Extrakcia kľúčových slov z výučbových dokumentov

Jozef Harinek
Marián Šimko

Motivácia

- Extrakcia RDT
- Ako zlepšiť výsledky?
- Používateľské anotácie – veľký potenciál, potreba analýzy

Metóda extrakcie

1. Existujúci ATR algoritmus na text
2. Extrahovanie RDT z anotácií + zohľadnenie používateľa
3. Kombinácia výsledkov

$$w_{final}(t,d) = (1-p) * w_{ATR}(t,d) + p * w_{annot}(t,d)$$

$$w_{annot}(t,d) = \frac{\sum_{a \in A} \sigma_a * rel_a(t,d)}{|A|}$$

$$rel_a(t,d) = \text{UserRanking(t)} * w'_{ATR}(t,d')$$

Overenie

- Hypotéza: zohľadnenie anotácií zlepší výsledky ATR algortimov
- Dataset (štatistiky)
  - PSI (ALEF)
  - 180 LO
  - 1000 používateľov
  - Priemerne 170 anotácií na 1 LO

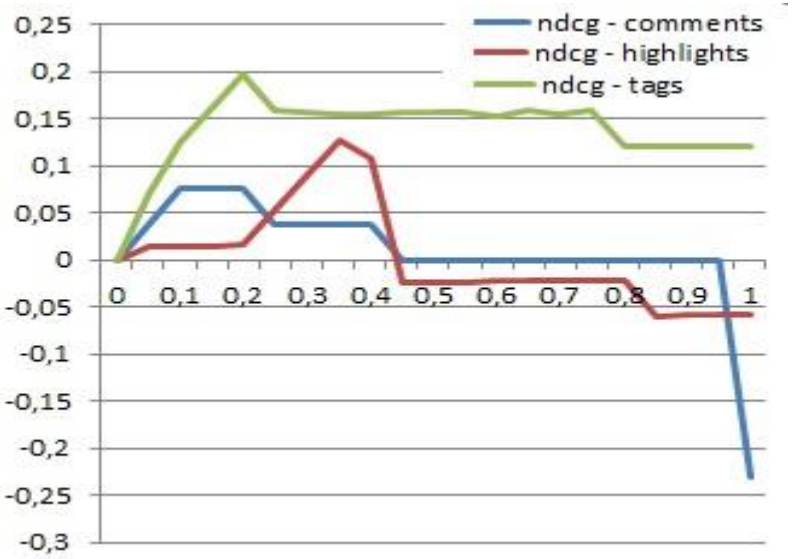
Overenie

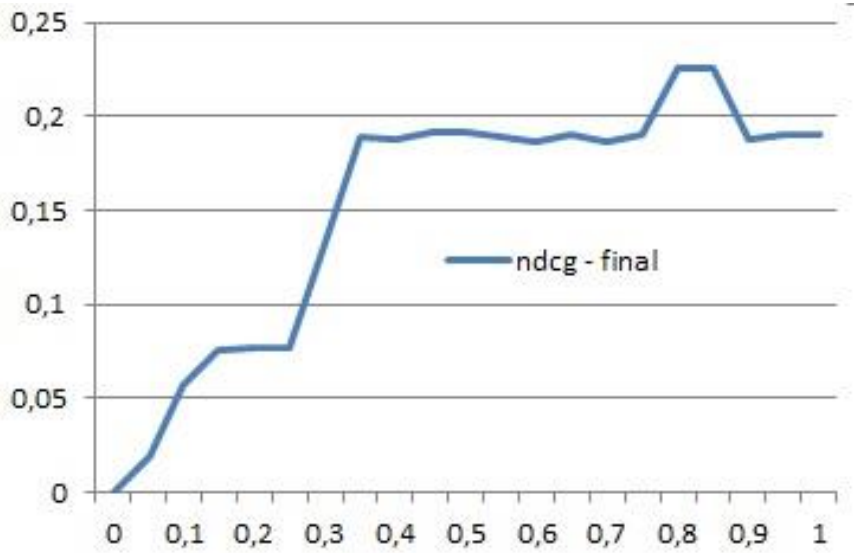
- Experimenty
  - kvantitatívne („zlatý“ štandard)
  - Cieľ: nastaviť parametre
  - Relevancia jednotlivých typov anotácií
    - Tagy, zvýraznenia, označené texty pri komentároch (zatiaľ nie obsah)
  - Vylepšenie za použitia anotácií
- Metriky
  - Dôležité je poradie vygenerovaných RDT
  - NDCG, Average Precision, F-measure

Overenie

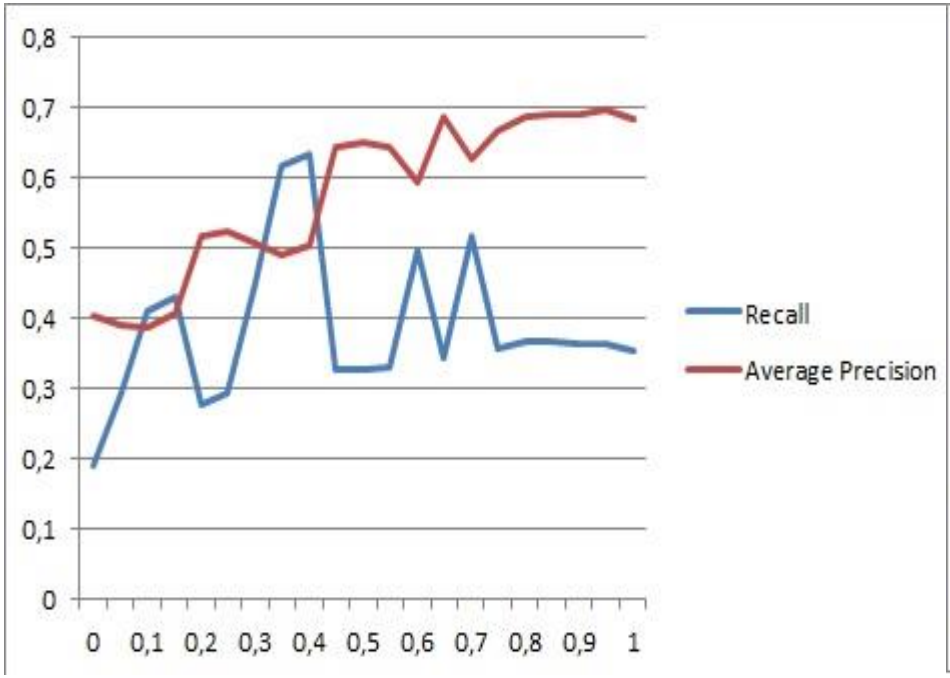
- Nastavenie
  - 180 LO („trénovacia“, testovacia časť)
- Výsledky
  - zlepšenie 19.8 % tagy, 12.5 % zvýraznenia, 7.8 % komentáre
  - 22,6 % celkové zlepšenie

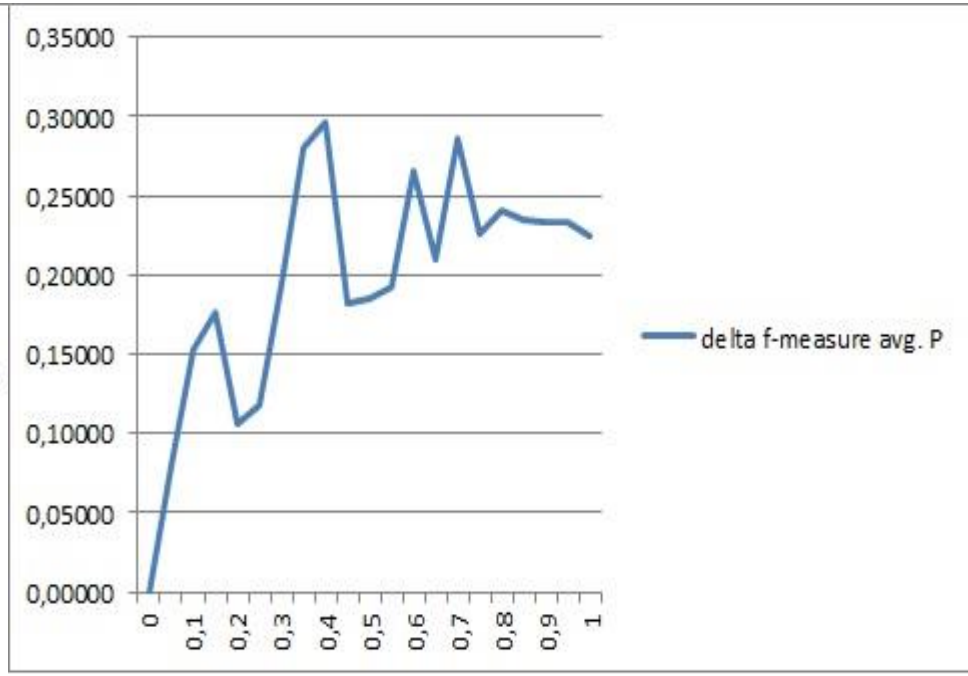
Výsledky


0,25
0,2
0,15
0,1
0,05
0
-0,05
-0,1
-0,15
-0,2
-0,25
-0,3
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1
ndcg - comments
ndcg - highlights
ndcg - tags


0,25
0,2
0,15
0,1
0,05
0
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1
ndcg - final

Výsledky


0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0 0,1 0,2 0,3 0,4 0,5 0,6 0,7 0,8 0,9 1
Recall
Average Precision


0,35000
0,30000
0,25000
0,20000
0,15000
0,10000
0,05000
0,00000
0 0,1 0,2 0,3 0,4 0,5 0,6 0,7 0,8 0,9 1
delta f-measure avg. P

Plány

- Skúmať viacero typov ATR
- Skúmať obsah komentárov
- Kvalitatívne overenie (experti – vyučujúci)
- Variácie User Ranking
- Využiť Metallurgy
- Využitie: Come^{2}t, ALEF