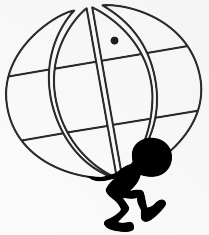
STU FIIT



PeWe@FIIT
personalized web group

# Hľadanie a získavanie metadát z webových sídiel

Bc. Milan Lučanský
Ing. Marián Šimko, PhD.

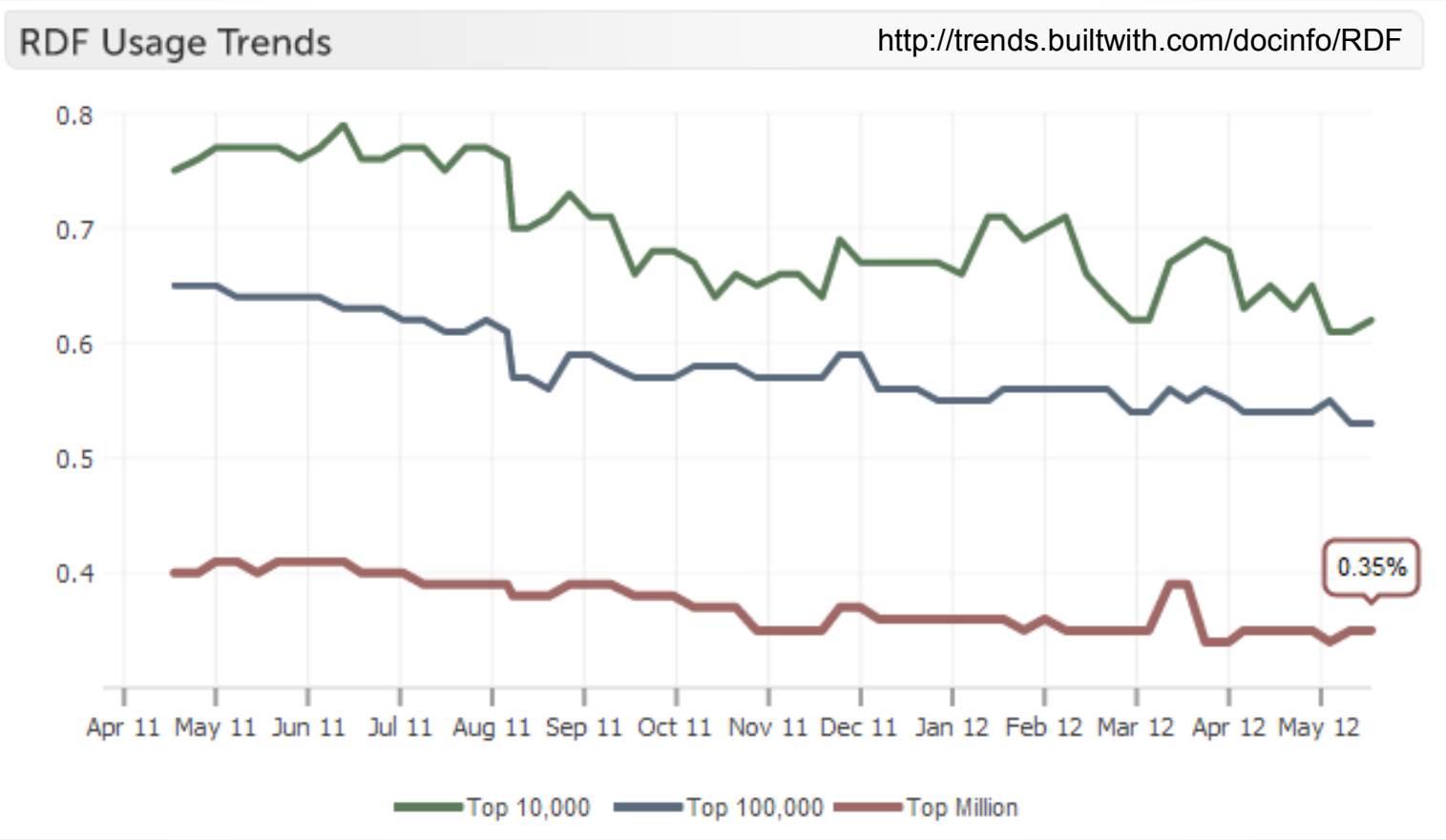
7. jún 2012

# Definovanie problému

Veľa dokumentov a málo sémantiky.

- 556 mil. registrovaných domén v decembri 2011 [5].

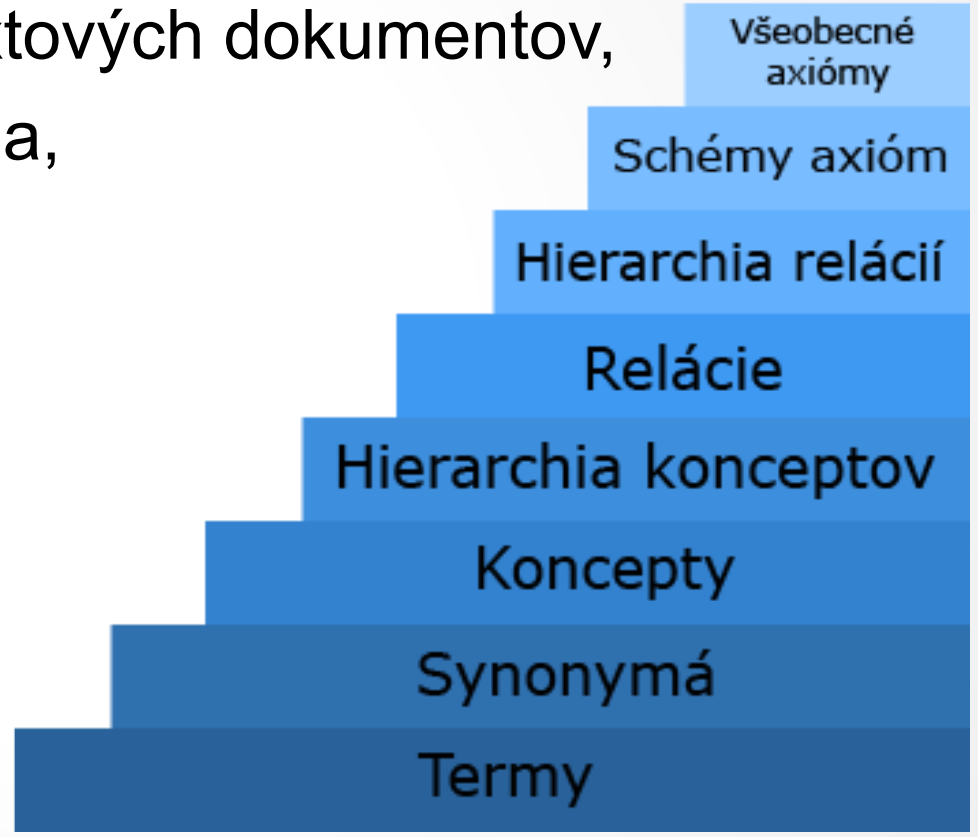
Existujúce technológie nie sú rozšírené.



# Definovanie problému

Základom pokročilého spracovania dokumentov sú termy:

- vytváranie ontológie z textových dokumentov,
- klasifikácia a kategorizácia,
- vyhľadávanie,
- indexovanie,
- a iné.




Ontologický koláč [1]

# Spracovanie dokumentov – ATR algoritmy

## Automatic Term Recognition algorithms

- Nástroj na spracovanie textových dokumentov.
- Veľmi dobré výsledky na špecializovaných korpusoch [2], napr. medicínske, biochemické články.
- Kvalita termov závisí aj od dĺžky dokumentov [2].

## Webové dokumenty majú špecifiká

- všeobecná doména,
- formátovanie dokumentov – štruktúra, **štýly**,
- prepojenie pomocou odkazov.

# Hypotéza

*ATR algoritmy v kombinácii s využitím implicitnej sémantiky ukrytej vo vizuálnej informácii kaskádových štýlov dokážu extrahovať z webu lepšie kľúčové slová, ako keby boli použité samostatne.*

# Vymedzenie voči bakalárskej práci

## Bakalárska práca

- podobný problém,
- využitie vybraných HTML značiek,
- syntetický test na malej množine stránok.

## Diplomová práca

- inšpirácia v Bc. práci,
- využitie kaskádových štýlov,
- prepracovanejší návrh metódy,
- overenie na "divokom webe",
- vyhodnotenie experimentu za pomoci strojového učenia,
- sledovanie atribútov, ktoré majú vplyv na úspešnosť metódy.

# Návrh metódy

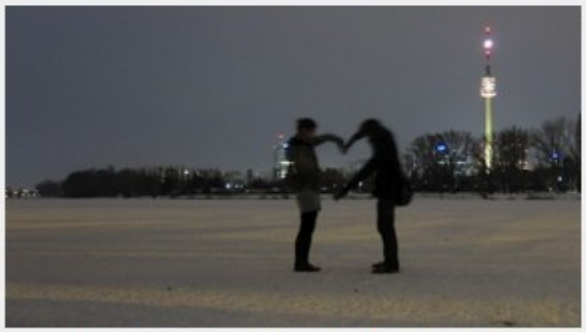
# Kaskádové štýly a ich potenciál

80,39 % stránok používa kaskádové štýly (2008) [4].

Lea Hajner grew up in Vienna, Austria with the beautiful Schönbrunn Palace on her doorstep and parents who dragged her to museums and plays. But it took a trip around the world to truly appreciate what her home town has to offer. Now Lea makes her living working for tripwolf.com, an online and mobile travel guide (see their 24 hours in Vienna). Check out her don't-miss picks for this central European hotspot. Follow Lea's adventures on Twitter **@vanilleah** and get great travel tips on the Trip Wolf blog.

**Vienna is My City**

**The first place I take a visitor from out of town is to** Stephansdom (Saint Stephen's Cathedral). From the top you have a great view of the city — and a lot of the best sights are in walking distance.



They heart Vienna. (Photo: Sebastian Fuchs)

**When I crave a** good cup of coffee **I go to** Café Hawelka, (Dorotheergasse 6, just off Graben). The interior, which was designed by a student of Adolf Loos, still looks the same as in the mid-1900s when actor Oskar Werner or artist Friedensreich Hunderwasser were regular guests. Their traditional coffee, "Wiener Melange," is the best in town and their pastries are heavenly delicious!

**To escape** my own four walls on Sunday afternoons **I head** into town for a leisurely stroll and some window shopping along the glamorous Kohlmarkt, followed by a visit to one of the museums' current exhibitions.

If you're like us here at Traveler, there was a moment — a single moment —...

**#FriFotos: Color Me _____**

For this week's #FriFotos* theme, COLORFUL, we chose this photo of a Kathakali performer applying his...

## Recent Digital Nomad Posts

- Sweet Sunday
- Welcome to Malawi
- Daze at Sea
- A Very Rare Bird
- Accessible

## Find Us on Facebook

**National Geographic Traveler** on Facebook

Like

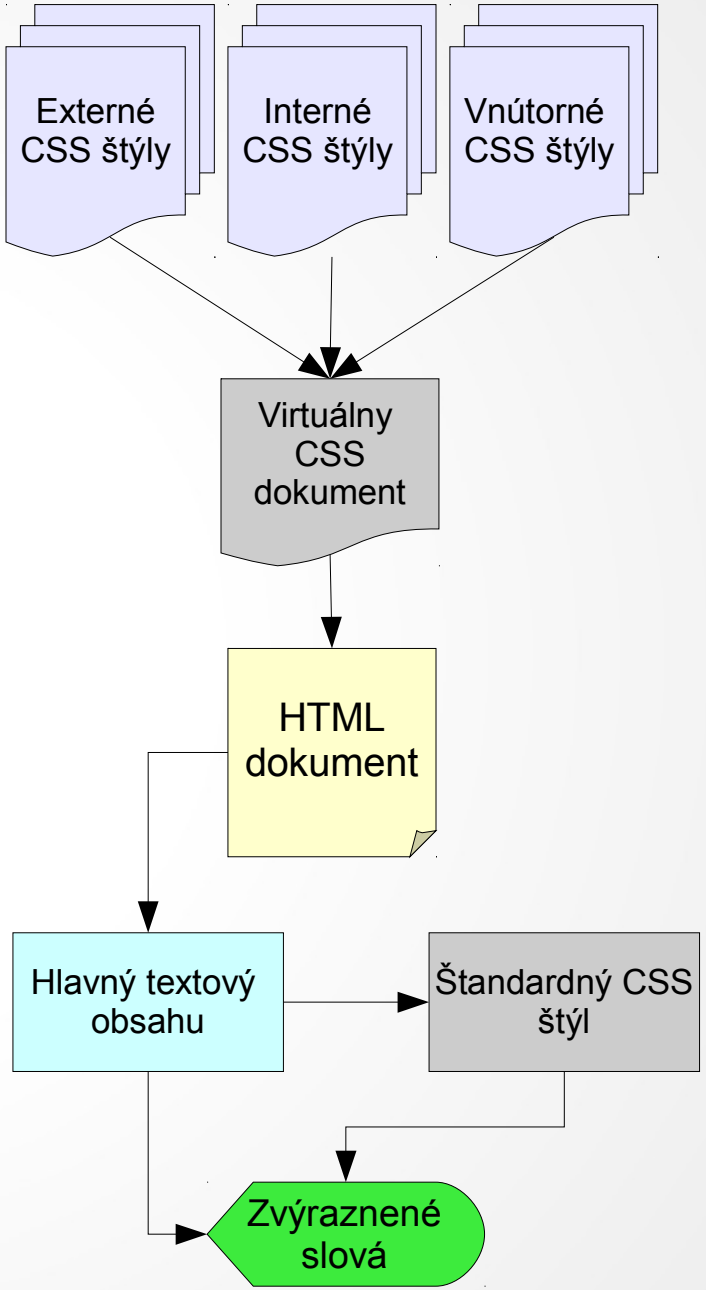
204,079 people like **National Geographic Traveler.**

Jettie Faeth Xinmu Haviez Shatzie

# Kaskádové štýly a ich spracovanie

Vybrané CSS atribúty pre formátovanie textu:

- font-size,
- font-style,
- font-variant,
- font-weight,
- color,
- background-color,
- text-decoration,
- text-transform.



# Model pre kombináciu ATR a CSS

Kombinácia pomocou úpravy váh:

$$w_F(k) = w'(k) + Css\,\mathrm{Re}\,l(k') \cdot p \qquad p = \frac{t}{|k'|}$$

$w_F(k)$ – výsledná váha kandidáta $k$

$k$ – kandidát získaný pomocou ATR

$w'(k)$ – váha kandidáta priradená ATR algoritmom

$CssRel(k')$ – koeficient

$k'$ – fráza, ktorá sa formátovaním odlišuje od štandardného štýlu

$t$ – počet spoločných slov pre $k$ a $k'$

$p$ – prienik fráz $k$ a $k'$

# Koeficient CssRel

Hodnota koeficientu závisí od dvoch faktorov

- počet slov,
- viditeľnosť text.

$$Css\,\mathrm{Re}\,l(k') = \left(1 + \frac{1}{\log(wc)}\right) \cdot vf(k')$$

*wc* – počet slov v HTML elemente, kde sa *k'* nachádza

*vf(k')* – viditeľnosť frázy *k'*

Predpoklady

- Čím menej slov, tým lepšie.
- Čím viditeľnejší text, tým lepšie.

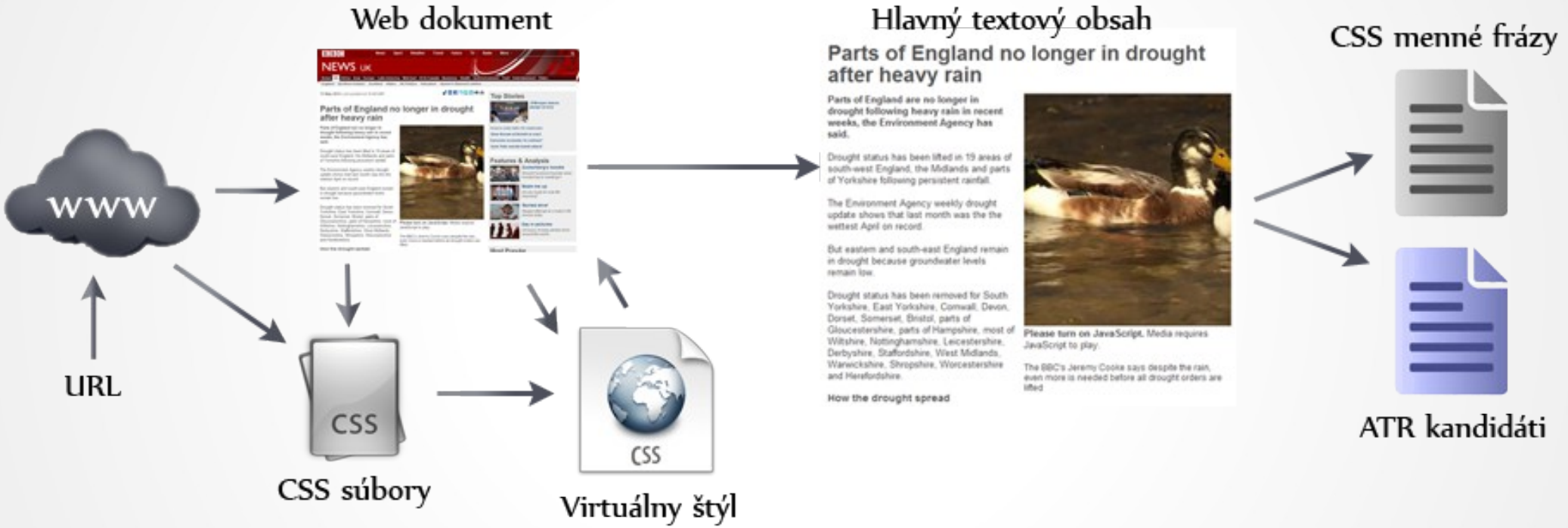
# Viditeľnosť textu

Miera, ako sa zvýraznený text odlišuje od majoritného textu.

$$vf(k) = 1 + fsz + fs + fw + fv + td + tt + lum$$

| Koeficient | Opis |
|---|---|
| *fsz* | koeficient pre font-size |
| *fs* | koeficient pre font-style |
| *fw* | koeficient pre font-weight |
| *fv* | koeficient pre font-variant |
| *td* | koeficient pre text-decoration |
| *tt* | koeficient pre text-transform |
| *lum* | „svietivosť“, závisí od fraby textu a pozadia |

# Abstraktný model prototypu



# Experimentálne overenie

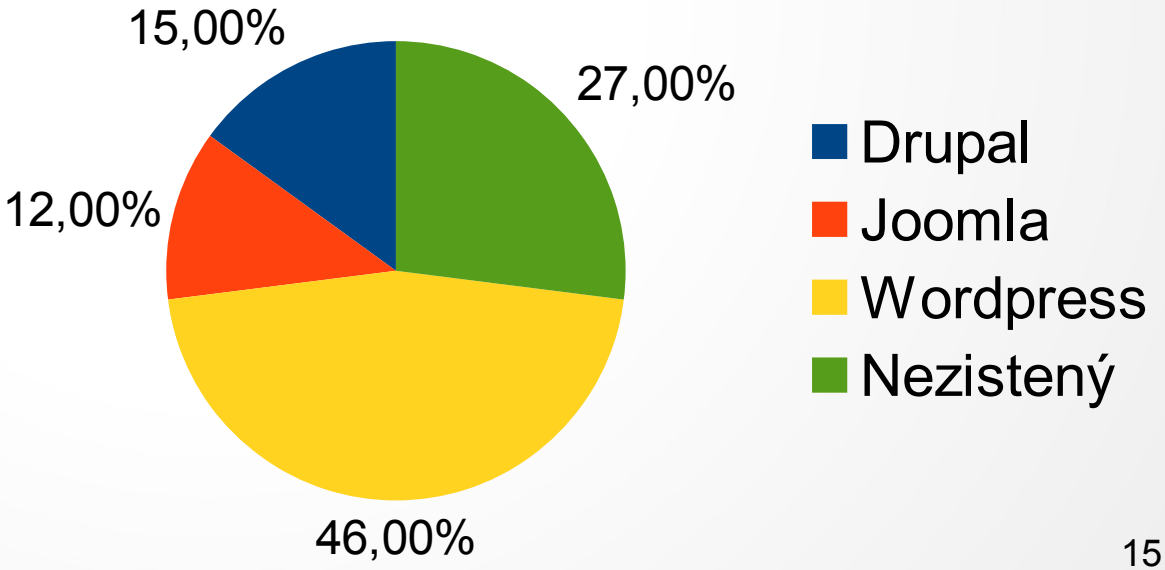
# Dátová vzorka

200 stránok z 13 rôznych portálov.

Blogy a novinové portály.

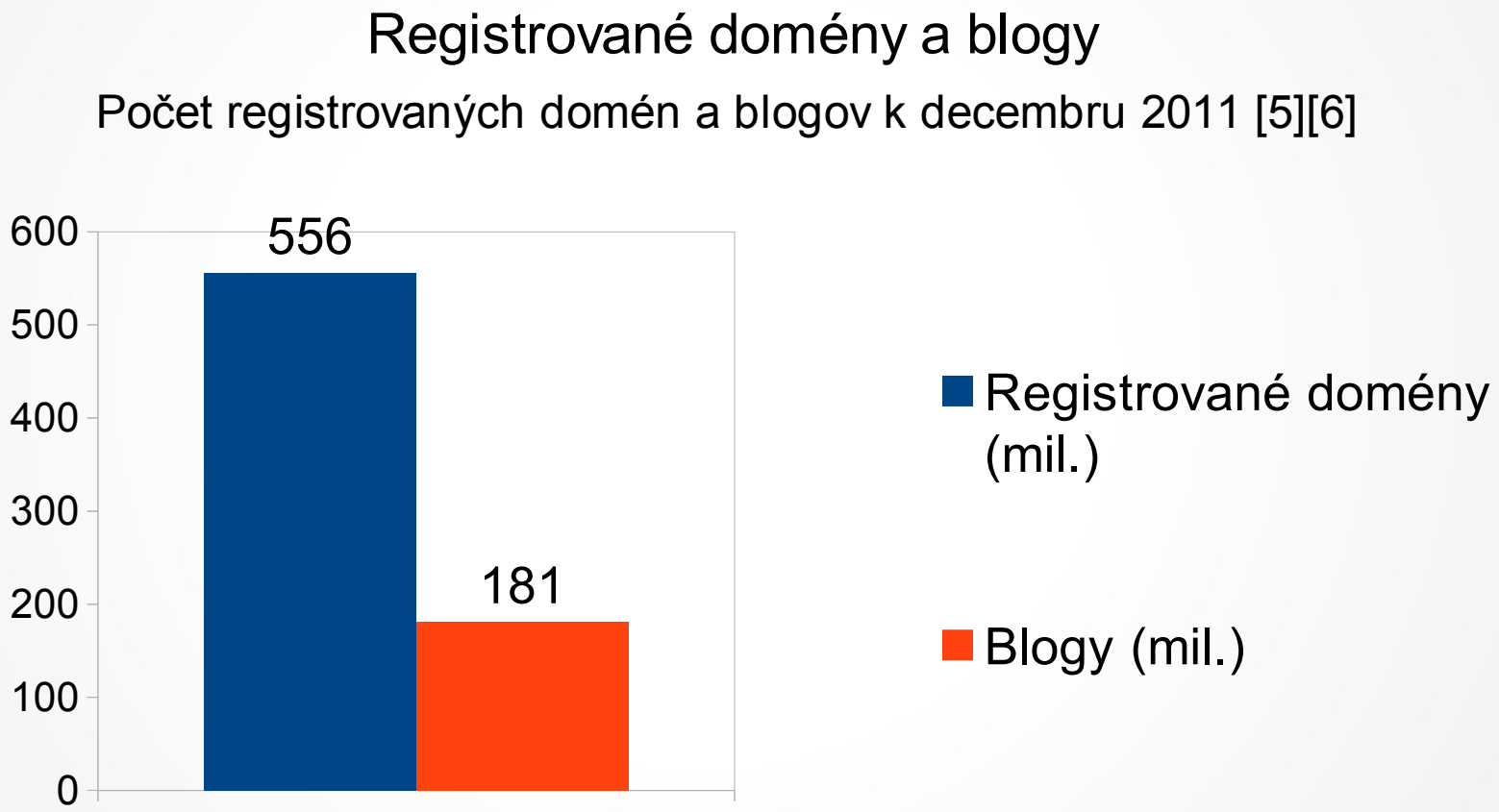
Redakčné systémy

- Drupal
- Joomla
- Wordpress
- Nezistený



# Blogy

Blogy tvoria veľký podiel stránok dnešného webu.



# Experiment

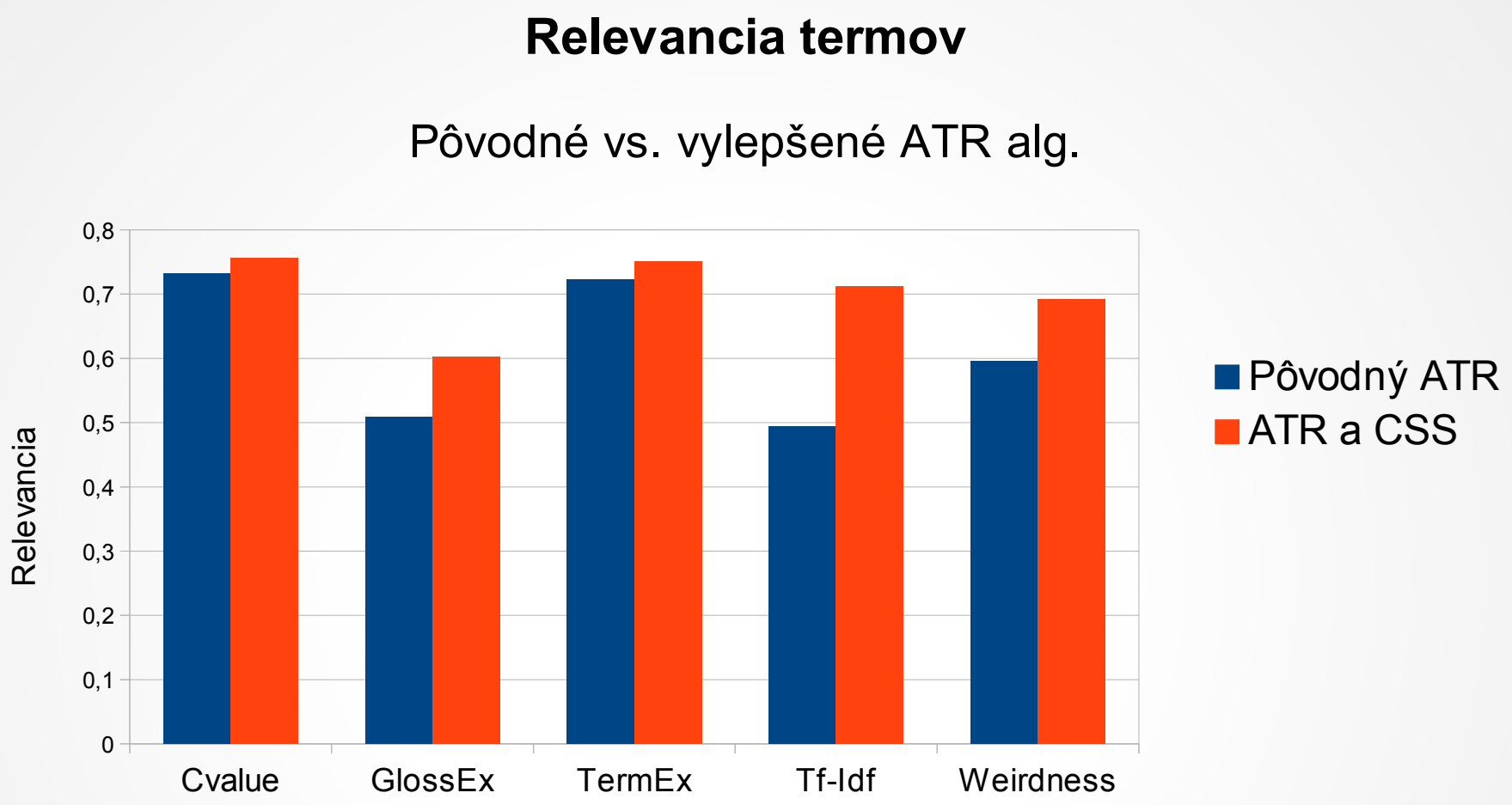
ATR algoritmy

- Cvalue, GlossEx, TermEx, Tf-Idf, Weirdness,
- 10 % najrelevantnejších termov.

55 používateľov hodnotilo získané termy

- spolu 21 394 slov a fráz,
- každá stránka mala aspoň 3 hodnotenia,
- hodnotenie prebiehalo cez webovú aplikáciu.

# Výsledky experimentu




**∆ zlepšenie (%)**

| Cvalue | GlossEx | TermEx | Tf-Idf | Weirdness |
|---|---|---|---|---|
| **2,46** | **9,34** | **2,73** | **21,84** | **9,62** |

# Zhoda „hodnotiteľov“

Metrika – Fleiss's kappa

"*Stupeň zhody pri klasifikácii očistený od zhody, ktorú hodnotitelia mohli dosiahnuť, ak by hodnotili náhodne.*"

Stupeň zhody: ~0,5 teda mierna zhoda.

| Kappa | Interpretation |
|---|---|
| < 0 | Poor agreement |
| 0.01 – 0.20 | Slight agreement |
| 0.21 – 0.40 | Fair agreement |
| 0.41 – 0.60 | Moderate agreement |
| 0.61 – 0.80 | Substantial agreement |
| 0.81 – 1.00 | Almost perfect agreement |

Interpretácia podľa [3].

# Klasifikácia pomocou C4.5

Klasifikovali sme CSS frázy na základe hodnotení používateľov.

Nástroj Weka v3.6

Rozhodovacie stromy C4.5 algoritmus.

- Vyhodnotenie pomocou 10 – násobnej krížovej validácie.

Výsledná presnosť klasifikácie ~75,17 %

# Vyhodnotenie CSS atribútov

Atribúty sme vyhodnotili pomocou informačného zisku, tzn. ktoré boli najužitočnejšie pri klasifikácii.

Nástroj Weka v3.6

Atribúty s najväčším informačným ziskom:

- font-size
- „svietivost“ textu
- font-style

# Video prezentácia

# Zhodnotenie

- Navrhli sme metódu, ktorá dokázala **zlepšiť relevanciu** kľúčových slov pre **všetky použité ATR algoritmy**.
- Overenie metódy prebehlo na vzorke 200 dokumentov z „divokého webu“ s reálnymi používateľmi.
- Overili sme **použiteľnosť** metódy pre **veľkú množinu dokumentov** (blogy a novinové portály) súčasného webu.
- Úspešnosť klasifikácie CSS fráz ~75 %.
- Identifikácia najužitočnejších CSS atribútov.
- Príspevok na konferencii IIT.SRC-2012.
- Ambícia publikovať výsledky na medzinárodnom fóre (SOFSEM 2013).

# Literatúra


- [1] Cimiano, P.: Ontology Learning and Population from Text Algorithms, Evaluation and Applications. Springer. University of Karlsruhe, Germany. 2006.
- [2] Feldman, R., Sanger J.: Text Mining Preprocessing Techniques. In: The text mining handbook Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, 2006.
- [3] Landis, J. R. and Koch, G. G. (1977) "The measurement of observer agreement for categorical data" in Biometrics. Vol. 33, pp. 159–174.
- [4] Wilson, Brian. 2008. MAMA: Key findings. http://dev.opera.com/articles/view/mama-key-findings/
- [5] Netcraft. 2011. December 2011 Web Server Survey. http://news.netcraft.com/archives/2011/12/09/december-2011-web-server-survey.html
- [6] Nmincite. 2012. Buzz in the Blogosphere: Millions more bloggers and blog readers. http://www.nmincite.com/?p=6531