Capstone II Milestone Report:
Topic Identification of Math Word problems using Unsupervised / Deep learning

Sukanya Chandramouli

May 2018

Introduction :

This report provides a summary of my work on the Capstone project (Topic Identification of Mathematical Word Problems using NLP / Unsupervised / Deep learning methods). I have used the Algebraic Question & Answer Dataset (DeepMind's **AQuA** dataset) that has math word problems. This dataset is available here

Project Goals:

The dataset consists of about 100,000 algebraic word problems with natural language rationales. Each problem is a 'json' object consisting of four parts:

- question - A natural language definition of the problem to solve
- options - 5 possible options (A, B, C, D and E), among which one is correct
- rationale - A natural language description of the solution to the problem
- correct - The correct option

In this project I have used the "question" in the dataset to perform semantic analysis and to categorize/cluster the questions using NLP and ML methods.

The motivation is to identify the underlying mathematical concepts so that a learning system can be built for teaching math. Such a classification will help in building an adaptive learning system to reinforce the math concepts through video tutorials / progressive testing.

Tools Used

Pandas,JSON	Data Wrangling
nltk , string , re	Text Pre processing , regular expression
Scikit learn text	Count Vectorizer , TFIDF Vectorizer
Scikit learn	KMeans Clustering , TSNE, TruncatedSVD
Scikit learn Metrics	Silhouette Score
Matplotlib & Seaborn	Data visualization
Keras Text	Tokenizer
Keras Layers	Sequential , Dense, Flatten, Conv1D, MaxPooling1D,Embedding
Word Embedding	GloVe

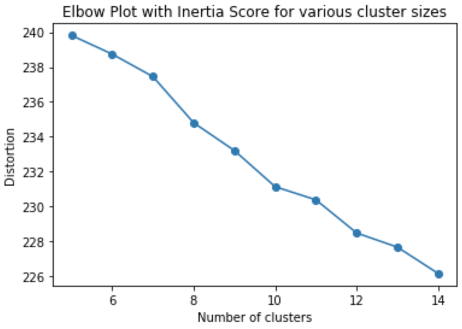
Text Pre-Processing

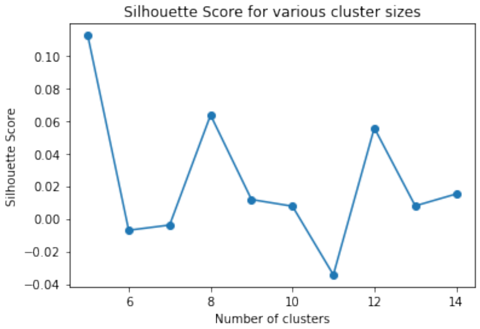
To build a vector representation of the questions in the dataset , I have applied the following pre-processing steps

- Stemming : reduce word variations to simpler forms to derive the roots . This helps to increase the coverage of NLP utilities.
- Named Entity Removal – identify and remove named entities (proper nouns) to simplify downstream processing.
- Tokenization/ Bag of Words –Break up the text into words and create a unique list of words

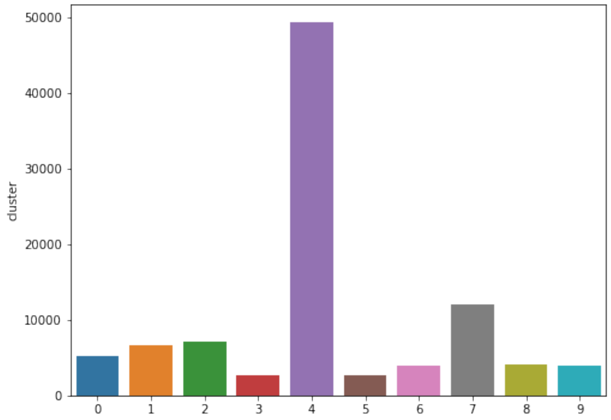
Vectorization with TFIDF Matrix, Clustering with KMeans

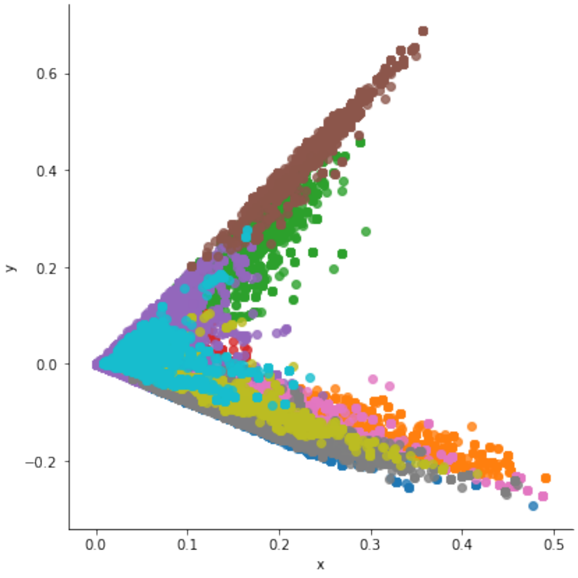
In the first model I have used the sci-kit-learn count vectorizer (term frequency matrix) and tfidf-vectorizer (relative term frequency matrix). This matrix model of the text is then used as input to the k-means clustering algorithm.Using the conventional ML/Clustering metrics , we do not get the perfect number for the cluster from the elbow plot / silhouette score plot .





Setting the number of clusters to10 , we have the following cluster map on the full dataset of 100K word data problems .Looking at the cluster map (after reducing the dimension of TF-IDF matrix using SVD) , we do not see a clear separation between the various clusters)





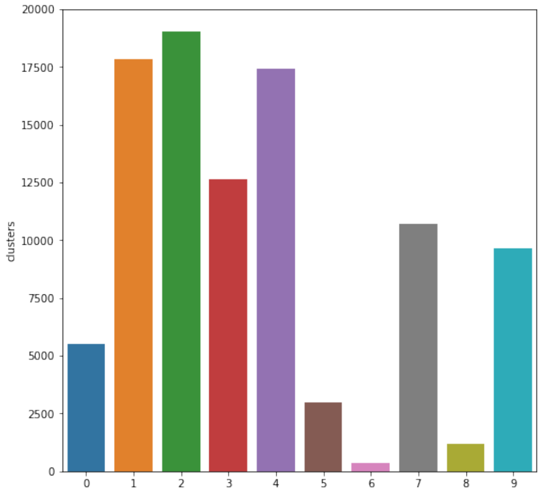
Looking at the clusters individually, we see that while the ML model has done a reasonable job of clustering , there is significant scope for improving the model.

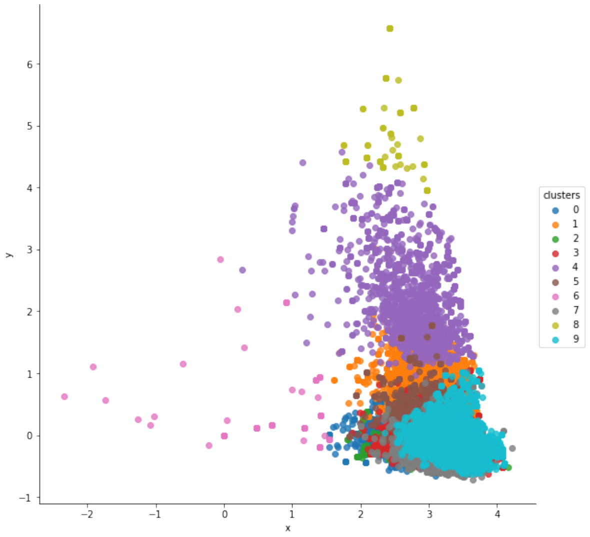
Jay is selling his shoes and making a 15% gain from it. His friend wants to buy them so he offers him a 5% discount. At what price did Jay originally buy his shoes, if he sold it to his friend for $218.50?
The prices of a scooter and a television set are in the ratio 3:2 . If a scooter costs Rs.6000 more than the television set, the price of the scooter is :
Divide Rs.32000 in the ratio 3:5?
In 1978, a kg of paper was sold at Rs25/-. If the paper rate increases at 1.5% more than the inflation rate which is 6.5% a year, then what will be the cost of a kg of paper after 2 years?
A salesman makes a 10% commission on the selling price for each light switch he sells. If he sells 220 switches and the selling price of each switch is $6, what is his total commission?
If a jewelry store wants to sell a necklace for $179.95 next week at a 50% off sale, how much is the price of the necklace this week?

GloVe Word Embeddings, Clustering with KMeans

In the second model , I have used Word embedding to create a vector representation .Word embedding creates a better vector representation by taking the *context* of words and is semantically more meaningful. GloVe (Global Vectors of Word Representation from Stanford) provides a suite of pre-trained word embeddings.

Using Keras Tokenizer we convert text to sequences and map the tokens to vectors using the GloVe embeddings. We next build the features by averaging word vectors for all words in a question .This feature vector is then then used as input to the k-means clustering algorithm. Setting the number of clusters to 10 , we have the following cluster map on the full dataset of 100K word data problems





Comparing the cluster map (after reducing the dimension of feature vector using SVD) , we see this model provides a better separation between the clusters and also performs much better than the previous model .

Solution A has 5% salt concentration and remaining water. After heating 50 litres of Solution A at a certain temperature, due to evaporation, the salt concentration increases to 10%. The amount of water remaining in the solution would be
In a mixture, the ratio of spirit and water is 3:2. If the amount of spirit is 3 litre more than amount of water, calculate the amount of spirit in mixture?
A cube shaped pool is half full of water. If the water is 36 inches deep, how much would the water in the pool weigh if the pool were filled to the brim? (1 cubic foot weighs 56 pounds)
In a mixture of milk and water, there is only 26% water. After replacing the mixture with 7 liters of pure milk, the percentage of milk in the mixture become 76%. The quantity of the mixture is:
Vegetables contains 68% water and green vegetables contains 20% water. How much green vegetables can be obtained from 100 kg of Vegetables ?
x contains 85% water and 15% oil; how many more liters of water than liters of oil are in 200 liters of solution x?
A sink contains exactly 11 liters of water. If water is drained from the sink until it holds exactly 5 liters of water less than the quantity drained away, how many liters of water were drained away?
John needs to mix a cleaning solution in the following ratio: 1 part bleach for every 4 parts water. When mixing the solution, John makes a mistake and mixes in half as much bleach as he ought to have. The total solution consists of 36 ml. How many ml of bleach did John put into the solution?
Three quarts of a bleaching chemical, Minum, contains 5 percent hydrogen peroxide and water. A different type of bleaching chemical, Maxim, which contains 20 percent hydrogen peroxide, will be mixed with the three quarts of Minum. How much of type Maxim should be added to the three quarts of Minum so that the resulting mixture contains 10 percent hydrogen peroxide?
The total weight of a tin and the cookies it contains is 2 pounds. After 3/4 of the cookies are eaten, the tin and the remaining cookies weigh 0.8 pounds. What is the weight of the empty tin in pounds?

Classification with Keras Deep Learning Model

In this, I have built a deep learning model to classify the problem set using a smaller subset of mathematical word problems with the math topics added to the DeepMind dataset (dev.json). This modified dataset (dev-class.csv) has been used to train deep learning model.

As in the previous two models , the first step in the process is to pre process the text and then create a vector representation of the vocabulary as a word embedding matrix. I have used the GloVe embedding for the word vectors . This embedding matrix is used as the Embedding Layer in the convolutional neural network as shown in (https://blog.keras.io/using-pre-trained-word-embeddings-in-a-keras-model.html) . We see a 65% accuracy in this classification which can be further improved by adding more layers to the CNN model.

Conclusion

In summary, using NLP and vectorization with word frequency , TFIDF and then applying the clustering algorithm like KMeans , we can categorize the mathematical word problems based on the topic with reasonable accuracy. We can further improve the performance using word embedding such as GloVe.We can apply deep learning methods using Keras/CNN model to perform Topic Identification. We could build more complex , multi layer model to improve the performance of the classification .