

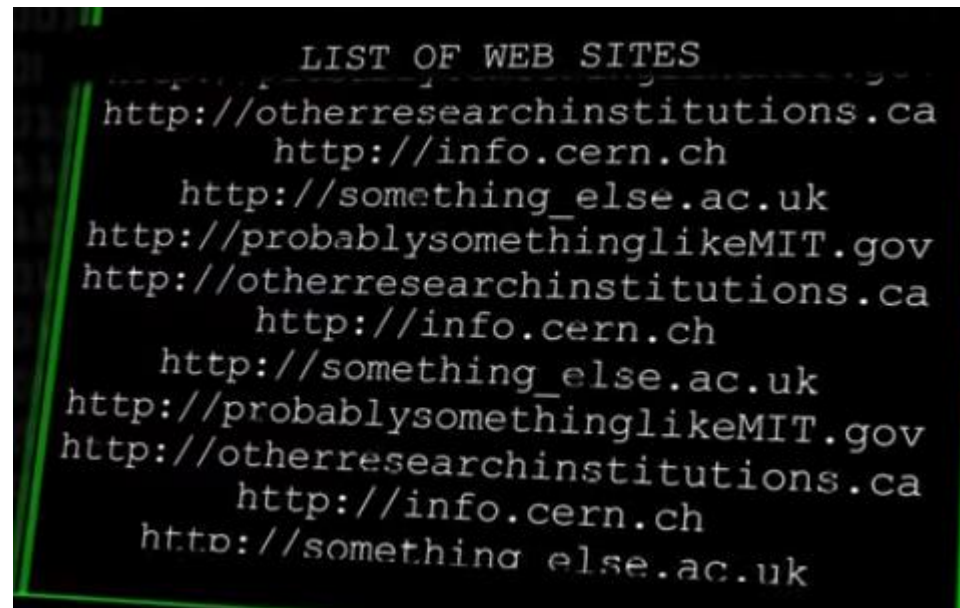
Vyhledávání z pohledu zpracování přirozeného jazyka

Vojtěch Prokop
pro0255

<https://github.com/pro0255/SE>

Úvod do vyhledávání

- Seznam webových stránek
- Sofistikovanější přístup -> **Fulltextové vyhledávání na Internetu**



```
LIST OF WEB SITES
http://otherresearchinstitutions.ca
http://info.cern.ch
http://something_else.ac.uk
http://probablysomethinglikeMIT.gov
http://otherresearchinstitutions.ca
http://info.cern.ch
http://something_else.ac.uk
http://probablysomethinglikeMIT.gov
http://otherresearchinstitutions.ca
http://info.cern.ch
http://something else.ac.uk
```

Struktura výstupu

- Stránka výsledku - SERP
- Popis obrázku
 1. Titulek
 2. URL
 3. Popisek
 4. Rychlé odkazy
 5. Doba publikování
 6. Osnova
- Komponenty
 - Miniaplikace
 - Odpovídač
 - Chytré karty

Medvěd lední

Lední medvěd



Medvěd lední (*Ursus maritimus*), označovaný též jako medvěd polární, je velký druh medvěda typický pro severní polární oblast. Oproti ostatním medvědům využívá užší ekologickou niku, na niž se... [Wikipedie](#)

Kmen	Strunatci
Třída	Savci
Řád	Šelmy
Ohroženost	2 – Zranitelné druhy

Něco se mi nezdařilo

Kam do zoo



Zoologická zahrada města Brno



Zoo Praha

1 + 1 =

2

Rad | Deg

xl

(

)

%

AC

Inv

sin

ln

7

8

9

÷

π

cos

log

4

5

6

×

e

tan

√

1

2

3

−

Ans

EXP

x^y

0

.

=

+

sqrt(4)

×

Q

Internet

Obrázky

Zboží

Mapy

Videa


Zprávy

Firmy

Slovník

√(4) =

2



Colours of Ostrava - Domů

1

https://colours.cz

2

Pořádání hudebního festivalu Colours of Ostrava.

3

Praktické

O festivalu

Novinky

Vstupenky

Předešlé ročníky

Sál Gong v bývalém plynojemu

4

Koronavirus: otevřely se první mateřské školky - Idnes.cz

idnes.cz · Před 17 dny

5



Města začínají otevírat mateřské školy. Aktivní byl v pondělí například Nymburk. Zařízení ale rodiče žádají, aby děti do školek posílali jen v...

Guláš – Wikipedie

https://cs.wikipedia.org/wiki/guláš

Slovo je odvozeno z maďarského „gulya“ (stádo hovězího dobytka), ze kterého pochází maďarský název gulyás (výslovnost [gujáš]) a z toho dále české guláš.


Další varianty · Další významy · Gulášový festival · Literatura · Reference · Externí odkazy

6

Přibližný počet výsledků: 77 200 000 (0,56 s)

https://cs.wikipedia.org/wiki/medvěd_lední

Medvěd lední - Wikipedie



Lední medvěd náleží do rodu *Ursus* (medvěd) a řadí se do čeledi *Ursidae* (medvědovití). Od ostatních savců řádu *Carnivora* (šelmy) se liší robustními šelmy.

Říše: *Živočišné (Animalia)* Podřád: *protivní (Carnifomia)*
 Řád: *medvěd (Ursus)* Řád: *šelmy (Carnivora)*
 Fyzické vlastnosti · Biologie a ekologie · Ohrožení · Vztah s lidmi

Hlavní události

Seznam zpráv

Studio Jablunkov: Lední medvědi neexistují, spadají pod ruské medvědy

před 2 dny

Další zprávy

https://www.zoopraha.cz/lexikon-zvrat/d-221-med-...
Medvěd lední - lexikon zvířat - Zoo Praha
 Medvěd lední (*Ursus maritimus*) - Má typický vzhled medvěda - statné tělo, silné nohy s mohutnými nezatažitelnými drápy, protáhý čenich a krátký ocas. Sít na...
 Potrava: obratlovci, zdechliny (mořští živočichové) · Rozšíření: Asie, Severní Amerika, Evropa...
 Biotope: moře, tundra (tundra, zamrzlé i volné) · Rozměry: délka těla 2-2,5 m, délka oc...

Medvěd lední

Živočišné

Medvěd lední, označovaný též jako medvěd polární, je velký druh medvěda typický pro severní polární oblast. Oproti ostatním medvědům využívá užší ekologickou niku, na niž se výborně adaptoval. Měsíčními vlastnostmi uzpůsobenými na nízké teploty, na pohyb na sněhu, po ledu a v neposlední řadě na plavání v chladné vodě. [Wikipedie](#)

Vědecký název: *Ursus maritimus*
Řád: šelmy (*Carnivora*)
Kmen: strunatci (*Chordata*)
Třída: savci (*Mammalia*)
Hmotnost: 450 kg (dospělý), populace v Beaufortově moři, 150 – 250 kg (dospělý)
Výška: 1,8 – 2,4 m (dospělý), na zadních nohách, 1,3 m (dospělý), po ramena)
Délka: 2,4 – 3 m (dospělý), 1,8 – 2,4 m (dospělý)

Lidé také hledají [Zobrazit další \(více než 10\)](#)

Proces vyhledávání

- Crawler
- Index
- Dotaz
- SERP

Co je to pavouk?

- Robot, pavouk, SeznamBot
- Kde začít?
- Robots.txt
- **Strukturovaná data**
 - schema.org
 - Open Graph

```
{
  "@context": "http://schema.org",
  "@type": "Product",
  "name": "iPhone 13 256GB černá",
  "description": "Mobilní telefon - 6,1" Super Retina XDR 2532 x 1170, procesor Apple A15 Bionic 6jádrový, RAM 4 GB, interní paměť 256 GB, zadní fotoaparát 12 Mpx (f/1,6) + 12 Mpx (f/2,4), přední fotoaparát 12 Mpx, optická stabilizace, GPS, Glonass, NFC, LTE, 5G, Lightning port, voděodolný dle IP68, single SIM + eSIM, neblokován, rychlé nabíjení 20W, bezdrátové nabíjení 15W, baterie 3240 mAh, iOS 15",
  "image": "https://cdn.alza.cz/Foto/f5/RI/RI036c2.jpg",
  "aggregateRating": {
    "@type": "AggregateRating",
    "ratingValue": "4.9",
    "ratingCount": "98"
  },
  "offers": {
    "@type": "Offer",
    "priceCurrency": "CZK",
    "price": "22290",
    "url": "https://www.alza.cz/iphone-13?dq=6731116",
    "itemCondition": "http://schema.org/NewCondition",
    "availability": "http://schema.org/InStock"
  },
  "brand": "Apple",
  "sku": "RI036c2",
  "mpn": "MLQ63CH/A",
  "gtin13": "194252708422"
} == $0
```

User-agent: SeznamBot
Disallow: /admin

User-agent: psbot
Disallow: /obrazky

iPhone 12 - Mobilní telefon | Alza.cz

alza.cz/iphone-12



Mobilní telefon iPhone 12 na www.alza.cz. Bezpečný nákup. Veškeré informace o produktu. Vhodné příslušenství. Hodnocení a recenze iPhone 12 od...

4.8 ★★★★★ (99+) · Cena 22 990 Kč · **Skladem**

Indexování

- Správná konfigurace stránky
- Automatická indexace
- Kontrola dostupnosti v indexu
- Přidat stránku do vyhledávání
- IndexNow

```
<meta name="robots" content="noindex">
```

site:alza.cz

Internet

Obrázky

Zboží

Mapy

Videa

Zprávy

Firmy

Slovník

Alza.cz – nakupujte bezpečně z pohodlí domova | Alza.cz

alza.cz/

Největší obchod s počítači a elektronikou Přes 30 prodejen a více než 270 Alzaboxů PC sestavy, notebooky, mobily, monitory, televize Otevřeno i o víken...

Košík

Chytré hodinky

Mobilní telefony

Notebooky

Mobily, chytré hodinky...

Jak nakoupit

- <https://search.seznam.cz/wt/pridej-stranku>

Pavouk mě dostal co teď?

- Sestavení SERP
- Soutěž o nejlepší umístění
- Balík algoritmů, dle kterých se výsledky řadí
 - PageRank
 - TF-IDF
 - Kosinová podobnost
- Data ovlivňující výstup
 - Historie prohlížení
 - Lokace
 - Položený dotaz
 - Kvalita hledané stránky

Okénko do zpracování přirozeného jazyka

- **Dotaz:** Můj kůň
- BoW vs TF-IDF
- Normalizace
- Invertovaný index
- **Pomalý výpočet**
 - Předzpracování dat
 - Přepočítání hodnot

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

<search engines>		term frequency (tf)
document 1	my: 25 horse: 00	25
document 2	my: 05 horse: 00	05
document 3	my: 10 horse: 10	20
document 4	my: 18 horse: 05	23
document 5	my: 15 horse: 00	15

<search engines>		inverse document frequency		
		term: my	term: horse	
<p>ACHILLES: Come here about me, you my horse horse. Mark what I say, I'll send me when I write. Bide not a while, but keep provisions in hand; When I hear the bloody Hector bound, Engage him with your weapons round about, In fabled manner execute your aims. <small>Exeunt one way and another, respectively cry.</small></p>	document 1 my: 25 horse: 00	$\frac{25}{5} = 5$	$\frac{00}{2} = 0$	5
<p>What want'st to you? CAESAR: No more than my meeting here at Rome Might be to you in Egypt. Yet, if you these Did practice on my state, your being in Egypt Might be my horse. ANTONY: How need'st you practice? my CAESAR: You may be pleas'd to catch at mine intent <small>Exeunt, one way being kill'd one, then calls and threatens.</small></p>	document 2 my: 05 horse: 00	$\frac{05}{5} = 1$	$\frac{00}{2} = 0$	1
<p>Please, her hand, or show the day is lost. Alarum. Enter KING RICHARD for a KING RICHARD: A horse + horse + horse for a CATEBY: Withdraw, and I'll fly you to my horse. KING RICHARD: Stow, I have said my horse upon a word And I will stand the hazard of the day.</p>	document 3 my: 10 horse: 10	$\frac{10}{5} = 2$	$\frac{10}{2} = 5$	7
<p>comes with CAPTAINS and soldiers. To them a MESSENGER MARCUUS: Yonder comes news, a messenger. They have met. LARTIUS: Is my horse to you? my MARCUUS: 'Tis done. LARTIUS: Agreed.</p>	document 4 my: 18 horse: 05	$\frac{18}{5} = 3.6$	$\frac{05}{2} = 2.5$	6.1
<p>ill every my horse, were there twenty brothers behind a Horse to my father in me as you, about I confess your coming before me is nearer to his reverence. OLIVER: What do'st OR ANDRO: Come, come, soldier brother, you are late.</p>	document 5 my: 15 horse: 00	$\frac{15}{5} = 3$	$\frac{00}{2} = 0$	3

Stopslova

- **Dotaz:** Můj kůň
- Irelevantní
- Představa vypuštění „můj“
- Časová složitost, komprese

a	2.11470	0.11143	2.22613	5,006
v	1.60243	0.23087	1.83330	12,593
se	1.50427	0.00423	1.50850	0,281
na	1.24863	0.10347	1.35210	7,653
je	0.74507	0.06108	0.80615	7,577
že	0.70938	0.00557	0.71495	0,779
o	0.58508	0.04603	0.63111	7,293
s	0.58867	0.03908	0.62775	6,225
z	0.53856	0.04787	0.58643	8,164
do	0.46202	0.02509	0.48711	5,151
i	0.43144	0.04308	0.47452	9,078
to	0.41652	0.05711	0.47363	12,057
k	0.33589	0.03961	0.37550	10,549

- <https://gist.github.com/sebleier/554280>

what
which
who
whom
this
that
these
those
am
is
are
was
were
be
been
being
have
has
had
having
do
does
did
doing
a
an
the
and
but
if
or

Lemmatizace a Stemmatizace

- **Dotaz:** Můj kůň
- Vyřešení derivátů
- Pokrytí většího počtu dokumentů
- Úprava dotazů a dokumentů
- Slovní augmentace

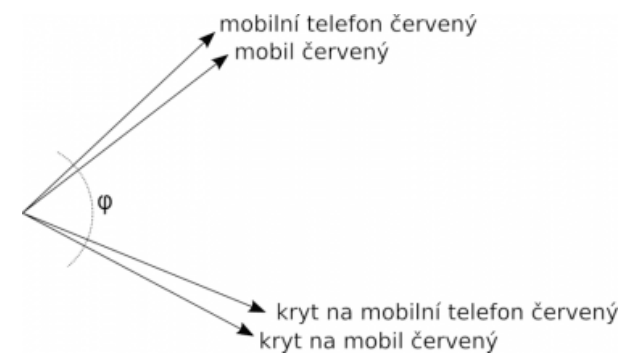
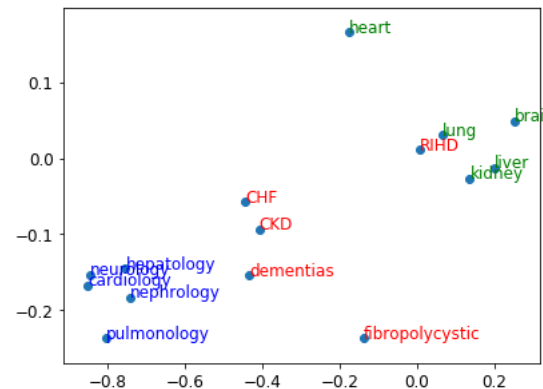
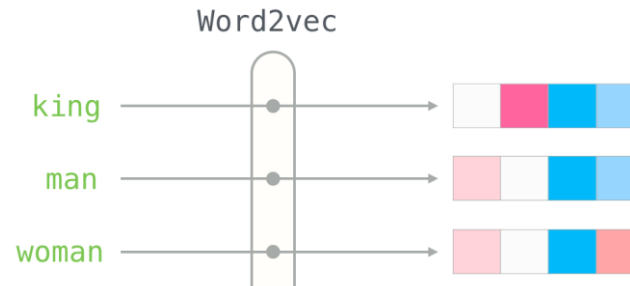
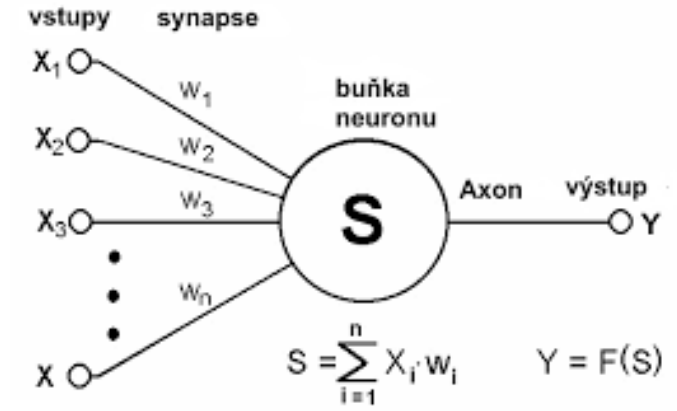
Word	Stemming	Lemmatization
information	inform	information
informative	inform	informative
computers	comput	computer
feet	feet	foot

Pozice slov

- **Dotaz:** Můj kůň
- Ztráta vědomosti o dané pozici v rámci dokumentu
- Bonusové ohodnocení za bližší výskyt
- **Příklad**
 - **Můj kůň** a já
 - **Můj** milovaný **kůň**
 - **Můj kamarád** vypadá jako **kůň**
- Neuronové sítě

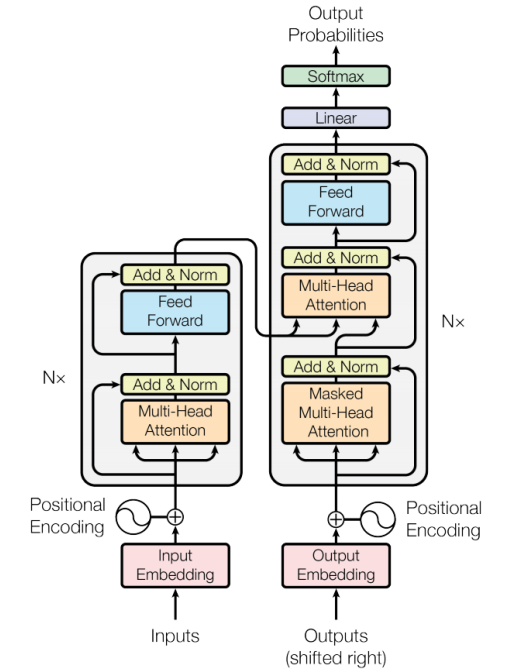
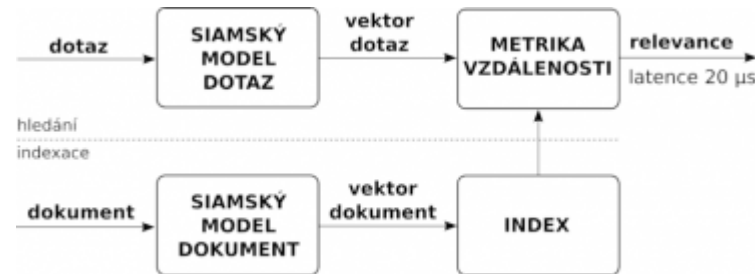
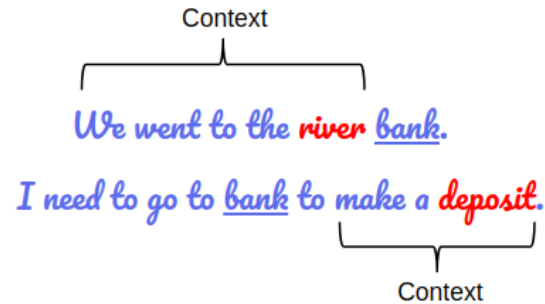
Neuronové sítě

- Metody založené na frekvenci výskytu slova
- Vnoření slov (Word2Vec, GloVe)
- Modelování přirozeného jazyka
- Kosinová podobnost



Nejnovější poznatky

- Transformers
- Nejlepší jazykový model
- Jaguár a polysémie



- Smolíček a HuggingFace (<https://huggingface.co/Seznam/small-e-czech>)

Chci být první!

- **Optimalizace pro vyhledávače** (*Search engine optimization*)
 - SEO
 - Faktory na stránce (on-page)
 - Faktory mimo stránku (off-page)
 - Přidat stránku do vyhledávání
 - IndexNow
-
- <https://napoveda.seznam.cz/cz/fulltext-hledani-v-internetu/optimalizace-webu/>
 - <https://developers.google.com/search/docs/beginner/seo-starter-guide>

Zajímavé odkazy

- **Seznam vyhledávání**

- <https://blog.seznam.cz/stitek/vyhledavani/>
- <https://blog.seznam.cz/2021/10/diky-neuronove-siti-jsme-zlepsili-vysledky-vyhledavani-a-detekujeme-clickbaitove-titulky/>
- <https://www.root.cz/clanky/rychla-oprava-dotazu-ve-vyhledavaci-pomoci-neuronovych-siti/>
- <https://www.root.cz/clanky/jazykove-modely-pro-vyhledavani-naucte-stroj-chapat-vyznam-jazyka/>

- **Evoluce zpracování přirozeného jazyka**

- [*https://medium.com/analytics-vidhya/evolution-of-natural-language-processing-nlp-ac941b6523e9*](https://medium.com/analytics-vidhya/evolution-of-natural-language-processing-nlp-ac941b6523e9)

Otázky?

Moc děkuji za pozornost!