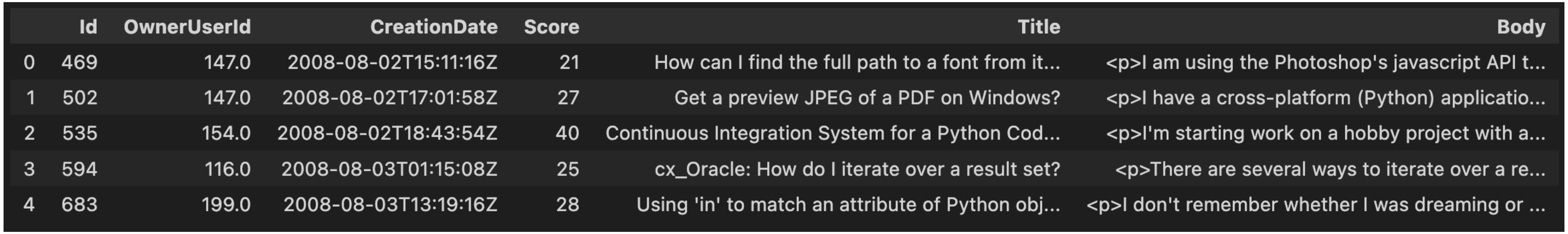
POC: Multilabel Text Classification For StackOverflow Questions

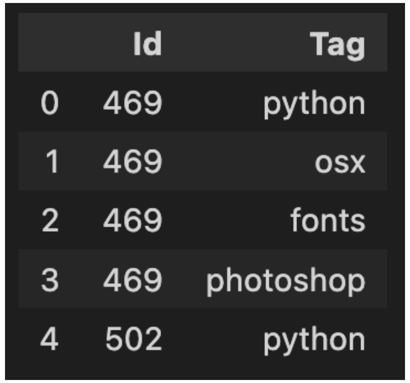
Metadata

questions.csv contains ~607k entries:



	Id	OwnerUserId	CreationDate	Score	Title	Body
0	469	147.0	2008-08-02T15:11:16Z	21	How can I find the full path to a font from it...	<p>I am using the Photoshop's javascript API t...
1	502	147.0	2008-08-02T17:01:58Z	27	Get a preview JPEG of a PDF on Windows?	<p>I have a cross-platform (Python) applicatio...
2	535	154.0	2008-08-02T18:43:54Z	40	Continuous Integration System for a Python Cod...	<p>I'm starting work on a hobby project with a...
3	594	116.0	2008-08-03T01:15:08Z	25	cx_Oracle: How do I iterate over a result set?	<p>There are several ways to iterate over a re...
4	683	199.0	2008-08-03T13:19:16Z	28	Using 'in' to match an attribute of Python obj...	<p>I don't remember whether I was dreaming or ...

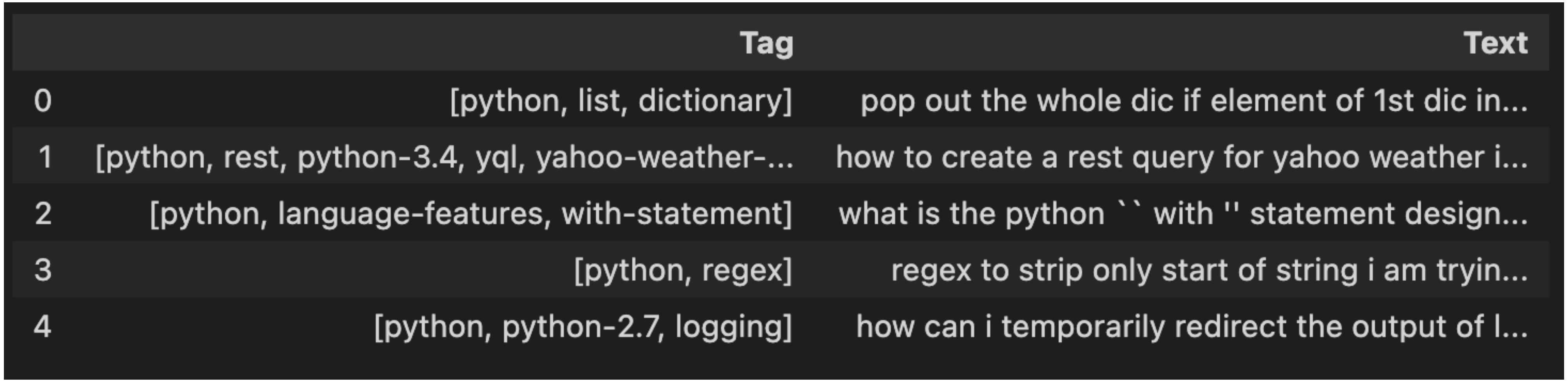
tags.csv contains ~1.8m entries:



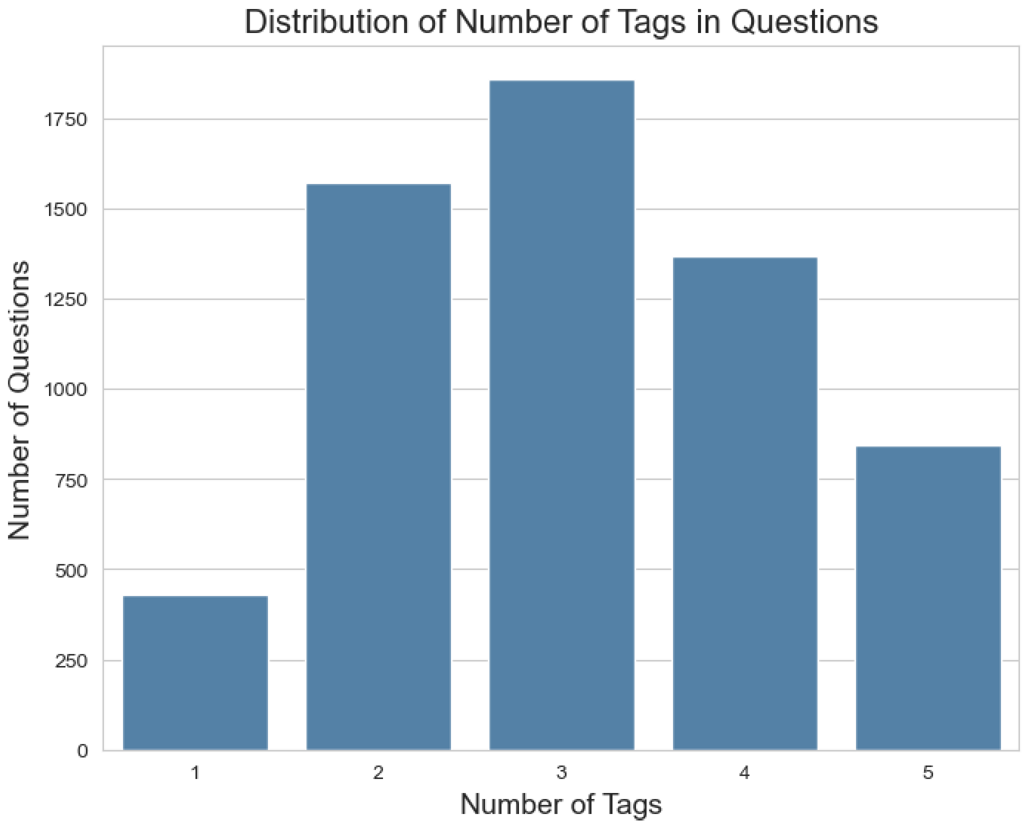
	Id	Tag
0	469	python
1	469	osx
2	469	fonts
3	469	photoshop
4	502	python

Data Pre-Processing

- Remove irrelevant columns, rows with Nan and duplicates
- Merge two tables on the Id
- Put all tags in a list(GroupBy)
- Sampleing subset from the whole dataset(1%)
- Remove HTML tags, punctuations and stopwords
- Lower cases
- Concatenate Title and Body



	Tag	Text
0	[python, list, dictionary]	pop out the whole dic if element of 1st dic in...
1	[python, rest, python-3.4, yql, yahoo-weather-...	how to create a rest query for yahoo weather i...
2	[python, language-features, with-statement]	what is the python `` with '' statement design...
3	[python, regex]	regex to strip only start of string i am tryin...
4	[python, python-2.7, logging]	how can i temporarily redirect the output of l...

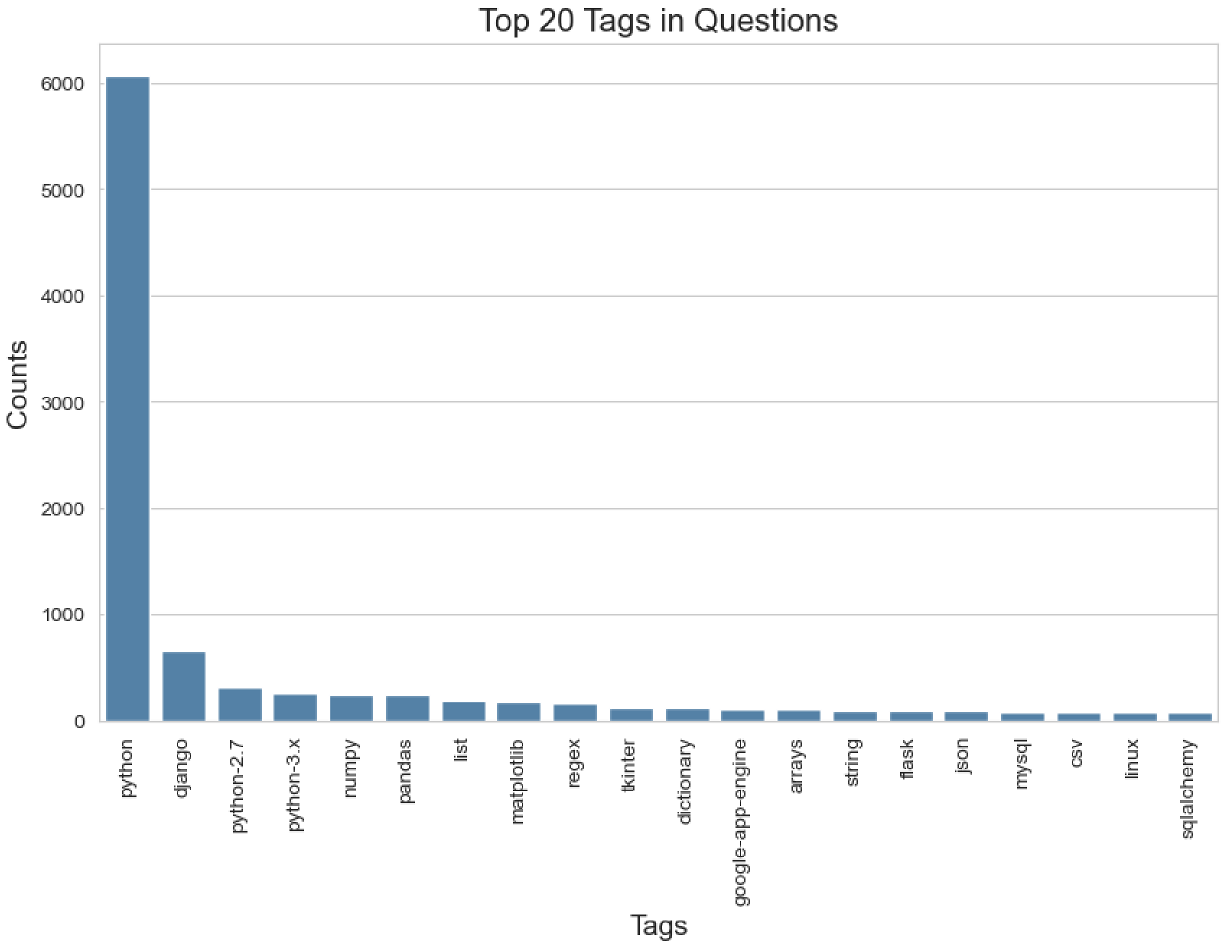


Data Analysis

~6k questions

~4k unique tags

~2k tags appeared only once



Data Analysis

- Issues
  - Imbalanced dataset
  - Many tags appeared only once
- Solutions (Future Work)
  - Reduce tags complexity
  - Data augmentation

Modeling

- Machine Learning Models:
  - SGDClassifier
  - LogisticRegression
  - LinearSVC
- Deep Learning Models(Future Work):
  - BERT
  - StackOverflowBERT

Evaluation Metrics

- Micro F1 Score: Calculate F1 globally.
- Macro F1 Score: Calculate F1 for each label, and find their unweighted mean.
- Hamming Loss: The Hamming loss is the fraction of labels that are incorrectly predicted.

Model Performance

	MultinomialNB	LR	SVC	BERT	StackOverflowBERT
F1 Micro	**64.83%**	62.86%	64.34%	/	/
F1 Macro	**62.00%**	53.82%	60.94%	/	/
Humming	0.07%	0.07%	0.07%	/	/

Explainability(Future Work)

Feature attribution methods like **integrated gradients**, **SHAP** and **attentions score**(transformer-based LM) can be used to explain the model's prediction.

Explainability tools not only build trust in our model, but also help us to generate useful labels for continuous training.

Development(Future Work)

- Production Enviroment
- Data Engineering
- Performance Optimization
- Monitoring and Logging
- Human in the Loop(HITL)

In summary, CI(testing and validating data and models), CD(training a pipeline and automatically deploy a model prediction service), and CT(automatic model retraining whenever the set model threshold is breached) of MLOps.