Mini Project2 Liu Yue

Project Goal

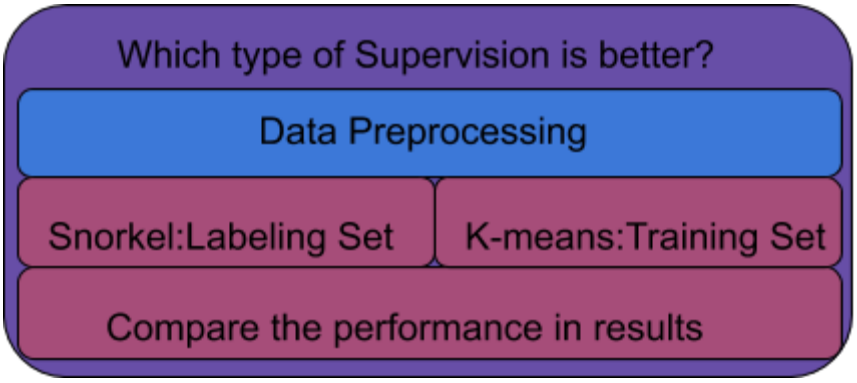
1. Analysis about the state of the art of a particular topic (typically, topic you did not take classes in)
2. In group discussion of your results
3. Joint presentations to class by the group

MVP & User Stories

1. Minimum Viable Product(MVP):
   (1) Labeling the data by using the Snorkel API;
   (2) Clustering the tweets based on the similarity.
2. User Story:
   (1) As a tweeter, I want to filter the fake tweets of disasters;
   (2) As a student in computational linguistics, I want to know users' understanding about the words describing a real disaster;

Modular Design

In this project, we focus on the mechanism of fake news detection and sentiment analysis and manipulate a larger range of data for the further precise visualization in network and map.



Introduction to the topic

In supervised learning, we need to get a collection of labeled target data, or at least some scoring system. However, in unsupervised learning, we don't have a specific goal. Unsupervised learning can be considered as a way to reduce the dimensions. Most of the unsupervised learning is a form of cluster analysis: some objects in the clusters are very similar, while others are distinct.

It depends, we can even apply classification in the unsupervised learning, as long as the goal is to identify similarities between the input. For example, we can use k-Nearest Neighbors (k-nn) to classify the similar groups of inputs, if we did not pick up an output label.

K-means clustering method is a frequently used method, which contains four steps: (1) select k "classifier" points at random; (2) classify according to closest classifier point; (3) Replacing the classifier points by the centroids; (4) Repeating steps 2 and 3 until set membership stabilizes.

In the real world, we can use k-means clustering to site cell phone towers; we can also use Fuzzy c-Means Clustering to analyze Gene Expression Data. Using a self-organizing map, neural network based clustering can transform a dataset into a topology-preserving 2D map. Dimension reduction can be used to lower the complexity, and it will be good to preprocessing the data for supervised learning.

Summary of references

Semi-supervised learning still cannot get rid of the structural assumptions. However, weak supervision can successfully avoid it by using subject matter experts (SMEs). We can achieve weak supervision in three ways: (1) providing higher-level, less precise supervision; (2) giving cheaper, lower-quality supervision; (3) taking existing resources.

New Snorkel(Snorkelv0.9) has three important functions. The goal of snorkel is to finish the operations of practitioners in the training data.

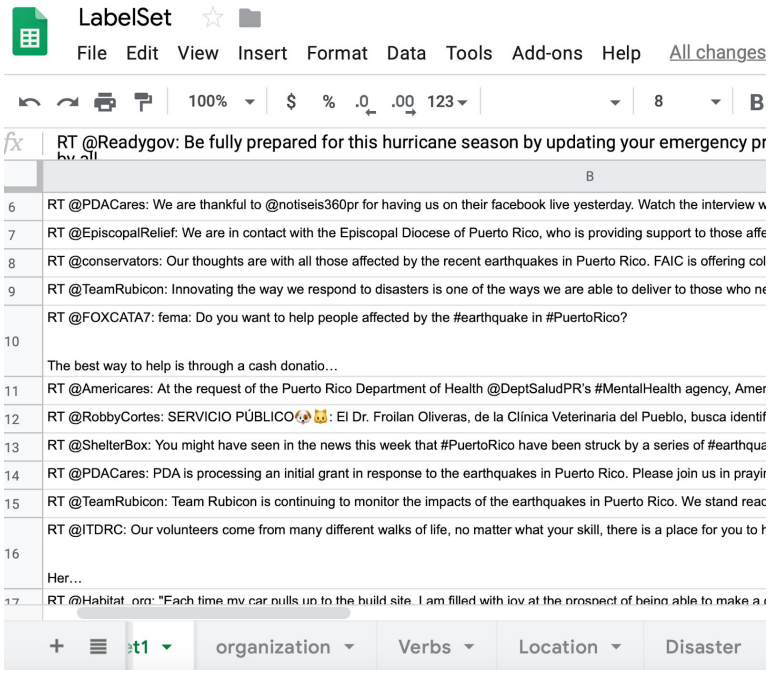
Functions	**Techniques**
Labeling Function (LF)	Labeling Training Data

Transformation Functions (TF)	Data Augmentation
Slicing Function (SF)	Monitoring Critical Data Subsets

Unlike the previous version, TF is the most important function because it enables the user to apply the machine learning models.

Snorkel can be used in two contrast applications: Natural Language & Images. Hancock, Braden et al. (2018) introduced the Babble Labble in natural language, a framework with semantic parser converting explanations into executable functions. While Varma Paroma et al. (2017) presented Coral, a paradigm is allowing users to write heuristics to label training data to interpret the raw pixels.

Analysis of results including pros and cons



The labeling dataset is collected from published tweets, using the application in the Mini Project1. The label set covers all the necessary columns for this project and will be used to get word lists in the LF API.

```
Module1_tfidf_preprocessing
Empty DataFrame
Columns: [Unnamed: 0, Tweets, id, keyword, location, text, target]
Index: []
200
(200, 7)
Empty DataFrame
Columns: [id, keyword, location, text, target]
Index: []
7613
(7613, 5)
Empty DataFrame
Columns: [id, keyword, location, text]
Index: []
3263
(3263, 4)
```

The training dataset and testing dataset are from a Kaggle ongoing competition, *Real or Not? NLP with Disaster Tweets*. There are five columns in the training set: id, keyword, location, text ,and target. We don't have the column of 'target' in the training set. Then, we use nltk library and sklearn library to split the phrases into root word and convert them into numbers for further calculation.

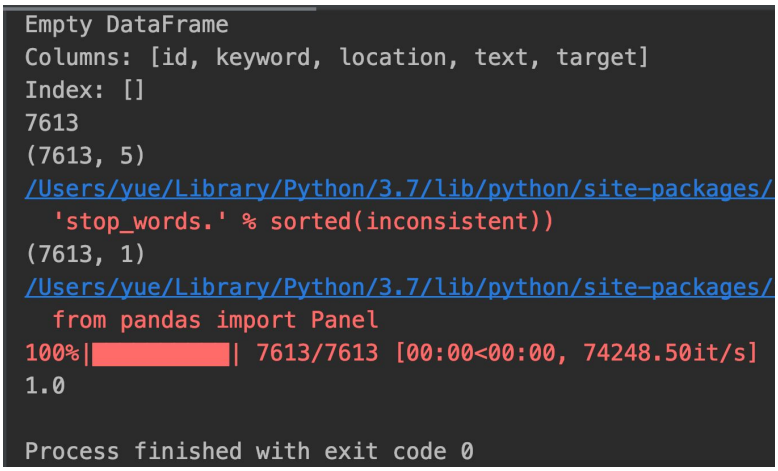
```
  'stop_words.' % sorted(inconsistent))
```

We will ignore the warning at this stage; it is not an error. After that, we will begin the weak supervision by using Snorkel.

```
apply.dask.DaskLFApplier
LFAnalysis
LFApplier
LabelModel
LabelingFunction
MajorityClassVoter
MajorityLabelVoter
lf.nlp.NLPLabelingFunction
PandasLFApplier
apply.dask.PandasParallelLFApplier
RandomVoter
apply.spark.SparkLFApplier
lf.nlp_spark.SparkNLPLabelingFunction
filter_unlabeled_dataframe
labeling_function
lf.nlp.nlp_labeling_function
lf.nlp_spark.spark_nlp_labeling_function
```

Label package in new Snorkel

I pick up the LableModel and labeling_fucntion to do the first step of labeling, with a result of 1.0 in the coverage score(the percentage of all candidates being labeled).



It is easy to label the data by Snorkel because there is no need to worry about the complex steps in building a professional knowledge base, and a scientific report will be easy to do. For now, it is better than Google Auto ML because we can use the API in Python in a large dataset.

Recommendations for a person who want to develop or use such systems

1. Use a third-party Knowledgebase after being familiar with the labeling. It is also recommended by many experts in machine learning, as well as the Snorkel workshop. As a beginner, I should follow this way later.
2. Try the dataset in your native language. It is worthwhile to test the accuracy of the linguists, who believe in the existence of a general syntax. Without letting the two groups of people get to know the difference between the industry and domain, it is hard to improve the model.
3. Add more complex models. My learning route is to begin with the most basic algorithm first, for example, k-nn classification and k-means clustering. You can't imagine how many mistakes will appear without repeating in different datasets on your own. Also, there are so many variants of the algorithms, which are highly related to optimization.

Conclusions

Machine Learning Type	Model/Algorithms	Pros	Cons
Weak Supervision	Snorkel	Less time in labeling the data; The model result is more explanatory; A balance between expert labeling and technician modeling	Labeling itself needs a balanced skill in domain and technology; Version Control: the users need to know the latest methods, and what are deprecated.
Unsupervised Learning	K-means Clustering	Easy to call a complex library to solve the problem	Hard to explain the best choice of the k value.

Group Discussion

Report Title
Future Languages in Machine Learning
State of the art of object segmentation methods based on machine learning algorithms
Machine Learning and Artificial Intelligence
A survey of Dimensionality Reduction Techniques
Multi-View 3D Object detection network for autonomous driving

References

[1] Marsland, S. (2014). *Machine learning: an algorithmic perspective*. Chapman and Hall/CRC.

[2] Ratne, Alex., Bach, Stephen., Varma, Paroma., Ré, Chris (2017-07-16). An Overview of Weak Supervision [Web blog post]. Retrieved from https://www.snorkel.org/blog/weak-supervisionL.

[3] Mastering Machine Learning: A Step-by-Step Guide with MATLAB Retrieved from https://www.mathworks.com/discovery/unsupervised-learning.html.

[4] Bach, S. H., Rodriguez, D., Liu, Y., Luo, C., Shao, H., Xia, C., ... & Kuchhal, R. (2019, June). Snorkel drybell: A case study in deploying weak supervision at industrial scale. In *Proceedings of the 2019 International Conference on Management of Data*(pp. 362-375). ACM.

[5] Hancock, Braden., Liang, Percy., Ré, Chris (2018-05-15). Training with Natural Language [Web blog post]. Retrieved from https://www.snorkel.org/blog/babble.

[6] Ratne, Alex., Bach, Stephen., Varma, Paroma., Ré, Chris (2017-07-16). An Overview of Weak Supervision [Web blog post]. Retrieved from https://www.snorkel.org/blog/weak-supervision .

[7] Ratne, Alex., Bach, Stephen., Ré, Chris (2017-12-01). Programming Training Data [Web blog post]. Retrieved from https://www.snorkel.org/blog/snorkel-programming .

[8] The Snorkel Team., (2019-08-14). Introduction to new Snorkel [Web blog post]. Retrieved from https://www.snorkel.org/blog/hello-world-v-0-9 .

[9] Varma, Paroma., He, Bryan., Ré, Chris (2017-09-14). Snorkel for Image Data [Web blog post]. Retrieved from https://www.snorkel.org/blog/coral.