Doménové kolokace

Jiří Materna

Centrum zpracování přirozeného jazyka, FI MU Brno
Seznam.cz

4. prosinec 2009

Křišťálová lupa: 1. cena v kategorii Vyhledávače a databáze

Motivace

- Zpracování dotazů ve fulltextovém vyhledávání Seznam.cz.
- Proximita – míra kolokability dvou tokenů (nebo obecně n-gramu).
- Reálné číslo z <0, 1> (0 značí absolutně nesouvisející, nikoliv neutrální).
- K dispozici texty všech zaindexovaných dokumentů (cca. 198G pozic v korpusu).

Standardní přístupy

- Počet výskytů každého tokenu t v korpusu: $f(t)$
- Počet výskytů všech bigramů (t_1, t_2) v korpusu: $f(t_1, t_2)$
- Počet pozic korpusu: n
- MI^x-score, t-score, dice, logdice,...
- Převod hodnot skóre lineárně normalizací maximální hodnotou skóre.
- Nejlepší výsledky MI^2-score, logdice.

Výhody a nevýhody klasické metody

- + Ověřené časem.
- + Snadný a rychlý výpočet.
- - Nerozpozná kolokace, týkající se specializované domény v případě, že jsou jednotlivé termy frekventované.

Příklady

slovní spojení	**proximita**
jízdní řády	0.952
karlovy vary	0.983
a ale	0.295
zelené myšlenky	0.278
rozhodovací strom	0.363
třecí síla	0.441
cz 75	0.291

Doménové kolokace – motivace

- Problému dochází když (a, b) je sice kolokace, ale "a" i "b" jsou velmi frekventované nebo nejednoznačné termy.
- "cz 75" je stěží kolokace, přesto vyžadujeme vysokou proximitu.
- **Řešení:** počítat specializované kolokace pouze nad omezenou doménou dokumentů.
- **Problémy:** výběr domén, identifikace specializovaného výrazu, přiřazení domény,...

Doménové kolokace – výpočet

- Použití osvědčených technik (MI^2-score, log dice).
- Uvažují se pouze dokumenty, ve kterých se současně vyskytují všechny konstituenty výrazu.
- Náročnější na výpočet (vyžadující speciální datové struktury).
- Vždy platí DOM_MI^2((a,b)) >= MI^2((a,b)).

$f((a, b))$ = počet výskytů bigramu (a, b).
$f_b(a)$ = počet výskytů a v dokumentech, obsahujících i b.
$f_a(b)$ = počet výskytů b v dokumentech, obsahujících i a.

Příklady

slovní spojení	proximita	dom. proximita
jízdní řády	0.952	0.992
karlovy vary	0.983	0.995
a ale	0.295	0.319
zelené myšlenky	0.278	0.286
rozhodovací strom	0.363	0.684
třecí síla	0.441	0.820
cz 75	0.291	0.516

Další zdroje dat

- Pro některé výrazy není ani doménová metoda dostatečná (cz 75).
- Titulky z české Wikipedie.
- Fráze ze slovníku (LangSoft).
- Kombinace jména a příjmení (data z ispellu).
- Whitelist (Názvy firem, místopisné názvy,...)

Kombinace zdrojů dat

- Výsledná proximita získána kombinací uvedených zdrojů.
- Dodatečný zdroj je do výpočtu zahrnut pouze v případě, že se v něm bigram vyskytuje (s hodnotou proximity 1).
- Celková proximita jako vážený aritmetický průměr.
- Váhy nastaveny ruční kalibrací:

MI^2	**wiki**	**dict**	**names**	**whitelist**
1	2	1.5	1.5	2

Příklady

slovní spojení	prox.	dom. prox.	komb. prox.
jízdní řády	0.952	0.992	0.999
karlovy vary	0.983	0.995	0.998
a ale	0.295	0.319	0.319
zelené myšlenky	0.278	0.286	0.286
rozhodovací strom	0.363	0.684	0.793
třecí síla	0.441	0.820	0.988
cz 75	0.291	0.516	0.724

Děkuji za pozornost.