**MS MARCO Dataset Notes**

**Queries**

* 1,010,916 query.
* Real user Bing searches.
* Filtered to only be questions, using a previously trained ML classifier.
* Broken down further into categories.
* Annoted by humans to see if they are answerable or not.

**Passages (context)**

* 8,841,823 passages.
* ~10 passages most relevant to the query.
* Selected using internal passage retrieval system at Bing.
* Human selects the passages they used to answer the query.
* is\_selected field indicates that a human selected the passage.

**Answers**

* 1,026,758 unique answers.
* Manually human-composed answers to queries based on passages.
* There can be zero or more answers to a question.
* 0 indicates the query could not be answered using the passages.
  + In our set, this appears as the list [‘No Answer Present.’]
* 1 indicates there was a valid answer to the query using the passages.
* >1 indicates there were multiple valid answers to the query using the passages.
* Humans instructed to read and understand the questions, inspect the ~10 passages, and then synthesize a natural language answer with the correct information.
* Answers were continually audited to ensure responses are of high quality.

**Well-Formed Answers**

* 182,669 well-formed answers.
* Some answers were re-written by an independent editor if:
  + The answer had improper grammar.
  + There is an indication the answer was copied from one of the passages.
  + The answer is not context-free (requires the query or passages as context to make sense).
* These are not present for all the answers, only as necessary.
  + In our set, this appears as the string ‘[]’
  + Despite empty being a string, non-empty entries are lists of strings.
* These effectively ‘override’ the contents of answers for this query.

**Documents**

* 3,563,535 web documents.
* Source documents for the selected ~10 passages. Includes:
  + URL
  + Body text
  + Title
* Roughly 300,000 documents could not be retrieved, because the web is not static.

**Question Type:**

* Each question was categorized into a type.
* Numeric: Query about some numeric information.
* Entity: Query about some non-person/location entity.
* Location: Query about some location.
* Person: Query about a person.
* Description: Query asking for a description of something.

**Data Structure:**

* Data is stored in a dict for each category.
* Each category is a dict, with the index of each element as string key.
* Eval set is a limited dataset, with only passages, query, query\_id, and query\_type categories. Passages category is also missing is\_selected.
* Dev and Training sets are full datasets, with the following categories: answers, wellFormedAnswers, passages, query, query\_id and query\_type. Passages category contains additional is\_selected flag.
* Dev is a subset of Training, with only 101,093 elements, as opposed to 808,731.