**Definition**:

**Dataset: all\_sessiondata\_2014**

**Session**: 一个session是一个topic(topic ID)和user(user ID)的组合，每当topiID或者userID产生变化时，则说明进入了一个新的session。Session level的data cleaning以session为单位：即，每一行（每一条record）代表一个session。(原理：topic和user是多对多的关系，一个user可以做多个topic，一个topic也可以有多个user做: 每一个unique combination of user and topic指向了一个session)。

**Query segment**：在一个session内，如果query(interaction.query)列发生了变化，则说明进入了一个新的query segment。一个session可能有一个query segment，也可能有多个query segments。 Query level的分析以每一个query segment为单位。

**Relevance judgment**: 在计算number of relevant documents, click precision等变量时，需要判断一个document的relevance。将interaction.results.result.clueweb12id中的ID和relevance judgment记事本的倒数第二列匹配，然后提取最后一列的relevance score。计算时relevance score要提前转化为binary (<=0为0，not relevant; >=1 为1，relevant)。

**注意**：在每个query segment中，数据将所有present在结果页上的page都作为record展示在了table中（interaction.results.result.Attribute:rank），然而只有在interaction.clicked.click.docno中的document才是被点击的（clicked）。一个query segment中，user可能没有click（则interaction.clicked.click.docno完全空白），可能click一个document，也可能click多个document。因此，clicked documentID会被重复或循环多遍。重复的ID只能算作同1个document (计算number of key documents, time to the first key document时除外)。

Topic product, topic goal, 以及task type三个指标需要将all\_sessiondata\_2014表格和TREC session track\_task classification\_2014表格相匹配获得。在最终的表格中直接放入原始的文本型变量即可，无需转化成数值型。

**OUTPUT needed:**

两个Table （csv格式即可，无需json），一个query level的table，一个session level的table。每个table需要分别包换对应的所有变量（如下）。注意：在query level的table中，每个query segment（每一行）所对应的所有session level的变量也需要被附加上。所以，对于属于同一个session的不同query segment，其session level的变量值应该完全相同。

**Query level**

**Topic number (topic.attribute.num)**

**Topic product (refer to TREC session track\_task classification\_2014 table)**

**Topic goal (refer to TREC session track\_task classification\_2014 table)**

**Task type (refer to TREC session track\_task classification\_2014 table)**

**Query (interaction.query)**

**Order number of the query under a given topic** (e.g., 1 represents the first query, 2 represents the second query)

**Query length (number of terms/words in the query, NOT number of characters!!!)**

**Query dwell time** (interaction.attribute.starttime of the next query- interaction.attribute.starttime)

**Number of clicked pages/documents in the query segment**

**Number of relevant documents clicked in the query segment**

**Number of key documents clicked in the query segment (key document: relevance score>1)**

**~~Average dwell time on every clicked document~~** ~~(for each clicked page, how to calculate dwell time: interaction.clicked.click.attribute.endtime – interaction.clicked.click.attribute.starttime)~~

**Total dwell time on clicked documents**

**Average ranking of clicks:** SUM of the rankings of all clicks/number of clicks (=0 if no click in this query segment) （refer to the column: interaction.clicked.click.rank）

**Click depth**: the ranking of the “deepest result”: max(interaction.click.click.rank) in the current query

**Dwell time on search engine result page** (query dwell time – Total dwell time on all clicked pages)

**PCTR: page/document click through rate**

* **PCTR@3**
* **PCTR@10**

**Click precision**: number of relevant**[[1]](#footnote-1)** documents clicked/number of pages clicked in total （relevance judgment: refer to relevance judgment document）

* **Click precision@3**
* **Click precision@10**

**Time to first click**: time delta between the start of the query segment and the first click (= query dwell time if there is no click)

**Time to last click**: time delta between the start of the query segment and the last click

**Exploitation/exploration ratio**: Total time spent on clicked documents/time spent on search engine result page

**~~Reciprocal Rank~~**~~: The rank position of the first relevant document for the current query.~~

**Discounted cumulative gain DCG@3 (next query segment)**

**DCG@10 (next query segment)**

**Rank biased precision RBP (0.1) (next query segment)**

**RBP (0.5) Next query segment)**

**RBP (0.8) (next query segment)**

**RBP (0.95) (next query segment)**

**Session level: (up to the current query segment: for example, if I am predicting the 4th query segment behavior, then I should calculate session-based information based on query segment 1-3).**

**Topic number** (topic.attribute.num)

**Topic product (refer to TREC session track\_task classification)**

**Topic goal (refer to TREC session track\_task classification)**

**Task type (refer to TREC session track\_task classification)**

**Number of queries in the session (total number of query segment in the session)**

**Average query length (all words used in all queries/number of queries in this session)**

**Max query length**

**Min query length**

**Average query similarity**

**Total query length** (all words used in all queries, NOT characters!!)

**Total number of unique terms in queries**

**Query vocabulary richness (QVR) = number of unique terms/total number of queries**

**Session completion time** (total dwell time on the whole session/all queries)

**Dwell time per query** (session completion time/number of queries)

**Max. query dwell time**

**Min. query dwell time**

**Total number of clicked documents**

**Number of relevant documents clicked in the session**

**Number of key documents clicked in the session (key document: relevance score>1)**

**Number of clicked documents per query**

**Max number of clicked documents in query**

**Min number of clicked documents in query**

**Number of relevant documents clicked per query** (Number of relevant documents clicked in the session/number of queries in the session)

**Average ranking of clicks:** SUM (average ranking of clicks in all queries)/number of queries

**Average click depth in the session (total values of depths in all queries/number of query)**

**Max click depth in query**

**Min click depth in query**

**Average dwell time on Search engine result page per query**

**Max dwell time on search engine result page in query**

**Min dwell time on search engine result page in query**

**Average dwell time on every clicked document**

**Max dwell time on clicked document**

**Min dwell time on clicked document**

**Average dwell time on clicked documents per query (total dwell time on content documents/number of queries)**

**Max dwell time on clicked documents in query**

**Min dwell time on clicked documents in query**

**Average Time to first click per query**

**Max time to first click in query**

**Min time to first click in query**

**average time to last click per query**

**Max time to first click in query**

**Min time to first click in query**

**Time to the first relevant document** (elapsed time from the beginning of the session to the first time when a user clicked a **relevant document (relevance score>0)**) (如果整个session没有relevant document，则=session completion time)

**Time to the first key document** (elapsed time from the beginning of the session to the first time when a user clicked a **document with the relevance score>1**) (如果整个session没有key document，则=session completion time)

**Average click precision (total score of click precisions/number of queries)**

**Max click precision in query**

**Min click precision in query**

**Prediction Ground Truth (as a prediction problem) Classifier: Lasso (Least Absolute Shrinkage and Selection Operator)**

Query length

Query similarity (between the last query and current, being predicted query)

Exploitation/exploration ratio

Query dwell time

Click depth

**(as a classification problem) Classifier: SVM, Random Forest, Naïve Bayes**

Query abandonment (click or no click)

**Reference-dependent model:**

Reference level 1: first query segment

Reference level 2: peak point up to the current segment

Reference level 3: average value up to the current segment

Reference level 4: the most recent current segment/last query segment

Reference level 5: (ref 2 + ref 4)/2

1. **Relevance score<=0: irrelevant, >0: relevant.** [↑](#footnote-ref-1)