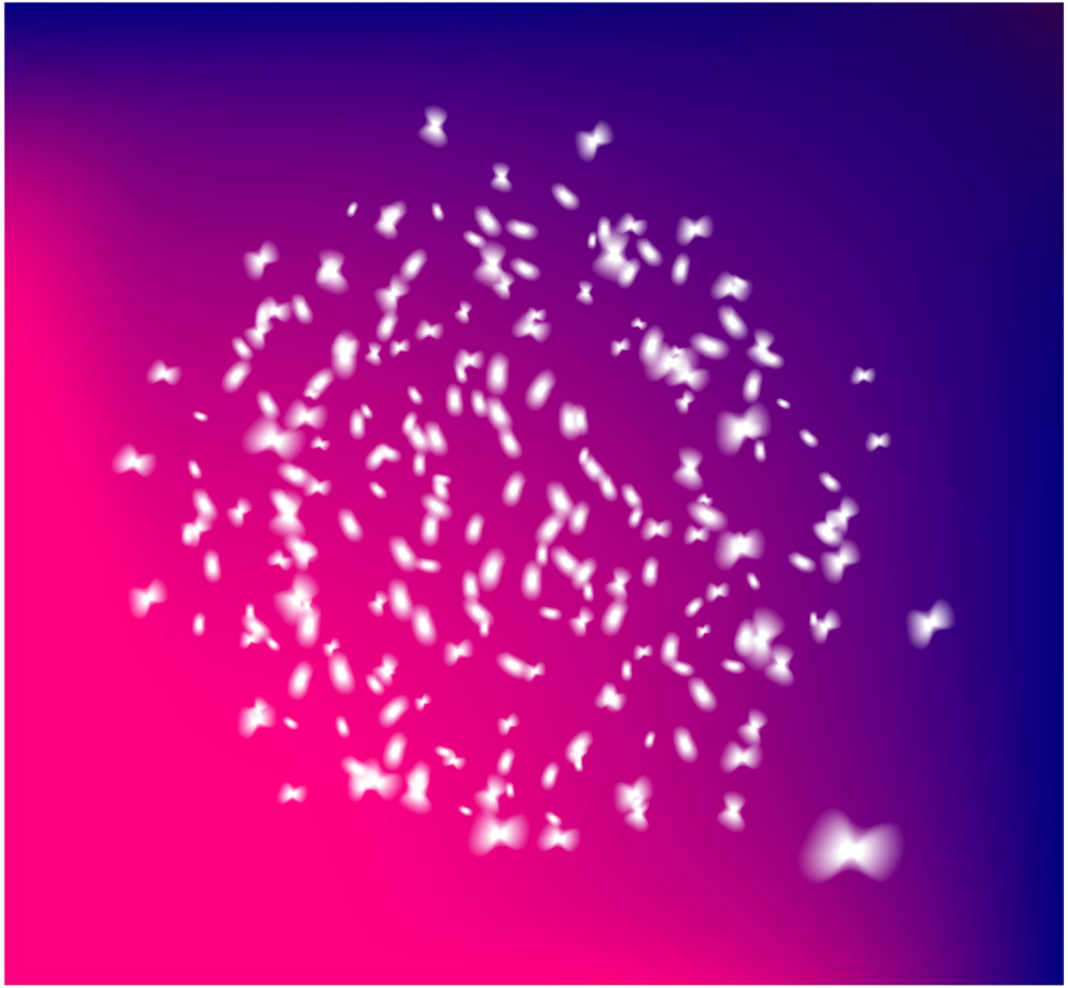


PyData Bristol

POLECAT™

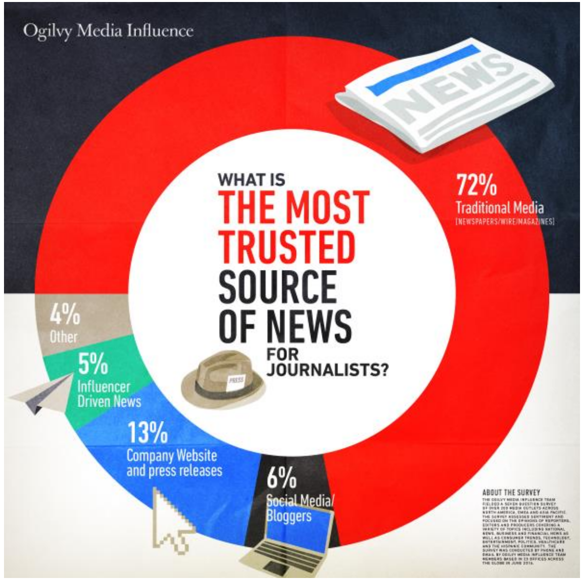
Data Science Pipeline Using DVC and VSCode

Alon Samuel

11 October 2022

Why is it needed - Polecat.com

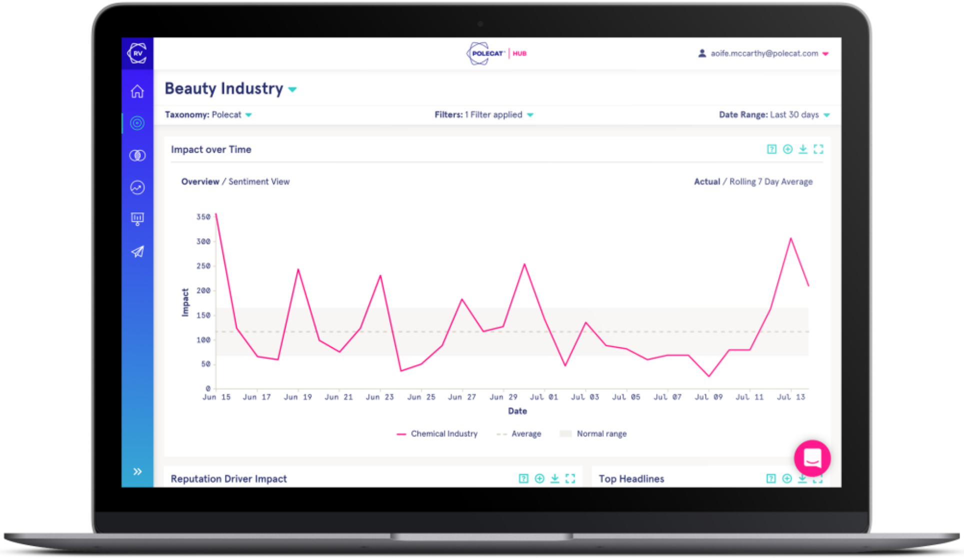
POLECAT™



ABOUT THE SURVEY

THE OGILVY MEDIA INFLUENCE TEAM FIELDED A SEVEN QUESTION SURVEY OF OVER 200 MEDIA OUTLETS ACROSS NORTH AMERICA, EMEA AND ASIA PACIFIC. THE SURVEY ASSESSED SENTIMENT AND FOCUSED ON THE OPINIONS OF REPORTERS, EDITORS AND PRODUCERS COVERING A VARIETY OF TOPICS INCLUDING NATIONAL NEWS, BUSINESS AND FINANCIAL NEWS AS WELL AS CONSUMER TRENDS, TECHNOLOGY, ENTERTAINMENT, POLITICS, HEALTHCARE AND THE HISPANIC COMMUNITY. THE SURVEY WAS CONDUCTED BY PHONE AND EMAIL BY OGILVY MEDIA INFLUENCE TEAM MEMBERS BASED IN 23 OFFICES ACROSS THE GLOBE IN JUNE 2016.

This Photo by Unknown author is licensed under CC BY-SA-NC.



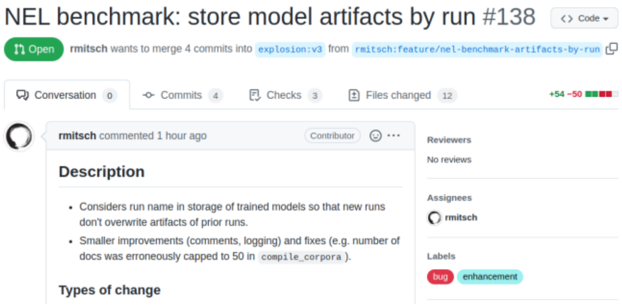
Taxonomy: Polecat

Filters: 1 Filter applied

Date Range: Last 30 days

Example Task

Classify whether a GitHub issue title is about documentation



NEL benchmark: store model artifacts by run #138

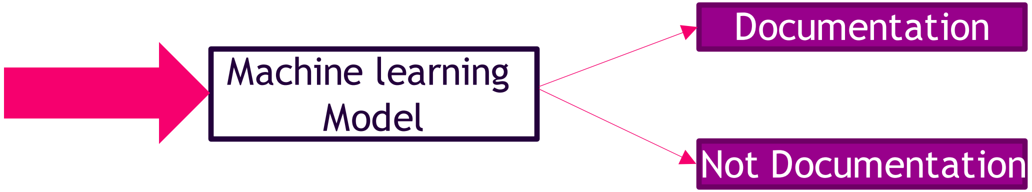
rmitsch wants to merge 4 commits into `explosion:v3` from `rmitsch:feature/nel-benchmark-artifacts-by-run`

rmitsch commented 1 hour ago

Description

- Considers run name in storage of trained models so that new runs don't overwrite artifacts of prior runs.
- Smaller improvements (comments, logging) and fixes (e.g. number of docs was erroneously capped to 50 in `compile_corpora`).

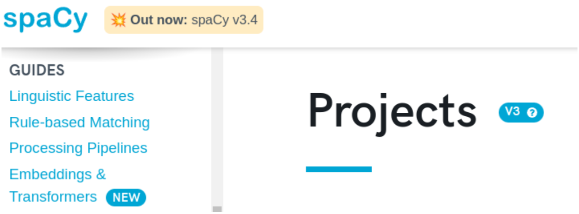
Types of change



Machine learning Model

Documentation

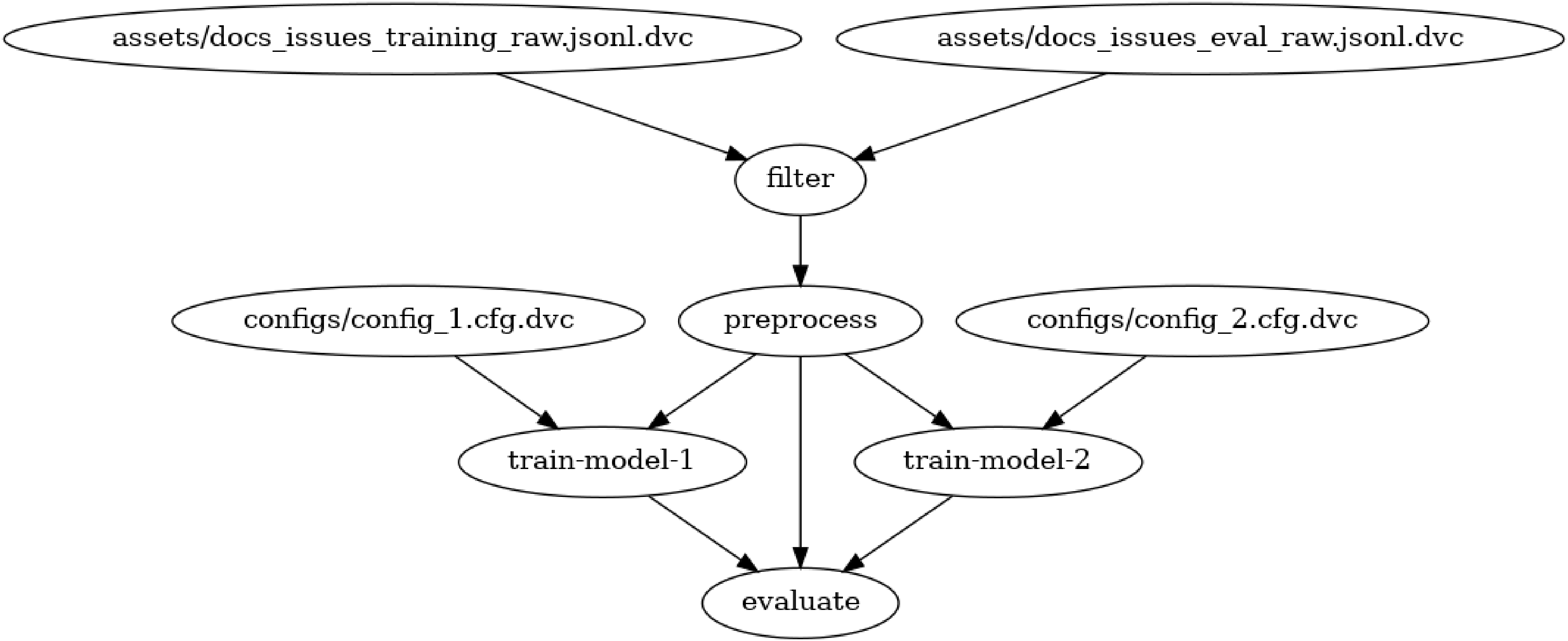
Not Documentation



This Photo by Unknown author is licensed under CC BY-SA.

DVC .org OPEN-SOURCE MACHINE LEARNING VERSION CONTROL

POLECAT™



assets/docs_issues_training_raw.jsonl.dvc

assets/docs_issues_eval_raw.jsonl.dvc

filter

configs/config_1.cfg.dvc

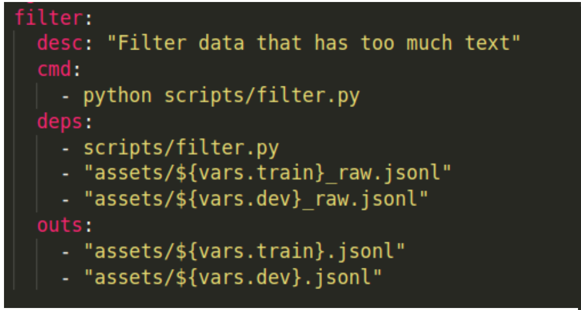
preprocess

configs/config_2.cfg.dvc

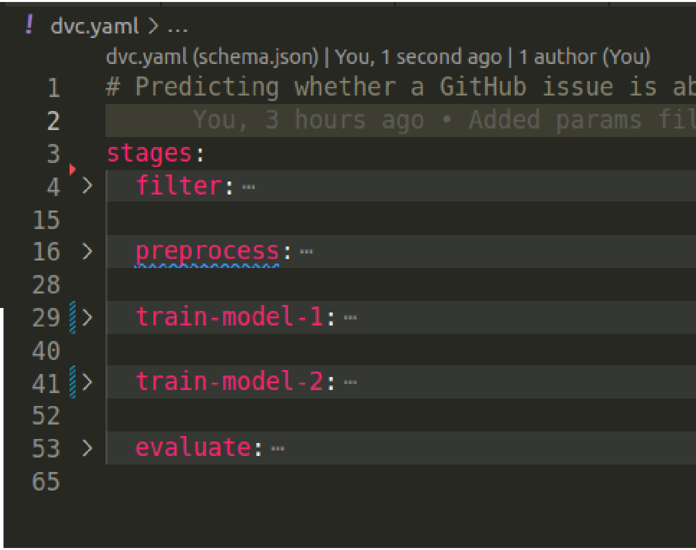
train-model-1

train-model-2

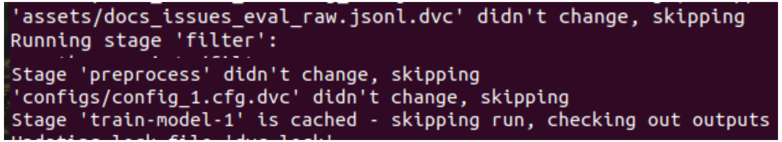
evaluate



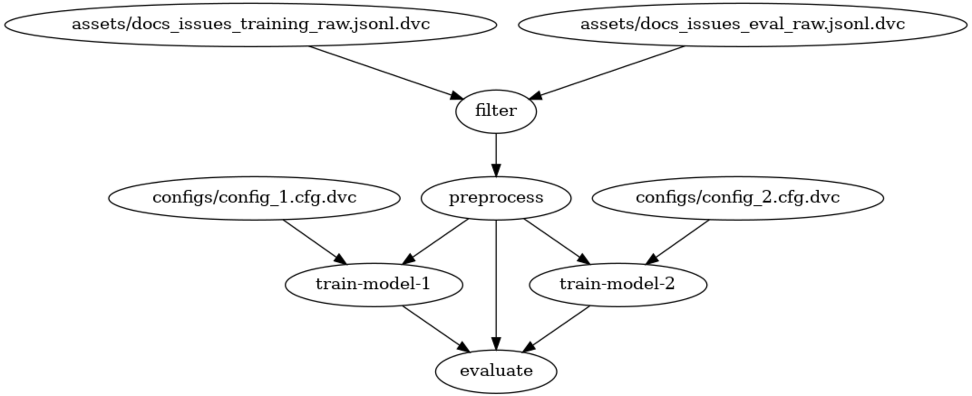
```yaml
filter:
  desc: "Filter data that has too much text"
  cmd:
    - python scripts/filter.py
  deps:
    - scripts/filter.py
    - "assets/${vars.train}_raw.jsonl"
    - "assets/${vars.dev}_raw.jsonl"
  outs:
    - "assets/${vars.train}.jsonl"
    - "assets/${vars.dev}.jsonl"
```



```yaml
# Predicting whether a GitHub issue is ab
        You, 3 hours ago • Added params fil
stages:
  filter: ⋯

  preprocess: ⋯

  train-model-1: ⋯

  train-model-2: ⋯

  evaluate: ⋯
```



```
'assets/docs_issues_eval_raw.jsonl.dvc' didn't change, skipping
Running stage 'filter':
Stage 'preprocess' didn't change, skipping
'configs/config_1.cfg.dvc' didn't change, skipping
Stage 'train-model-1' is cached - skipping run, checking out outputs
```



Live demo - adding a model

Track results

VSCode Notebook

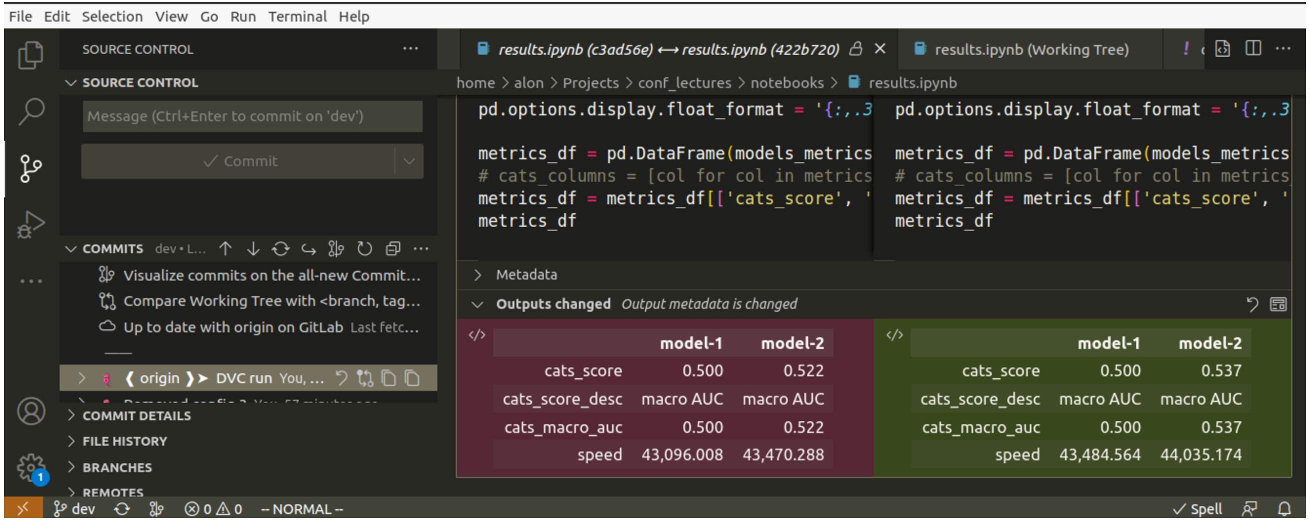
My first notebook in VS Code!

```
import matplotlib.pyplot as plt
import numpy as np
```

Histogram with 5 bins

```
x = [21,22,23,4,5,6,77,8,9,10,31,32,33,34,35,36,37,18,49,50,100]
num_bins = 5
plt.hist(x, num_bins)
plt.show()
```

Histogram with 10 bins



File Edit Selection View Go Run Terminal Help
SOURCE CONTROL
SOURCE CONTROL
Message (Ctrl+Enter to commit on 'dev')
✓ Commit
COMMITS dev • L...
Visualize commits on the all-new Commit...
Compare Working Tree with <branch, tag...
Up to date with origin on GitLab Last fetc...
❬ origin ❭ ➤ DVC run You, ...
COMMIT DETAILS
FILE HISTORY
BRANCHES
REMOTES
results.ipynb (c3ad56e) ⟷ results.ipynb (422b720)
results.ipynb (Working Tree)
home > alon > Projects > conf_lectures > notebooks > results.ipynb
Metadata
Outputs changed Output metadata is changed
dev
⊗ 0 ⚠ 0
-- NORMAL --
✓ Spell

```
pd.options.display.float_format = '{:,.3

metrics_df = pd.DataFrame(models_metrics
# cats_columns = [col for col in metrics
metrics_df = metrics_df[['cats_score', '
metrics_df
```

```
pd.options.display.float_format = '{:,.3

metrics_df = pd.DataFrame(models_metrics
# cats_columns = [col for col in metrics
metrics_df = metrics_df[['cats_score', '
metrics_df
```

	model-1	model-2
cats_score	0.500	0.522
cats_score_desc	macro AUC	macro AUC
cats_macro_auc	0.500	0.522
speed	43,096.008	43,470.288

	model-1	model-2
cats_score	0.500	0.537
cats_score_desc	macro AUC	macro AUC
cats_macro_auc	0.500	0.537
speed	43,484.564	44,035.174

Live demo - visualise results

Links

- Example project: https://github.com/explosion/projects/tree/v3/tutorials/textcat_docs_issues

- Materials: https://github.com/polecat-dev/dvc_vscode

- DVC repo: https://github.com/iterative/dvc

This Photo by Unknown author is licensed under CC BY-SA-NC.

POLECAT™

For further information, please contact:

✉ Alon.samuel@polecat.com