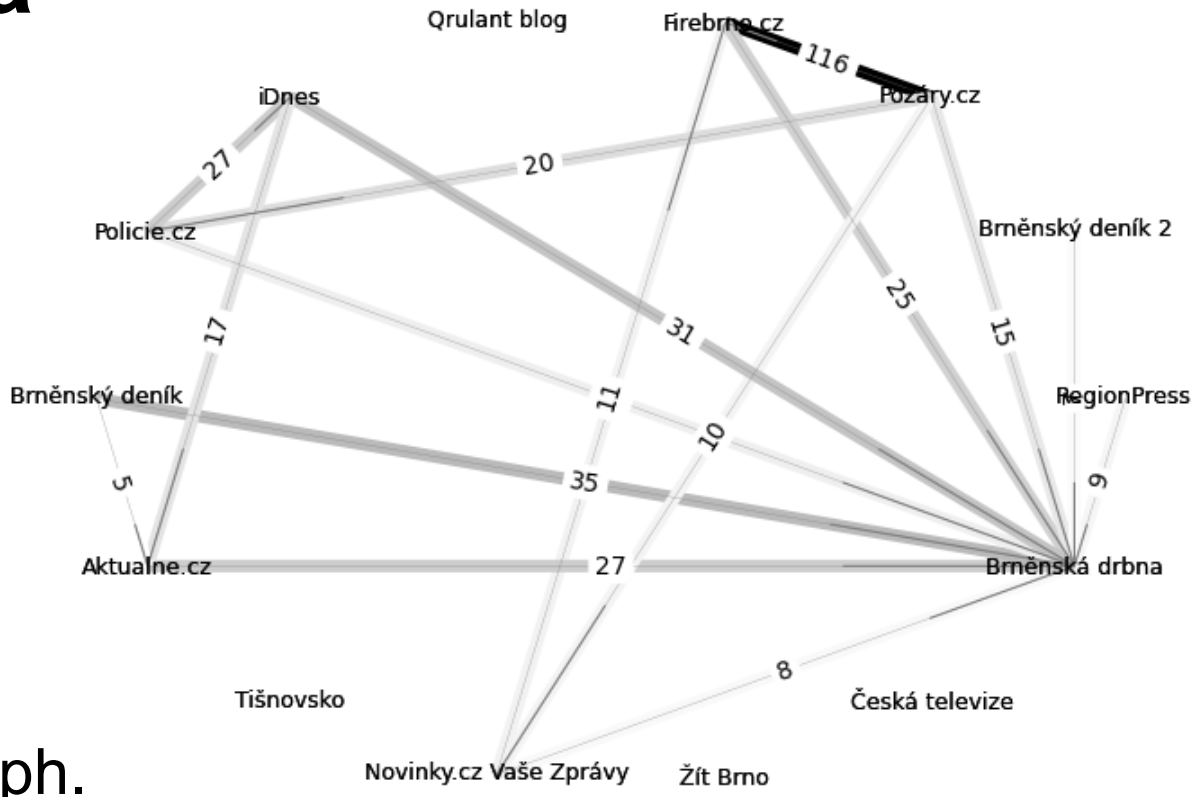
Influence of Czech news agency to text reuse is weak, especially for text with highly similarity degree in comparison with regional articles.

~ 20% of similar articles
~ 10% of very similar articles

Chang-Keon R. is right



Cross media content similarity



Just interesting graph.

Conclusion

Approximately 10% of all articles are 80% or more similar to another one published by different server which makes Chang-Keon Ryu's statement valid in Czech environment considering the national dataset.

The ČTK is responsible for 20-40% of text-reuse cases on national level, but other causes have not been revealed. On the regional level the ratio of similar text is five times lower and the most similar articles are created by ČTK.

Questions?

Ondřej Veselý -- www.orwen.org