Clef–Ip 2010
Submission Guidelines

Florina Piroi

Information Retrieval Facility
Vienna

Please refer to the Track's Guidelines for a description of the document collection, topic file structures, etc. available on the track's web and download sites [1,2].

The format of the Clef–Ip submissions is similar to the Trec format.

For all tasks, a submission consists of a single ASCII text file containing at most 1000 lines per topic, in the standard format used for most Trec submissions: five columns separated by whitespace.

Example of a (part of) Pac[1] task submission file:

```
PAC-100       Q0       EP-0000001-A1     1   1012
PAC-100       Q0       EP-0000002-A2     2   1011
PAC-100       Q0       EP-0000002-B2     3    999
```

Example of a (part of) Cls[2] task submission file:

```
CLS-100       Q0       A20K     1   3010
CLS-100       Q0       B32L     2   3008
CLS-100       Q0       B46A     3   2985
```

where the values shown in the third columns are only made–up examples, and not actual results. In the submission files

- the first column is the topic number (Pac-, respectively Cls-, followed by a number);
- the second column is the query number within that topic. This is to be set to Q0;
- the third column is
  a) for the Pac task, the document number of the retrieved document, *including* the kind code; this is the value of the `ucid` attribute in the `patent-document` Xml tag;

[1] Prior Art Candidates search task
[2] Classification task

b) for the Cls task, the subclass symbol to which the topic is assumed to belong to;

- the fourth column is the rank of the document retrieved, respectively the rank of the subclass symbol (maximum 1000, starting from 1);
- the fifth column shows the score (integer or floating point) that generated the ranking. This score must be in decreasing order.

A run is uniquely identified by the participant id, the method used, the task type and, in for Pac submissions, the size of the topic set on which the run was executed (all or small).

*participantID*_*method and/or runID*_`PAC`_`{all|small}`.*extension*
*participantID*_*method and/or runID*_`CLS`.*extension*

The maximum number of runs per task for a participant is 8. Submissions will be uploaded to an ftp server, access credential are to be communicated at a later time. Together with the run files, and for each of them, we require the participants to upload a concise description of the retrieval methods and document fields used to obtain the respective runs. The naming of the description files should follow the naming of the run files, where the group id, method and task type are part of the file name.

Evaluation Measures. We will mainly use the the `trec_eval` implementations of the precision, recall, F1, Map and nDcg measures. Further, we plan to try out new measures like, for example, Pres (see [3]). The relevance judgements to be used during the evaluation will be extracted automatically from the patent documents used to create the topics.

References

1. http://www.ir-facility.org/research/evaluation/clef-ip-10
2. https://download.ir-facility.org/clef/2010/
3. W. Magdy and G.J.F. Jones. Pres*: A Score Metric for Evaluating Recall–Oriented Information Retrieval Applications*. In Proceedings of Sigir'10, Geneva, Switzerland.