Rudolf Rosa
rosa@ufal.mff.cuni.cz

Jak pracuje internetový vyhledávač

Univerzita Karlova
Matematicko-fyzikální fakulta
Ústav formální a aplikované lingvistiky

Den otevřených dveří MFF UK, Praha, 23. 11. 2016

...pokud zrovna pracuje

[The IT Crowd]

Jak najít informaci na stránce?

- hledám: *ptakopysk*
- projdu stránku
  od začátku do konce,
  hledám slovo „ptakopysk“

Jak najít informaci na internetu?

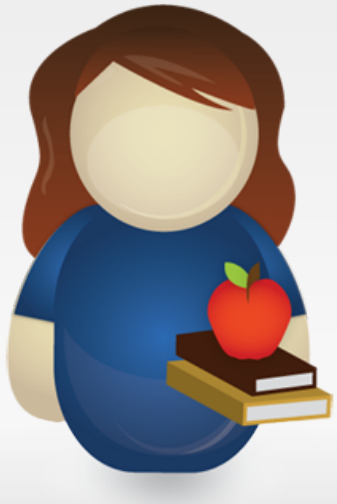
- hledám: *ptakopysk*
- projdu internet
  od začátku do konce,
  hledám slovo „ptakopysk“

Jak najít informaci na internetu?

- hledám: *ptakopysk*
- projdu internet
  od začátku do konce,
  hledám slovo „ptakopysk“
- za 30 let mám výsledek 

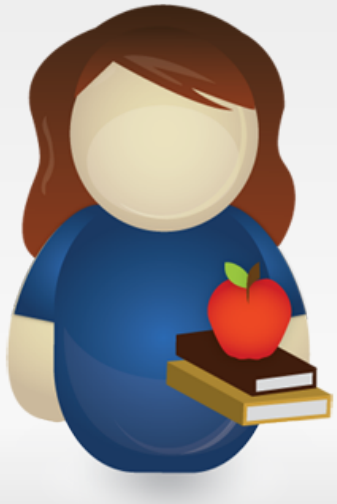
Jak najít informaci v knize?

- hledám: *ptakopysk*
- projdu knihu
  od začátku do konce,
  hledám slovo „ptakopysk“



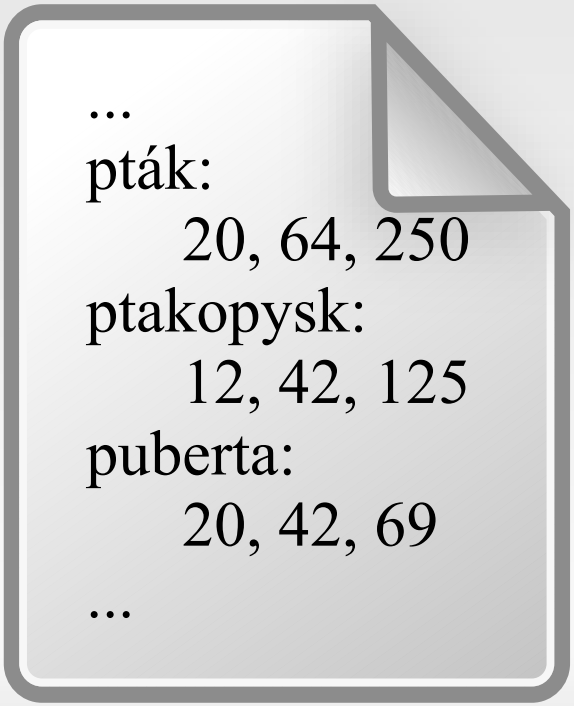
Jak najít informaci v knize?

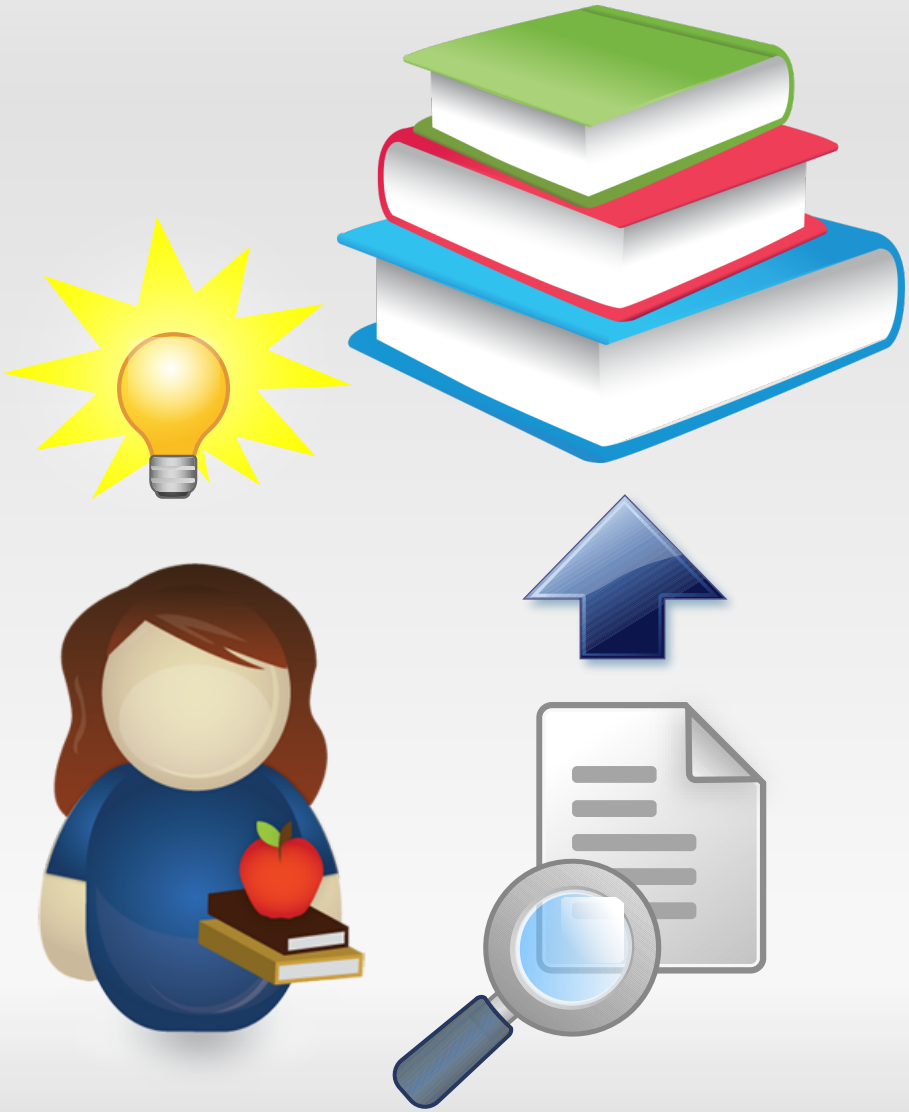
- hledám: *ptakopysk*
- projdu knihu
  od začátku do konce,
  hledám slovo „ptakopysk“
- za týden mám výsledek



Jak najít informaci v knize?

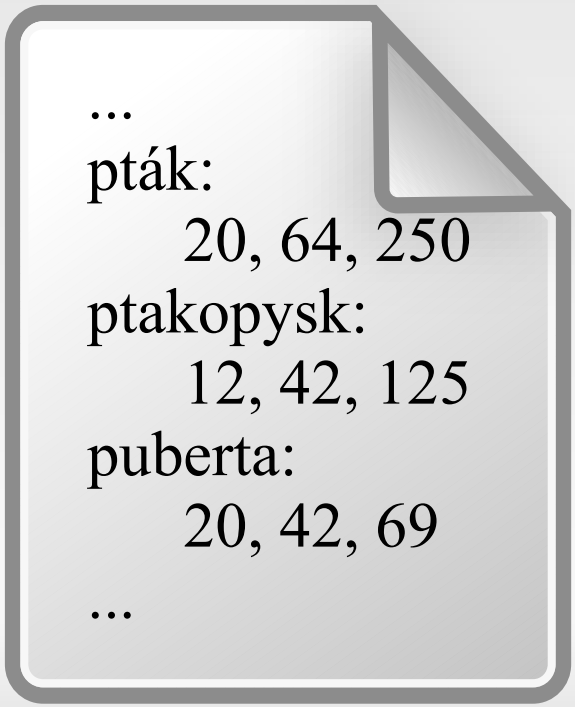
- hledám: *ptakopysk*
- kniha mívá obsah, nebo dokonce **rejstřík**!





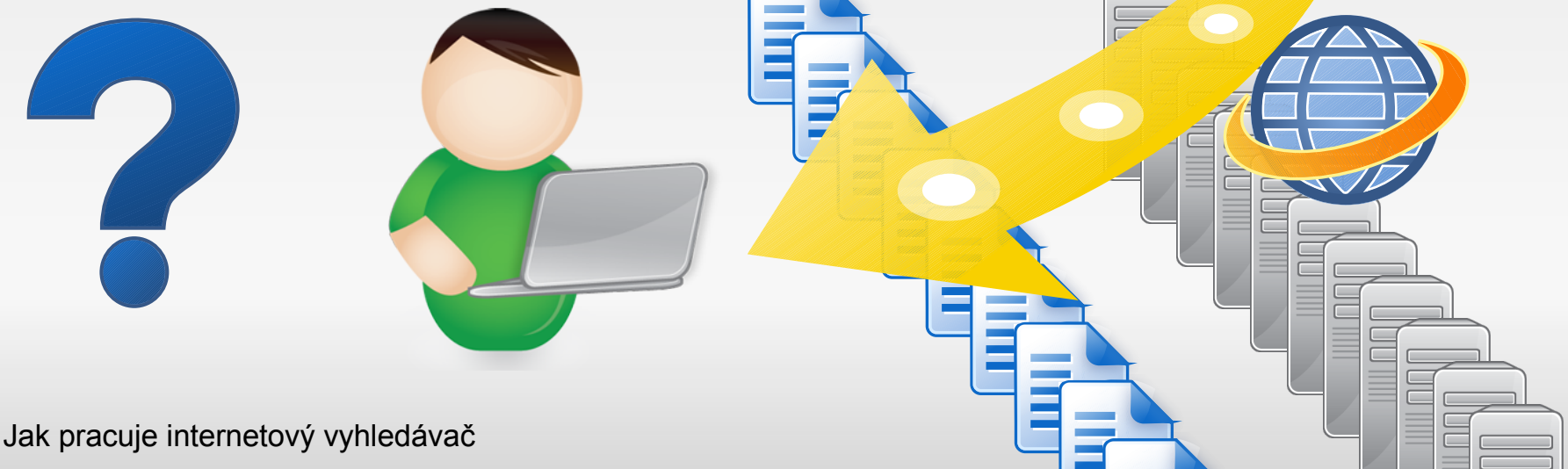
Jak najít informaci na internetu?

1. příprava
   - uložím si internet na disk (jen text)
   - udělám si pro něj „rejstřík“ (*index*)
   - „číslo stránky“: URL odkaz na stránku
   - abecední uspořádání → rychlejší hledání



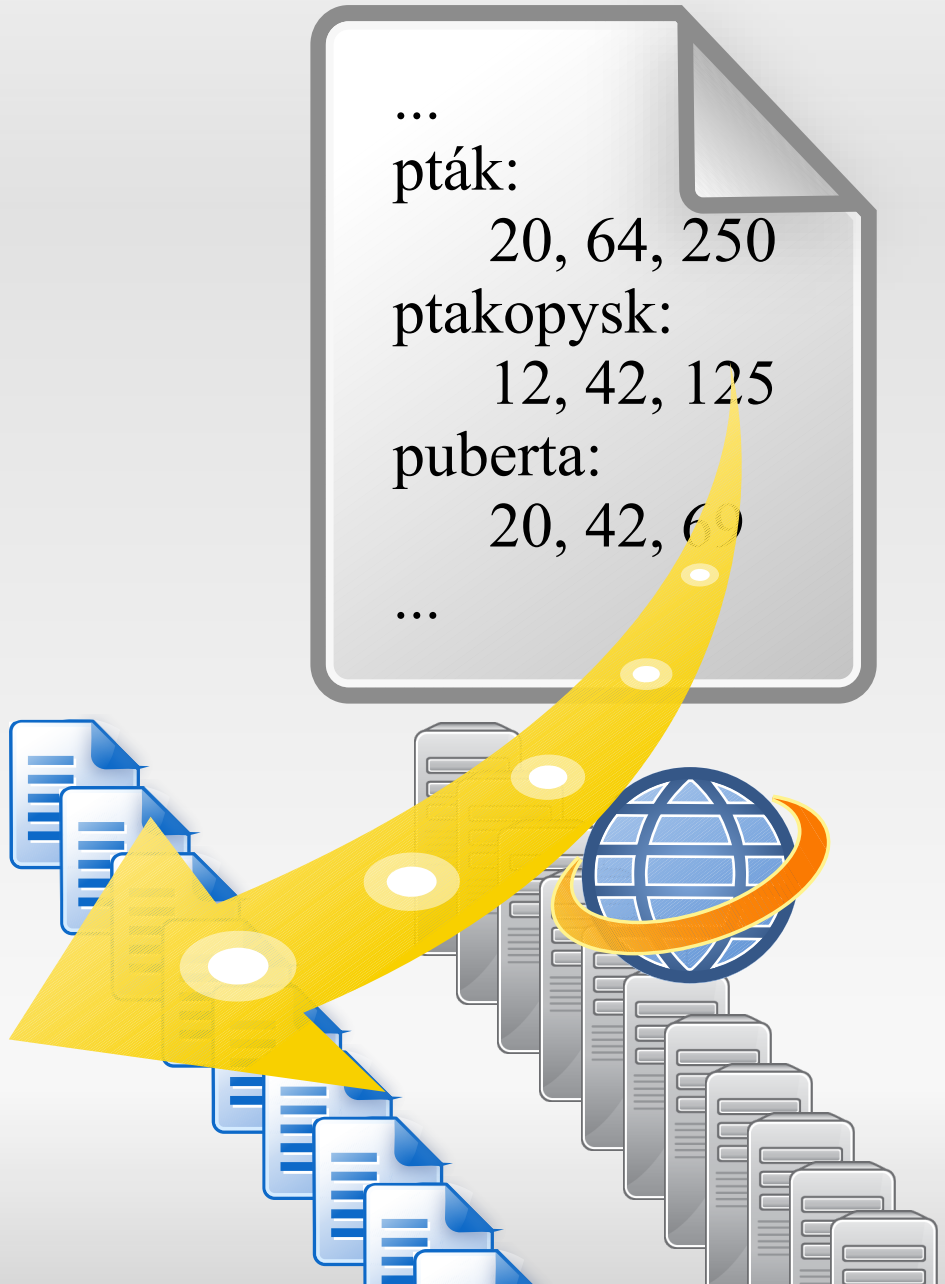
Jak najít informaci na internetu?

2. hledání
   - hledám: *ptakopysk*
   - najdu „ptakopysk“ v indexu
   - dostanu 50 000 odkazů
   - přečtu si všech 50 000 stránek



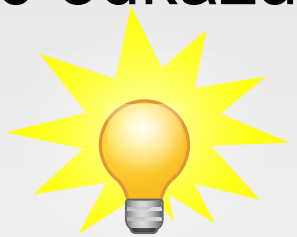
Jak najít informaci na internetu?

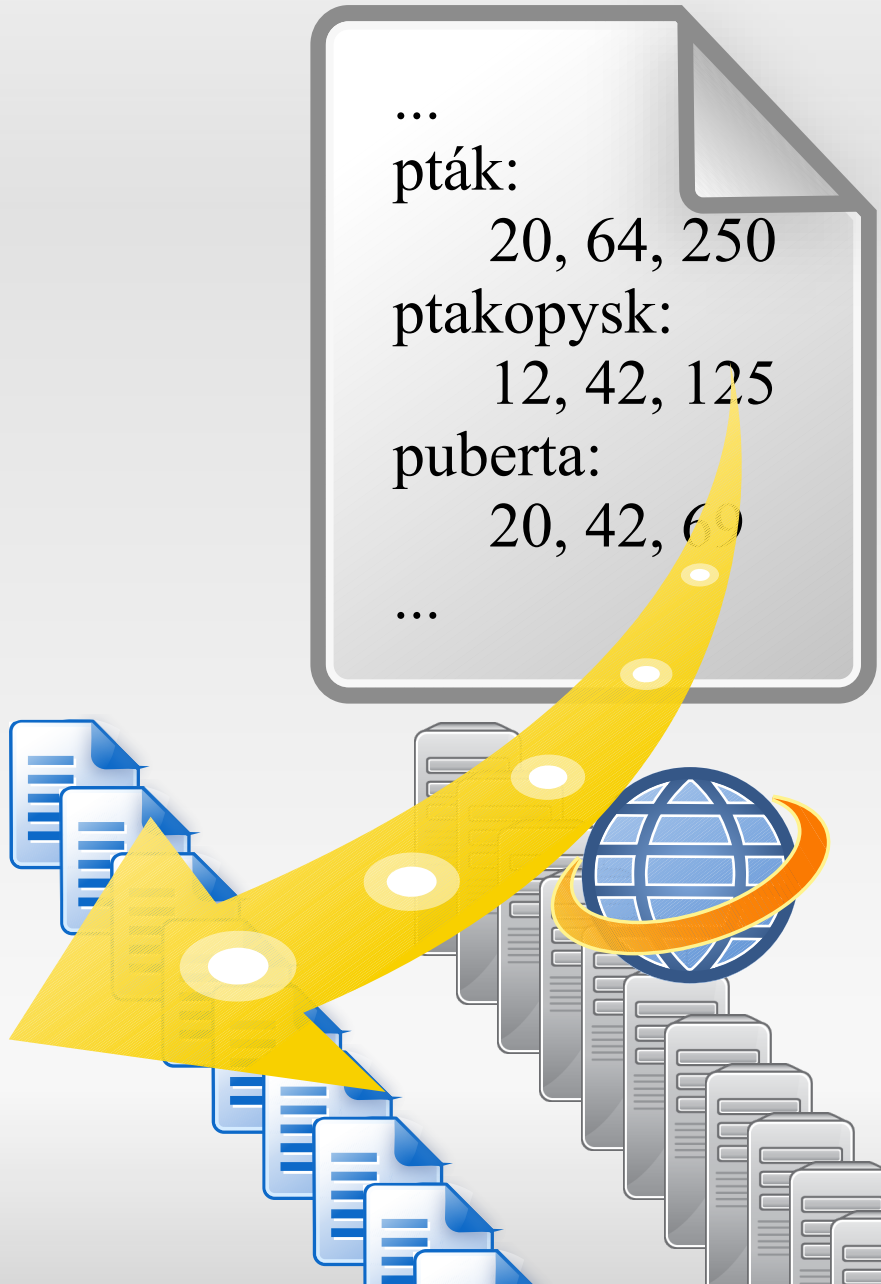
2. hledání
   - hledám: *ptakopysk*
   - najdu „ptakopysk“ v indexu
   - dostanu 50 000 odkazů
   - přečtu si všech 50 000 stránek



Jak najít informaci na internetu?

2. hledání
   - hledám: *ptakopysk*
   - najdu „ptakopysk“ v indexu
   - dostanu 50 000 odkazů
   - přečtu si jen 5 prvních!
   - ale které z těch 50 000 jsou první?





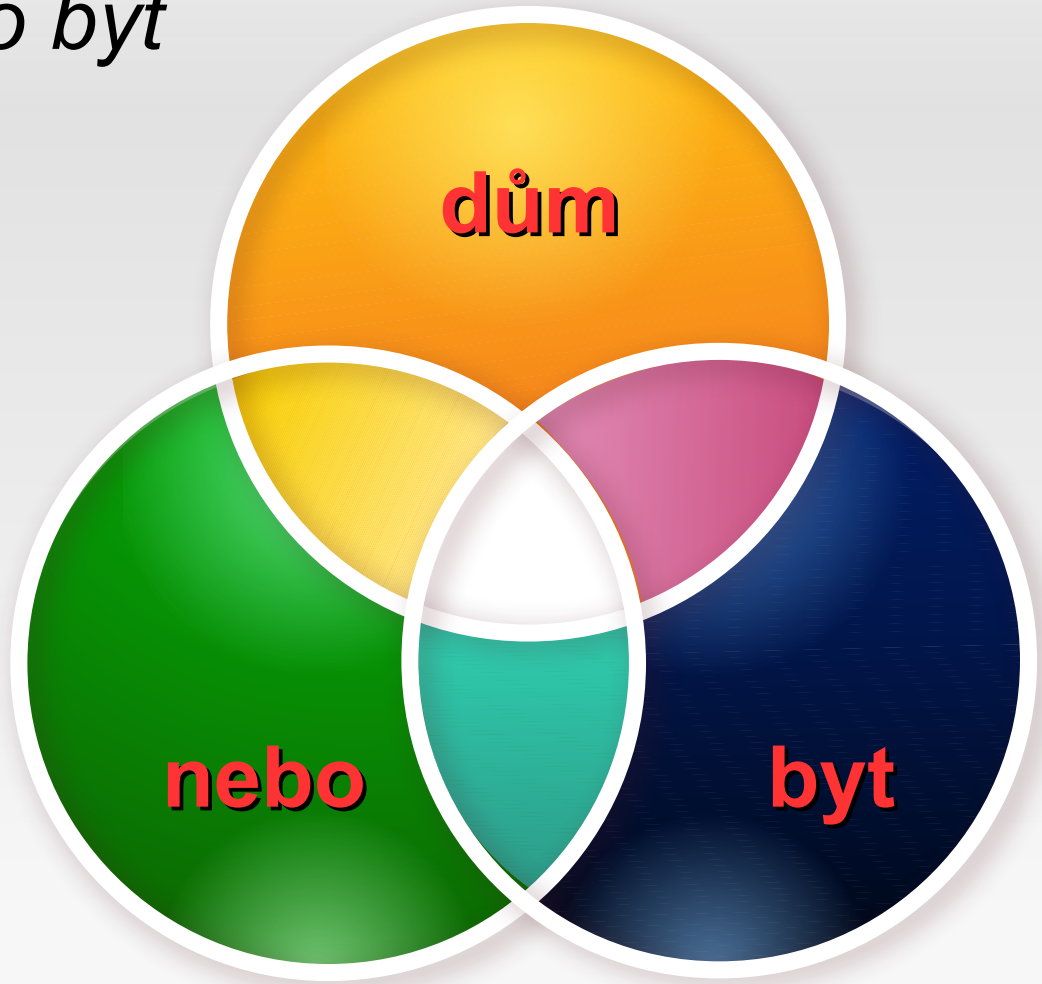
Řazení výsledků vyhledávání

- nestačí jen najít výsledky:
  ještě je potřeba je seřadit od nejlepších!
- relevance vzhledem k hledání
  - frekvence hledaného slova, významnost umístění hledaného slova (nadpis)...
- obecná kvalita stránky
  - kvalita obsahu, oblíbenost, aktuálnost, důvěryhodnost, bezpečnost...

Pepův
supr ptakopysk
ptakopysk ptakopysk
ptakopysk ptakopysk
ptakopysk ptakopysk
ptakopysk ptakopysk
ptakopysk jéééé!!!

Víceslovné dotazy

- hledám: *dům nebo byt*
- průnik výsledků pro jednotlivá slova



Víceslovné dotazy

- hledám: *dům nebo byt*
- dobré výsledky:
  - ***Dům nebo byt***
  - ***Byt nebo dům***
  - ***Dům***, ***nebo*** *raději* ***byt****?*
  - *Chcete* ***dům****?* ***Nebo*** *se pro vás hodí spíše* ***byt****?*
  - *…*

Víceslovné dotazy

- hledám: *dům nebo byt*
- špatný výsledek:
  - 1x dům, 1x byt, 6x nebo
- běžná slova nedůležitá

Petr nebo Pavel?
Nevím, jestli si vybrat Petra nebo Pavla. Pavel nemá tak pěkný **dům nebo** auto jako Petr, ale to je snad jedno, **nebo** ne? A líbí se mi, jak se na mě Pavel dívá **nebo** jak mě hladí. Ale vadí mi, když pije **nebo** kouří. A moc si neuklízí svůj **byt**, ale to vlastně Petr taky ne... **Nebo** že bych si nechala oba...?

Víceslovné dotazy

- hledám: *dům nebo byt*
- špatný výsledek:
  - 1x dům, 1x byt, 6x nebo
- běžná slova nedůležitá
  - stoplist:
    dům ~~nebo~~ byt

Petr nebo Pavel?
Nevím, jestli si vybrat Petra nebo Pavla. Pavel nemá tak pěkný **dům nebo** auto jako Petr, ale to je snad jedno, **nebo** ne? A líbí se mi, jak se na mě Pavel dívá **nebo** jak mě hladí. Ale vadí mi, když pije **nebo** kouří. A moc si neuklízí svůj **byt**, ale to vlastně Petr taky ne... **Nebo** že bych si nechala oba...?

Víceslovné dotazy

- hledám: *dům nebo byt*
- špatný výsledek:
  - 1x dům, 1x byt, 6x nebo
- běžná slova nedůležitá
  - stoplist:
    dům ~~nebo~~ byt
  - důležitost slova **s**:

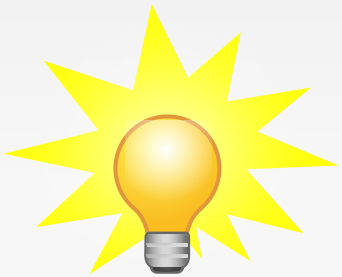
$$\frac{\textit{frekvence } \mathbf{s} \textit{ v dokumentu}}{\textit{frekvence } \mathbf{s} \textit{ v celé databázi}}$$

Petr nebo Pavel?

Nevím, jestli si vybrat Petra nebo Pavla. Pavel nemá tak pěkný **dům nebo** auto jako Petr, ale to je snad jedno, **nebo** ne? A líbí se mi, jak se na mě Pavel dívá **nebo** jak mě hladí. Ale vadí mi, když pije **nebo** kouří. A moc si neuklízí svůj **byt**, ale to vlastně Petr taky ne... **Nebo** že bych si nechala oba...?

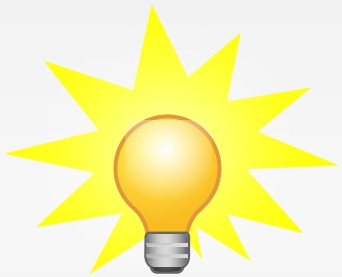
Rozšířené hledání

- skloňování
  - *dům → dům, domu, domem, domy, domům...*
  - nebo naopak v dokumentech: *domu, domy... → dům*
    - *Rádi bychom bydleli ve velkém domě →*
      rád by bydlet v velký dům



Rozšířené hledání

- skloňování
  - *dům → dům, domu, domem, domy, domům...*
  - nebo naopak v dokumentech: *domu, domy... → dům*
    - *Rádi bychom bydleli ve velkém domě →*
      rád by bydlet v velký dům
- synonyma
  - *dům → dům, domek...*
  - *přinutit → přinutit, donutit, přimět...*



Pojmenované entity

- rozlišení vlastní jméno/obecné slovo
  - hledám: *miroslav donutil*
    -  *miroslav **donutil** již nehraje v národním divadle*
    -  *miroslav **přinutil** řidiče zastavit*

Pojmenované entity

- rozlišení vlastní jméno/obecné slovo
  - hledám: *miroslav donutil*
    - 🙂 *miroslav **donutil** již nehraje v národním divadle*
    - 😟 *miroslav **přinutil** řidiče zastavit*
  - hledám: *dny nato*
    - 🙂 *dny **nato** se konají v ostravě*
    - 😟 *den **nato** se rozešli*

Pojmenované entity

- rozlišení vlastní jméno/obecné slovo
  - hledám: *miroslav donutil*
    - 😊 *miroslav **donutil** již nehraje v národním divadle*
    - 😟 *miroslav **přinutil** řidiče zastavit*
  - hledám: *dny nato*
    - 😊 *dny **nato** se konají v ostravě*
    - 😟 *den **nato** se rozešli*
  - hledám: *ministr chovanec*
    - 😊 *vyjádření ministra milana **chovance***
    - 😟 *ministr navštívil **chovance** v ústavu*

Shrnutí

- příprava databáze
  - sbírka dokumentů, sestavení indexu
- hledání slova/více slov
  - hledání v indexu, průnik výsledků
  - skloňování, synonyma, pojmenované entity...
- řazení výsledků
  - frekvence v dokumentu / frekvence v celé databázi
  - umístění (nadpis)
  - kvalita dokumentu, aktuálnost, důvěryhodnost...

Děkuji za pozornost

Univerzita Karlova
Matematicko-fyzikální fakulta
Ústav formální a aplikované lingvistiky

Počítačové zpracování přirozeného jazyka

překlad z angličtiny do češtiny *ovládání počítače hlasem*

větný rozbor *určování slovních druhů*

vyhledávání v textech *textový popis obrázku*

Studium počítačové lingvistiky na Matyfzu

- **Bc.**: Obecná informatika, zaměření Matematická lingvistika
- **Mgr.**, **PhD**: Informatika, obor Matematická lingvistika

http://ufal.mff.cuni.cz/ http://ufal.cz/rudolf-rosa/