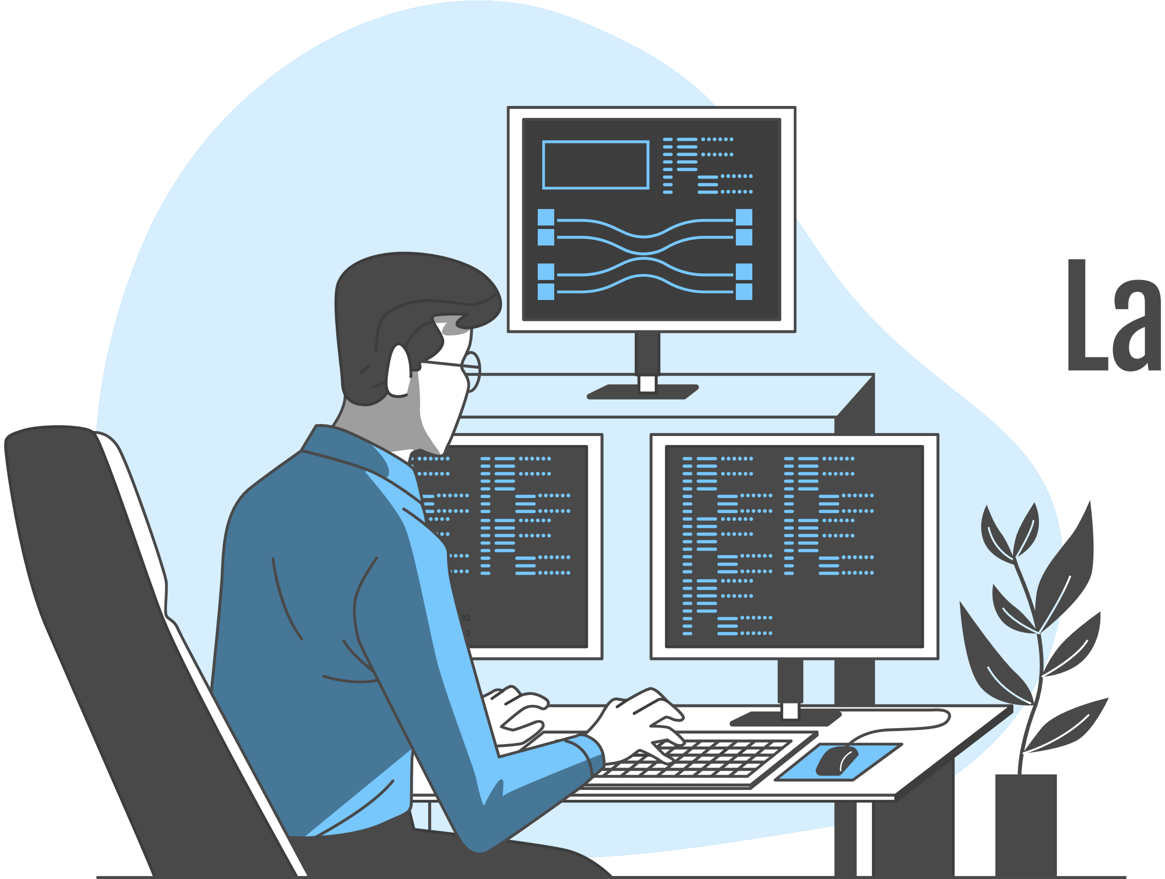


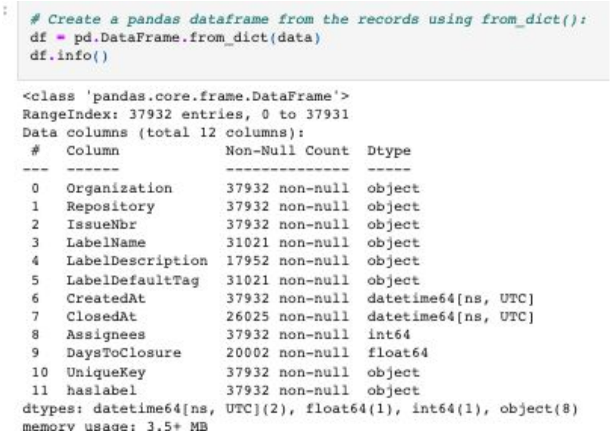
Labels Survey Analysis

Hack for LA Data Science

Project Background

The Hack4La Github repository has seen the usage of many labels in order to categorize the various tasks and projects taken on by our organization. Through years of projects and their labels we have found that our labels lack a clear and standard structure and that it would be much more convenient for our future endeavors if we had a standardized system of labels for our assignments. Thus the Data Science Community of Practice has taken on a project in order to analyze the current labels used in our system and give our recommendation for what should be standard practice moving forward.

Project Description: This dataset has been scrapped from the Hack4La repository on github and then converted into pandas DataFrame for further analysis. The purpose of project is to analyze the labels used throughout the hack4LA repositories. In this dataset the date attributes(ClosedAt,CreatedAt) are converted from str into datetime and DaystoClosure is calculated based on CreatedAt and ClosedAt dates of repositories. UniqueKey has been generated in order to uniquely identify the issues. In order to understand that how labeling an issue could be effective, we have visualize the dataset in different ways and answered few questions.



```
# Create a pandas dataframe from the records using from_dict():
df = pd.DataFrame.from_dict(data)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 37932 entries, 0 to 37931
Data columns (total 12 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Organization      37932 non-null  object
 1   Repository        37932 non-null  object
 2   IssueNbr          37932 non-null  object
 3   LabelName         31021 non-null  object
 4   LabelDescription  17952 non-null  object
 5   LabelDefaultTag   31021 non-null  object
 6   CreatedAt         37932 non-null  datetime64[ns, UTC]
 7   ClosedAt          26025 non-null  datetime64[ns, UTC]
 8   Assignees         37932 non-null  int64
 9   DaysToClosure     20002 non-null  float64
 10  UniqueKey         37932 non-null  object
 11  haslabel          37932 non-null  object
dtypes: datetime64[ns, UTC](2), float64(1), int64(1), object(8)
memory usage: 3.5+ MB
```

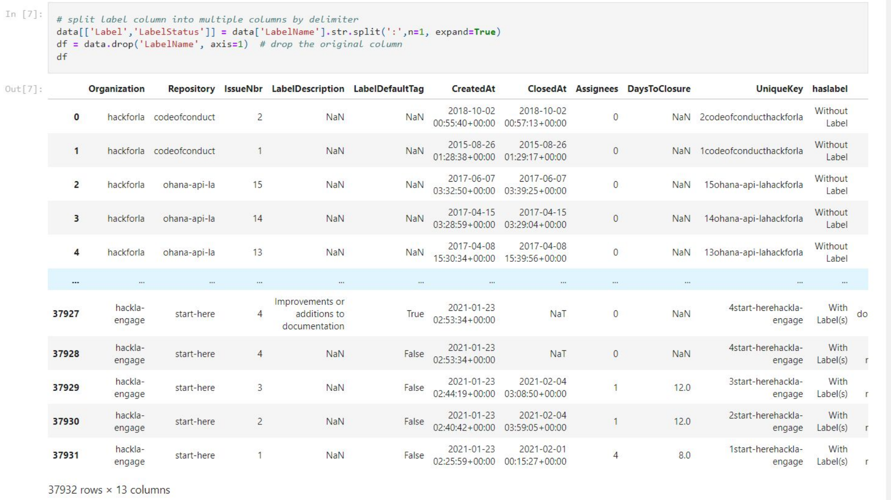
Project Data Manipulation Overview:

- DaysToClosure attribute is added to analyze how the usage of label helps resolving the issue faster.
- Assigned uniqueKey to identify the issues uniquely.
- Flagging the issues without labels (haslabel)
- LabelName column is split into Label + LabelStatus in-order to plot the labels in categorical way.
- Identify the null values and handled the case sensitive label for projecting.

Packages Utilized:

- pandas
- numpy
- multidict
- wordcloud

Structure of Labels



In [7]:

```python
# split label column into multiple columns by delimiter
data[['Label','LabelStatus']] = data['LabelName'].str.split(':',n=1, expand=True)
df = data.drop('LabelName', axis=1)  # drop the original column
df
```

Out[7]:

	Organization	Repository	IssueNbr	LabelDescription	LabelDefaultTag	CreatedAt	ClosedAt	Assignees	DaysToClosure	UniqueKey	haslabel	
0	hackforla	codeofconduct	2	NaN	NaN	2018-10-02 00:55:40+00:00	2018-10-02 00:57:13+00:00	0	NaN	2codeofconducthackforla	Without Label	
1	hackforla	codeofconduct	1	NaN	NaN	2015-08-26 01:28:38+00:00	2015-08-26 01:29:17+00:00	0	NaN	1codeofconducthackforla	Without Label	
2	hackforla	ohana-api-la	15	NaN	NaN	2017-06-07 03:32:50+00:00	2017-06-07 03:39:25+00:00	0	NaN	15ohana-api-lahackforla	Without Label	
3	hackforla	ohana-api-la	14	NaN	NaN	2017-04-15 03:28:59+00:00	2017-04-15 03:29:04+00:00	0	NaN	14ohana-api-lahackforla	Without Label	
4	hackforla	ohana-api-la	13	NaN	NaN	2017-04-08 15:30:34+00:00	2017-04-08 15:39:56+00:00	0	NaN	13ohana-api-lahackforla	Without Label	
...	...	...	...	...	...	...	...	...	...	...	...	
37927	hackla-engage	start-here	4	Improvements or additions to documentation	True	2021-01-23 02:53:34+00:00	NaT	0	NaN	4start-herehackla-engage	With Label(s)	do
37928	hackla-engage	start-here	4	NaN	False	2021-01-23 02:53:34+00:00	NaT	0	NaN	4start-herehackla-engage	With Label(s)	r
37929	hackla-engage	start-here	3	NaN	False	2021-01-23 02:44:19+00:00	2021-02-04 03:08:50+00:00	1	12.0	3start-herehackla-engage	With Label(s)	r
37930	hackla-engage	start-here	2	NaN	False	2021-01-23 02:40:42+00:00	2021-02-04 03:59:05+00:00	1	12.0	2start-herehackla-engage	With Label(s)	r
37931	hackla-engage	start-here	1	NaN	False	2021-01-23 02:25:59+00:00	2021-02-01 00:15:27+00:00	4	8.0	1start-herehackla-engage	With Label(s)	r

37932 rows × 13 columns

Labels often are in the form of “Label: Status.” For example, many projects have a size label in order to signify how complex or large an assignment is (ie “Size: Large”). The most common parent labels are:

- feature
- p-feature
- size
- role

In order to scrape and standardize these labels, our label category will be split whenever label is written in this format.

Some Interesting Statistics

40.2%

Of issues are labeled

This leaves 59.8% of issues to have no labels at all.

326

Distinct Labels

Across all of H4LA's repositories we've used a total of 326 unique labels

10+

Labels used

Most projects actually see the use of over 10 different labels.

71

Days until Closure

There is an average of 71 days until closure for projects that use 10+ labels

Standardizing the Labels

In order to rationalize and do automation as well as org-wide audits, we need a standard system of labels that can be implemented for our current and future projects. Many labels that are ‘unique’ in the data are actually different ways of writing similar or even the same label (ie size: Large, size: L, size: 8 pts.). Each large parent label has been scraped into different sheets with their statuses.

I believe that focusing on these **three topics** can help with these issues:

Capitalization

...

Compound Statuses

...

Spacing and Numbers

...

Capitalization

In order to preserve simplicity, all labels/statuses will keep a strictly lowercase format. This excludes exceptions where the name itself contains capital letters (DevOps, UI/UX, CoP)

role: Data Science

↓

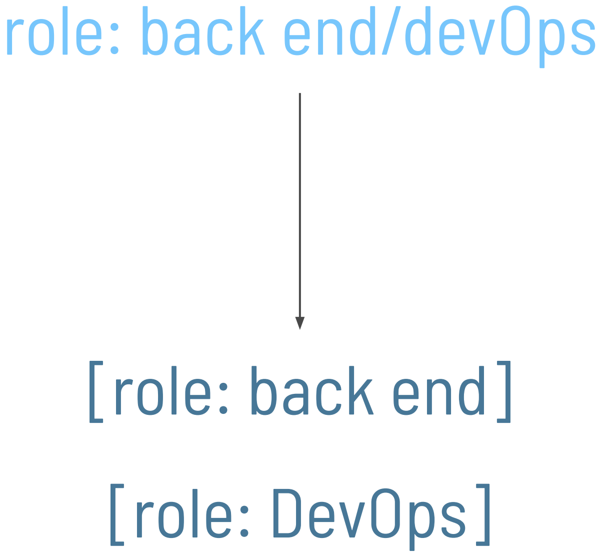
role: data science

... ...

Compound Statuses

Many labels are a combination of two existing statuses (eg. role :backend/DevOps). This causes clutter and an oversaturated amount of unique labels. Instead, any labels of this form should take on multiple role labels instead. A master list of these unique labels should be implemented based on our data (in order to easily implement a script to detect these compound statuses.

Note: devOps was also changed to DevOps due to our capitalization rules



role: back end/devOps

[role: back end]

[role: DevOps]

Spacing and Numbers

Normal spacing will be used instead of combining words. In terms of the size labels, numbers will be utilized instead of shirt sizes or rarities, these will take the format

“size: n pts”

[role: DataScience/front-end]
[size: Epic]

↓

[role: data science]
[role: front end]
[size: 8 pts]

... ...

Further Comments

- Potentially get rid of labels or features that are only used once or that are too specific. Try to fit these labels into a more general umbrella term that could be used for more projects
- Create a master list of labels and features
  - Added labels must be a part of this sheet
- If you would like to add a feature that is not a part of the established sheet → implement a system to document and record these as well
  - After review/analysis of these labels, ones with high demand can be officially implemented
- Label sheet with manually translated size labels (and others) can be found here: Labels Sheet
- Deciding whether or not a label is worth keeping can be subjective and tricky, feel free to refer to our jupyter notebook analysis or consult the Data Science CoP!