I have a list of companies (not classified by vertical) and their investments. I also have another list of companies with just their description and classification. I want to find out which verticals energy companies are investing in to say where corporations foresee the future of the industry. Are they investing in companies in the same verticals, complimentary verticals or completely different verticals.

### Training Data

- **Table A** - Energy companies and their descriptions
- **Table B** - Non-energy companies and their descriptions
- **Table C** - Unclassified companies with their descriptions and investments

### Features:
- We used TFIDF to get the features. The size of the vector is 300. I tried it at different levels and above 300 were marginal improvements so I decided to stick it out with 300. The vocab size is 130,977 for the entire dataset and it was 98,349 for the training set.

### Explanation:
- I ran both regular classifiers and also neural nets. The training accuracy and testing accuracy so I was overfitting to the data. When I increased the dropout rate all the way to 0.8, then the difference became more reasonable. I used sigmoid since it was a binary classification problem.

### My work

I experimented with different dropout rates even. Even at 0.5, there was a significant difference between training accuracy and testing accuracy so I was overfitting to the data. When I increased the dropout rate all the way to 0.8, then the difference became more reasonable.

I performed mini-batch gradient descent with a batch_size of 40 and at 30 epochs. The optimal accuracy was hit by the 3rd and 12th epoch so it wasn’t necessary but I let it run since it was going pretty quickly anyways.

With all of this we were able to achieve:

- Training accuracy: 0.9788
- Testing accuracy: 0.9711

I had more than 1 dense layer. I performed mini-batch gradient descent with a batch_size of 40 and at 30 epochs. The optimal accuracy was hit by the 3rd and 12th epoch so it wasn’t necessary but I let it run since it was going pretty quickly anyways.

With all of this we were able to achieve:

- Training accuracy: 0.9788
- Testing accuracy: 0.9711

### Machine Learning

2nd model:

- **Layer**: Output Shape
- **F-score**:
- **CNN**:
- **Passive-Aggressive**: 96%
- **SVM (Linear SVC)**: 96%
- **Logistic Regression**: 97%
- **Ridge Classifier**: 96%
- **Ridge Classifier**: 96%
- **Logistic Regression**: 97%
- **CNN**: 99%
- **CNN**: 99%

### Related Work:

- Yoon Kim’s CNN but I used sigmoid for the last sense layer because it is a binary classifier and I had more than 1 dense layer.

These changes were made through trial and error and thankfully brought us to an even higher accuracy than I thought was possible. At the moment, I am still overfitting a bit but I increased the dropout to 0.6. This allowed us to achieve an accuracy of 0.9732.

### Discussion:

To the right is the visualization of our results, energy companies invest the most in productivity software, energy production and alternative energy equipment verticals the most. It makes sense that energy incumbents would primarily invest in energy production and alternative energy equipment but it was surprising to see many of them making investments in the productivity software space. Our results were really good as you can see with the accuracy of the various classifiers which we used.