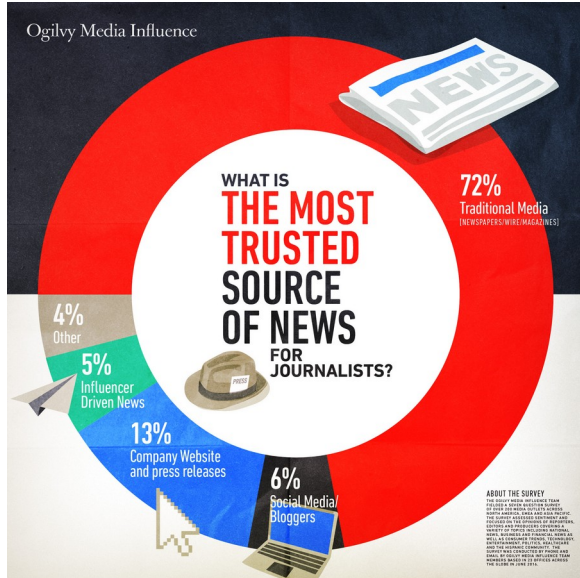


Data Science Pipeline Using DVC and VSCode

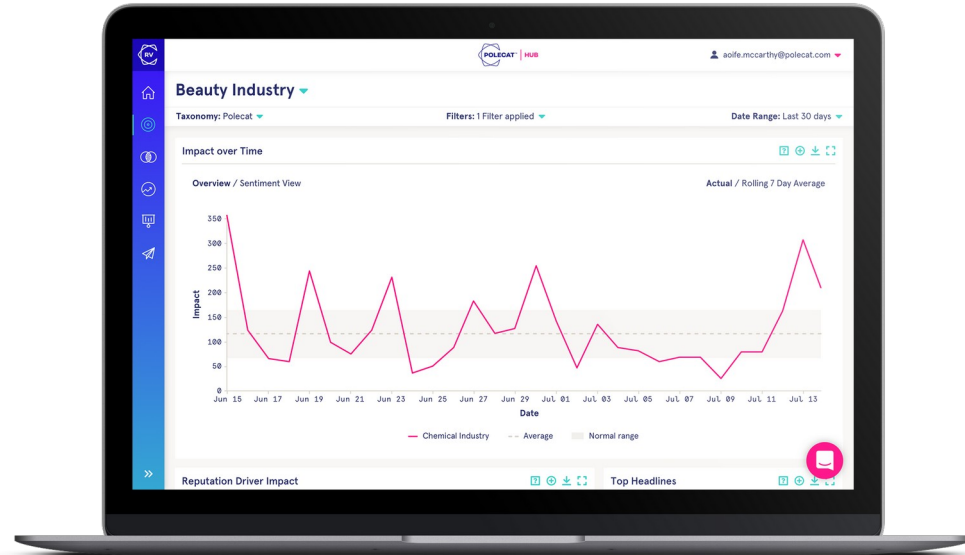
Alon Samuel

11 October 2022

Why is it needed - Polecat.com



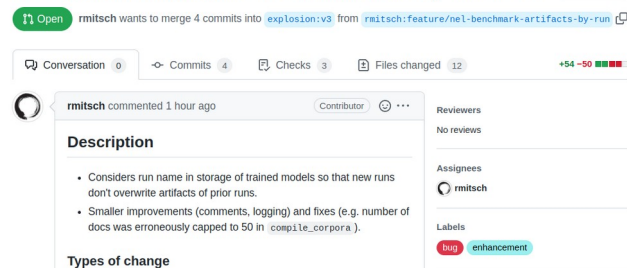
This Photo by Unknown author is licensed under [CC BY-SA-NC](#).



Example Task

Classify whether a GitHub issue title is about documentation

NEL benchmark: store model artifacts by run #138



Machine
learning
Model

Documentation

Not
Documentation

spaCy

Out now: spaCy v3.4

GUIDES

Linguistic Features
Rule-based Matching
Processing Pipelines
Embeddings &
Transformers **NEW**

Projects

V3

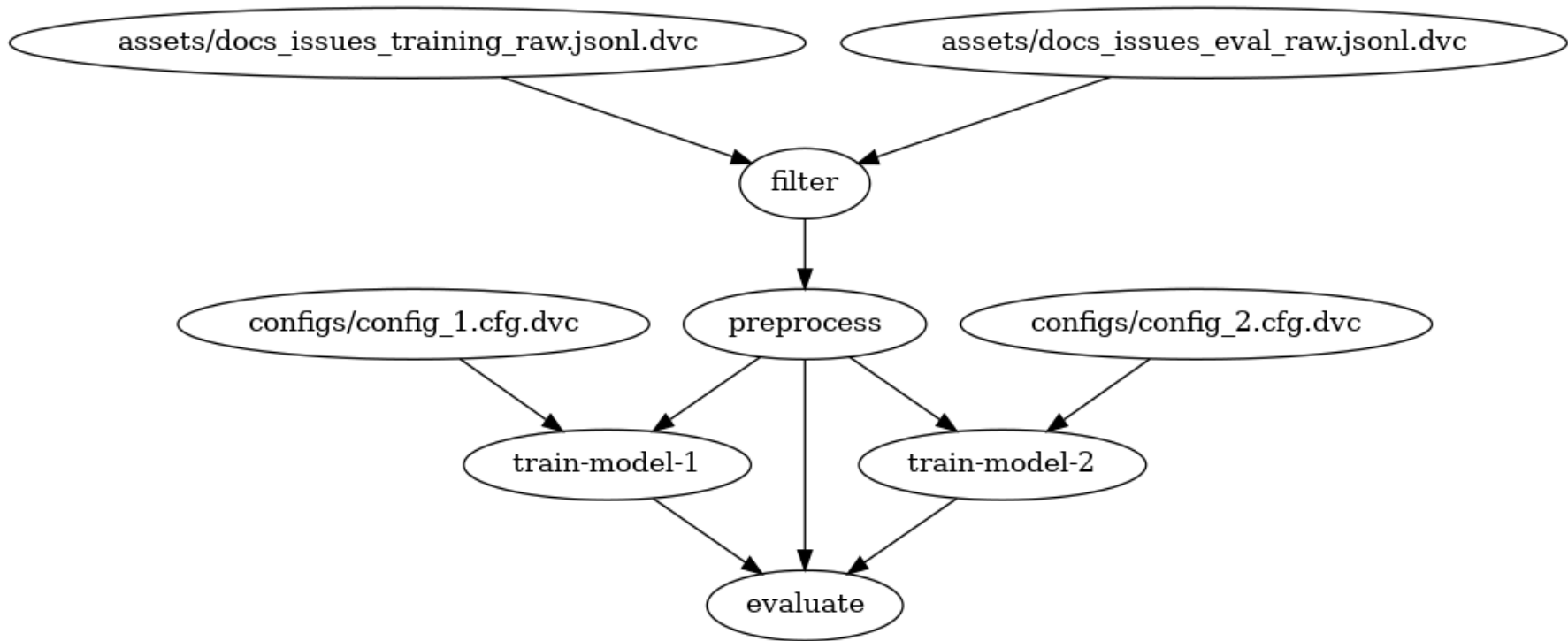


ISSUE IS CRITICAL



OPEN-SOURCE

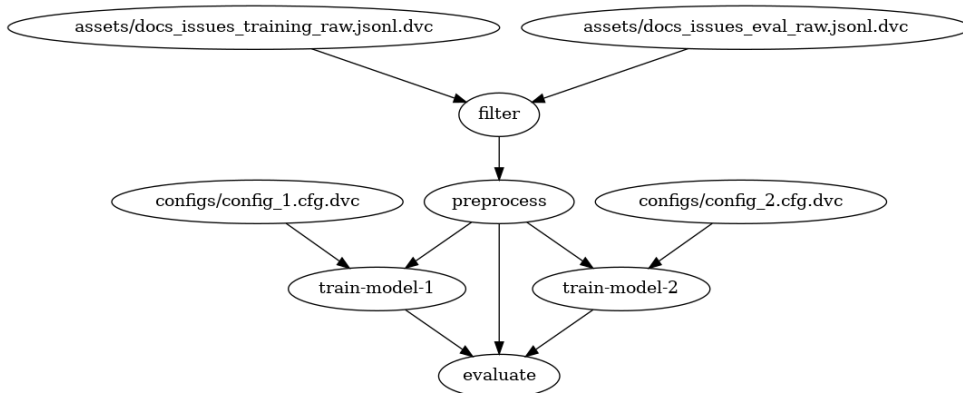
MACHINE LEARNING
VERSION CONTROL



```
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"
```

```
! dvc.yml > ...
  dvc.yml (schema.json) | You, 1 second ago | 1 author (You)
  1 # Predicting whether a GitHub issue is ab
  2   You, 3 hours ago • Added params fil
  3 stages:
  4 > filter: ...
  15
  16 > preprocess: ...
  28
  29 > train-model-1: ...
  40
  41 > train-model-2: ...
  52
  53 > evaluate: ...
  65
```

```
'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
```

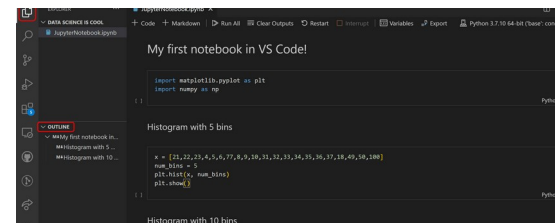




Live demo - adding a model

Track results

VSCode Notebook



File Edit Selection View Go Run Terminal Help

SOURCE CONTROL

Message (Ctrl+Enter to commit on 'dev')

Commit

COMMITTS 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

REMITES

dev 0 0 --NORMAL--

results.ipynb (c3ad56e) ↔ results.ipynb (422b720) results.ipynb (Working Tree)

home > alon > Projects > conf_lectures > notebooks > results.ipynb

```
pd.options.display.float_format = '{:,.3f}'
metrics_df = pd.DataFrame(models_metrics)
# cats_columns = [col for col in metrics
metrics_df = metrics_df[['cats_score', '
metrics_df
```

> Metadata

Outputs changed Output metadata is changed

	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

✓ Spell



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](#).



For further information, please
contact:

✉ Alon.samuel@polecat.com