Tech Debt Transformer Pipeline

David Gao

Recap: Transformers!

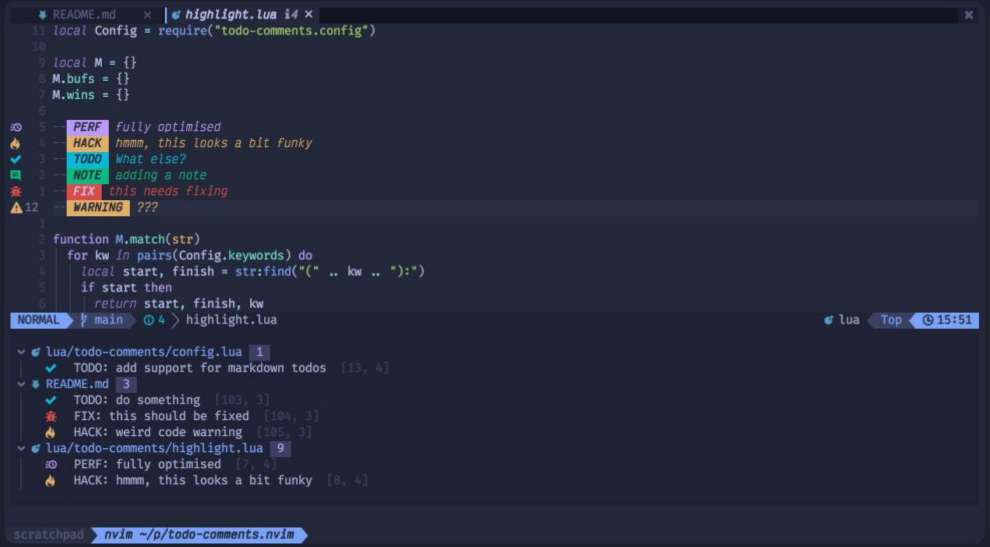
A good way to capture a lot of information!

Classification

Generation

Recap: Technical Debt

Definition: “Implied cost of future reworking required when choosing an easy but limited solution instead of a better approach that could take more time” Wikipedia.



```
local Config = require("todo-comments.config")

local M = {}
M.bufs = {}
M.wins = {}

-- PERF fully optimised
-- HACK hmmm, this looks a bit funky
-- TODO What else?
-- NOTE adding a note
-- FIX this needs fixing
-- WARNING ???

function M.match(str)
  for kw in pairs(Config.keywords) do
    local start, finish = str:find("(" .. kw .. "):")
    if start then
      return start, finish, kw
```

```
lua/todo-comments/config.lua 1
    TODO: add support for markdown todos [13, 4]
README.md 3
    TODO: do something [103, 3]
    FIX: this should be fixed [104, 3]
    HACK: weird code warning [105, 3]
lua/todo-comments/highlight.lua 9
    PERF: fully optimised [7, 4]
    HACK: hmmm, this looks a bit funky [8, 4]
```

Existing Solutions, Their Limitations, and Novelty

- Existing solutions usually utilize rules to find technical debt (more have started to use machine learning recently)
- Only one transformer based approach: finding technical debt in issue trackers. (Skryseth, Daniel, et al.)

- This study will be focused on directly looking at code from git commits to get insights into technical debt!

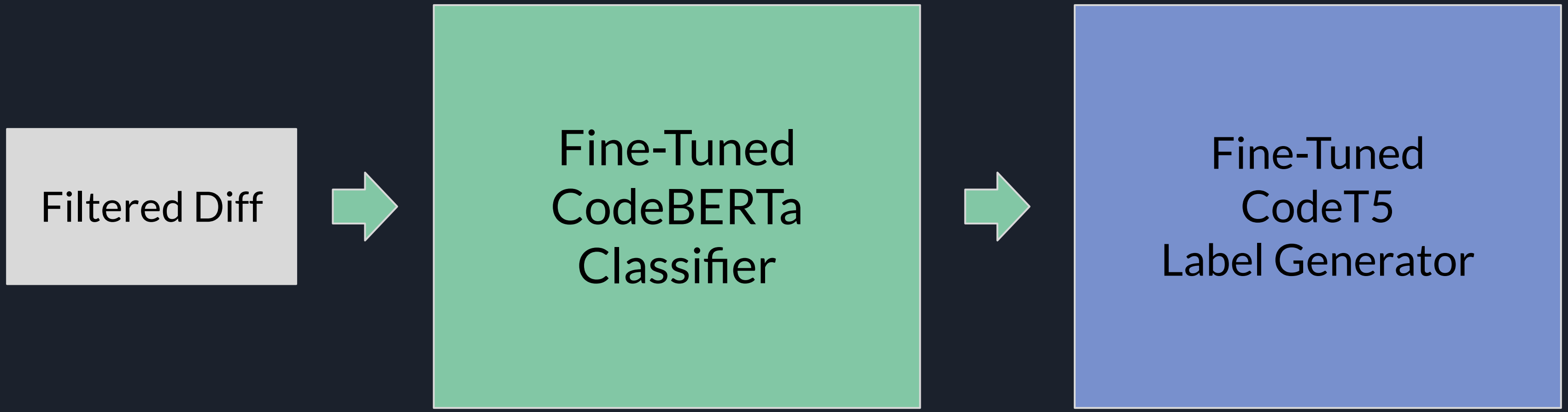
Problem Statement

In open source projects, managing technical debt can be a challenging process. This study explores how fine-tuned transformer models can help improve the identification and classification of these issues.

Research Questions

1. How well can a fine-tuned transformer model classify technical debt, given code from a commit?
2. How well can a fine-tuned transformer model generate labels for technical debt?

Proposed Architecture


Filtered Diff
Fine-Tuned CodeBERTa Classifier
Fine-Tuned CodeT5 Label Generator

The Technical Debt Dataset

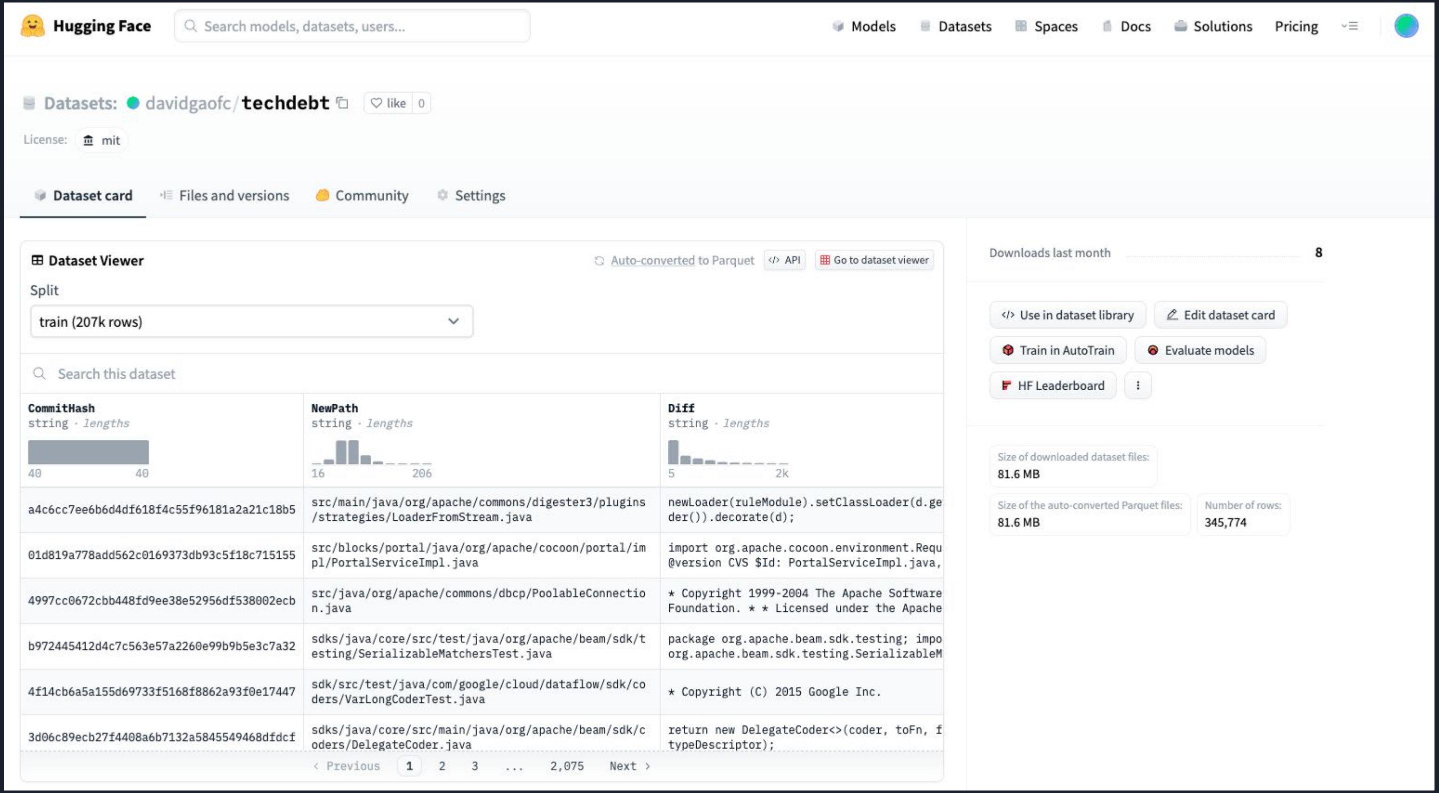
- 78K commits from 33 Java Projects
- SZZ algorithm annotations
  - find the fix, trace backwards
- SonarQube
  - static code analysis, rule based

Classification Task

How well can a fine-tuned transformer model classify technical debt, given code from a commit?

Metrics: Accuracy, F1 Score

Preliminary Work



Datasets: davidgaofc/techdebt ♡ like 0

License: mit

Dataset Viewer

Split: train (207k rows)

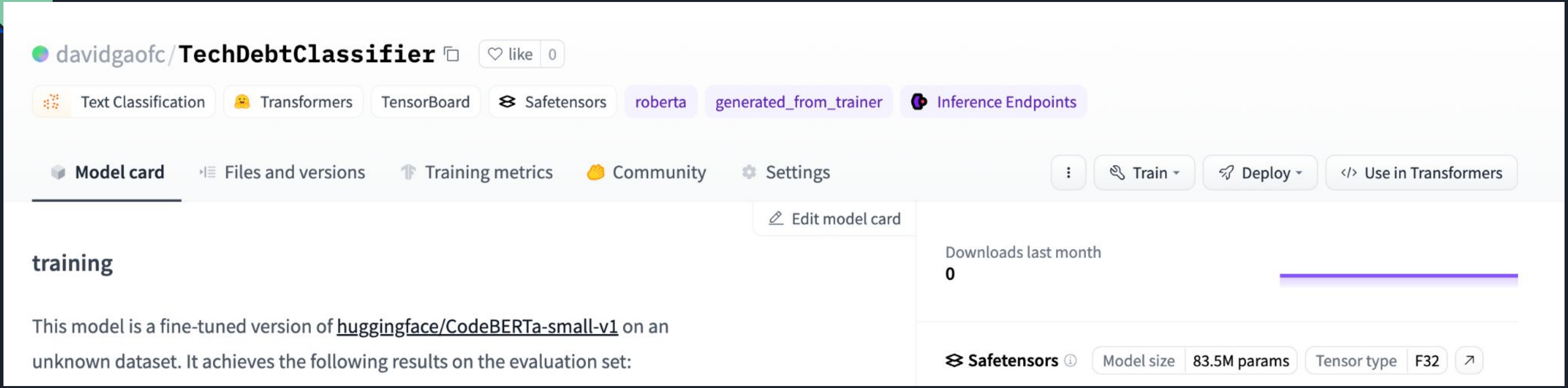
CommitHash string · lengths 40 – 40	NewPath string · lengths 16 – 206	Diff string · lengths 5 – 2k
a4c6cc7ee6b6d4df618f4c55f96181a2a21c18b5	src/main/java/org/apache/commons/digester3/plugins/strategies/LoaderFromStream.java	newLoader(ruleModule).setClassLoader(d.ge der()).decorate(d);
01d819a778add562c0169373db93c5f18c715155	src/blocks/portal/java/org/apache/cocoon/portal/impl/PortalServiceImpl.java	import org.apache.cocoon.environment.Requ @version CVS $Id: PortalServiceImpl.java,
4997cc0672cbb448fd9ee38e52956df538002ecb	src/java/org/apache/commons/dbcp/PoolableConnection.java	* Copyright 1999-2004 The Apache Software Foundation. * * Licensed under the Apache
b972445412d4c7c563e57a2260e99b9b5e3c7a32	sdks/java/core/src/test/java/org/apache/beam/sdk/testing/SerializableMatchersTest.java	package org.apache.beam.sdk.testing; impo org.apache.beam.sdk.testing.SerializableM
4f14cb6a5a155d69733f5168f8862a93f0e17447	sdk/src/test/java/com/google/cloud/dataflow/sdk/coders/VarLongCoderTest.java	* Copyright (C) 2015 Google Inc.
3d06c89ecb27f4408a6b7132a5845549468dfdcf	sdks/java/core/src/main/java/org/apache/beam/sdk/coders/DelegateCoder.java	return new DelegateCoder<>(coder, toFn, f typeDescriptor);

‹ Previous 1 2 3 ... 2,075 Next ›

Downloads last month 8

Size of downloaded dataset files:	81.6 MB
Size of the auto-converted Parquet files:	81.6 MB
Number of rows:	345,774

Preliminary Work (continued)



davidgaofc/**TechDebtClassifier** like 0

Text Classification | Transformers | TensorBoard | Safetensors | roberta | generated_from_trainer | Inference Endpoints

Model card | Files and versions | Training metrics | Community | Settings | Train | Deploy | Use in Transformers

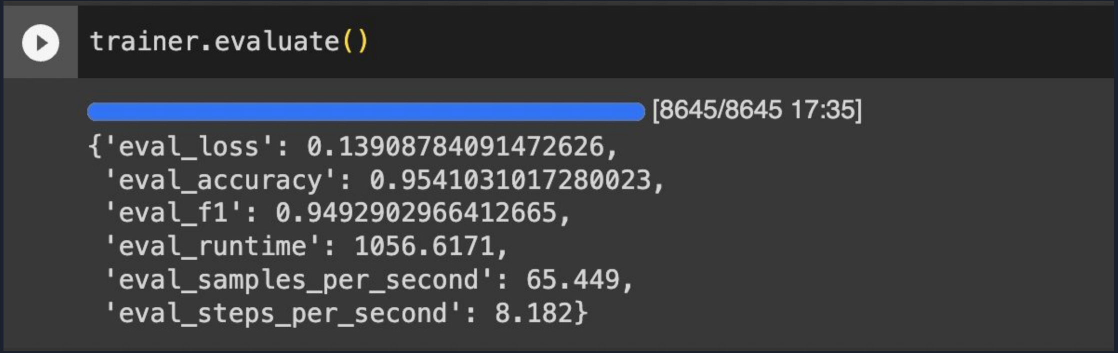
Edit model card

training

This model is a fine-tuned version of huggingface/CodeBERTa-small-v1 on an unknown dataset. It achieves the following results on the evaluation set:

Downloads last month
0

Safetensors | Model size 83.5M params | Tensor type F32



```
trainer.evaluate()
```

```
[8645/8645 17:35]
{'eval_loss': 0.13908784091472626,
 'eval_accuracy': 0.9541031017280023,
 'eval_f1': 0.9492902966412665,
 'eval_runtime': 1056.6171,
 'eval_samples_per_second': 65.449,
 'eval_steps_per_second': 8.182}
```

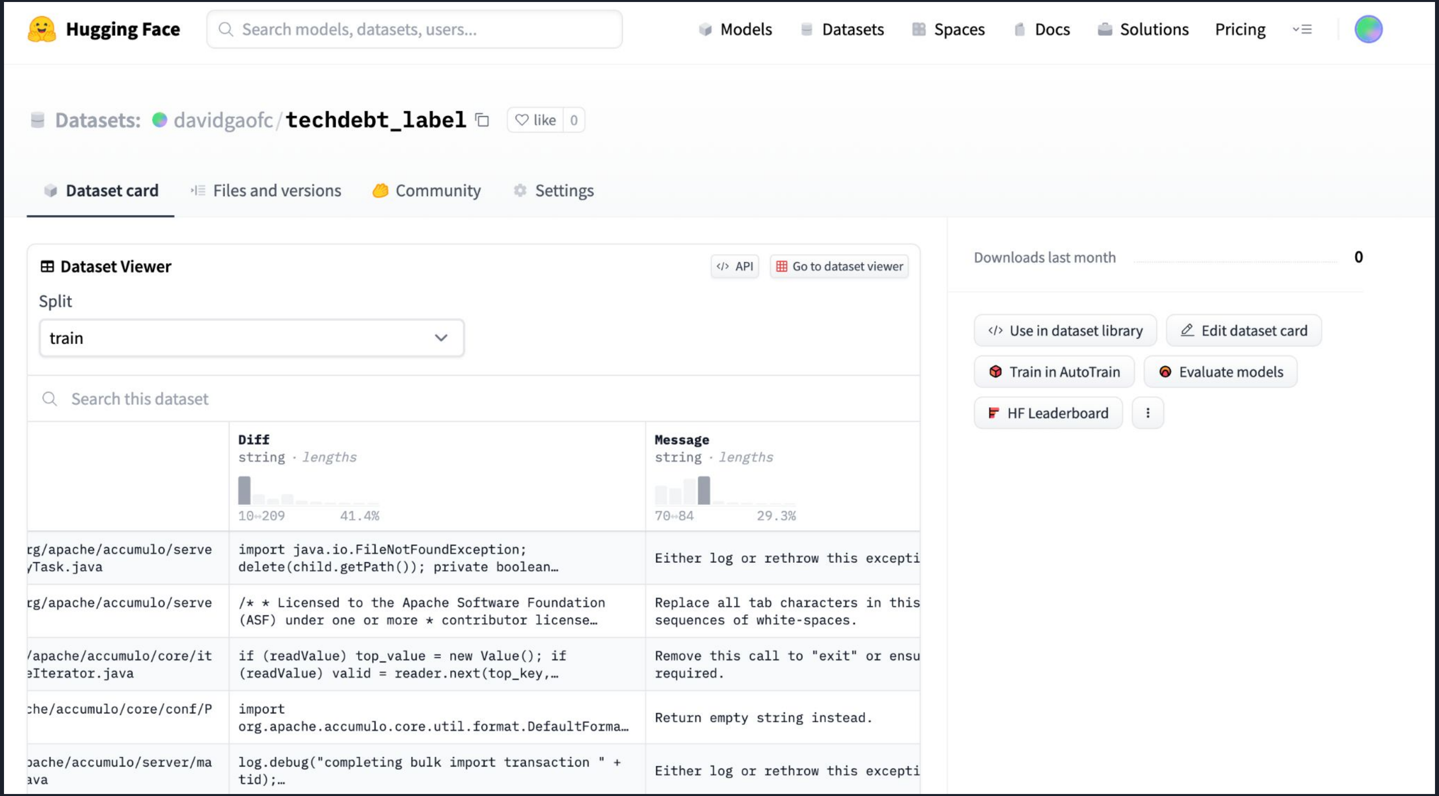
$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Generation Task

How well can a fine-tuned transformer model generate labels for technical debt?

Metrics: BLEU, ROUGE

Preliminary Work



Datasets: davidgaofc/**techdebt_label**

Dataset Viewer

Split

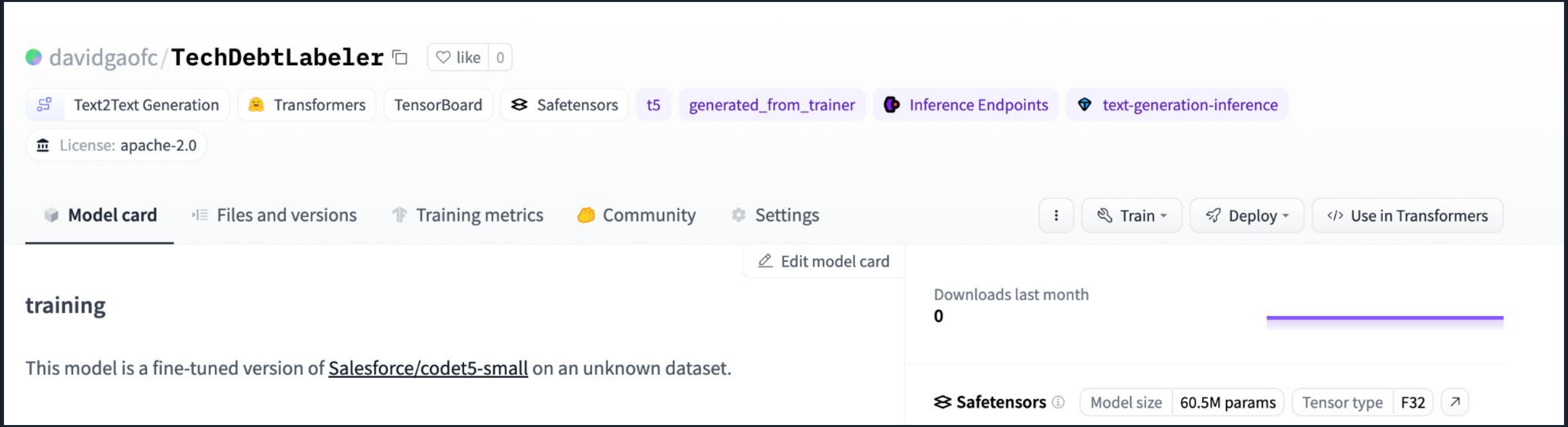
train

Search this dataset

Downloads last month 0

	Diff string · *lengths* 10↔209 41.4%	Message string · *lengths* 70↔84 29.3%
rg/apache/accumulo/serve yTask.java	import java.io.FileNotFoundException; delete(child.getPath()); private boolean…	Either log or rethrow this excepti
rg/apache/accumulo/serve	/* * Licensed to the Apache Software Foundation (ASF) under one or more * contributor license…	Replace all tab characters in this sequences of white-spaces.
/apache/accumulo/core/it eIterator.java	if (readValue) top_value = new Value(); if (readValue) valid = reader.next(top_key,…	Remove this call to "exit" or ensu required.
che/accumulo/core/conf/P	import org.apache.accumulo.core.util.format.DefaultForma…	Return empty string instead.
pache/accumulo/server/ma ava	log.debug("completing bulk import transaction " + tid);…	Either log or rethrow this excepti

Preliminary Work (continued)



davidgaofc/TechDebtLabeler

like 0

Text2Text Generation · Transformers · TensorBoard · Safetensors · t5 · generated_from_trainer · Inference Endpoints · text-generation-inference

License: apache-2.0

Model card · Files and versions · Training metrics · Community · Settings · Train · Deploy · Use in Transformers

Edit model card

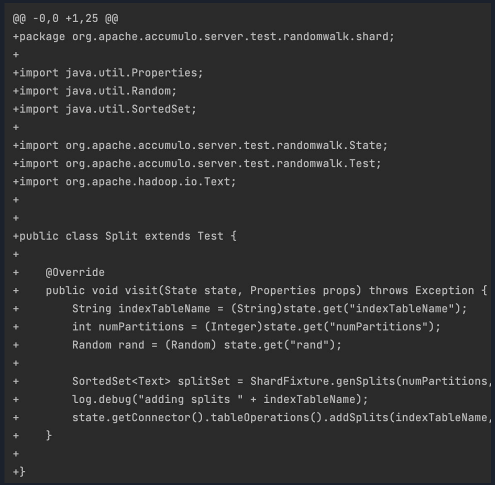
training

This model is a fine-tuned version of Salesforce/codet5-small on an unknown dataset.

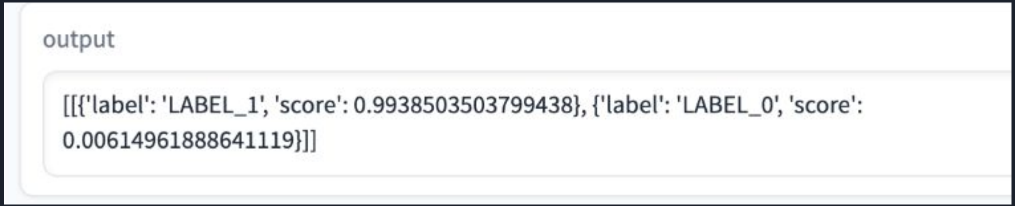
Downloads last month
0

Safetensors · Model size 60.5M params · Tensor type F32

Pipeline

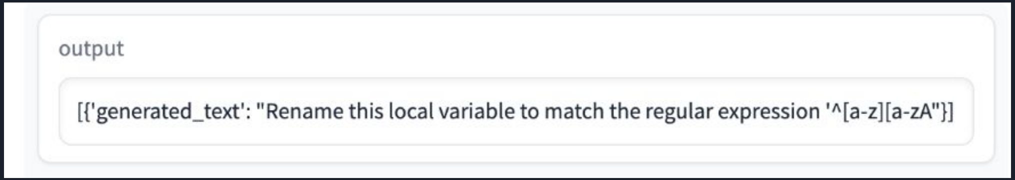


```
@@ -0,0 +1,25 @@
+package org.apache.accumulo.server.test.randomwalk.shard;
+
+import java.util.Properties;
+import java.util.Random;
+import java.util.SortedSet;
+
+import org.apache.accumulo.server.test.randomwalk.State;
+import org.apache.accumulo.server.test.randomwalk.Test;
+import org.apache.hadoop.io.Text;
+
+
+public class Split extends Test {
+
+    @Override
+    public void visit(State state, Properties props) throws Exception {
+        String indexTableName = (String)state.get("indexTableName");
+        int numPartitions = (Integer)state.get("numPartitions");
+        Random rand = (Random) state.get("rand");
+
+        SortedSet<Text> splitSet = ShardFixture.genSplits(numPartitions,
+        log.debug("adding splits " + indexTableName);
+        state.getConnector().tableOperations().addSplits(indexTableName,
+    }
+
+}
```



output

```
[[{'label': 'LABEL_1', 'score': 0.9938503503799438}, {'label': 'LABEL_0', 'score':
0.006149618886411119}]]
```



output

```
[{'generated_text': "Rename this local variable to match the regular expression '^[a-z][a-zA"}]
```

Questions/Comments?

References

Valentina Lenarduzzi, Nyyti Saarimäki, and Davide Taibi. 2019. The Technical Debt Dataset. In Proceedings of the Fifteenth International Conference on Predictive Models and Data Analytics in Software Engineering (PROMISE'19). Association for Computing Machinery, New York, NY, USA, 2–11. https://doi.org/10.1145/3345629.3345630