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

Sukanya Chandramouli
May 2018

Introduction :

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

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 "questions" in the dataset to perform semantic analysis and to categorize/cluster the questions using NLP and ML

The motivation of this project 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.

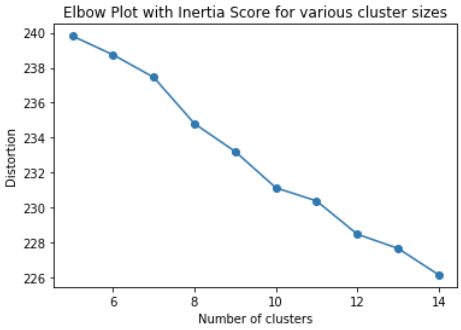
Text Pre-Processing

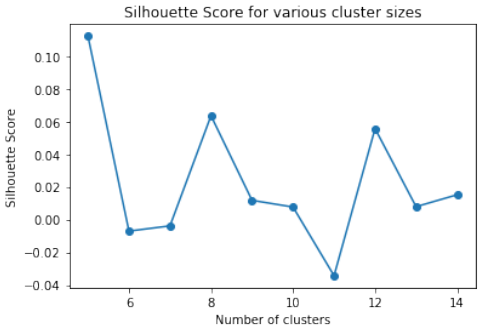
To build a vector representation of the questions in the dataset , I have applied the following pre-processing steps

- 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

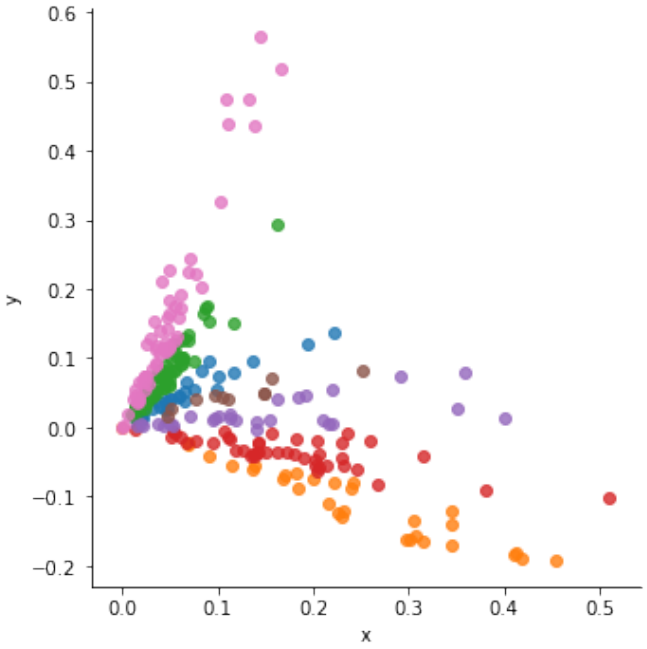
Vectorization & Clustering

To model the text into a vector space , I have used the sci-kit-learn count vectorizer (term frequency matrix) and tfidf-vectorizer (relative term frequency matrix). This matrix model of 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 / silhoutte plot .





Looking at the cluster map (after reducing the dimension of tfidf matrix using SVD) , we do not see a clear separation between the various clusters)



Lets look at some of questions in one of the clusters. From the subjective analysis 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?

Next Steps

As this is an unsupervised learning, one of the challenges of this project is to find the objective metrics to evaluate the performance of the vector space conversion and the clustering model.As a next step I would like to improve the vector space conversion (through use of Word2Vec and GloVe nueral nets) and improve the performance of the clustering with different algorithms and Cross Validation techniques