Formal Validation of Bag of Words LOINC Program

Mizzou, MCW: Names

Introduction

Before the program results can be validated, make sure you have properly run the program on your files and the resulting CSV file is stored in an appropriate table. If you haven't run the program, consult the GitHub page for all the relevant information.

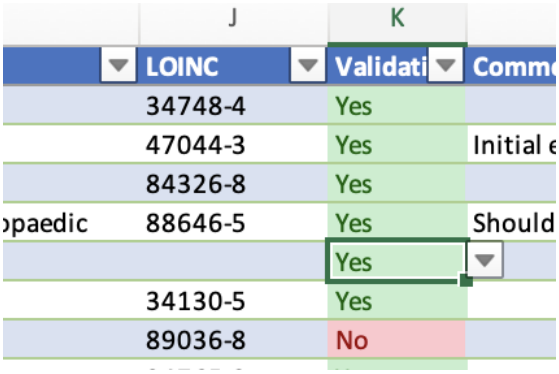
For this validation, run the SQL code provided below. It should look at the resulting table from the program and randomly extract 200 rows from among the top 1000 most common metadata column combinations. Run the query and sum the *Count* column in order to ensure sufficient document count (around 10-20 million). Rerun the query until a high enough document count is reached.

SQL CODE

SELECT * FROM (SELECT * FROM *TABLE_WHERE_PROGRAM_OUTPUT_IS_STORED* ORDER BY COUNT DESC LIMIT 1000) SAMPLE (200 ROWS) ORDER BY COUNT DESC;

Guided Steps

1. Run the program on clinical note's metadata.
2. Store resulting CSV file as a table in SQL database.
3. Run SQL Code provided above (changing the placeholder table name)
    a. Sum all values in Count column to ensure sufficient document count.
    b. Repeat if document count is too low, otherwise move on to next step.
4. Download the resulting table as a CSV/Excel file.
5. Open the file using Excel and add additional columns at the end for validation.
    a. Add a "Validation" column, used for storing the decision for validation
    b. Add a "Comment" column. This may be used for notes or to make comments for discussion and future fixes.
    c. You may use Excel's validation functionality to add dropdowns and syntax highlighting for easier visualization (See image below)



6. Optionally, create another set of validation/comment columns for a second validator.

7. Go down the individual rows and compare the LOINC Code/LOINC Long Name to the actual metadata and consider the following:
    a. Is it correct or incorrect...
        i. Some Criteria for incorrect mapping are as follows:
            1. The LOINC Long Name includes information that isn't present or cannot be derived based on the note's metadata
            2. The Metadata contains multiple Subject Matter Domain or multiple Note types, but the LOINC Code only selected one of them. In this case, both SMDs should be skipped for a more general LOINC Code.
            3. Incorrect priority given to certain information. i.e., a 'telephone note' should not be labeled as "operative surgery notes" even if the information may be found in the metadata columns
            4. More possible patterns will be added as they are found.
    b. **If it is too general, the current decision is to mark it as correct**. We can also make a note in the comments section where we can either:
        i. Update the synonymy list to improve precision.
        ii. Update the hierarchy list for priority. (*Hospital Note* should have lower priority than *telephone note*, even though both have same number of dimensions filled... etc...)
8. Once done, a second pass may be made by a second approver.
9. Lastly, calculate the accuracy of validation using the validation columns.
    a. Validate based on the number of *Notes*.
        i. Get the total number of notes covered by the 200 rows.
        ii. Get the total number of notes using the valid rows.
        iii. Calculate the percentage of valid notes.
    b. Validate also based on the number of *Rows*.
        i. Get the count of the number of valid rows based on validation column.
        ii. Simply divide the number of valid rows by total rows (200 rows)

University of Missouri Validation Efforts

For the validation efforts at the University of Missouri, Columbia, our 200 rows covered around 15.2 million notes. After a two people validation effort, the values were consolidated. **167 of the 200 rows** were valid, covering around **13.9 million** notes. This is around **83.5%** based on row count and **91.6%** based on notes count. We also calculated precision and recall values, which came to **0.954** and **0.913**, respectively. The F1 score was **0.933**.

Medical College of Wisconsin Validation Efforts

For the validation efforts at the ___ , our 200 rows covered around ___ million notes. After a two people validation effort, the values were consolidated. ___ **of the 200 rows** were valid, covering around ___ notes. This is around ___ based on row count and ___ based on notes count. We also calculated precision and recall values, which came to ___ and ___ , respectively. The F1 score was ___.