Identifying food and diet-related queries in large query samples

Thema:
Identifying food and diet-related queries in large query samples
Art:
BA
BetreuerIn:
Prof. Dr. David Elsweiler, Dr. David E. Losada
BearbeiterIn:
Matthias Herrmann
ErstgutachterIn:
N.N.
ZweitgutachterIn:
N.N.
Status:
abgeschlossen
Stichworte:
query expansion and classification, Natural language processing, food and diet
angelegt:
2019-06-17
Beginn:
2019-07-31
Antrittsvortrag:
2019-07-22
Textlizenz:
Unbekannt
Codelizenz:
Unbekannt

Hintergrund

By classifying user queries the “effectiveness, efficiency, and revenue potential in general-purpose web search systems“ can be improved (Beitzel et al., 2005). Categorizing queries can also be used to create a text collection containing only relevant topics, which is interesting for data miners and also for reducing mismatches between search engine results and what the user wants to receive (Li, Zheng, & Dai, 2007). However, the classification is challenging because the web is heterogeneous, and many user queries are short (Kang & Kim, 2003). The winning team of the 2005 KDD cup, which “is the annual Data Mining and Knowledge Discovery competition organized by ACM Special Interest Group on Knowledge Discovery and Data Mining“ , enriched user queries by running a web search to provide a broader context and to achieve better classification results (Shen et al., 2005). We will adopt the approach to expand queries for finding food and diet-related queries from a large list of query samples. Identifying food and diet-related web searches can help understand what users ask search engines for when they have an information need regarding this topic. Further, we can build a food and diet corpus based on the user queries which could help researchers. There are many applications possible, e.g. studying coherence between nutrition and depression, that are not part of

the bachelor thesis.

Zielsetzung der Arbeit

We want to identify food and diet-related queries from a large set of query samples. Therefore a tool is created that can expand and classify queries using Google Custom Search and Google Classify. The organizers of the KDD cup provided 800 queries which are labeled by human editors and can be used for evaluating the classification results (Li et al., 2007). Different strategies can be employed for classifying user queries with the help of Google Classify. The query results can be concatenated to one text and then send to the Google Classify API, or every single snippet text of each hit can be classified seperately. The performance of those different classification strategies can be measured and compared.

Konkrete Aufgaben

- Implementierung eines Tools zur Klassifizierung von Suchanfragen
- Query expansion mit Google Custom Search
- Klassifizierung mit Google Cloud Natural Language API
- Evaluierung des Tools im Kontext der Aufgabenstellung Essens- und Ernährungsrelevante Suchanfragen zu finden - unter Anwendung unterschiedlicher Strategien

Erwartete Vorkenntnisse

Keine

Weiterführende Quellen

Beitzel, S. M., Jensen, E. C., Frieder, O., Grossman, D., Lewis, D. D., Chowdhury, A., & Kolcz, A. (2005). Automatic web query classification using labeled and unlabeled training data. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’05 (p. 581). New York, New York, USA: ACM Press. https://doi.org/10.1145/1076034.1076138 Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/J.PATREC.2005.10.010 Kang, I.-H., & Kim, G. (2003). Query type classification for web document retrieval. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval - SIGIR ’03 (p. 64). New York, New York, USA: ACM Press. https://doi.org/10.1145/860435.860449 Li, Y., Zheng, Z., & Dai, H. (Kathy). (2007). KDD CUP-2005 report. ACM SIGKDD Explorations Newsletter, 7(2), 91–99. https://doi.org/10.1145/1117454.1117466 Shen, D., Pan, R., Sun, J.-T., Pan, J. J., Wu, K., Yin, J., & Yang, Q. (2005). Q 2 C@UST. ACM SIGKDD Explorations Newsletter, 7(2), 100–110. https://doi.org/10.1145/1117454.1117467 API Schnittstellen: https://cloud.google.com/natural-language/docs/classifying-text https://developers.google.com/custom-search/