Query Segmentation
REST-API Documentation

February 27, 2020

Contents

1 Segmentation Approaches

The following segmentation approaches can be used via the REST-API (see Table 1):

Table 1: Segmentation Approaches

Approach	Identifier	Reference	Link
WT-Baseline	wt-baseline	[HPBS12]	[CIKM 2012]
WT-SNP-Baseline	wt-snp-baseline	[HPBS12]	[CIKM 2012]
HYB-A	hyb-a	[HPBS12]	[CIKM 2012]
HYB-B	hyb-b	[HPBS12]	[CIKM 2012]
HYB-I	hyb-i	[HPBS12]	[CIKM 2012]
Wiki-Based Approach	wiki-based	[HPSB11]	[WWW 2011]
Naive Approach	naive	[HPSB10]	[SIGIR 2010]

Remark: Different from the original publication [HPBS12], we use POSTaggerME from the OpenNLP toolkit in the demo.

2 Segmentation of A Single Query

The REST-API provides the functionality to send one query to the server and to get the segmentation of all approaches listed in Table 1 in form of a JSON string.

2.1 cURL Command

```
curl -v -X POST -d "query"
host:port/query-segmentation-server/query
```

2.2 Example

POST Request

```
curl -v -X POST -d "this is a test query"
https://demo.webis.de/query-segmentation/query
```

GET Request

```
curl https://demo.webis.de/query-segmentation/
query?text=this%20is%20a%20test%20query
```

Response

```
{
    "hyb-i":"this is a test|query",
    "hyb-b":"this is a test|query",
    "hyb-a":"this is a test|query",
    "wiki-based":"this is a test|query",
    "wt-baseline":"this is a test|query",
    "wt-snp-baseline":"this is a test|query",
    "naive":"this|is a|test query"
}
```

3 Segmentation of a Set of Queries

The REST-API provides the functionality to send a set of queries in a file to the server in order to get the segmentation for a specific approach.

3.1 File Format

The queries file contains an id and a query separated by a tab per line:

```
id<tab>query
```

The id is a number, which will be casted to a long value, and the query is a string.

Example File

```
1004073900        graffiti fonts alphabet
1004593125        stainless steel chest freezers
1004933775        rutgers online graduate classes
1005728827        review on breezes
```

Limitations

The files that are send to the REST-API have to fulfill the constraints listed in Table 2. If a constraint is violated the response contains the HTTP status code 500 (Internal Server Error) and a error message, which is also listed in the table below.

Table 2: Constraints

Constraint	Error Message
Maximum 10 KB size	org.apache.tomcat.util.http.fileupload.FileUpload-Base$FileSizeLimitExceededException: The field file exceeds its maximum permitted size of 10240 bytes.
Maximum 100 lines	Too many lines in file: {file}. Files with more than 100 lines, i.e. queries, will not be processed.
Predefined format	Failed to parse file: {file}. Please check the file format.

3.2 cURL Command

```
curl -v -X POST -F "file=@path-to-queries-file"
host:port/query-segmentation-server/approach/approach-identifier
```

The `approach-identifier` for the available segmentation approaches can be found in Table 1.

3.3 Example

Request

Consider the example queries file from above (see Section 3.1) to be stored on your disk at `/home/test/queries.txt`. The cURL request for getting segmentation of the *Naive Approach* (see Table 1) will look like this:

```
curl -v -X POST -F "file=@/home/test/queries.txt"
https://demo.webis.de/query-segmentation/approach/naive
```

Response

```
1004073900      graffiti|fonts alphabet
1004593125      stainless steel|chest freezers
1004933775      rutgers|online graduate classes
1005728827      review on breezes
```

References

[HPBS12] Matthias Hagen, Martin Potthast, Anna Beyer, and Benno Stein. Towards Optimum Query Segmentation: In Doubt Without. In Xuewen Chen, Guy Lebanon, Haixun Wang, and Mohammed J. Zaki, editors, *21st ACM International Conference on Information and Knowledge Management (CIKM 12)*, pages 1015–1024. ACM, October 2012.

[HPSB10] Matthias Hagen, Martin Potthast, Benno Stein, and Christof Bräutigam. The Power of Naïve Query Segmentation. In Fabio Crestani, Stéphane Marchand-Maillet, Hsin-Hsi Chen, Efthimis N. Efthimiadis, and Jacques Savoy, editors, *33rd International ACM Conference on Research and Development in Information Retrieval (SIGIR 10)*, pages 797–798. ACM, July 2010.

[HPSB11] Matthias Hagen, Martin Potthast, Benno Stein, and Christof Bräutigam. Query Segmentation Revisited. In Sadagopan Srinivasan, Krithi Ramamritham, Arun Kumar, M. P. Ravindra, Elisa Bertino, and Ravi Kumar, editors, *20th International Conference on World Wide Web (WWW 11)*, pages 97–106. ACM, March 2011.