Fulltext pro MUNI

Jakub Černý, Ph.D.

MUNI Brno, 9.12.2009

Co dnes servírujeme?

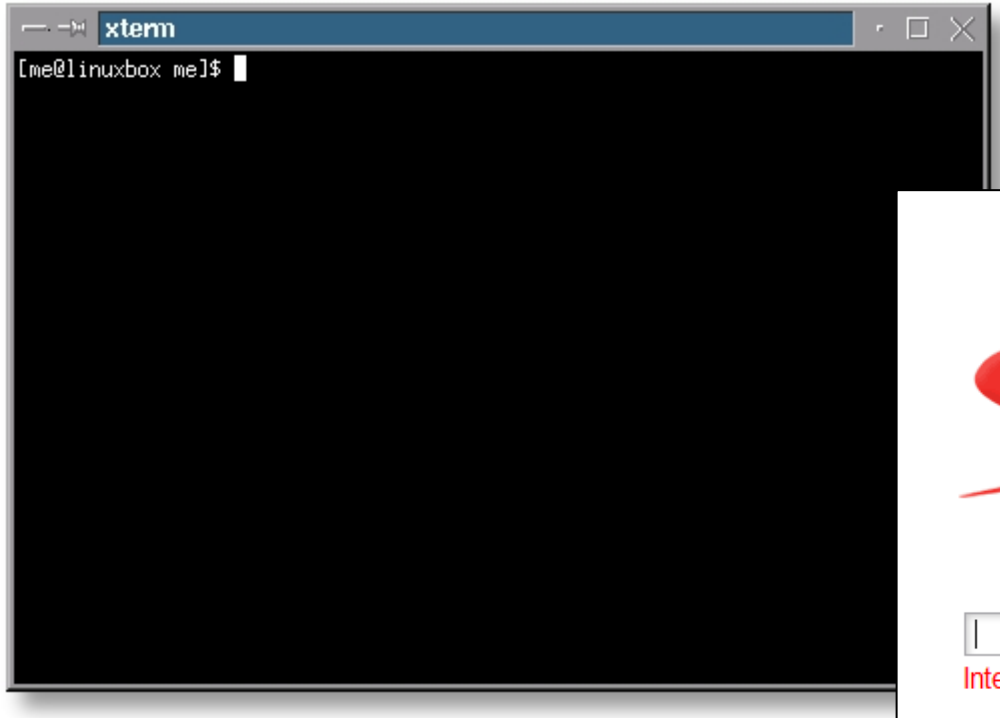
- Jak měřit kvalitu fulltextu?
  Jak se srovnávat s konkurencí?
  Jak nastavovat parametry algoritmu hledání?
- Jak funguje textový signál relevance?
- SEO pro běžné uživatele z pohledu lidí,
  co píší fulltext

Co byste chtěli slyšet vy?

Jak tečou uživatelé internetem?

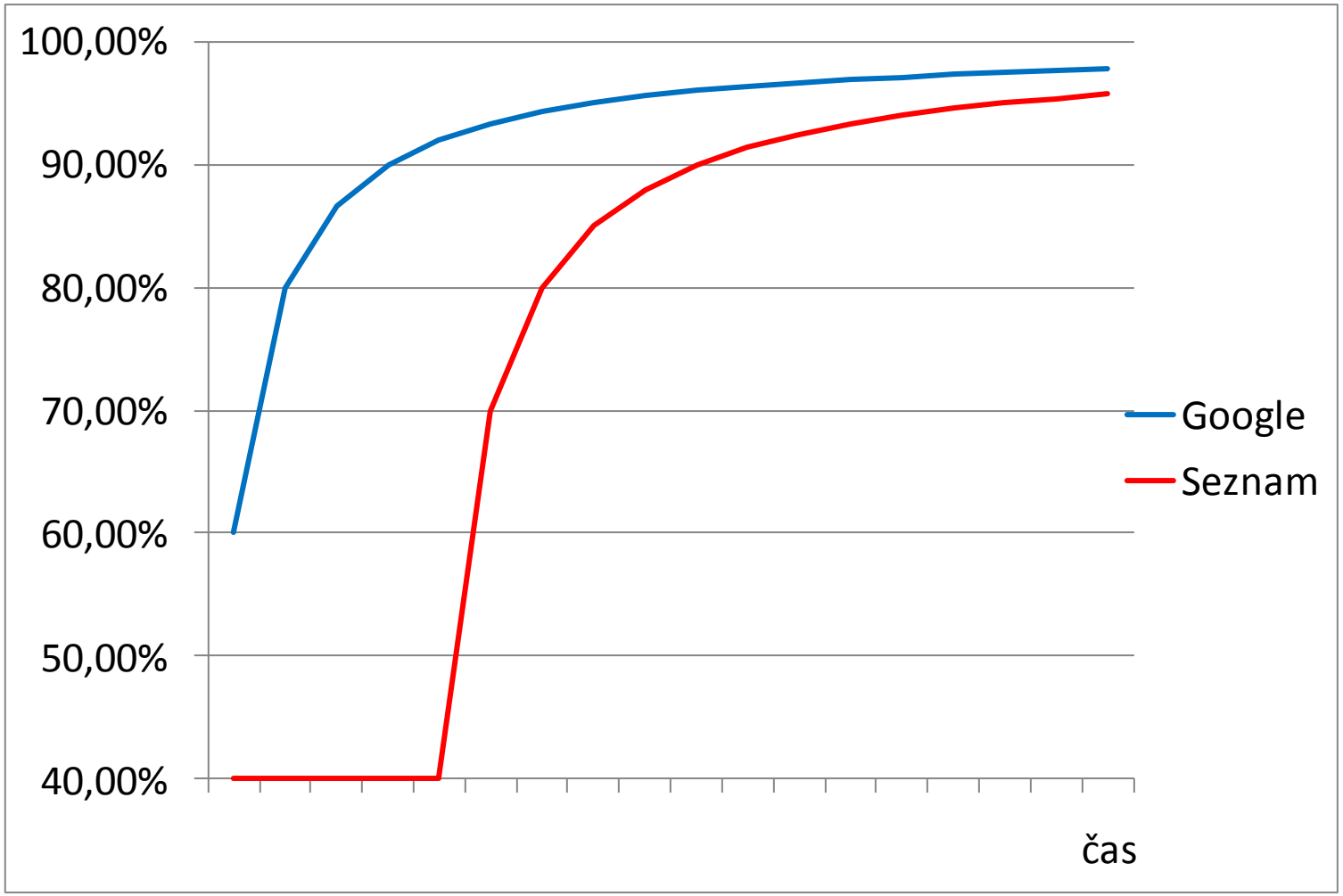
- Internet a odkazy jsou jako dálnice
  – co dělá běžný uživatel z pohledu mimozemšťana?
- Kde každý začíná?
  – homepage, fulltext, znám adresu
- Máte webový portál, kde sehnat návštěvníky?
  – postavit lepší přípojku z dálnice (SEO)
  – reklama

Znovu objevení kola



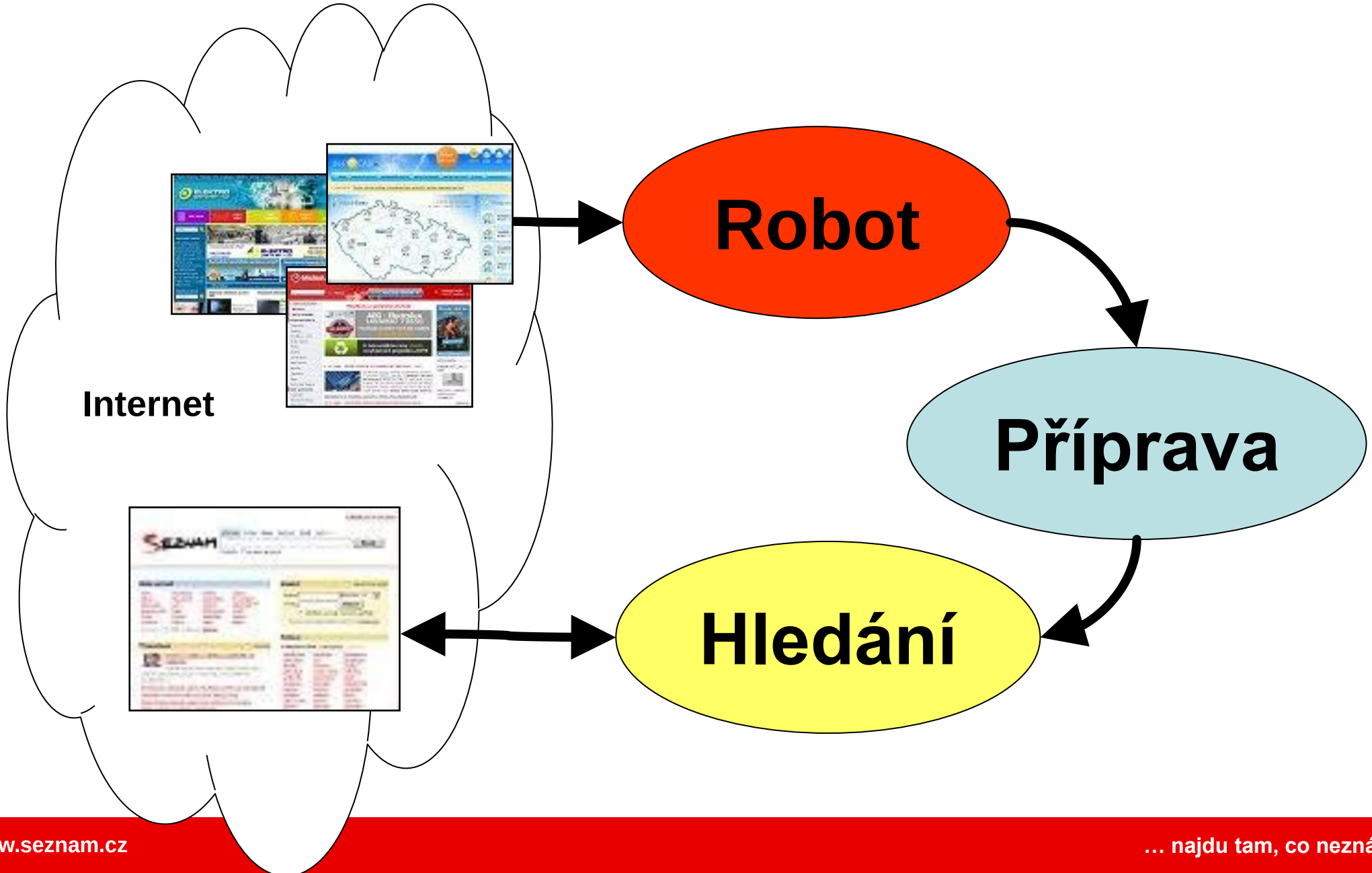
Do roka to bude řádka s URL v prohlížeči.

Seznam vs. Google



Proč Seznam vydrží?

Opakování: Jak funguje Fulltext



Jak měřit úspěch?

Proč? Co chceme?

- Měření kvality vyhledávačů
- Srovnání Seznamu s konkurencí
  - Kdo je lepší?
  - Na kterých kategoriích?
  - Na kterých dotazech?
  - Jak popsat skupinu dotazů, kde se to děje?
- Dostaneme tip, co zlepšovat
- Měřitelnost toho, jak jsme se zlepšili (SMART)

Otázka pro vás:

Jak měřit kvalitu výsledků fulltextového hledání?

- Čistě pořadí výsledků, ne rychlost hledání, či kvalitu webovky, snippetů

Kalibrace

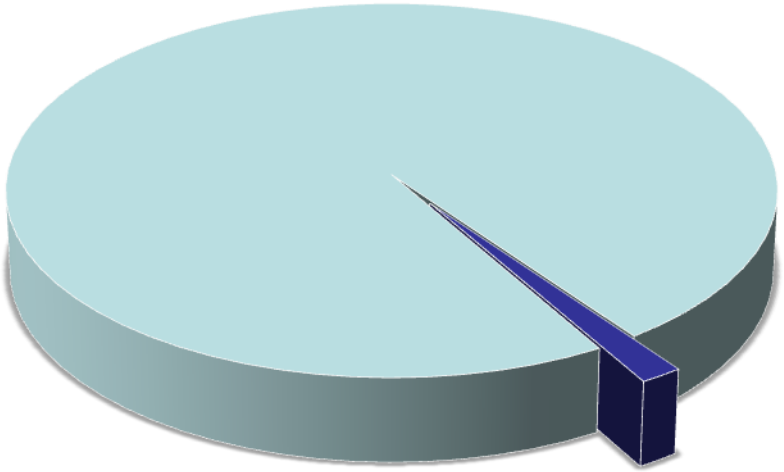
- Vital
- Usefull
- Relevant
- Non-relevant
- Off-topic

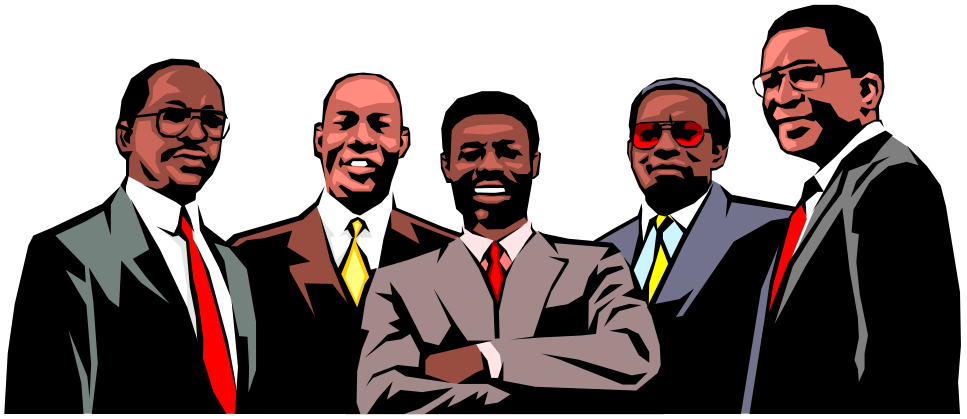
Kalibrace

Vital	**(navigační výsledek)** Dotaz má jasnou interpretaci a stránka je oficiální stránkou (jedinečnost). q=youtube … youtube.cz
Usefull	**(užitečný výsledek)** Stránka je hodně uspokojující, vyčerpávající výklad, vysoká kvalita,důvěryhodný zdroj. q=houby … atlashub.cz
Relevant	**(dobrý výsledek)** q=harry potter … knihy.cz/prodej/harry-potter
Non-relevant	**(blbý výsledek)** Sice je to k tématu, ale není užitečné (málo informací, staré info, příliš obecné). q=praha … zoopraha.cz
Off-topic	**(výsledek mimo mísu)** Výsledek obsahuje hledaná slova, ale tématicky je mimo. q=houby … „je to na houby“

Kalibrace

Výběr dotazů





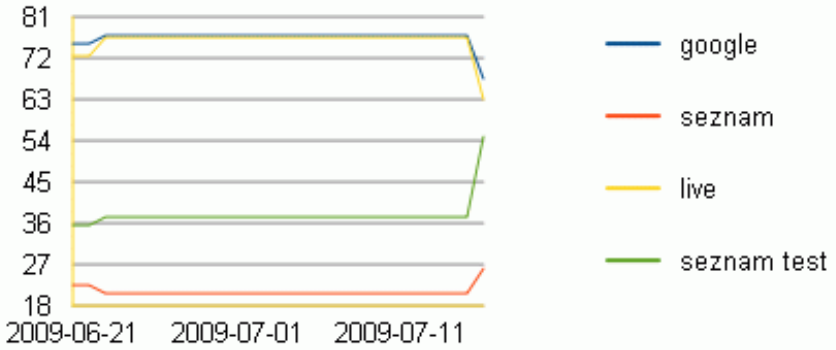
Sociodemo kalibrátorů

- Porozumění dotazu
- Kvalifikace pro zhodnocení kvality
- Muži vs. ženy (fotbal x parfémy)
- Puberťáci vs. důchodci (q=hudba)

Kvalita dotazu: bazén podolí

Graf kvality | Graf spolehlivosti | Graf fitness kvality

Celé období | **Měsíc** | Týden



Fulltext	Kvalita	Spolehlivost
seznam	26%	82%
seznam test	55%	82%
google	68%	59%
live	63%	87%

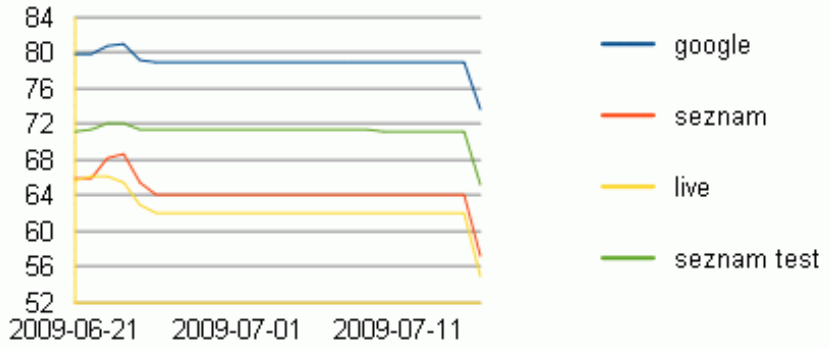
Výpis výsledků

	Pořadí	URL	Dotaz	Box	Akce
☐	1	http://www.pspodoli.cz	bazén podolí	vital	
☐	2	http://www.pspodoli.cz/zarizeni.htm	bazén podolí	useful	
☐	3	http://www.bazenpodoli.cz/bazeny-podoli	bazén podolí	useful	
☐	4	http://cs.wikipedia.org/wiki/Plaveck%C3%BD_stadion_Podol%C3%AD	bazén podolí	useful	
☐	5	http://expedice.rps.cz/lokality/12388-plavecky-stadion-podoli-bazen.html	bazén podolí	relevant	
☐	6	http://zuzikwww.blog.cz/0904/jeste-krasnejsi-nez-bazen-v-praze-4-podoli	bazén podolí	relevant	
☐	7	http://www.nelso.cz/cz/place/8597	bazén podolí	relevant	
☐	8	http://www.pragueout.cz/sport/bazeny/plaveckystadionpodoli	bazén podolí	relevant	
☐	9	http://www.vitalia.cz/katalog/bazeny/plavecky-stadion-podoli-cstv	bazén podolí	relevant	
☐	10	http://naturista.cz/drupal/?q=lokality/praha_podoli	bazén podolí	relevant	
☐	11	http://www.zaket.cz/8x4p_mista.php?akce=9	bazén podolí	relevant	
☐	12	http://sechtl-vosecek.ucw.cz/en/cml/35mm/film35mm1516.html	bazén podolí	non-relevant	
☐	13	http://6rbtata.com/view/hRhkHC2Lt_0/Hu%C4%8D%C3%ADnovi_-_baz%C3	bazén podolí	non-relevant	

Kvalita kategorie: Víceslovné

Graf kvality | Graf spolehlivosti | Graf fitness kvality

Celé období | **Měsíc** | Týden



Fulltext	Kvalita	Spolehlivost
google	74%	78%
seznam	57%	78%
live	55%	45%
seznam test	65%	55%

Výpis

	Pořadí	Název	Úplnost	Seznam		Google			Live			Seznam test			Akce
				Q	Sp	Q	Sp	+/-	Q	Sp	+/-	Q	Sp	+/-	
☐	1	bazén podolí	100%	26%	82%	67%	59%	+30.5	63%	86%	+33.4	55%	82%	+26.1	
☐	2	karnevalové masky a kostýmy	100%	66%	93%	73%	87%	+4.6	55%	80%	-10.9	100%	0%	-72.3	
☐	3	over ball žlutý	100%	50%	67%	81%	69%	+24.4	100%	0%	-54.4	84%	43%	+17.4	
☐	4	nokia servis	100%	53%	93%	72%	92%	+15.5	54%	98%	+1.7	100%	0%	-61.5	
☐	5	výpočet čisté mzdy	100%	78%	86%	92%	85%	+10.5	100%	0%	-80.0	72%	82%	-5.4	
☐	6	levandule wiki	100%	33%	58%	99%	64%	+50.6	100%	0%	-38.7	29%	62%	-2.8	
☐	7	přesun brněnského nádraží	100%	39%	98%	66%	98%	+23.2	60%	98%	+18.6	100%	0%	-49.7	
☐	8	psí útulek	100%	67%	88%	66%	82%	-1.4	100%	0%	-71.9	62%	78%	-5.9	
☐	9	odstředivá síla vzorec	100%	27%	98%	60%	98%	+30.4	40%	98%	+13.3	100%	0%	-37.7	
☐	10	ústava české republiky	100%	43%	57%	87%	71%	+36.6	100%	0%	-46.7	75%	33%	+14.8	

☐ Označit vše Smazat označené Provést

Předchozí **1** 2 3 4 5 6 7 Následující (celkem 251 záznamů)

Srovnání výsledků fulltextů

Dotaz: avon

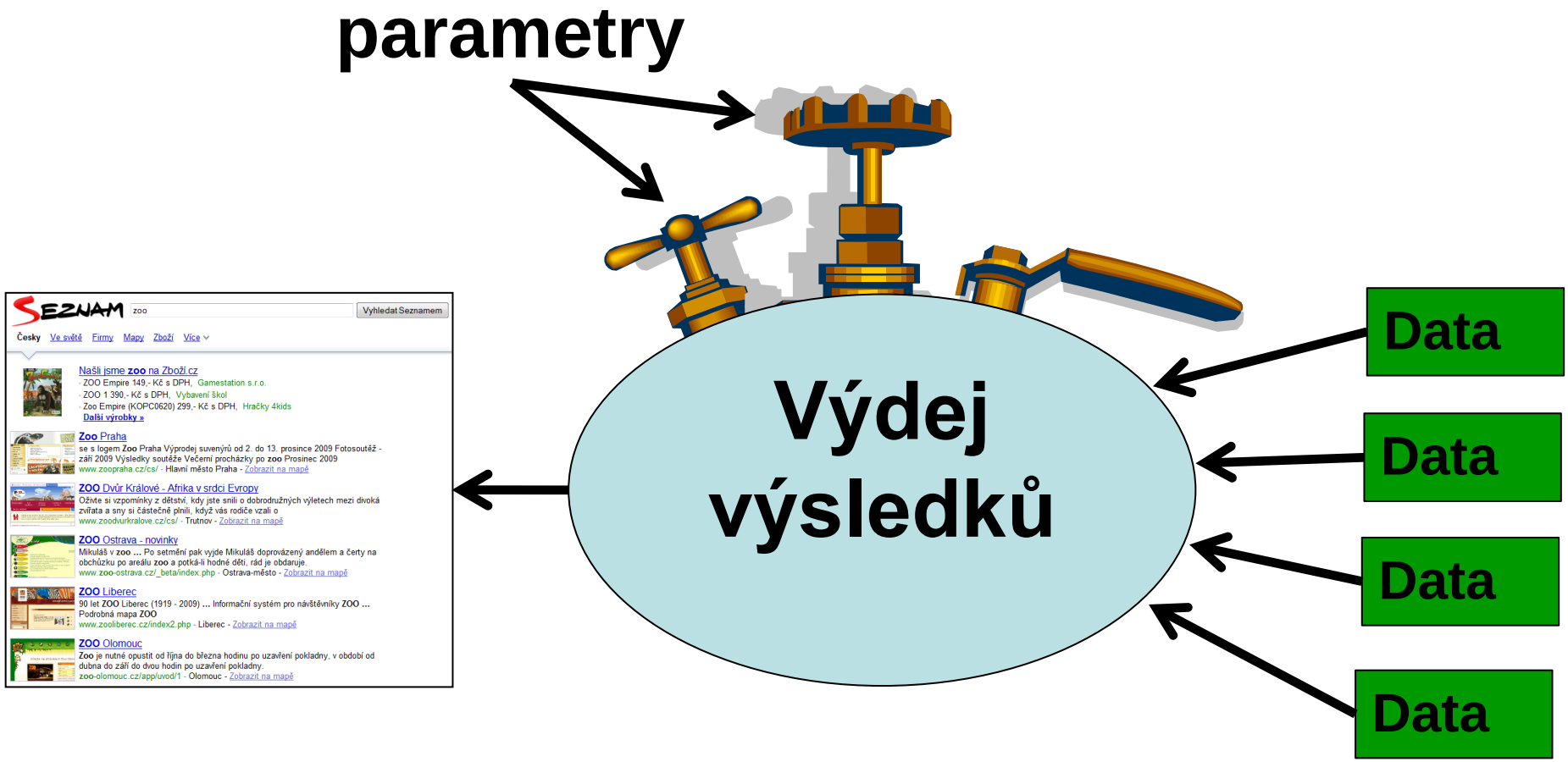
Pořadí	google	seznam	live	seznam test
Kvalita / Spolehlivost:	81.4% / 68.2%	84.3% / 91.8%	42.1% / 82.1%	84.3% / 91.8%
1	www.avoncosmetics.cz	www.avon-kosmetika.cz	www.avon.com	www.avon-kosmetika.cz
2	www.avoncosmetics.cz	www.avoncosmetics.cz	www.avoncosmetics.cz	www.avoncosmetics.cz
3	www.avon.cz	www.kosmetika-avon.cz	www.avon.com.au	www.kosmetika-avon.cz
4	www.avon-plus.cz	www.avon.cz	www.avon.ca	www.avon.cz
5	www.krasa.cz	www.avon-eshop.com	www.avon.cz	www.avon-eshop.com
6	www.krasa.cz	www.avon-kosmetika.eu	www.avon.org	www.avon-kosmetika.eu
7	www.avon-kosmetika.cz	www.avon-plus.cz	www.avon-plus.cz	www.avon-plus.cz
8	zena.centrum.cz	www.online-avon.cz	www.ar.avon.com	www.online-avon.cz
9	zena.centrum.cz	www.vuneprotebe.cz	www.pl.avon.com	www.vuneprotebe.cz
10	www.zdravaprsa.cz	www.avon-styl.cz	www.avon.ru	www.avon-styl.cz
11	www.avon-online.sk	www.avonlady-online.com	www.avon.co.nz	www.avonlady-online.com
12	vltava2000.cz	www.avon-eshop.eu	www.br.avon.com	www.avon-eshop.eu
13	www.firmy.cz	www.muj-avon.cz	www.avon.bg	www.muj-avon.cz
14	cs.wikipedia.org	www.avonland.cz	www.avon.it	www.avonland.cz
15	www.mammahelp.cz	www.kosmetika-avon.biz	www.avon.fi	www.kosmetika-avon.biz
16	www.estav.cz	www.krasa.cz	www.avon-kosmetika.cz	www.krasa.cz
17	www.gemoney.cz	www.avon-relax-centrum.com	www.avon.com.tr	www.avon-relax-centrum.com
18	tn.nova.cz	www.krasnadama.cz	www.avon.gen.tr	www.krasnadama.cz
19	avon.heureka.cz	www.avon-centrum.cz	www.avon.lt	www.avon-centrum.cz
20	www.lekarna.cz	avon-land.euweb.cz	www.cl.avon.com	avon-land.euweb.cz

Přínosy

- Možnost automatického nastavování parametrů fulltextu
- Rozhodování se na základě reálných dat
- Rychlejší vývoj a testování změn relevance fulltextu (prototypy úprav).
- Přenesení práce na externí kalibrátory
- Bonzování, co jsou nepovedené dotazy a jejich následné sledování -- víme na co se zaměřit
- Včas zjistíme, jak se zlepšila konkurence, co provedli -- můžeme je včas dohnat

Automatické ladění parametrů fulltextu

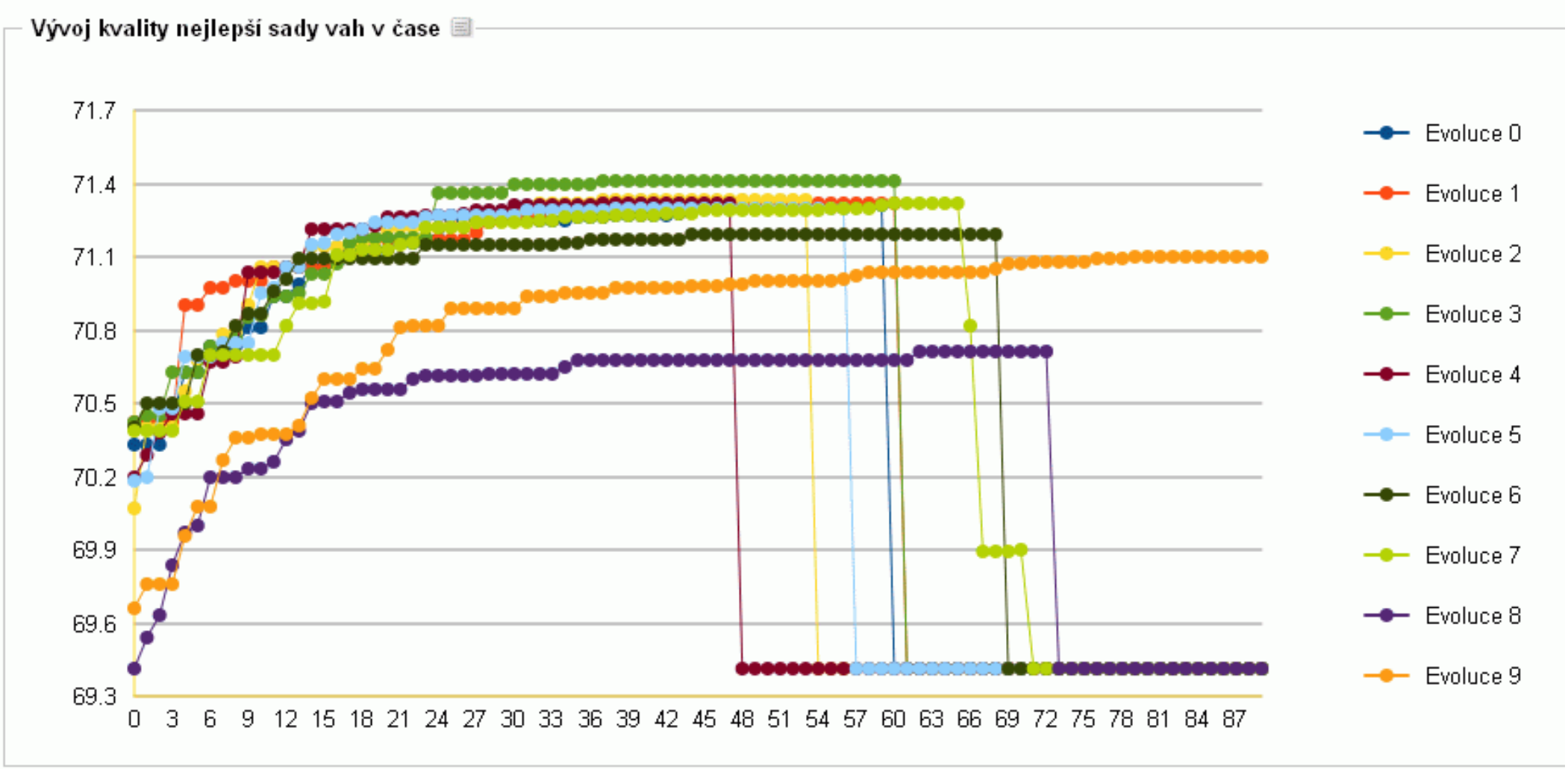
Jak nastavit parametry na optimum?



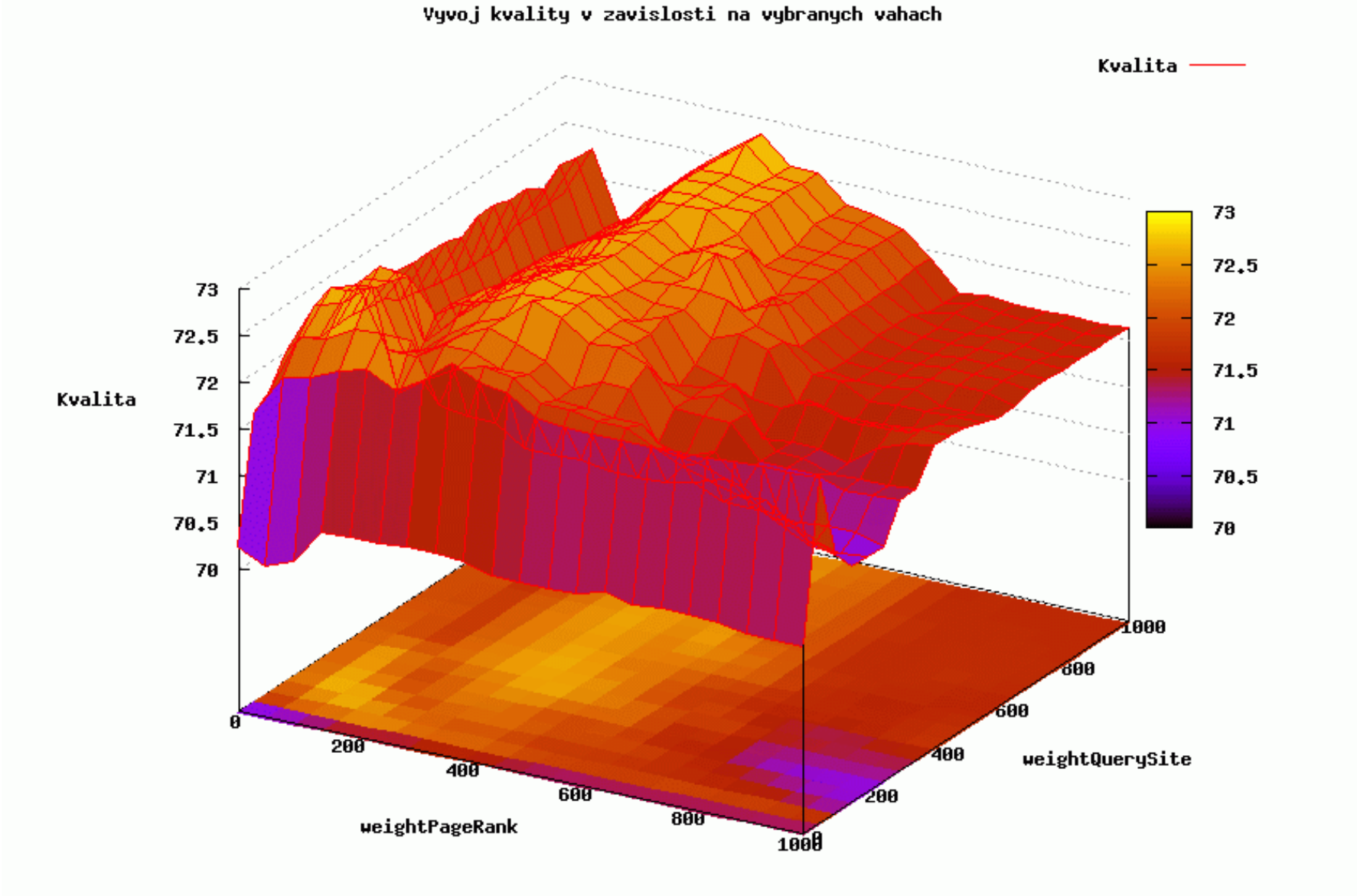
Historie ladění parametrů v Seznamu

- Od oka
  - nějak nastavit parametry a pak to nějak zkoumat
  - ve více lidech od oka, pak se hádáme
  - každý dodá dotazy, kde jsme lepší, horší, beze změny
- Využití kalibrací a měření kvality fulltextu
  - Ručně nastavovat, ale hned vidím kvalitu (i dotazy, na kterých to drhne)
- Automatické nastavování vah

Nastavovače vah



Nastavovače vah

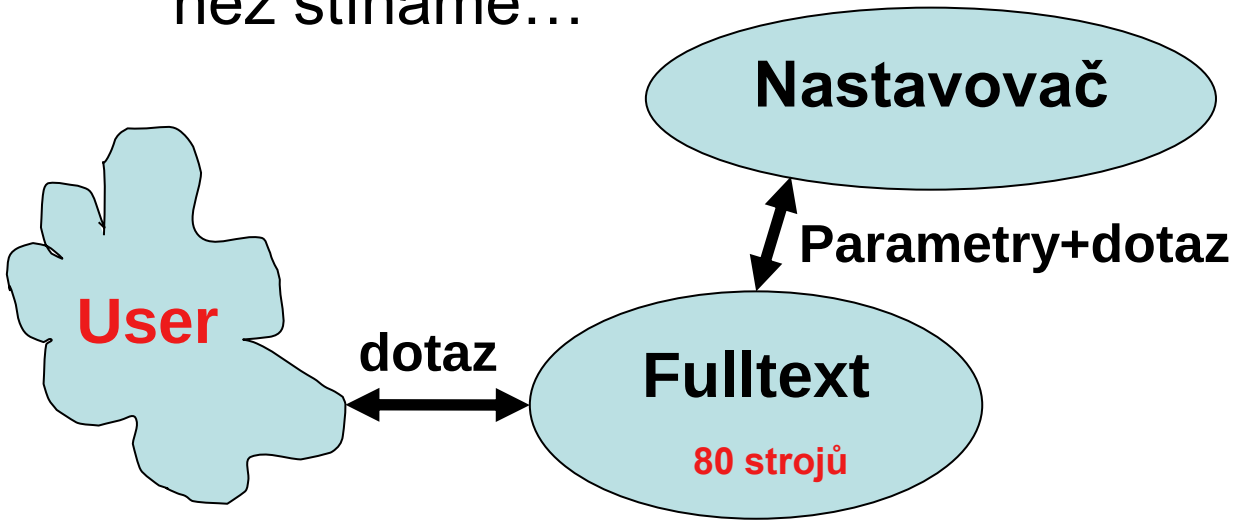


Otázka pro vás:

Jak odstranit bottle neck?

Když změníme parametry, tak se musíme pro všechny nakalibrované dotazy zeptat fulltextu na nové pořadí výsledků. Podle toho poznáme, jestli jsme si pomohli…

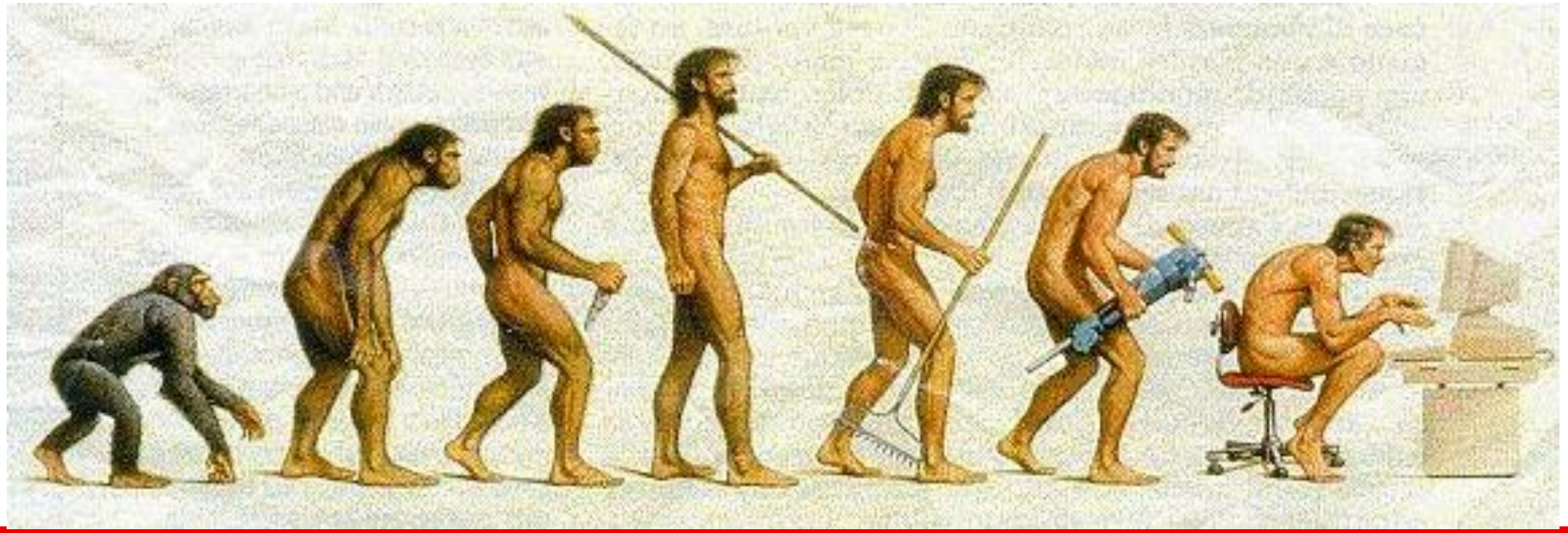
Potřebujeme se ptát mnohem více než stíháme…



Textový signál relevance

Textový signál relevance

- Je to názorná ukázka evoluce 1 signálu
- …jak probíhá výzkum
- Uslyšíte, jak funguje hledání v textech
  (to můžete na vašich stránkách ovlivnit)



Vývojové generace TXT signálu

- Jen slova z dotazu, přesná shoda tvaru
  - Jen 50% relevantních dokumentů obsahuje slova z dotazu.

Příklad: Dotaz „ČNB“, ale relevantní stránka obsahuje jen „oficiální úroková míra v České národní bance“.

Vývojové generace TXT signálu

- Přidání lemmatizace slov
- Různé váhy slov podle výskytiště (H1, URL, Title, odstavec, bold, …)

- Příklady vtipné lematizace:
  - Stát, ženu, lov lína, barum, jizdní rady, dog

Vývojové generace TXT signálu

- Různé váhy slov podle jejich korpusové četnosti
  - tf x idf
  - vynechávání slov

Příklad dotazů: Petr a Pavel, Jak se odstraňuje vosí hnízdo?

Otázka pro vás:

3-slovné dotazy: Máme zvýhodňovat výsledky, kde se slova z dotazu najdou blíže u sebe? Nebo je to jedno?

Vývojové generace TXT signálu

- Proximita a pořadí slov z dotazu

- Příklady:
  - Jakub Černý x Černý Jakub
  - Václav Klaus video
  - Já do lesa nepojedu, já do lesa nepůjdu
- Kolokace
  - Velký vůz, černý Petr, Česká republika

Vývojové generace TXT signálu

- Předzpracování dotazu
  - Poslechnu si uživatele a přeložím to do jazyka, ve kterém fulltext umí vyhledávat.
  - Nastavení proximity, …

- Příklady:
  - VŠE, MŽP, IE8 (ale i naopak)
  - Kdy vyhořelo Národní divadlo?
  - (běžné otázky jako na kamaráda)

Vývojové generace TXT signálu

- Doplňování slov odjinud
  - ze zpětných odkazů (bazén podolí)
  - anonymní termy
    - jméno, datum, místo, video
    - pro odpovědi na otázky: Kdo? Kdy? Kde?

- Příklady:
  - Václav Klaus video
  - Kdy vyhořelo Národní divadlo?

Další okolnosti kolem TXT signálu

- Body text extraction (BTE)
- Site-wide texty (SWT)
  - rozpoznání důležitosti slov podle vzhledu site
  - odstranění neopodstatněných nároků na důležitost
    - Všechny texty v H1 apod.

- Různé chování pro různé kategorie dotazů:
  - Navigační
  - Informační
  - Transakční

Další okolnosti kolem TXT signálu

- Desambigulace
  - Vyloučení nejednoznačnosti

  - Řekněte mi něco o německých tancích?
  - Hrách vs. (o počítačových) hrách

SEO

Jak to funguje ve fulltextu

Uvidíte, že SEOptimalizátoři někdy vaří z vody a tvrdí blbosti.
(ale hlavně že zákazník zaplatí).

Úkoly SEO:

- Pořadí výsledků
  - být v první 10ce výsledků
- Snippety
  - kvalitní popisek u výsledku
  - ovlivňuje to CTR výsledku
- Robot a rychlá indexace

Pozor na náklady!

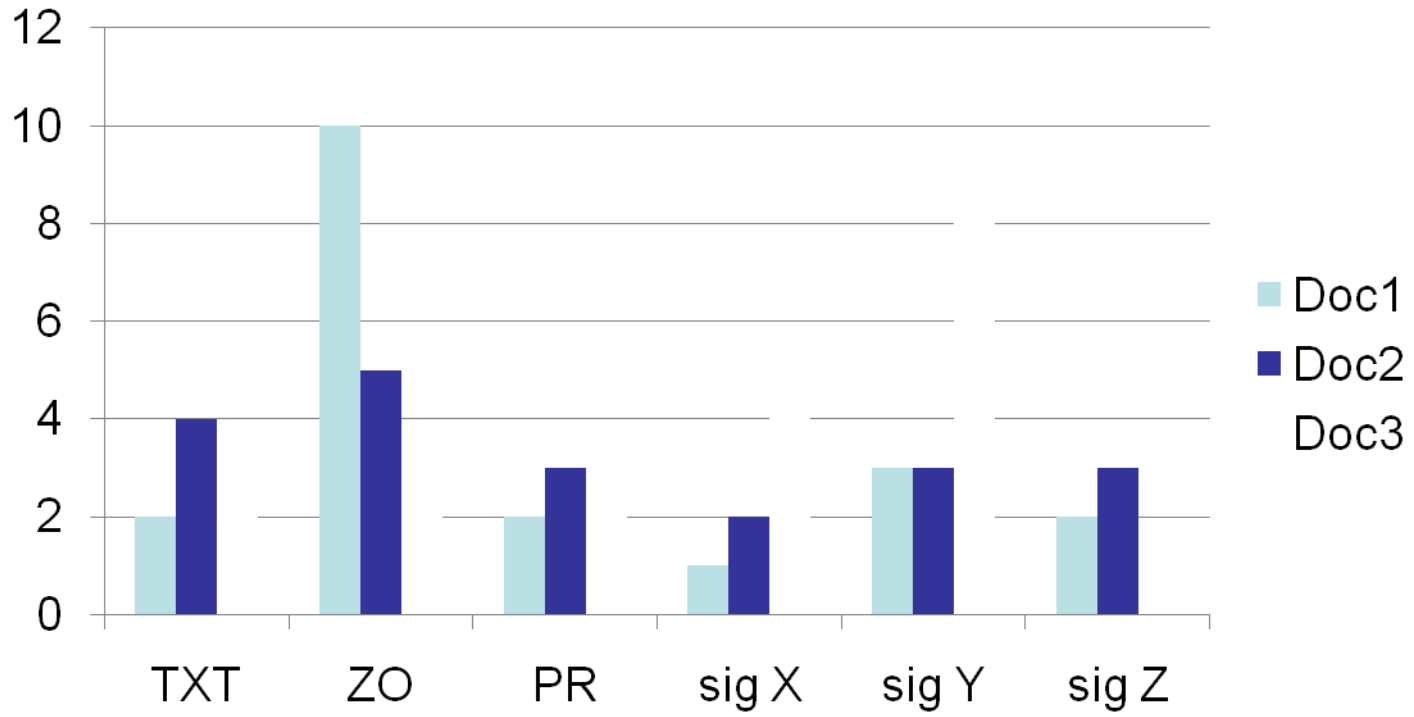
Krásný web neznamená nalezitelný web! Stroje mají jiné oči.

Proces hledání

- keywords
- Být v 1. desítce
- Podle čeho si uživatel vybere, na co klikne?
- Uživatel už je skoro lapen.
- Neztratí se?
- Neuteče nám?
- Hurá, dosáhli jsme cíle.

Pořadí výsledků

Mixování signálů relevance:



Kdo je lepší? Jak to míchat?

Signály relevance

	On page	Off page	User
obecné	Doména, historie, struktura stránky	Page Rank	???
tématické (k dotazu)	TXT	Zpětné odkazy	???

Úspěch záleží i na samotných uživatelých

- Sice je nalákáte na svůj web, ale když se jim tam nebude líbit, tak utečou.
  - Relevantní obsah
  - Jasná navigace
  - Výzva k akci
  - Identifikace webu: kontaktní údaje

Doporučení:

- *Steve Krug: Web* design: Nenuťte uživatele přemýšlet!

SEO - Rada

Všeho s mírou. Každá rada jde přehnat a zprasit.

Pak je to často naškodu.

Pište dobrý a užitečný web, vykašlete se na podvody. (Uznejte, že někdo může být lepší).

Rozdíl mezi SEO a praSEo.

SEO – On page faktory

- Volba klíčových slov
  - Nástroje pro analýzu klíčových slov (Sklik, AdWords, …)
  - Statistiky Seznamu
  - Long tail
- Copywriting

- Titulek, URL, nadpisy, alt
- Meta description, katalogový popisek

SEO – Off page faktory

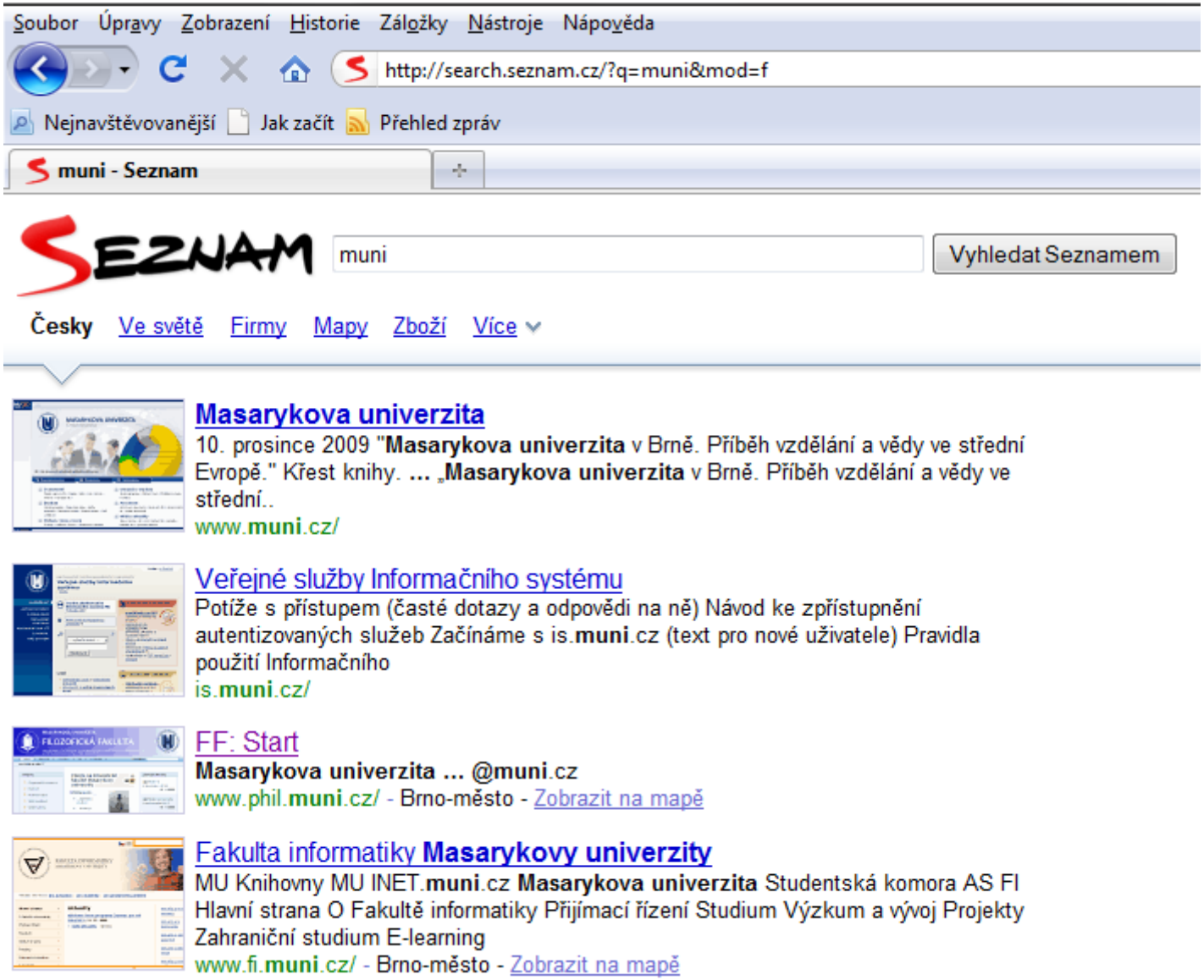
- Zpětné odkazy
  - Interní x externí
  - Důležitý je text odkazu, zohledňuje se i okolí
  - Tématická podobnost odkazované stránky

- Odkazová síť
  - Page rank

SEO - Snippety

- Ovlivňují proklikovost výsledku
- Jejich cíl:
  - Ve dvou větách ukázat, o čem je váš web
  - Ukázat v jakých souvislostech se nalezla slova z dotazu
- Kde se berou texty snippetu?
  - Title, URL (hesla popisující stránku)
  - Meta description (popis stránky ve 2 větách;ne na celé site stejné!)
  - Text stránky

Seo – snippety (příklad)



SEO - Robot

- 1. krok je, aby se vaše stránka dostala do indexu.
- Přidání URL do hledání
  (formulář na webu fulltextu)

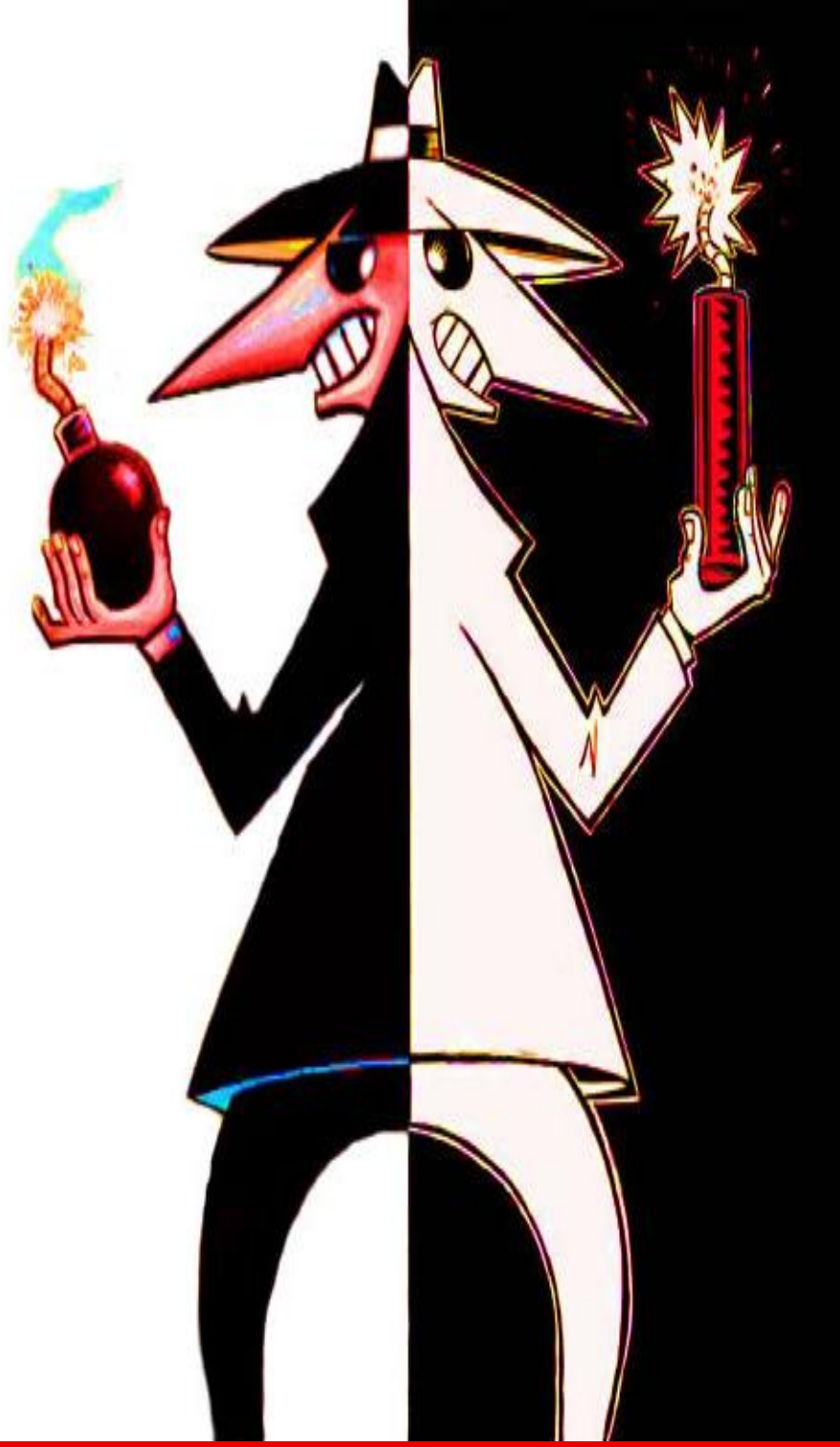
- Jak pomoci robotům?
  – Sitemap.xml
- Jak jim něco povolit a něco zakázat?
  – Robots.txt
- Redirekty, 404, …(dodržet jedinečnost URL)

Black hat SEO

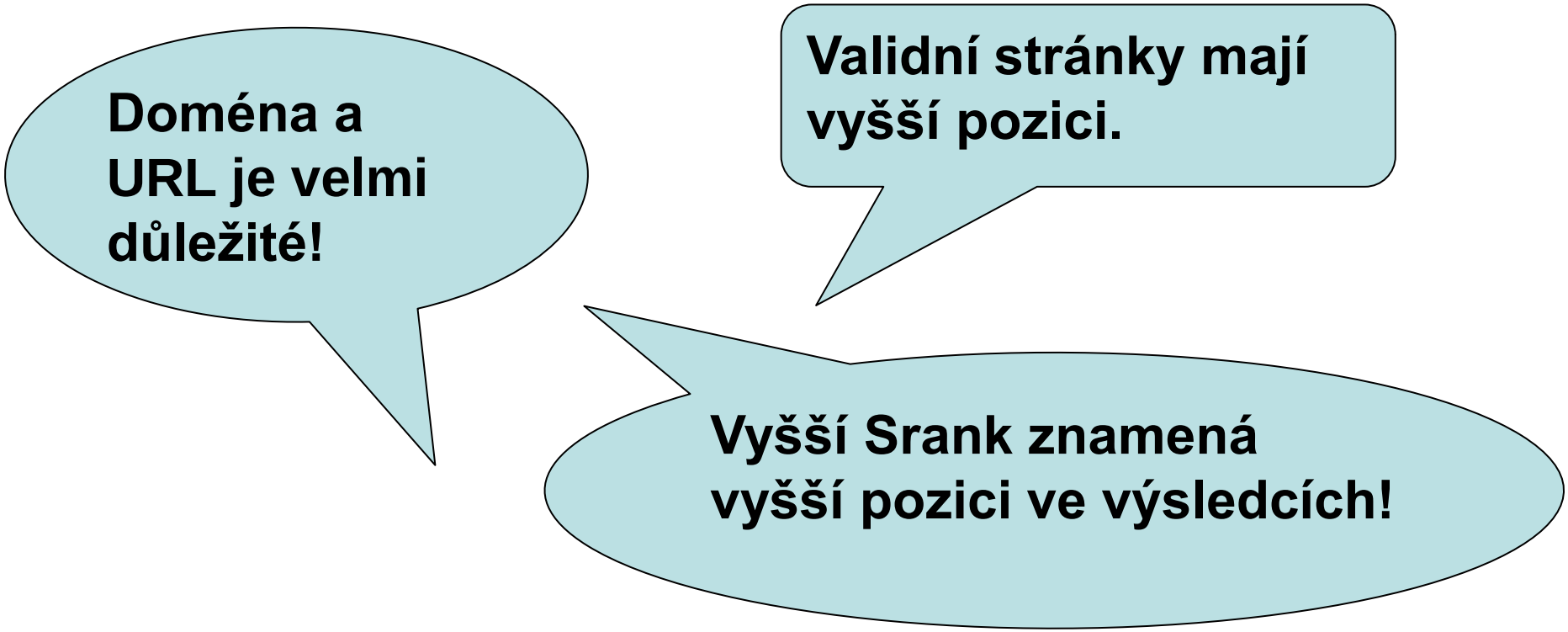
Za účelem podvádět (spam)

- Skryté odkazy a texty
- MFA
- Doorway pages
- Link farmy
- Krádeže obsahu
- Další

Hrozí penalizace.



SEO - báchorky



- A co se říká dál? Ptejte se.

Děkuji za pozornost.