### Using NLP for Anonymizing Telco Data

#### Detailed workflow

<table>
<thead>
<tr>
<th>Week</th>
<th>Task</th>
<th>Status</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-May</td>
<td>Study Work: State of art on the models, optimization and Evaluation</td>
<td>Done</td>
<td>Look for optimization techniques, how they evaluate anonymization models.</td>
</tr>
<tr>
<td>27-May</td>
<td>Finalizing Dataset and Libraries to use -- suppression/\rename/ .. etc.</td>
<td>Done</td>
<td>Kubernetes logs/Metrics, Openstack logs/metrics .. any data that has PII information</td>
</tr>
<tr>
<td>3-June</td>
<td>Anonymization Impact on the Model’s utility</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10-June</td>
<td>Containeration and the APIs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17-June</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24-June</td>
<td>Automation using Python</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-July</td>
<td>Testing of the containerized Architecture</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8-July</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15-July</td>
<td>NLP Model for anonymizing Telco Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22-July</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29-July</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-Aug</td>
<td>Evaluation of the Model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12-Aug</td>
<td>Integration of the developed model with the architecture</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19-Aug</td>
<td>Documentation and release of the code</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26-Aug</td>
<td>[BUFFER]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Proposed architecture:

![Proposed architecture diagram]

#### API end-points:

1. Tell where the Raw-Data exists (file-path, url, etc.)
2. Start the anonymization process.
3. Tell where to put the anonymized data (file-path, url, etc.)
4. Receive notification once anonymization is complete (SUCCESS or ERROR) 1 and 4 can be just configuration.

#### Commonly existing data anonymization techniques:
State of the art models for anonymizing textual data:

1. Named Entity Recognition (NER) based models:
   - These models are trained to identify and classify named entities within text, such as people's names, locations, and organizations. Popular frameworks include spaCy and NLTK.
   - Once identified, PII entities can be replaced with anonymized tokens (like "[NAME]") or masked with techniques like character-level redaction (e.g., "Jo** ***th").

2. Rule-based systems:
   - While simpler, rule-based systems can be effective for specific use cases.
   - These systems rely on predefined rules and regular expressions to identify PII based on patterns (e.g., phone number formats, email address structures).

3. Presidio:
   - Provides a user-friendly interface for defining custom PII analyzers.
   - It can then be anonymized using the pre-built anonymization pipeline.

There are dozens of softwares and APIs in the market for anonymization working on these three techniques under the hood.

Ways of evaluating anonymization models:

There are 2 basic methods for evaluation of anonymization models namely, the degree of anonymization and the decrease in the utility of the text.

1. Precision and Recall: These metrics are commonly used to assess the performance of NLP models in text anonymization. Precision measures the proportion of correctly anonymized information among all the information that the model labeled as sensitive, while recall measures the proportion of correctly anonymized information among all the sensitive information present in the text.
2. F1 Score: The F1 score provides a balanced evaluation of the model's performance in anonymizing text data. It considers both false positives and false negatives, offering an assessment of the model's effectiveness.
3. But we need to have the ground truth for testing the validity of the models using the above methods.
4. To test the decrease in the utility of the text, one way is to train a model before anonymization and to train again after anonymization to check the difference in the performance. Lesser the difference, better the anonymization process.
5. Human Evaluations: Human evaluations involve experts assessing the anonymized documents for re-identification risks and data utility preservation.

Reference Research papers:

1. https://aclanthology.org/2021.acl-long.323.pdf (Showcases the problems and the evaluation methodology for anonymization models)
Datasets:

Key-points:

- Although, the logs data in themselves do not contain too much PIIs, but when combined with datas of equal size can yield a well suited data for anonymization problem.
- I found a supermarket dataset consisting of nearly all possible PIIs that exists and another factor for choosing it was the feasibility of evaluating the depreciation in the model’s prediction and performance. [https://data.world/2918diy/global-superstore](https://data.world/2918diy/global-superstore) Evaluation can be done via predictive models such as:
  - Segment, target high-value customers.
  - Predict future sales, optimize pricing.
  - Recommend products, personalize experience.
- I also found many tele communications and relevant dataset that can be taken into consideration for anonymization, but with introduction of certain PIIs:
  - [https://www.kaggle.com/datasets/shimul/graphical-mapping-dataset](https://www.kaggle.com/datasets/shimul/graphical-mapping-dataset): Location specific data
  - [https://www.kaggle.com/datasets/shimul/graphical-mapping-dataset](https://www.kaggle.com/datasets/shimul/graphical-mapping-dataset): Location specific data
  - [https://www.kaggle.com/datasets/uciml/adult-census-income](https://www.kaggle.com/datasets/uciml/adult-census-income): name, age, relation, race, education, occupation, income ( ideal for evaluation)

Libraries and Methods:

Methods:

- **Suppression**: This removes sensitive information entirely.
  - **Advantages**: Simple, strong anonymization.
  - **Disadvantages**: Data loss, may affect analysis depending on what's removed.
  - **Impact on Models**: Significant degradation, especially if removing features crucial for prediction.

- **Pseudonymization**: Replaces sensitive data with fictitious identifiers.
  - **Advantages**: Preserves data structure, allows some analysis.
  - **Disadvantages**: Not truly anonymous, re-identification risk with complex data.
  - **Impact on Models**: Varies depending on replaced data. May require model retraining.

- **Generalization**: Replaces specific details with broader categories. (“John” -> “Male”).
  - **Advantages**: Balances privacy and usability, less data loss than suppression.
  - **Disadvantages**: May introduce bias or reduce information value for models.
  - **Impact on Models**: Moderate degradation depending on the level of generalization. Retraining might be needed.

- **Tokenization with Masking**: Replaces sensitive tokens (words/phrases) with symbols (****).
  - **Advantages**: Easy to implement, protects specific data points.
  - **Disadvantages**: Limited protection for contextual information, may affect readability.
  - **Impact on Models**: Varies depending on masked tokens. May require feature engineering for models.

- **Differential Privacy**: Adds controlled noise to data to achieve statistical protection.
  - **Advantages**: Strong privacy guaranteed, allows some analysis with provable privacy bounds.
  - **Disadvantages**: Complex implementation, can significantly impact data utility for models.
  - **Impact on Models**: High potential for degradation due to added noise. Models might require significant adjustments.

Libraries:

Here are some popular libraries that implement these methods:

- **Presidio (Python)**: Open-source library for identifying and anonymizing entities like names, locations, and dates. ([https://github.com/microsoft/presidio](https://github.com/microsoft/presidio))
- **spaCy (Python)**: Powerful NLP library with built-in named entity recognition capabilities for anonymization tasks. ([https://spacy.io/](https://spacy.io/))
- **Text Anonymizer (Python)**: Framework offering various anonymization techniques like suppression and generalization. ([https://medium.com/@openredact/anonymizer-a-framework-for-text-anonymization-499855f639d4](https://medium.com/@openredact/anonymizer-a-framework-for-text-anonymization-499855f639d4))
- **ARX (Java)**: Open-source suite for anonymizing various data types, including text, with features like k-anonymity. ([https://github.com/topics/data-anonymization?o=desc&s=updated](https://github.com/topics/data-anonymization?o=desc&s=updated)) Haven't explored this much yet.

Evaluation:

Evaluating the effectiveness of anonymization will involve a trade-off between:

- **Level of Anonymization**: How well identities are protected.
- **Data Utility**: How well the anonymized data retains its usefulness for predictive analysis or models.

1. Metrics like precision, recall, and F1-score can be used to assess how well the method identifies sensitive information.
   - [https://github.com/anonymous-NLP/anonymisation/blob/main/aggregated_annotations.pdf](https://github.com/anonymous-NLP/anonymisation/blob/main/aggregated_annotations.pdf) I also thought of to somehow compare the anonymization with the one given so as to have a valid approval for the model's performance.
2. However, the impact on models requires domain-specific evaluation. Some approaches that I will follow are:
   - a. **Compare model performance**: Train and test models on original and anonymized data to see the accuracy drop.
b. **Evaluate information loss**: Measure how much relevant information is lost due to anonymization.