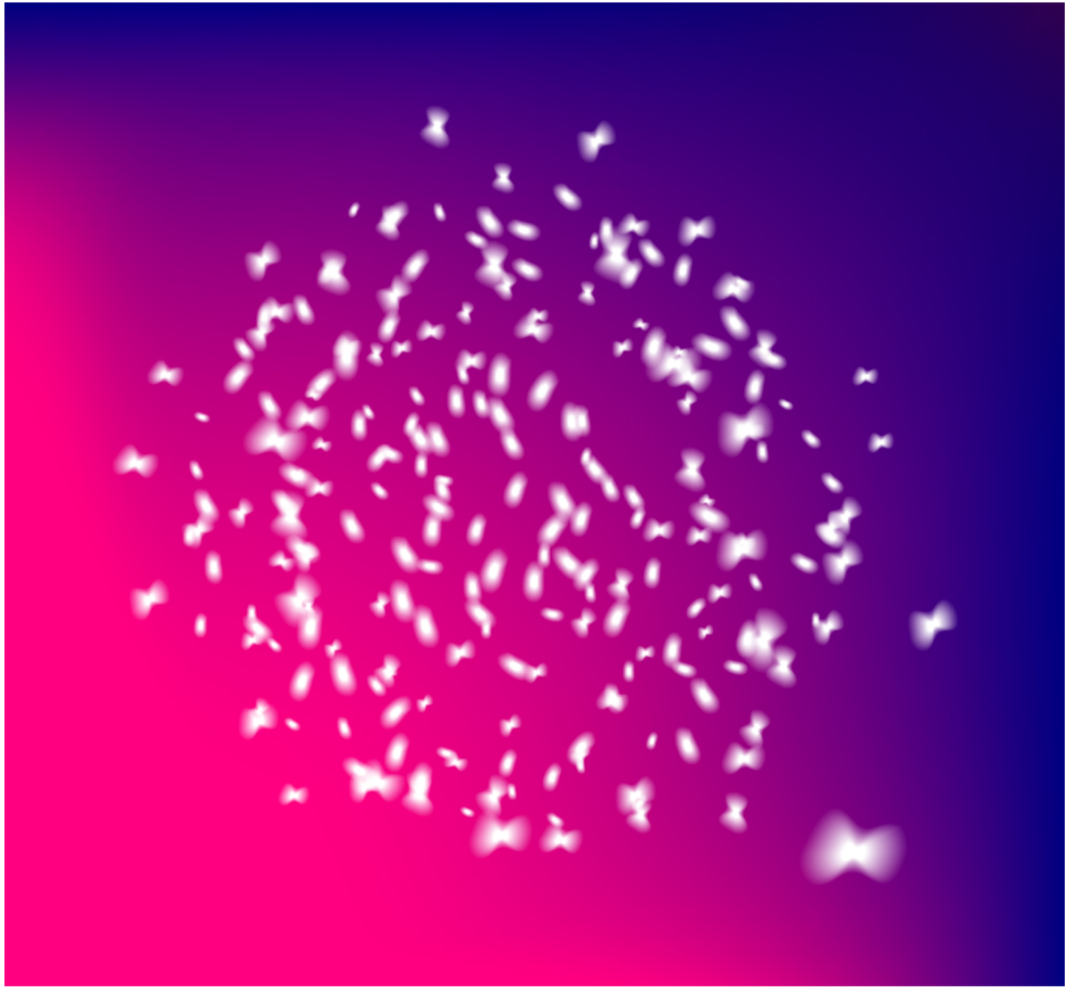


PyData Bristol

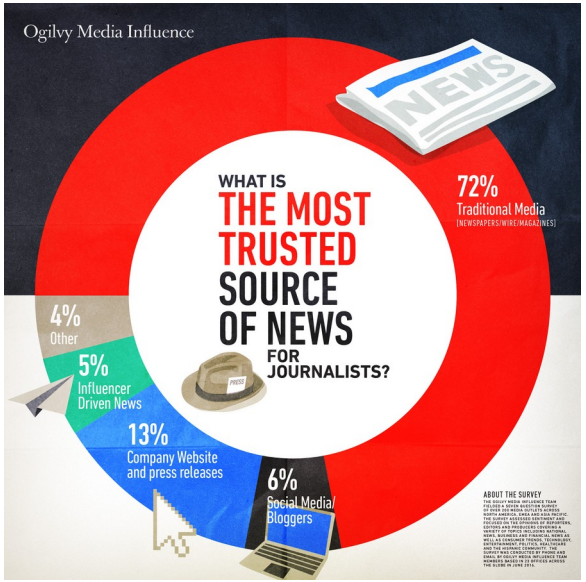
POLECAT™

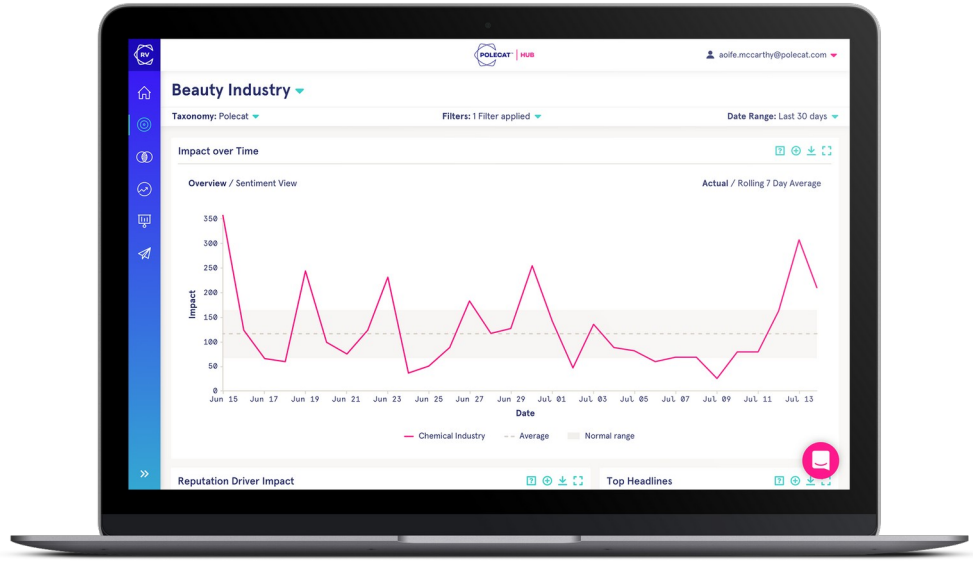
Data Science Pipeline Using DVC and VSCode

Alon Samuel

11 October 2022

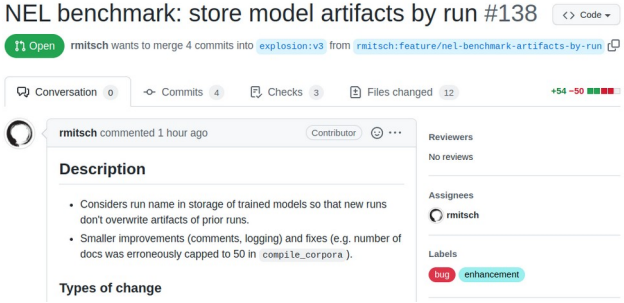
Why is it needed – Polecat.com

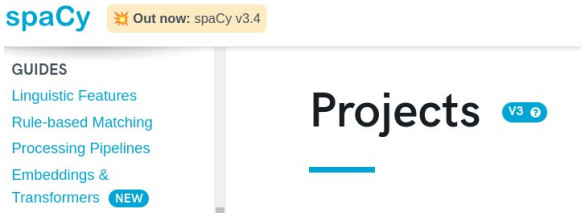




Example Task

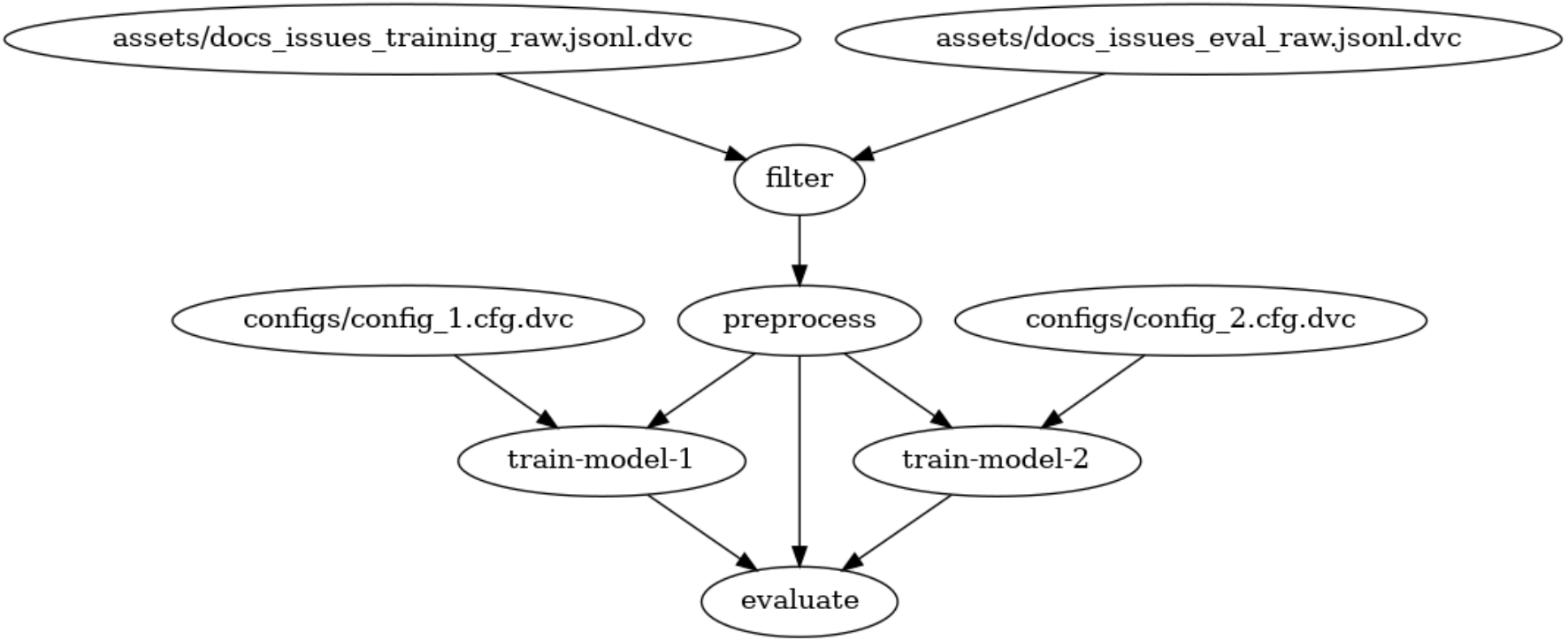
Classify whether a GitHub issue title is about documentation





OPEN-SOURCE
MACHINE LEARNING
VERSION CONTROL

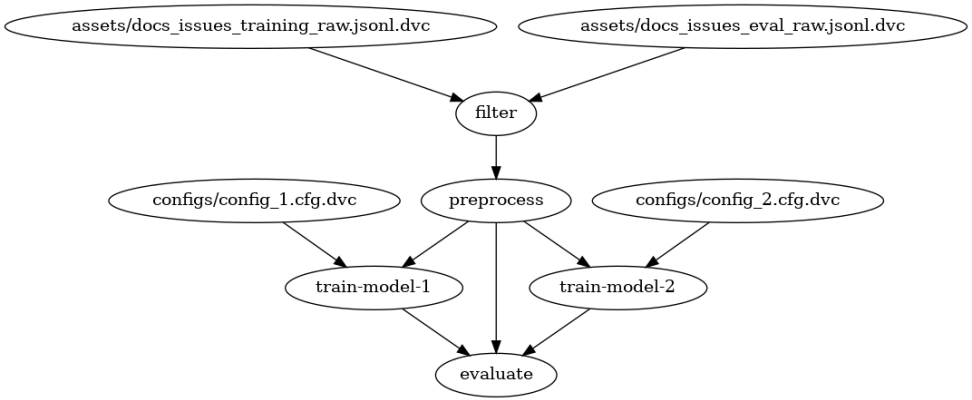
POLECAT™



```
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.yaml > ...
dvc.yaml (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
```

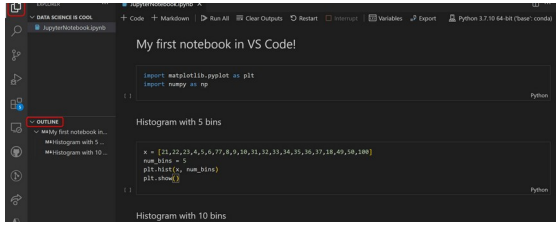
```
'assets/docs_issues_eval_raw.jsonl.dvc' didn't change, skipping
Running stage 'filter':
Stage 'preprocess' didn't change, skipping
'configs/config_1.cfg.dvc' didn't change, skipping
Stage 'train-model-1' is cached - skipping run, checking out outputs
```

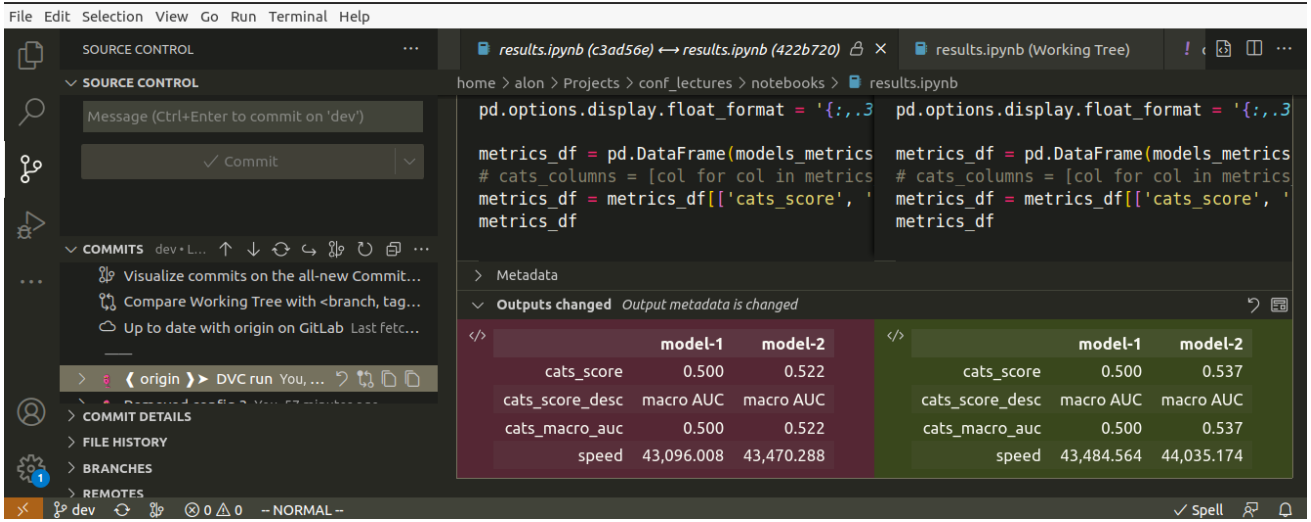


Live demo – adding a model

Track results

VSCode Notebook





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

POLECAT™

POLECAT™

For further information, please contact:

Alon.samuel@polecat.com