AI Explainability WG

September 2023

September 2023 updates

- TrustyAI service
    - KServe integration
    - Drift detection
- Core
    - MeanShift
- Python TrustyAI
    - Python 3.8 notebook images
    - TrustyAI 0.3.0 notebook images
- Community
    - TrustyAI console

Python TrustyAI

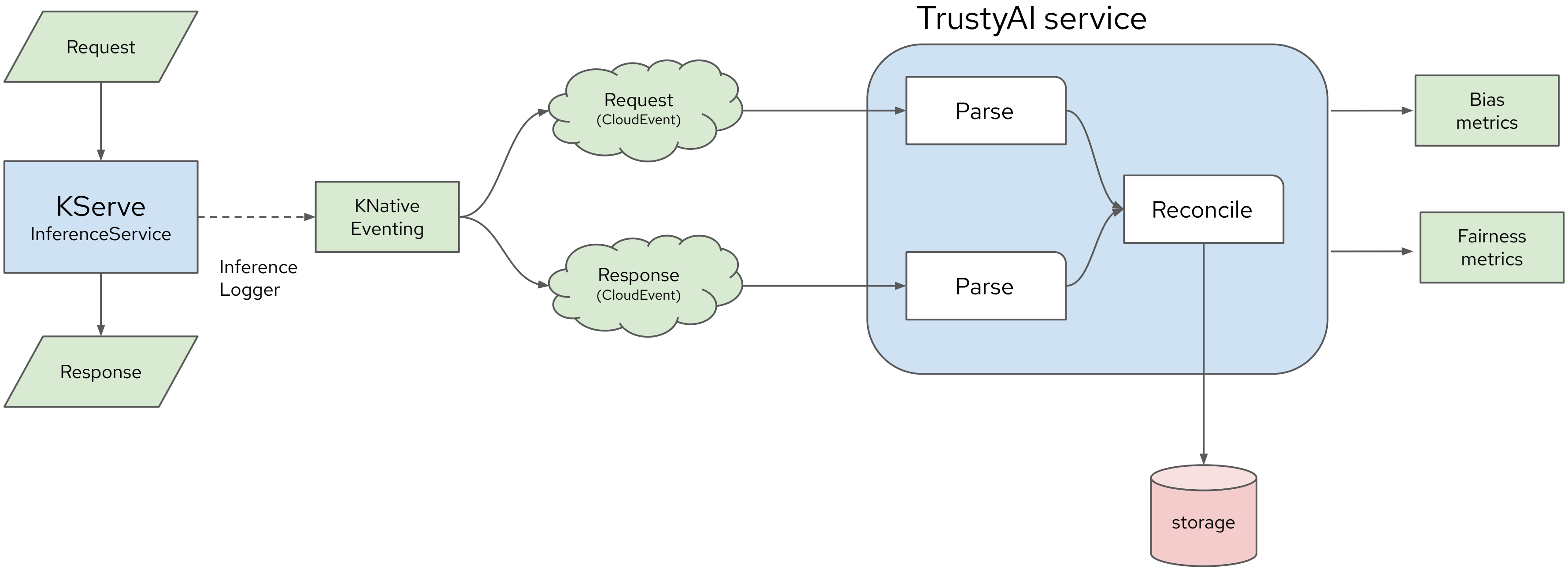
Python TrustyAI

- Python 3.8 ODH notebook images
- TrustyAI 0.3.0 notebook images[1]

[1]: https://github.com/opendatahub-io/notebooks/pull/209

TrustyAI core / service

KServe integration



KServe integration

```yaml
apiVersion: "serving.kserve.io/v1beta1"
kind: "InferenceService"
metadata:
  name: "my-model"
spec:
  predictor:
    logger:
      mode: all
      url: http://trustyai-service
    model:
      modelFormat:
        name: sklearn
      storageUri: …
```

Drift - MeanShift

- MeanShift t-test added to TrustyAI core[1]
- Endpoint in service to calculate MeanShit[2]
- Data upload endpoint
  - Tagging support
- Data download endpoint

  -
    ```
    {
      "modelId": "name-age-nationality",
      "matchAll": [
        {"columnName": "age", "operation": "BETWEEN", "values": [50, 75]},
        {"columnName": "nationality", "operation": "EQUALS", "values": ["Italian", "French"]},
      }], … }
    ```

[1]: https://github.com/trustyai-explainability/trustyai-explainability/pull/352
[2]: https://github.com/trustyai-explainability/trustyai-explainability/pull/342

TrustyAI Console

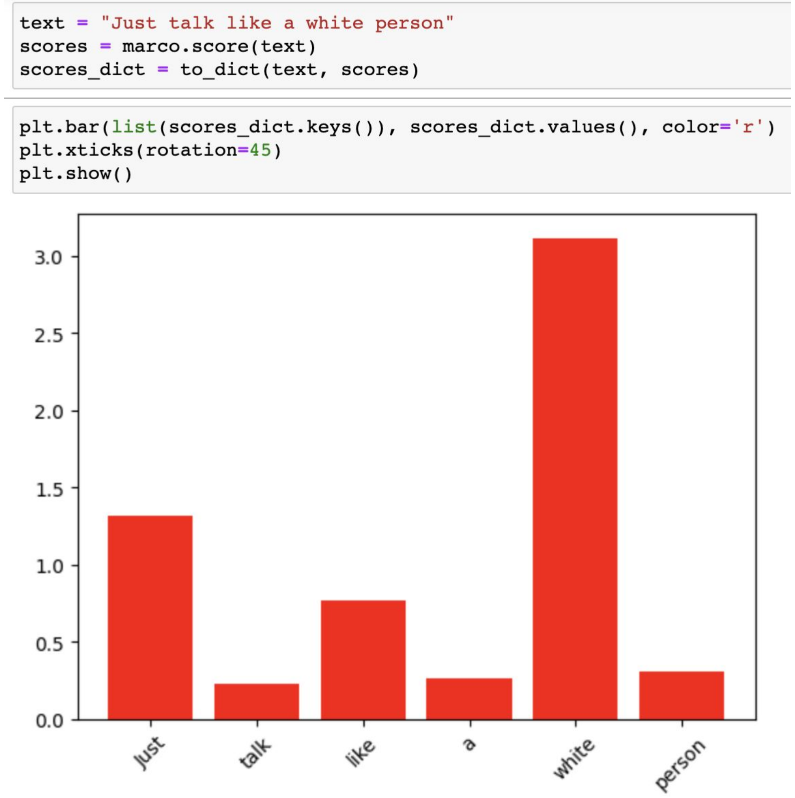
TrustyAI console

- Introducing a new way for users to interact with Trusty AI
- An open source GUI for configuring, monitoring and visualising XAI metrics
- Integrates with the TrustyAI Service and Grafana
- Initial features limited to displaying scheduled metrics configurations and charts
- Future plans include creation of scheduled metrics and more.

Toxic Content in (Large) Language Models

Toxic Content in LLMs

- Make it possible to just “score” tokens for experts’ disagreement
- More efficient logits generation
- Bi-directional rephrasing (in progress)
    - https://github.com/trustyai-explainability/trustyai-explainability/issues/346



```
text = "Just talk like a white person"
scores = marco.score(text)
scores_dict = to_dict(text, scores)
```

```
plt.bar(list(scores_dict.keys()), scores_dict.values(), color='r')
plt.xticks(rotation=45)
plt.show()
```

Community

Releases

- TrustyAI core / service
  - v0.4.0 released Sept 11th[1]
  - Available on Quay.io - `quay.io/trustyai/trustyai-service:0.4.0`
  - v0.5.0 planned release date - Sept 29th
- TrustyAI operator
  - v1.9.0 released Sept 11th[2]
  - Available on Quay.io - `quay.io/trustyai/trustyai-service-operator:1.9.0`
  - v1.10.0 planned release date - Sept 29th

[1]: https://github.com/trustyai-explainability/trustyai-explainability/releases/tag/v0.4.0
[2]: https://github.com/trustyai-explainability/trustyai-service-operator/releases/tag/v1.9.0

Roadmap

TrustyAI 2023 roadmap

July 2023

- *Explainers*
  - Support for explainers LIME, SHAP, CF at service level
- *Metrics*
  - Flexible scheduling/batching
  - Improve service metadata endpoints
    - Include available categories
- *Operator*
  - TrustyAI Operator v1
- *Explainers*
  - Support for external explainability libraries

September 2023

- *Storage*
  - **Wider storage support** (database backends)
- *Metrics*
  - **Additional metrics**
  - Metrics statistical tests
- **KServe integration**
- **Drift detection**
- **HAP/PII**

December 2023

- *Storage*
  - Wider storage support (database backends)
- *Explainers*
  - NLP explainability support
- *Metrics*
  - Support for user-defined historical windows