Pravidla anotování relevance

verze 4.0.0 2022.01

Denně uživatelé zadávají do Vyhledávání Seznam.cz miliony dotazů, aby našli informace, které potřebují. Vyhledávač se snaží pochopit, co uživatelé hledají, co se nachází na webových stránkách a následně vyhodnotit, jaké webové stránky jsou pro daný dotaz nejužitečnější. Využívá k tomu modely strojového učení, pro jejichž vytvoření je potřeba velké množství příkladů (anotací). K tomu potřebujeme vás, anotátory.

Tento dokument popisuje způsob vytváření anotací, které využijeme ke zlepšení relevance výsledků služby Vyhledávání Seznam.cz

1. Co je úkolem anotátora

Úkolem anotátora je určit, jak dobře jeden nalezený dokument ve výsledcích vyhledávání odpovídá tomu, co chtěl uživatel. Anotátor postupně anotuje jednotlivé dvojice dotaz-dokument. Každou dvojici může buď anotovat nebo odmítnout.

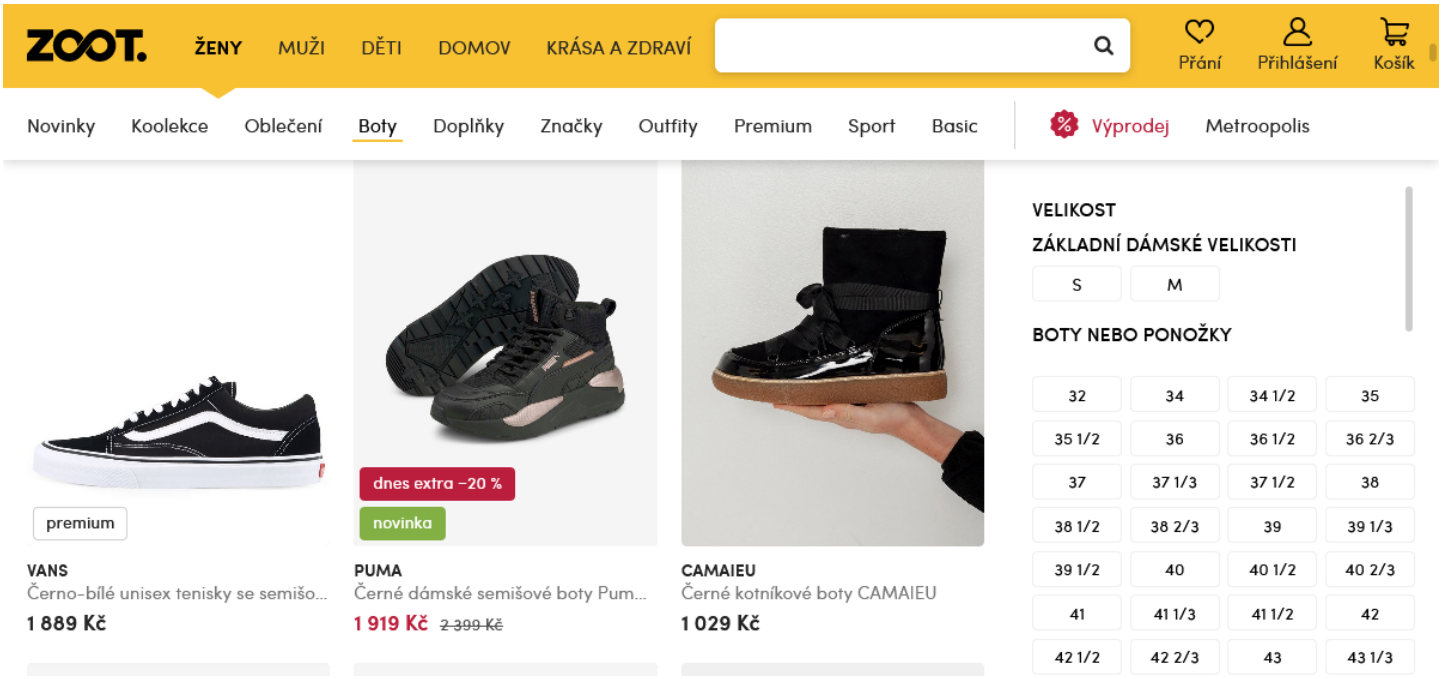
Uživatelský záměr

Uživatelé vyhledávají proto, aby se dozvěděli aktuální informace, získali zboží, kontaktovali nějakou firmu atd. Důvodů k zadání dotazu do vyhledávače existuje mnoho a říkáme jim uživatelské záměry.

Dotaz

Dotaz je text, který uživatel zadal do vyhledávače. Jde o reprezentaci uživatelského záměru. Dotazy uvádíme v hranatých závorkách.

1.1. Příklad anotace 1

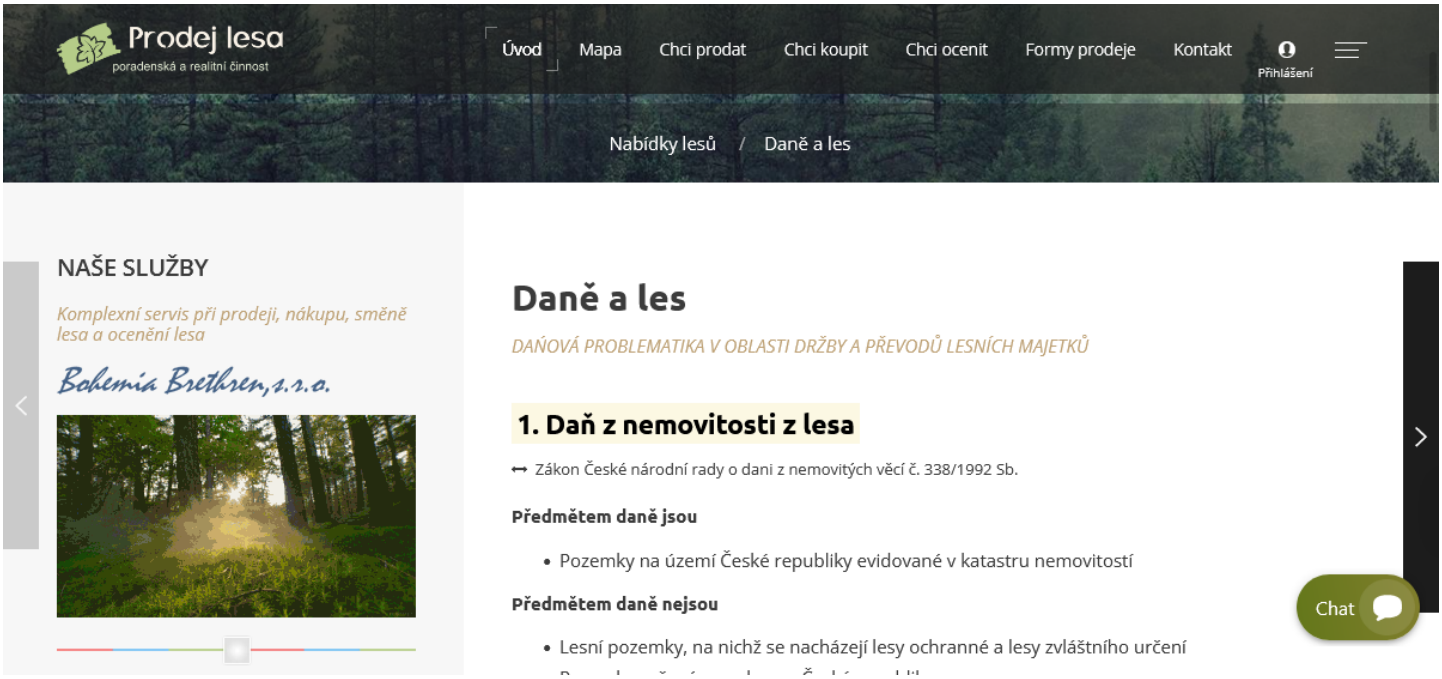


Dotaz je [**dámské boty**]. Uživatelský záměr je vybrat si a koupit dámské boty. Nalezený dokument je z e-shopu.

Anotace:

1. Dotazu i dokumentu rozumíme. Není důvod odmítnout anotaci.
2. Dotaz míří na konkrétní typ zboží, dokument je o daném typu zboží a o ničem jiném. Jedná se o přesnou odpověď.
3. Dokument by měl představovat užitečnou odpověď pro většinu uživatelů, kteří chtějí koupit dané zboží.

1.2. Příklad anotace 2



Dotaz je [**zdanění příjmu z lesa**]. Uživatel potřebuje poradit s daněmi. Nalezený dokument je článek.

Anotace:

1. Dotazu i dokumentu rozumíme. Není důvod odmítnout anotaci.
2. Nevíme jistě, s jakým záměrem uživatel dotaz zadal a odpověď není přesně o zdanění příjmu z lesa. Odpověď není přesná, s dotazem ale určitě souvisí.
3. Pro některé uživatele by dokument mohl představovat užitečnou odpověď.

2. Anotační proces

Anotační proces má tři kroky:

1. Rozumíme dotazu i dokumentu a lze hodnotit relevanci?
2. Co je předmětem dotazu, co je předmětem dokumentu a jak to spolu souvisí?
3. Jak moc užitečnou odpověď dokument představuje?

2.1. Odmítnutí anotace

Někdy je těžké pochopit, co dotaz znamená nebo není jasné, o čem je dokument.
V těchto případech dává smysl anotaci odmítnout.

Nejčastější důvody odmítnutí anotace:

- Dotaz je jen shlukem znaků bez významu.
- Dotaz je neúplný. Část dotazu důležitá pro pochopení záměru uživatele chybí.
- Dotaz je v jazyce, kterému nerozumíme.
- Dokument se nenačte nebo se jedná o chybovou stránku.
- Dokument má špatné kódování znaků nebo jiný technický problém.
- Dokument je v jazyce, kterému nerozumíme.
- Nechceme hodnotit porno.

Dotazy, na které odmítáme anotaci

[katastrální]
není zřejmé, co uživatel hledal (úřad, mapy, území, ...)

[dovolená itálie 2012]
dotaz měl smysl maximálně v roce 2012, nyní již smysl nedává, nevíme, co uživatel očekává

[nová škoda felicia]
poslední generace tohoto vozu se vyráběla v letech 1994 - 2001, proto dotaz nedává smysl

[samozřejmě]
není zřejmé, co uživatel očekává za odpověď (může jít o dotaz, který uživatel zadal jen proto, aby zjistil, jak se dané slovo píše)

[jaké i se píše ve slově slon]
nedává smysl, nelze očekávat, že existují rozumné výsledky

[výprodej]
není zřejmé, co uživatel očekává za odpověď

Dotazy, na které anotaci neodmítáme

Drobné nedostatky, jako je chybějící diakritika, nesprávná interpunkce, překlepy apod., nejsou důvodem k odmítnutí anotace, pokud dotaz chápeme.

[volby 2010]
v roce 2010 proběhly parlamentní volby a je legitimní hledat např. jejich výsledky

[ford focus combi 2012]
uživatele zajímá auto vyrobené v konkrétním roce

[mudr josef řehák]
chybějící tečka v akademickém titulu není důvod k odmítnutí

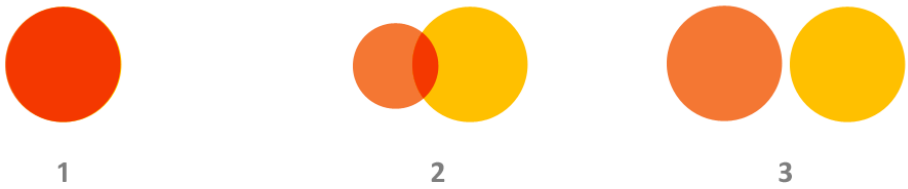
[psaci stul]
navzdory chybějící diakritice je zřejmý význam dotazu

[fejsbuk]
je zřejmé, že uživatel hledal facebook, jen to chybně napsal

[chaty achalupy]
v dotazu chybí mezera, je však zřejmé, co chtěl uživatel hledat

2.2. Anotace přesnosti odpovědi

Přesnost odpovědi vyjadřuje, do jaké míry pokrývá obsah dokumentu téma dotazu.



(1) Přesná odpověď

- zabývá se předmětem dotazu a ničím jiným
- je reprezentována celým dokumentem
- může se jednat o užitečnou odpověď (viz anotace užitečnosti odpovědi)

(2) Související odpověď

- zabývá se širší nebo užší oblastí, než jakou vymezuje dotaz, nebo se nějak jinak dotazu týká
- je reprezentována jen částí dokumentu (zbylá část dokumentu odpovídá na jiné dotazy)
- může se jednat o užitečnou odpověď

(3) Nesouvisející odpověď

- odpověď se zabývá něčím úplně jiným než oblastí vymezenou dotazem
- dokument je automaticky hodnocen jako neužitečný

Rozdíly mezi třídami přesnosti

[přilby]

- **(1)** Přesná: rozcestník v e-shopu odkazující na nabídku přileb pro různé sporty (na kolo, na lyže, ...)
- **(2)** Související: výpis přileb na kolo v e-shopu
- **(2)** Související: detail konkrétní přilby nějaké značky, velikosti a barvy v e-shopu
- **(3)** Nesouvisející: článek o nehodě motorkáře, který jel bez přilby

[divadla v praze]

- **(1)** Přesná: rozcetník na divadla v Praze
- **(2)** Související: hlavní stránka webu Národního divadla v Praze

[fobie z výšek]

- **(1)** Přesná: článek na wikipedii o fobii z výšek
- **(2)** Související: článek na wikipedii o fobiích, který obsahuje i informace o fobii z výšek
- **(2)** Související: zpravodajský článek o tom, že nějaká známá osobnost trpí fobií z výšek

2.4. Anotace užitečnosti odpovědi

2.4.1 Určení intentu

Někdy může být obtížné určit uživatelský záměr jen na základě znalosti dotazu:

- [**le mans 66**] - uživatel mohl chtít zjistit něco o slavném závodu, ale také si mohl chtít zakoupit film
- [**škoda octavia**] - uživatel mohl chtít přejít na oficiální stránku na webu skoda-auto.cz, ale také mohl chtít najít nabídku aut v bazaru nebo si přečíst článek na wikipedii o historii vozu

Abychom mohli v těchto případech rozhodnout a následně hodnotit užitečnost dokumentu, potřebujeme k tomu znát dodatečnou informaci - intent.

Intent

Některé uživatelské záměry jsou si podobné. Pokud chceme například koupit boty, je to podobné, jako když chceme koupit spacák. V obou případech očekáváme nabídku produktů v e-shopu. O takových dotazech pak říkáme, že mají stejný intent. Intent je typ uživatelského záměru. Příklady intentů:

- Přejít na web
- Získat zboží
- Najít firmu
- Dozvědět se více o entitě
- Poradit

Intentů může být velké množství v závislosti na tom, jak je definujeme. Jejich aktuální seznam s definicemi je součástí přílohy těchto pravidel.

Intent závisí na anotátorovi

1. **Anotátor si na začátku anotování vybere úlohu, ze které se mu následně vydávají jednotlivé dovojice dotaz-dokument. Název úlohy odpovídá intentu.** Pokud si například vybere úlohu »Získat (zboží)«, u všech dotazů pak předpokládá tento intent.
2. Existuje i varianta anotační úlohy, ve které intent není určen předem výběrem úlohy, ale anotátor jej určuje pro každý dotaz zvlášť. Intent se v tomto případě určuje vždy tak, aby se shodoval s typem právě anotovaného dokumentu. Příklad: obsahuje-li dokument výpis kategorie zboží v e-shopu, předpokládáme intent dotazu »Získat (zboží)«, ne »Dozvědět se více« apod.

2.4.2 Anotace užitečnosti

Užitečnost odpovídá tomu, kolik uživatelů, kteří hledají s daným uživatelským záměrem, odpověď může uspokojit.

(1) Užitečná odpověď

- uspokojí většinu uživatelů
- pro daný uživatelský záměr může být na první straně výsledků vyhledávání

(2) Trochu užitečná odpověď

- uspokojí některé uživatele
- pro daný uživatelský záměr může být ve výsledcích vyhledávání na druhé a další straně

(3) Skoro neužitečná odpověď

- většinu uživatelů neuspokojí, ale může existovat uživatel, kterému to stačí

(4) Neužitečná odpověď

- neuspokojí žádného uživatele

Rozdíly mezi třídami užitečnosti

[přilby na kolo] s intentem »Získat (zboží)«

- **(1)** Užitečná: kategorie zboží v e-shopu s přilbami na kolo, dostatek zboží
- **(2)** Trochu užitečná: kategorie zboží v e-shopu s přilbami na kolo, omezená nabídka
- **(3)** Skoro neužitečná: kategorie zboží v e-shopu s přilbami na kolo, velmi omezená nabídka
- **(4)** Neužitečná: kategorie zboží v e-shopu s přilbami na kolo, žádné odpovídající zboží
- **(4)** Neužitečná: článek s testy přileb na kolo

[přilby na kolo] s intentem »Dozvědět se více«

- **(1)** Užitečná: článek s testy přileb na kolo, velké množství informací
- **(2)** Trochu užitečná: článek, který obsahuje nějaké informace o přilbách na kolo
- **(3)** Skoro neužitečná: článek, který obsahuje malé množství informací k přilbám na kolo
- **(4)** Neužitečná: článek, který kromě nadpisu nemá žádný obsah
- **(4)** Neužitečná: kategorie zboží v e-shopu s přilbami na kolo

Užitečnost odpovědi je relativní k uživatelskému záměru

Na dotaz [pečení chleba] s intentem dozvědět se více, by dokument z e-shopu s nabídkou domácích pekáren a ničím jiným, představoval neužitečnou odpověď. S intentem získat zboží, by ale na daný dotaz stejný dokument mohl představovat užitečnou odpověď.

3. Tipy pro usnadnění anotování

- Pokud nerozumíme dotazu, můžeme zkusit zahledání libovolným vyhledávačem.

- Různá velikost písmen v dotazu může odhalit uživatelský záměr: NiFe - chemický vzorec, fiXa - hudební skupina, LaTeX - značkovací jazyk, MarS - výrobce padáků, apod.
- Při určování přesnosti odpovědi je dobré podívat se na URL, hlavní nadpis a titulek dokumentu. Nelze ale hodnotit jen na základě toho, protože tyto informace nemusí vždy odpovídat obsahu dokumentu.
- Pokud se nemůžeme rozhodnout, volíme vždy horší hodnocení.

4. Složitější případy

4.1. Čas

Čas může ovlivnit přesnost i užitečnost. Stránka o události, která proběhla v jiném roce, než byl zadán v dotazu, představuje nesouvisející (a tím pádem neužitečný) výsledek.

Příklady jak hodnotit:

- [tour de france 2016] - dokument s informacemi o Tour De France 2015 představuje (3) Nesouvisející, (4) Neužitečnou odpověď.
- [tour de france] - přesnost ani užitečnost odpovědi nezávisí na čase. Dokument s informacemi o Tour De France 2015 představuje stejně kvalitní odpověď, jako dokument s informacemi z roku 2020.

4.2. Lokalita

Lokalita může ovlivnit přesnost i užitečnost. Stránka restaurace se stejným názvem, ale v jiné lokalitě, než byla zadána v dotazu, představuje nesouvisející (a tím pádem neužitečný) výsledek.

Příklady, jak hodnotit:

- [restaurace u hrocha brno] - přesnost odpovědi u stránky restaurace s daným názvem v Brně hodnotíme (1) Přesná odpověď, u stránky restaurace s daným názvem v Praze jako (3) Nesouvisející, (4) Neužitečná.

- [restaurace u hrocha] - přesnost ani užitečnost odpovědi zde nezávisí na lokalitě. Stránka restaurace v Praze i v Brně může představovat přesnou a užitečnou odpověď.

4.3. Nejednoznačné dotazy

Některé dotazy mohou mít více významů, např. [koleje], [liška]. Hodnotíme podle významu, o kterém píše dokument, tedy např. dokument o lišce zvířeti hodnotíme podle lišky zvířeti, dokument o Liškovi herci hodnotíme podle Lišky herce.

V případě, že má dotaz jasný majoritní význam, hodnotíme podle tohoto významu. Existuje kominík Václav Havel v Táboře. Ale pro dotaz [Václav Havel] by dokument o něm byl nesouvisející. Jinak by to bylo samozřejmě u dotazu [vaclav havel tabor].

4.4. Jazyk

Pokud je dokument v jazyce, kterému nerozumíme, anotaci odmítneme.

4.5. Porno

Pokud nechceme anotovat, lze odmítat už na základě dotazu.

4.6. Kvalita obsahu dokumentu

Vlastnosti dokumentu při anotování ignorujeme. Patří mezi ně například:

- vzhled
- použitelnost
- pravopis
- pravdivost informací
- množství reklam
- množství obsahu

Kvalita obsahu dokumentu nemá vliv na užitečnost

I dokument, který je nepřehledný, obsahuje velké množství reklamy a pravopisné chyby, může obsahovat užitečnou odpověď na daný uživatelský záměr.

Pokud se dokument nenačte nebo je technicky zpracován tak, že nerozumíme jeho obsahu, anotaci odmítneme.

4.7. Typ obsahu dokumentu

4.7.1. Obsah vyžadující přihlášení

Vždy hodnotíme obsah, který dostaneme, aniž by bylo potřeba se přihlašovat.

4.7.2. Obsah vyžadující spuštění

Audio, video, hry, aplikace. Může jít i například o různé kalkulačky na dotazy typu [kolik eur je 5 000 Kč]. Nic nespouštíme. Vždy hodnotíme na základě informací, které dostaneme, aniž by bylo potřeba něco spouštět.

4.7.3. Obsah ke stažení

Soubory ke stažení nestahujeme a neotevíráme. Vždy předpokládáme, že to lze a že po otevření obsahují to, co říká dokument, který na ně odkazuje. Hodnotíme přesnost a užitečnost odkazujícího dokumentu.

4.7.4. Cookies hlášky, potvrzení věku, ...

Dialogová okna, bez jejichž zavření se nelze dostat k obsahu stránky, odklikáváme. Hodnotíme přesnost a užitečnost obsahu, který se zobrazí po jejich zavření.

4.7.5. Skrytý obsah

Někdy stránka zobrazuje jen část obsahu a vyžaduje dodatečnou interakci k tomu, aby uživatel získal všechen obsah. Typicky jde o tlačítko »Více« a skrytí delšího textu. Interakce, která je nutná pro získání obsahu stránky, je povolená.

4.7.6. Reklamy

Reklamy na stránce ignorujeme i v případě, že poskytují relevantní odpověď na dotaz.