

# **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	<a href="http://www.brnenskadrba.cz/rss/">http://www.brnenskadrba.cz/rss/</a>	Brněnská drbna
15	<a href="http://www.zitbrno.cz/feed">http://www.zitbrno.cz/feed</a>	Žít Brno
16	<a href="http://www.super.cz/rss2">http://www.super.cz/rss2</a>	Super.cz
17	<a href="http://www.denik.cz/rss/z_domova.html">http://www.denik.cz/rss/z_domova.html</a>	Deník.cz
18	<a href="http://zpravy.aktualne.cz/rss/">http://zpravy.aktualne.cz/rss/</a>	Aktuálně.cz
19	<a href="http://www.ceskatelevize.cz/ct24/rss/vsechny-zprav...">http://www.ceskatelevize.cz/ct24/rss/vsechny-zprav...</a>	ČT24.cz
20	<a href="http://idnes.cz.feedsportal.com/c/34387/f/625936/i...">http://idnes.cz.feedsportal.com/c/34387/f/625936/i...</a>	iDnes.cz
21	<a href="http://hn.ihned.cz/?p=500000_rss">http://hn.ihned.cz/?p=500000_rss</a>	Hospodářské noviny
23	<a href="http://servis.metro.cz/rss.aspx">http://servis.metro.cz/rss.aspx</a>	Metro
24	<a href="http://echo24.cz/rss/s">http://echo24.cz/rss/s</a>	Echo24
25	<a href="http://www.parlamentnilisty.cz/export/rss.aspx">http://www.parlamentnilisty.cz/export/rss.aspx</a>	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 chápání jako oficiální postoj německých sociálních demokratů. Berlín, 29.05.2015. Poškodily a... [Read more »](#)

### [Zápisky z cest – Holdingová divadelní společnost v Grazu](#)

2.6.2015 13:02

Více zdroje (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 mluvících...

url	perex	date	address	lat	lng	title	source	updated	hash	fulltext
<a href="http://www.plzenskenovinky.cz/sid=nis8q1cenjg779s8...">http://www.plzenskenovinky.cz/sid=nis8q1cenjg779s8...</a>	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><div><div align="justify"><p>Zapálením...

Cinema screen  
Full HD IPS monitor  
s HDMI LG 23MP65HQ

CINEMA SCREEN

LG  
Life's Good

Kup na alza.cz

4 199,-

**iDNES.cz** / Brno a jižní Morava

Průběh >

Průběh 10. srpna 2013, Úterek

[iDNES.cz](#)
[Zprávy](#)
[Rok](#)
[Sport](#)
[Kultura](#)
[Ekonomika](#)
[Finance](#)
[Žijeme](#)
[Galérie](#)
[Auto](#)
[Hobby](#)
[Mobi](#)
[Technik](#)
[Dna](#)
[Krimi](#)
[Revue](#)
[Blog](#)
[Hry](#)

# Napočtvrtě šéf brněnských strážníků k soudu přišel, obhajuje 31 skutků

Bývalý ředitel brněnských strážníků Jaroslav Příkryl napočtvrtě dorazil k soudu, který posuzuje i legálnost jeho pracovních postupů. Popírá, že by dopravní přestupky známých řáží pouhou domnělou namísto pokuty. Obžalován je z 31 takových skutků. Hrozí mu až rok vězení.

## Další články z rubriky

[Náměstek perfectoid česky, ale do Brna patří jen řba a veřejná kanalizace](#)

[Pět set dětí z Brna a okolí bez školy. A to kvůli tomu, že nechtějí](#)

[Drochovice posílají platit nebezpečný, tak občanům si domy, aby se neblýželi](#)

[VUT v oparce zasloužil fakultu, oběti to ale on, upravení přístřeší](#)

[dobře se najím.cz](#)
[a další restaurace](#)

**Restaurace Reblo**

**NORDSEE**

ve Václavské a Žitné

"Obžaloba je vytvářena v kontextu, je to nesourodý sjezd necenzurních nesourodých skutkových dějů. Musí se odpovědět na otázky, jak jsem se já k takové situaci vůbec dostal. Neřeší se přestupky. Tito lidé, kteří jsou zde uvedeni, se namí obraceli s otázkou, aby se zeptali, proč se strážníci chovají tak, jak se chovají. Považovali to za šikanu," tvrdil u soudu Příkryl.

Bývalý šéf strážníků uvedl, že vznikane pravidla, která se zahazování přestupků **poznává** známým. Tvrdí, že mnohé z těchto lidí vůbec nezná. Provozní se podle strážníků v rámci dopravního přestupku kvůli parkování na chodníku, namísto se zákazem stání nebo většinou blízkostí chodníků.

"Každý z těchto lidí (31 lidí) si přišel na městskou policii stěžovat nato, že jim hrozí pokuta. Policie jim orgán se při vyšetřování vůbec nezabýval podstatou věci," řekl u soudu Příkryl.

Mezi lidmi, jímž byl zaházen přestupek, je třeba brněnský zastupitel za ČSSD Pavel Šabrákovič.

**Jaroslav Příkryl**  
 Narodil se v roce 1954 v Brně.  
 Vyškolil se na policejní akademii v Praze, titul JUDr. si dopřel v Bratislavě a titul Bc. na VUT v Brně.

1

2

similarity processing

[blog.orwen.org/textmatcher](http://blog.orwen.org/textmatcher)

Mezi lidmi, jimž byl zahlazen přestupek, je třeba...

Podle Příkryla obžaloba obsahuje údaje "vytržené z kontextu, které jsou nesoudným slepencem několika skutkových dějů". Příkryl uvedl, že vznikla nepravdivá fáma, že zahlazováním přestupků pomáhal známým. Soudu při pondělním zahájení procesu řekl, že mnohé z těchto lidí vůbec nezná. Ti se podle strážníků v terénu dopustili přestupků kvůli parkování na chodníku, na místech se zákazem stání, nebo v těsné blízkosti křižovatek. "Každý z těchto lidí (31 řidičů) si přišel na městskou policii stěžovat na to, že jim hrozí pokuta. Policejní orgán se při vyšetřování vůbec nezabýval podstatou věci," řekl u soudu Příkryl. Mezi lidmi, jimž byl zahlazen přestupek, by měli být podnikatelé, nebo dva komunální

Bývalý ředitel brněnských strážníků Jaroslav Příkryl napočtvrté dorazil k soudu, který posuzuje legálnost jeho pracovních postupů. Popřel, že by dopravní přestupky známých řešil pouhou domluvou namísto pokuty. Obžalován je z 31 takových skutků. Hrozí mu až rok vězení. Podle Příkryla obžaloba obsahuje údaje "vytržené z kontextu, které jsou nesoudným slepencem několika skutkových dějů". Bývalý šéf strážníků uvedl, že vznikla nepravdivá fáma, že zahlazováním přestupků pomáhal známým. Soudu při pondělním zahájení procesu řekl, že mnohé z těchto lidí vůbec nezná. Ti se podle strážníků v terénu dopustili přestupků kvůli parkování na chodníku, na místech se zákazem stání, nebo v těsné blízkosti křižovatek. "Každý z těchto lidí (31 řidičů) si přišel na městskou policii stěžovat na to, že jim hrozí pokuta. Policejní orgán se při vyšetřování vůbec nezabýval podstatou věci," řekl u soudu Příkryl. Mezi lidmi, jimž byl zahlazen přestupek, by měli být podnikatelé, nebo dva komunální

81% match

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

1. how A is similar to B  $\Leftrightarrow s_{AB} = \text{sim}(A,B)$
2. how B is similar to A  $\Leftrightarrow s_{BA} = \text{sim}(B,A)$

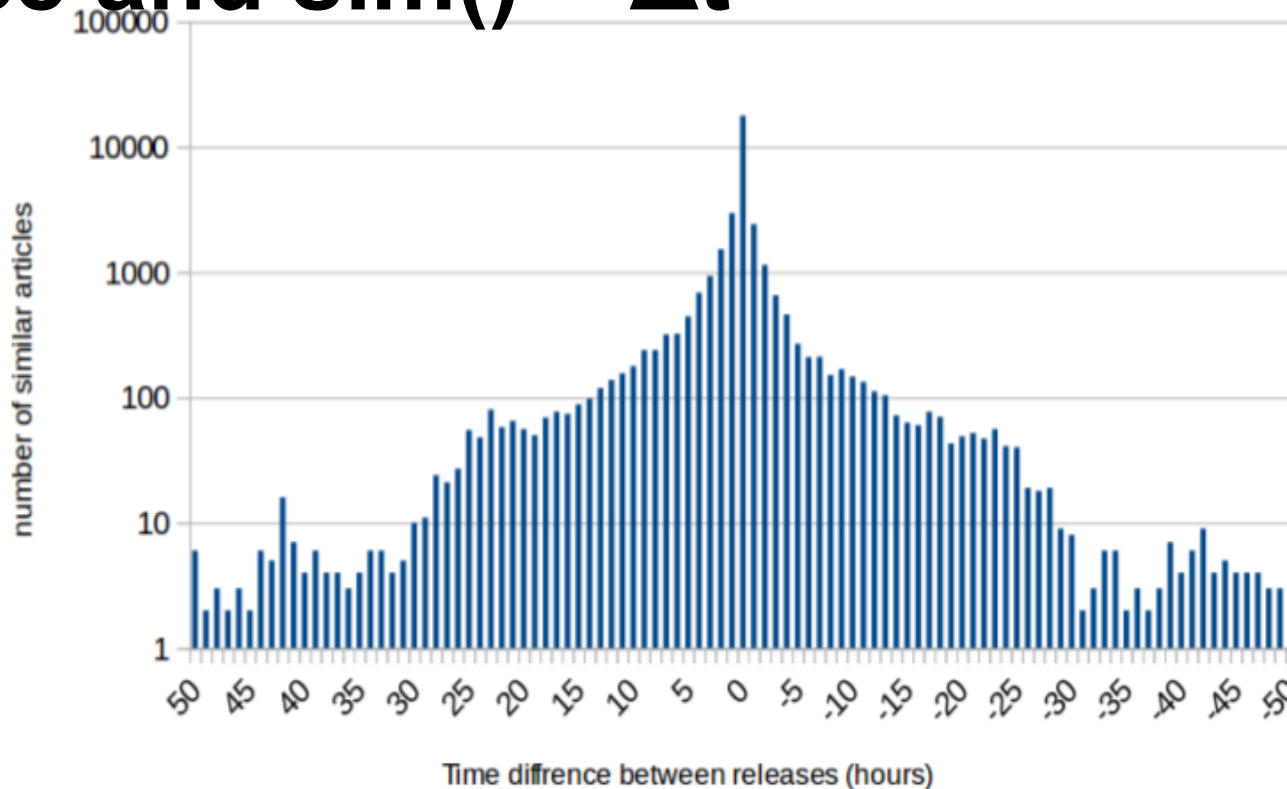
For example for texts A = "aaabbb" and B = "aaa", the results are

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

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

# Performance and $\text{sim}() \sim \Delta t$

- 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).*



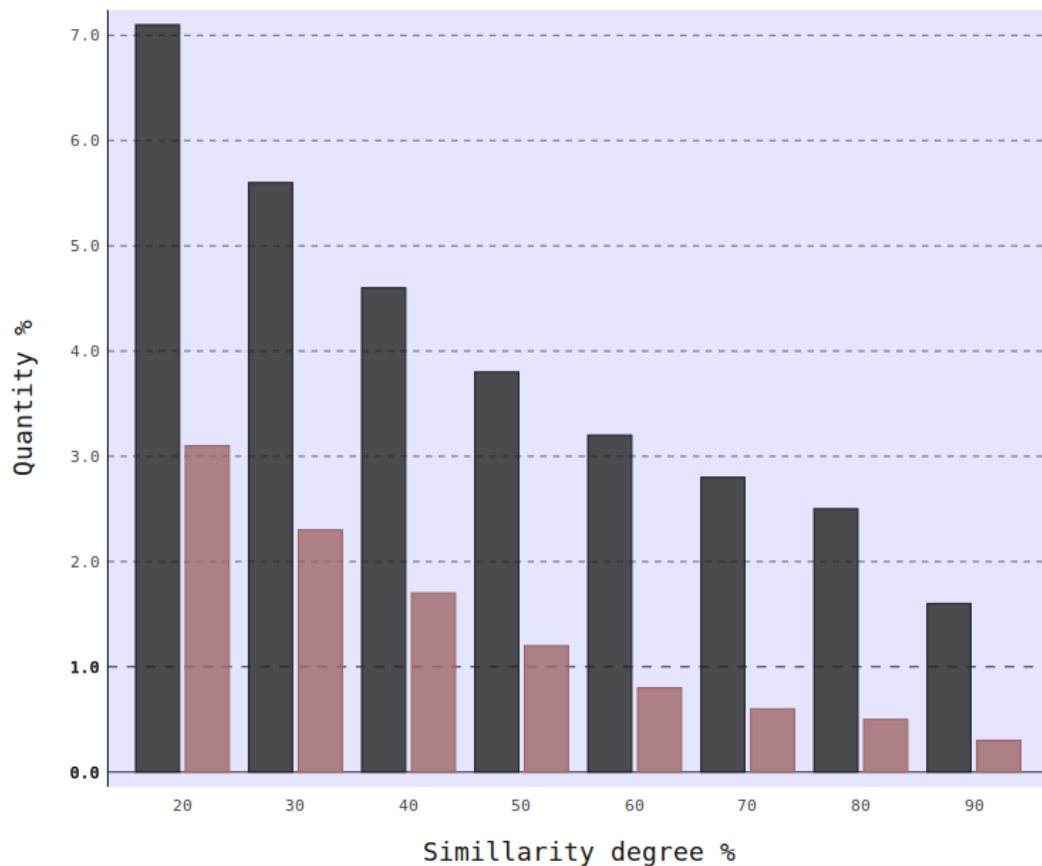
# Straight to results (regional media)

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**

■ All  
■ Without ČTK



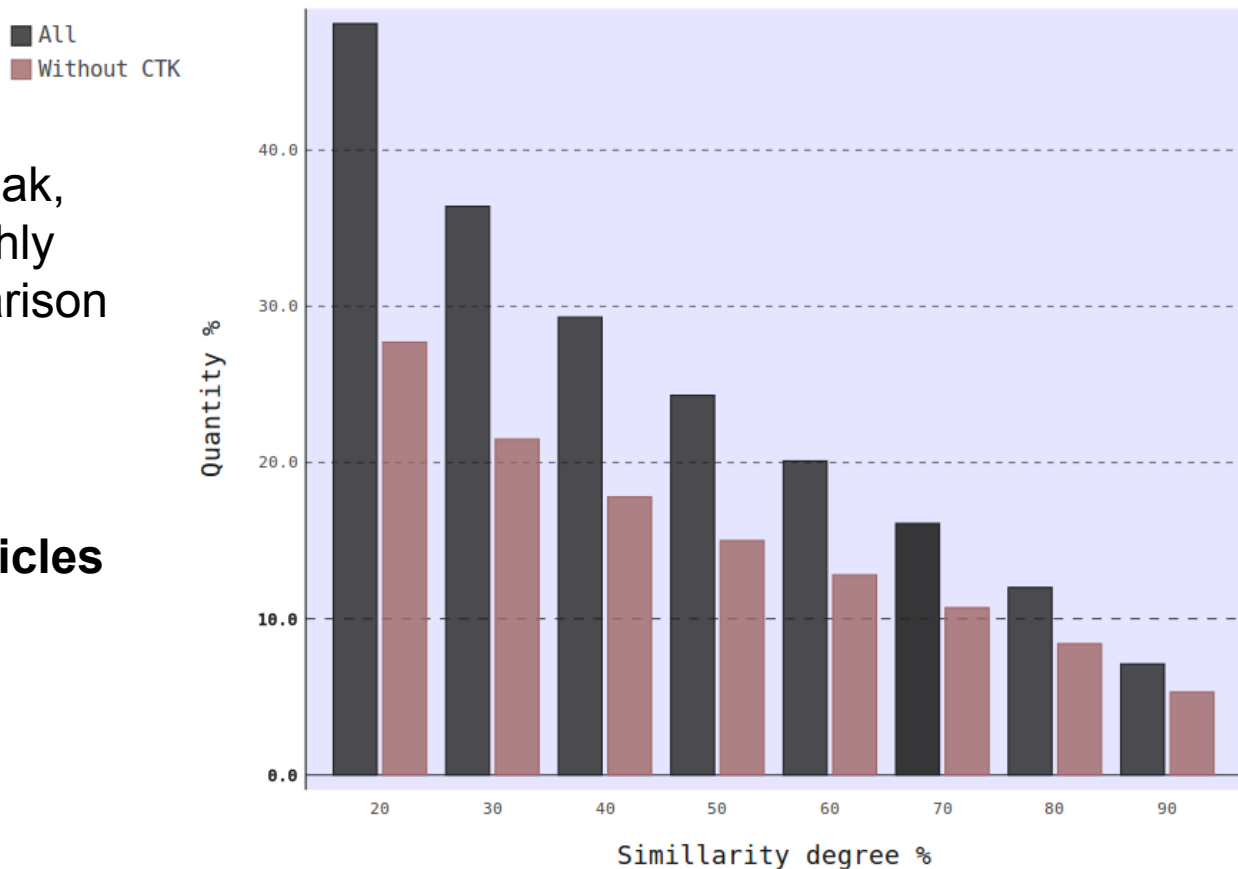
# Straight to results (national media)

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

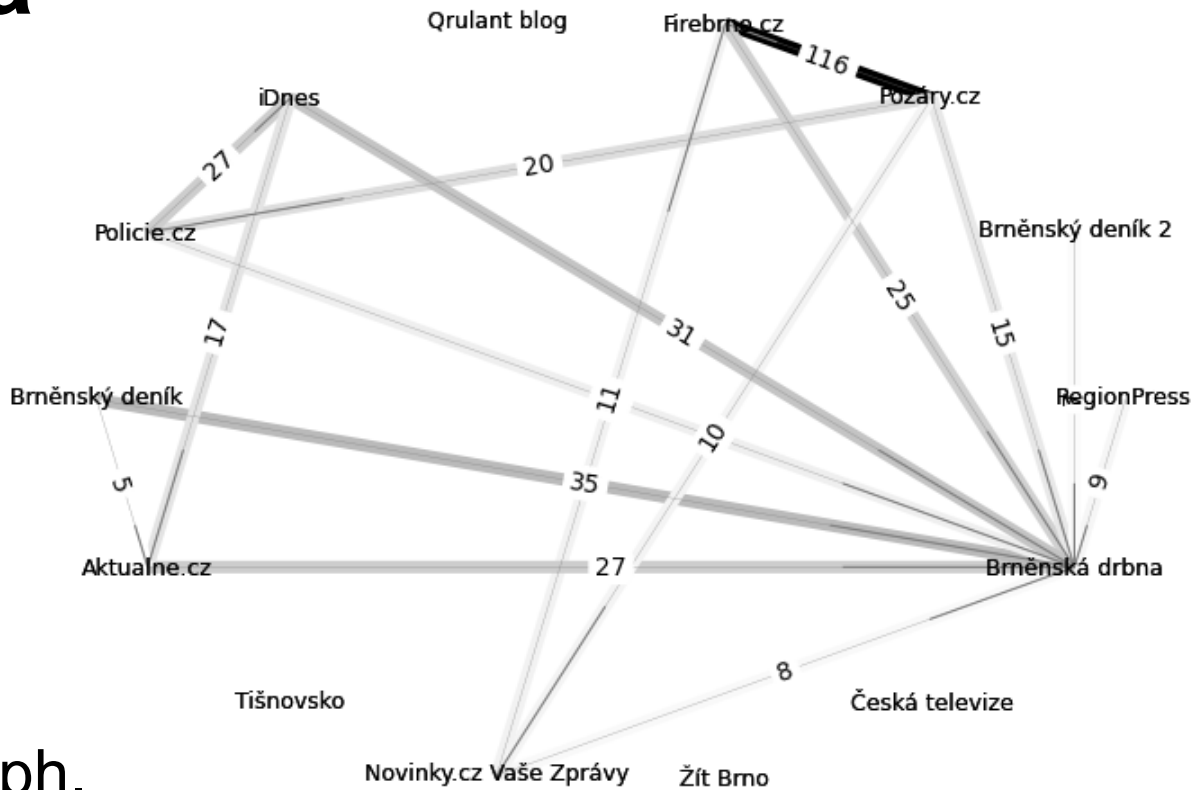
**~ 20% of similar articles**

**~ 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](http://www.orwen.org)