**# 9fin Machine Learning Engineer Technical Challenge**

Welcome to the challenge!

When you are ready to go through what you've done, drop me a line, and we can schedule the technical interview.

There is no time limit here but please don't spend more than **2-3 hours on this task**.

If you feel stuck for some time or there is an issue with the challenge, please stop and send me a quick email to help you out.

The aim of this challenge is to provide **something for us to talk** around in the interview.

We want to test your ability to **write software** to **clean/enrich/analyse data** and **implement a predictive model** on it.

We are also interested in your ability to work in a team so ability to deliver your results in a way in which is easy for us to review (e.g. **comments, clear writeup**).

We can always **discuss** any ideas you may have wanted to explore if there was more time to iterate on this.

Good luck, and may the data be with you.

**## Problem**

We have a large number of **news headlines** that come from **various sources**.

We would like to enrich these headlines with the labels **'Calendar'**, **'Press Release'**, or **'Results'** (mutually exclusive labels).

We've gone ahead and **annotated these samples** so that you can explore if there is a feasible solution.

**## Data**

I've included the following files with this challenge (in the `data` directory):

- titles.jsonl.gz

- db.sqlite3

The compressed jsonl file `**titles.jsonl.gz**` contains a **subset of our news titles** that the business team have happily **annotated** for us.

```shell

$ gzip -dc titles.jsonl.gz | head -n3

{"**title**": "Diversified Healthcare Trust: Diversified Healthcare Trust First Quarter 2021 Conference Call Scheduled for Thursday, May 6th", "**id**": "1b26de1d-0c7e-49c2-94d7-2db9d187aaed"}

{"title": "Nathan's Famous: Nathan's Famous, Inc. Reports Second Quarter Results", "id": "fdd95479-f76f-41e3-92f7-f9fcaa025931"}

{"title": "Dufry: Dufry wins new, ten-year duty-free contract at Martinique International Airport", "id": "d3207342-6972-4ec9-97f3-a4c5e3511ecf"}

```

The sqlite file `db.sqlite3` contains tables `**company**` and `**news\_meta**`. These tables might come in handy.

```shell

$ sqlite3 db.sqlite3 -table "SELECT \* FROM company LIMIT 3"

+------+------------------------------+-------------------------------------------+---------------+-------------------------------------------+

| id | **name** | **description** | **location** | **website** |

+------+------------------------------+-------------------------------------------+---------------+-------------------------------------------+

| 2236 | Diversified Healthcare Trust | Healthcare Real Estate Investment Trust | United States | https://www.dhcreit.com/home/default.aspx |

| 2748 | Nathan's Famous | Operates a chain of fast food restaurants | United States | https://www.nathansfamous.com/ |

| 663 | Dufry | Travel retailer | Switzerland | http://www.dufry.com/ |

+------+------------------------------+-------------------------------------------+---------------+-------------------------------------------+

$ sqlite3 db.sqlite3 -table "SELECT id, url, company\_id FROM news\_meta LIMIT 3"

+--------------------------------------+---------------------------------------------------------------------------------------------------------------------------------------------------------------------+---------------+------------+

| **id** | **url** | **label** | **company**\_id |

+--------------------------------------+---------------------------------------------------------------------------------------------------------------------------------------------------------------------+---------------+------------+

| 1b26de1d-0c7e-49c2-94d7-2db9d187aaed | https://www.dhcreit.com/news/press-release-details/2021/Diversified-Healthcare-Trust-First-Quarter-2021-Conference-Call-Scheduled-for-Thursday-May-6th/default.aspx | CALENDAR | 2236 |

| fdd95479-f76f-41e3-92f7-f9fcaa025931 | https://nathansfamous.gcs-web.com/news-releases/news-release-details/nathans-famous-inc-reports-second-quarter-results-18 | RESULTS | 2748 |

| d3207342-6972-4ec9-97f3-a4c5e3511ecf | https://www.dufry.com/en/press\_release/2021-06-10/dufry-wins-new-ten-year-duty-free-contract-martinique-international | PRESS RELEASE | 663 |

+--------------------------------------+---------------------------------------------------------------------------------------------------------------------------------------------------------------------+---------------+------------+

```

**## Task**

- Look at this as a mini Kaggle challenge.

- Perform any **cleaning, exploratory analysis, and/or visualizations** to use the provided data.

- Build a **predictive model** for this **multi-classification** problem.

- Provide some simple **metrics and visualisations** on model performance for the **stakeholders**.

- We would like this to be done in **Python** (sorry R fans!).

- This can be done in jupyter notebooks, scripts, or both.

- Feel free to chose any python **dependencies** you wish.

- We need to be able to reproduce the results so make sure to provide the python version and dependencies in some manner (doesn't matter if its using pure pip, or conda, poetry, docker,..etc.)

- Include **instructions on how to run** and reproduce with any additional comments in a markdown file.

- Commit your files to a **private git repo** and **share it with us** (or upload your code repo to the same shared drive). Up to you here!

**## Questions**

We've listed some of the questions that you can think about for the interview.

- Discuss your **approach**.

- What **insight** did you extract from this dataset?

- How would we go about **monitoring this model's performance** in production?

- If we receive another set of annotated samples, how would we go about **improving** our existing model? What about introducing an additional label to this problem?