Automatically Delegates Computation to Established Tools
(decoupling semantics and implementation)

AP(\text{rel}=2)@10 via \texttt{pytrec_eval}[1]
\texttt{trec_eval qrels run -m map\_cut.20 -l2}

RBP(\text{p}=0.3, \text{rel}=2) via \texttt{cwl_eval}[2]
\texttt{cwl_eval.ruler.RBPCWLMetric(0.3)} (+ other handling for \text{rel}=2)

Also supports measures from gdeval, msmarco-eval, ndeval, ranx, trectools, and others:
\text{alpha}_nDCG, \text{Accuracy}, \text{AP\_IA}, \text{BPM}, \text{Bpref}, \text{Compat}, \text{ERR@k}, \text{ERR\_IA}, \text{infAP}, \text{INSQ}, \text{INST}, \text{IPrec}, \text{Judged@k}, \text{MAP}, \text{nDCG}, \text{NERR10}, \text{NERR11}, \text{NERR8}, \text{NERR9}, \text{nERR\_IA}, \text{nNRBP}, \text{NRBP}, \text{NumQ}, \text{NumRel}, \text{NumRet}, \text{P@k}, \text{P\_IA}, \text{R@k}, \text{RBP}, \text{Rprec}, \text{RR}, \text{SDCG@k}, \text{SetAP}, \text{SetF}, \text{SetP}, \text{SetR}, \text{StRecall@k}, \text{Success@k}

Several Convenient Interfaces

Example: \texttt{ir\_measures} computing TREC DL’s official measures:

(With \texttt{trec_eval}, requires multiple invocations)

\text{PyTerrier}[3]

\begin{align*}
\text{Old way: trec_eval with multiple invocations} \\
\$ \text{trec_eval qrels run -m ndcg\_cut.10} \\
\text{ndcg\_cut.10} 0.5536 \\
\text{trec_eval qrels run -m map -m recip\_rank -l2} \\
\text{recip\_rank} 0.6996 \\
\text{map} 0.3684
\end{align*}

\text{natural expression of measures, right in Python}

\begin{align*}
\text{natural expression of measures, right in Python} \\
\text{ir\_measures.calc\_aggregate}([\text{nDCG@10 \ RR(\text{rel}=2) \ AP(\text{rel}=2)}], \text{qrels}, \text{run}) \\
\{\text{nDCG@10}: 0.5536, \\
\text{RR(\text{rel}=2)}: 0.6996, \\
\text{AP(\text{rel}=2)}: 0.3684\}
\end{align*}

Also Integrated with other tools, including ir-datasets, OpenNIR, Experimaestro, DiffIR, and Patapsco.

References


Explore Measures:
Make changes to a ranking list and see the effect it has on a set of measures in real-time.
https://demo.ir-measur.es/explore

Reverse Measures:
Find example ranking lists that produce scores for a given measure.
https://demo.ir-measur.es/reverse