Devon Ankar
3/26/2018

DSE 6000 Assignment 2 Write Up

The following fields were created on the fly. The top-level bullets are fields and the second-level bullets are copyFields. Under each field is the critical analysis for that field.

- by_statement
  - by_statement_str

The field type is text, which is correct because the by statement is usually the name of an author(s) or organization (university, company, etc.), which needs to be formatted as a text string.

Indexed, Tokenized, Stored, and Multivalued all show green (checked), meaning this field was created correctly. This is probably because the by_statement on this entry in OpenLibrary is typical and representative of what by_statements usually consist of. For example, a number was not mistakenly entered in the by_statement; had it been, it may have been created as a numerical field.

- description
  - description_str

The field type is text, which is correct because the description is a brief explanation text of what is contained within a book. The 256-character limit could be a problem for some descriptions, if a long summary is written for a given book.

Indexed, Tokenized, Stored, and Multivalued all show green (checked), meaning this field was created correctly. Again, this is probably because the field is consistent with what such fields usually consist of.

- first_sentence
  - first_sentence_str

The field type is text, which is correct because the first sentence of a book is text. Indexed, Tokenized, Stored, and Multivalued all show green (checked), meaning this field was created correctly. Again, this is probably because the field is consistent with what such fields usually consist of.

- isbn_10
  - isbn_10_str

The field type is text, which is correct because the ISBN can sometimes contain letters, so it should be formatted as text. If it is formatted as a number, there will be parsing errors if some book’s ISBN contains numbers.

Indexed, Tokenized, Stored, and Multivalued all show green (checked), meaning this field was created correctly. Again, this is probably because the field is consistent with what such fields usually consist of.

- notes
  - notes_str

Devon Ankar
3/26/2018

DSE 6000 Assignment 2 Write Up

The field type is text, which is correct because the notes are text. Indexed, Tokenized, Stored, and Multivalued all show green (checked), meaning this field was created correctly. Again, this is probably because the field is consistent with what such fields usually consist of.

- number_of_pages

The field type is LongPointField, which is a numerical value. Page numbers should always be integers, so it is correct that decimal places are not needed. However, page numbers are not usually more than three digits, so if there is a smaller integer field available, that might be a better option if it would take up less storage space.

- publish_date

The field type is LongPointField, which is a numerical value. Years (in which a book is published) should always be integers, so it is correct that decimal places are not needed. However, years are no longer than four digits, so if there is a smaller integer field available, that might be a better option if it would take up less storage space.

- publishers
  - publishers_str

The field type is text, which is correct because the name of a publisher is text. Indexed, Tokenized, Stored, and Multivalued all show green (checked), meaning this field was created correctly. Again, this is probably because the field is consistent with what such fields usually consist of.

- subjects
  - subjects_str

The field type is text, which is correct because the subjects of a book are text. Indexed, Tokenized, Stored, and Multivalued all show green (checked), meaning this field was created correctly. Again, this is probably because the field is consistent with what such fields usually consist of.

- title
  - title_str

The field type is text, which is correct because the title of a book is text. Indexed, Tokenized, Stored, and Multivalued all show green (checked), meaning this field was created correctly. Again, this is probably because the field is consistent with what such fields usually consist of.

- table_of_contents.label

The field type is LongPointField, a number value, because the label is the field used to order items in the table of contents (it sounds like a text field, but it's not). It is correct because it is supposed to be a number. Labels should always be integers, so it is correct that decimal places are not needed. However, labels are no longer than one or two digits, so if there is a smaller integer field available, that might be a better option if it would take up less storage space.

- table_of_contents.level

The field type is LongPointField, a number value, because the level is the level of nesting of an item in the table of contents. It is correct because it is supposed to be a number. Levels should always be integers, so it is correct that decimal places are not needed. However, levels are only one digit, so if there is a smaller integer field available, that might be a better option if it would take up less storage space.

- table_of_contents.pagenum
    - table_of_contents.pagenum_str

The field type is Text, which *seems* incorrect because pagenum should be a number (it's a page number). However, looking at the HTML website on OpenLibrary, almost all the page numbers are numbers, but roman numerals are used on one occasion, so the field needs to be a text field to accommodate roman numerals when they are used. It is good that the first line item happened to be a roman numeral, otherwise the field would have been automatically generated as numbers and there would have been a parsing error if there was a roman numeral later.

- table_of_contents.title
    - table_of_contents.title_str

The field type is text, which is correct because the title of a section of the table of contents is text. Indexed, Tokenized, Stored, and Multivalued all show green (checked), meaning this field was created correctly. Again, this is probably because the field is consistent with what such fields usually consist of.

- table_of_contents.type.key
    - table_of_contents.type.key_str

The field type is text, which is correct. Indexed, Tokenized, Stored, and Multivalued all show green (checked), meaning this field was created correctly. Again, this is probably because the field is consistent with what such fields usually consist of.