Semanticizing Search Engine Queries

The University of Amsterdam at the ERD 2014 Challenge

David Graus
University of Amsterdam
Daan Odijk
University of Amsterdam
Manos Tsagkias
University of Amsterdam
Wouter Weerkamp
904Labs
Maarten de Rijke
University of Amsterdam

d.p.graus@uva.nl
dvdgrswww.graus.co

Frequency of mention in first paragraph of entity’s document
Frequency of mention in first sentence of entity’s document
Frequency of mention in title of entity’s document
Position of the first occurrence in the first paragraph of entity’s document
True if mention contains the title of entity
True if title of entity contains mention
True if title of entity equals mention
CMNS Probability of entity being target of link with mention as anchor
Levenshtein Distance between mention and title of entity
Levenshtein Distance between normalized mention and entity
True if the title of entity begins with mention
True if the title of entity ends with mention
Number of shared words between title of entity and mention
Number of different words between title of entity and mention

Features

Mention Features

TF_paragraph Frequency of mention in first paragraph of entity’s document
TF_sentence Frequency of mention in first sentence of entity’s document
TF_title Frequency of mention in title of entity’s document
PM Frequency of mention in all Wikipedia pages
POS Probability that mention is used as an anchor text in Wikipedia (all occurrences)
NCT Number of articles whose title matches a sub-n-gram of mention
TCN Number of articles whose title equals a sub-n-gram of mention
TEN True if title of entity equals mention
CMNS Probability of entity being target of link with mention as anchor
LD Levenshtein Distance between mention and title of entity
ND Levenshtein Distance between normalized mention and entity
SS_MATCH_1 Number of shared words between title of entity and mention
SS_MATCH_2 Number of different words between title of entity and mention
W_MISS Number of different words between title of entity and mention

Entity Features

LEN Number of terms in mention
IDF_anchor Inverse Document Frequency of mention in all Wikipedia anchors
IDF_content Inverse Document Frequency of mention in all Wikipedia pages
IDF_title Inverse Document Frequency of mention in all Wikipedia titles
LENPM Number of articles whose title matches a sub-n-gram of mention
LENKL Number of articles whose title equals a sub-n-gram of mention

Contextual Features

DOCSIM (Normalized) retrieval score of q to entity content
RERANKER (Normalized) retrieval score of q to entity virtual document
SEMANTIZER (Normalized) retrieval score of q to entity virtual document

Results

1. SMAPR Team 0.6858
2. NTU-LIP 0.6797
3. Seznam Research 0.6693
4. Magnetic_IISAS 0.6557
5. WebSAIL 0.6414
6. UBC 0.6369
7. CS 0.6194
8. ExPoSe 0.6170
9. CI_RD 0.6130
10. UvA 0.6062
11. InriaBerlin 0.5746
12. XI-lab 0.5182
13. NTU-ULS 0.4947
14. whb_1Z 0.4866
15. TAIL-UPC 0.4508
16. UIUC-UGS 0.4003
17. clues_RD 0.3642
18. SIEL-ERD 0.2060
19. Pasquale 0.2040

Input: Query
new york daily news sports

DOCSRANKER
0.60 Daily News (New York)
0.25 News
0.17 News (band)
0.13 New York

DATA FUSION

Combine both lists

POST-PROCESS

Keep longest mention(s)
Cutoff candidate list

Final Output!

1.00 Daily News (New York)
0.11 New York
0.00 News (band)