AI Explainability WG

December 2023

December 2023 updates

- TrustyAI core / service
  - 0.9.0 release
  - https://github.com/trustyai-explainability/trustyai-explainability/releases/tag/v0.9.0
  - quay.io/trustyai/trustyai-service:v0.9.0
- TrustyAI operator
  - 1.14.0 release
  - https://github.com/trustyai-explainability/trustyai-service-operator/releases/tag/v1.14.0
  - quay.io/trustyai/trustyai-service-operator:v1.14.0
- TrustyAI Python

TrustyAI core / service

TrustyAI core / service

0.9.0 - What's new?

- Available on GitHub and Quay.io
  - https://github.com/trustyai-explainability/trustyai-explainability/releases/tag/v0.9.0
  - quay.io/trustyai/trustyai-service:v0.9.0
- Language performance metrics
  - Added Levenshtein, WEP, MER, WIL, WIP, BLEU, Fuzzy-Exact Match
  - Added Gaussian Anomaly Detection functions
- Fixes / improvements
  - Fix for KServe v2 protocol payload parsing
  - Include prediction identifiers, tags, timestamps in /data/download endpoint
  - Allow for multiple selections in bias metrics
  - Documentation updates

TrustyAI core / service

Language performance metrics

- Word Error Rate (WER)
  - Percentage of errors at the word level compared to a reference text
  - Related metrics
    - Match Error Rate (MER)
    - Word Information Lost (WIL)
    - Word Information Preserved (WIP)
- Bilingual Evaluation Understudy (BLEU)
  - Quantify similarity between model text output to a set of high-quality references
- Exact Match
  - Correctness metric for (fuzzy) exact matches
- Levenshtein distance
  - Token-based and character-based distance metric

TrustyAI core / service

Gaussian Anomaly Detection functions

- Bounded probability calculation
  - How likely is a particular feature value, +/- a window, to be drawn from the feature's distribution?
- Normalized Bounded probability calculation
  - How likely is a particular feature value, +/- a window, to be drawn from the feature's distribution, normalized by the maximum possible likelihood for that window size?
- Normalized Deviation
  - How many standard deviations away from the column mean is a particular feature value?

TrustyAI operator

TrustyAI operator

1.14.0 - What's new?

- Available on GitHub and Quay.io
  - https://github.com/trustyai-explainability/trustyai-service-operator/releases/tag/v1.14.0
  - quay.io/trustyai/trustyai-service-operator:v1.14.0
- Authentication
  - Added authentication to service's external endpoints (oauth-proxy)
- Dependency updates
  - Container base image update (ubi-minimal:8.7 → 8.9)
- Fixes / improvements
  - TrustyAI editor and viewer roles (non-admin user)
  - KServe event check
  - Expanded CI and unit tests

TrustyAI operator

Authentication

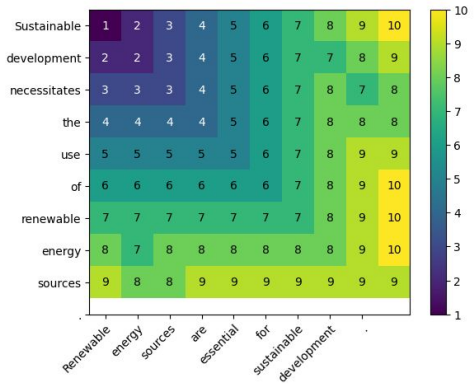
- OAuth authentication for external endpoints
  - `TOKEN=$(oc whoami -t)`
  - `curl -H “Authorization: Bearer ${TOKEN}” https://trustyai-service…`
- In progress
  - Support for self-signed certificates

Python TrustyAI

Python TrustyAI

Language performance metrics

- Added support for Levenshtein and WER
- Visualisation for distance matrices
  - ```
    from trustyai.metrics.distance import levenshtein
    A = "Renewable energy sources are essential for sustainable development."
    B = "Sustainable development necessitates the use of renewable energy sources."
    d = levenshtein(A, B)
    d.plot()
    ```


    ```

Python TrustyAI

Detoxification (HAP)

- Added *rephrase* + *self-reflection* capability
  - https://github.com/trustyai-explainability/trustyai-detoxify/pull/1
- Code hardening (WIP)
  - Custom models, more robust parsing, etc.
- Added **trustyai-detoxify** repo
  - https://github.com/trustyai-explainability/trustyai-detoxify/
  - Will soon move into **trustyai-explainability-python** as an optional package
    - `pip install trustyai[detoxify]`

Community

Documentation

- New TrustyAI documentation page
  - Central location for TrustyAI documentation
  - Antora (Asciidoc based)
  - https://trustyai-explainability.github.io/trustyai-site/0.8.0/main.html
  - Source
    - https://github.com/trustyai-explainability/trustyai-explainability.github.io

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
- ODH v2 onboarding

December 2023

- *Storage*
  - Wider storage support (database backends)
- *Explainers*
  - NLP explainability support
  - Language metrics (WER, BLUE, EM)
- *Detection*
  - HAP/PII
- *Metrics*
  - Support for user-defined historical windows