Masking customer data for data mining w/ GDPR concerns

Henry Zhu

Special thanks to my mentor Heheng Li and my manager David Chen

Background

On May 2018, the EU introduced GDPR, which protects a user's PII (personally identifiable information) from being accessed and used by companies.

My project reads in unmasked data containing information about Microsoft employees and protects the employees' PII.

Example: REDMOND\itatstsv > REDMOND\xavaeaez

Constraints

1. Keep non-PII part of the file unchanged
2. Ensure obfuscated data still looks similar to original
3. Ensure that PII cannot be predicted from masked data

Program Flow

1. Read masking specifications from .json file
2. Read unmasked data file and map unmasked to masked data
3. Replace unmasked data w/ masked data
4. Output masked

Masking Method: Radix

1. Keep track of a counter that increments each time a new unmasked string (e.x. REDMOND\itatsv) is encountered.
2. Add trailing zeros to the counter number so that it has the same length as the unmasked string (e.x. REDMOND\ita3sv > 00000000000001
3. Replace each character in the number w/ its corresponding letter in the alphabet.
4. Leave separators (\, /, -, _, etc.) as they are, and if the character of an unmasked string is a digit, replace it with a different digit
5. Final: REDMOND\ita3sv > aaaaaaa\aaa3ab

Masking Method: Random

1. Generate a random mapping of letters to letters and digits to digits.
2. Replace each letter in the string with its corresponding random letter.
3. Replace each digit in the string with its corresponding random digit.
4. Leave separators (\, /, -, _, etc.) as they are.
5. REDMOND\itatstsv > xostwvs\xavaeaez

Future Improvements

1. Partial masking (REDMOND is not PII)

   Example: REDMOND\itasv3 > REDMOND\aaaaa5

2. Maintain capitalization
3. Utilize foreign keys
4. Creating a GUI

Appendix

JSON Specifications: **categoryIndexes** indicate which columns are to be masked, and **method** indicates what strategy of masking is to be used.

```
{

    "categoryIndexes": [ 2, 3, 4, 5, 6, 7, 8, 9, 11, 16 ],

    "method": "random"

}
```