# Topic Identification of Mathematical Word Problems

Sukanya Chandramouli

Data Science

May 2018

# DeepMind AQuA Dataset

- Consists of 100,000 Math Questions (Word Problems)
- Used by DeepMind for Problem Solving/Rationale Generation
- Dataset is in JSON format
- 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

| | |
|---|---|
| “question”: | Two trains running in opposite directions cross a man standing on the platform in 27 seconds and 17 seconds respectively and they cross each other in 23 seconds. The ratio of their speeds is |
| “options”: | A) 3/7 B) 3/2 C) 3/88 D) 3/8 E) 2/2 |
| “rationale”: | Let the speeds of the two trains be x m/sec and y m/sec respectively. Then, length of the first train = 27x meters, and length of the second train = 17 y meters. (27x + 17y) / (x + y) = 23 → 27x + 17y = 23x + 23y → 4x = 6y → x/y = 3/2 |
| “correct option”: | B |

https://github.com/deepmind/AQuA

# Goals of this Capstone Project

- Categorize the word problems based on the underlying math concepts using Machine Learning Model
- Evaluate the performance of the Machine Learning model
- Build a Deep Learning model for Topic Identification using NLP and CNN

# Text Pre Processing

- Stemming/Lemmatization : 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.
- Stop words removal
- Tokenization/ Bag of Words –Break up the text into words and create a unique list of words

# Vectorization

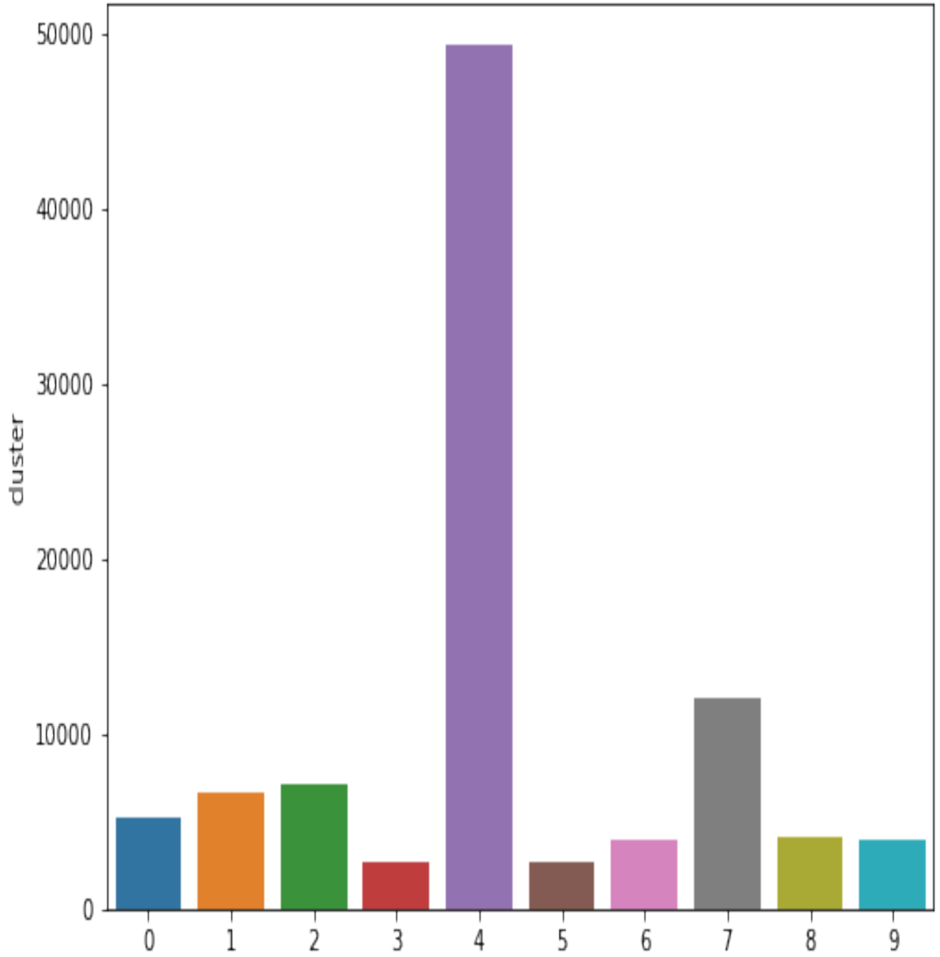
- Bag-of-Words
  - *Simplistic model based on word count (Count Vectorizer)*
- TF-IDF weighting
  - *weight of a term is the product of its tf weight and its idf weight.*

$$\mathrm{w}_{t,d} = (1 + \log \mathrm{tf}_{t,d}) \times \log \; N / \mathrm{df}_t$$

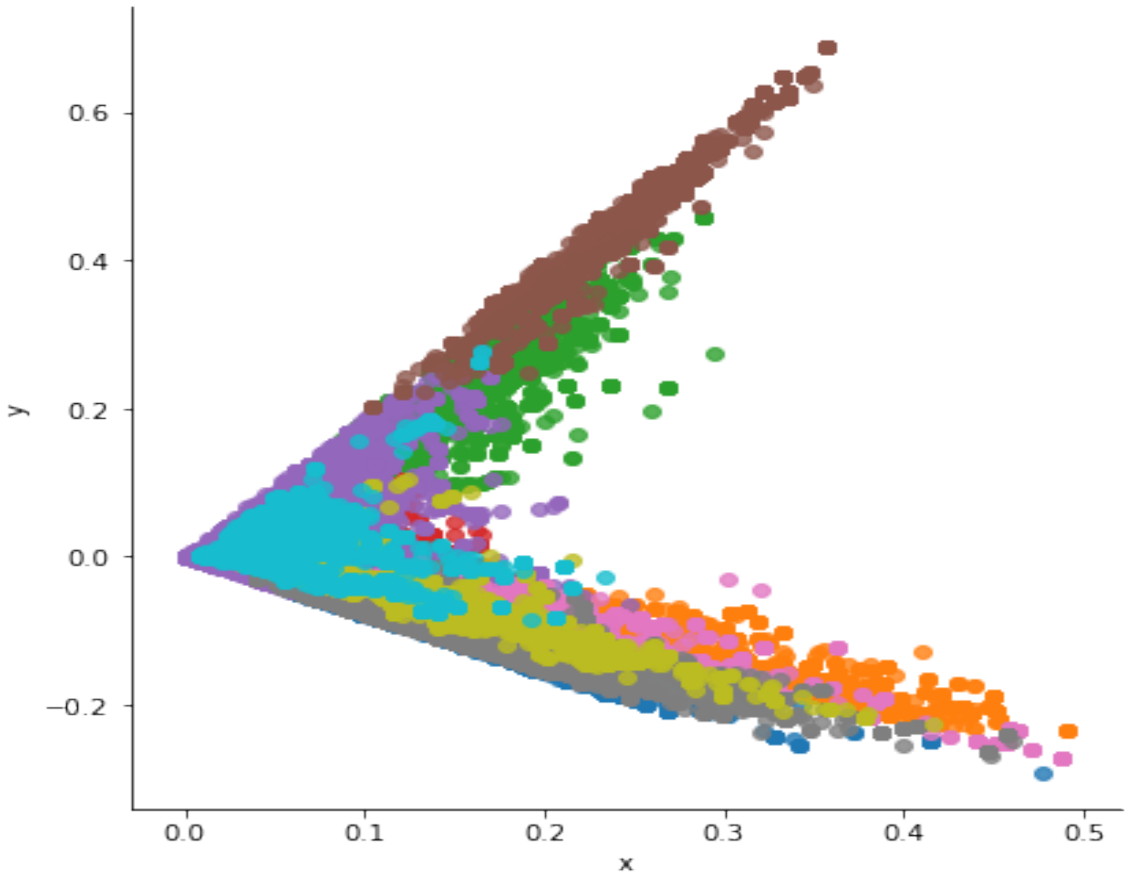
  - *Increases with the number of occurrences within a document*
  - *Increases with the rarity of the term in the collection*

# Clustering
## *Kmeans with TF-IDF matrix*

- Clustering with Kmeans , using Cosine similarity measure as the scoring function
- #Clusters set to 10
- We get the cluster breakup with as shown
- With roughly 50% of the dataset split on cluster-4



# Clustering
## *Kmeans with TF-IDF matrix*



Further analysis of the Cluster Map (with dimensionality reduction with SVD) shows poor separation between clusters

# Clustering

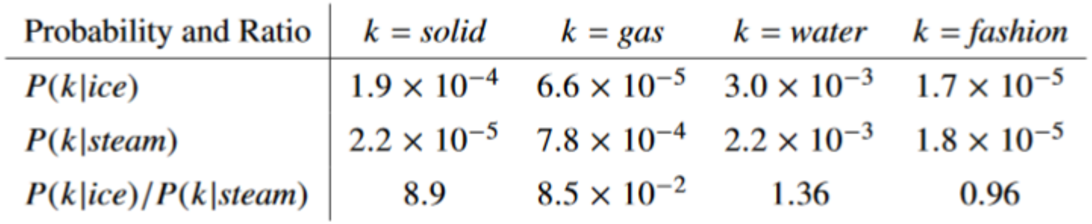
## *Kmeans with TF-IDF matrix*

Analyzing the cluster/questions we see a few wrong placements in cluster as shown below

| |
|---|
| 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? |

# Word Embedding

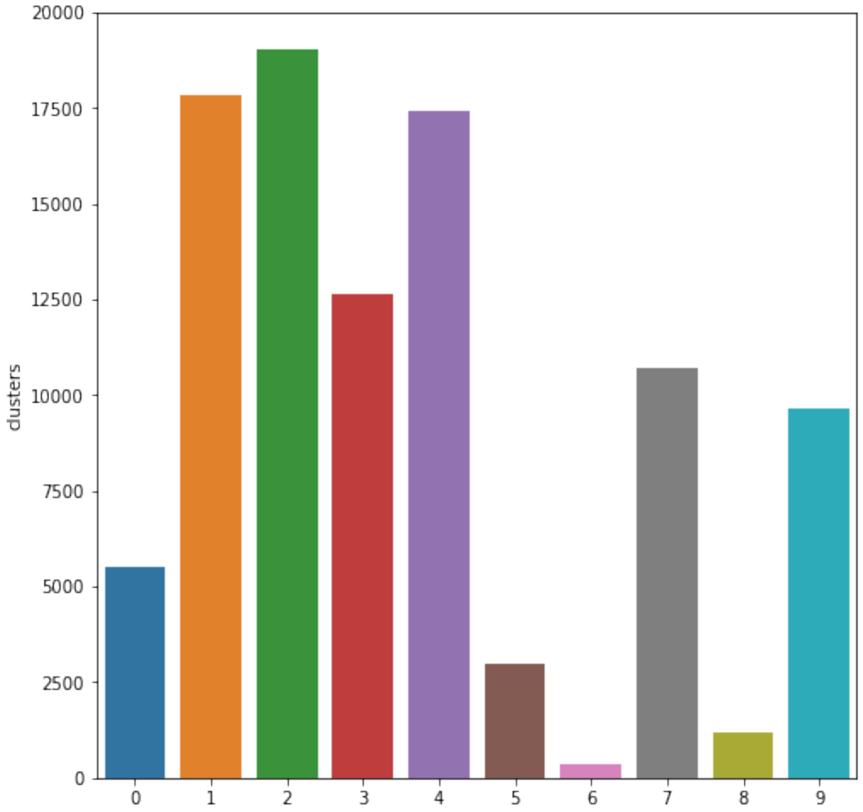
- Word Embedding
  - Dense Vector representation based on word meanings and relationship between words
  - word2vec (neural network based) and GloVe (unsupervised)
- GloVe
  - Global Vectors for Word Representation
  - Uses ratios of co-occurrence probabilities
  - Provides dense vectors in 50,100,200,300 sizes



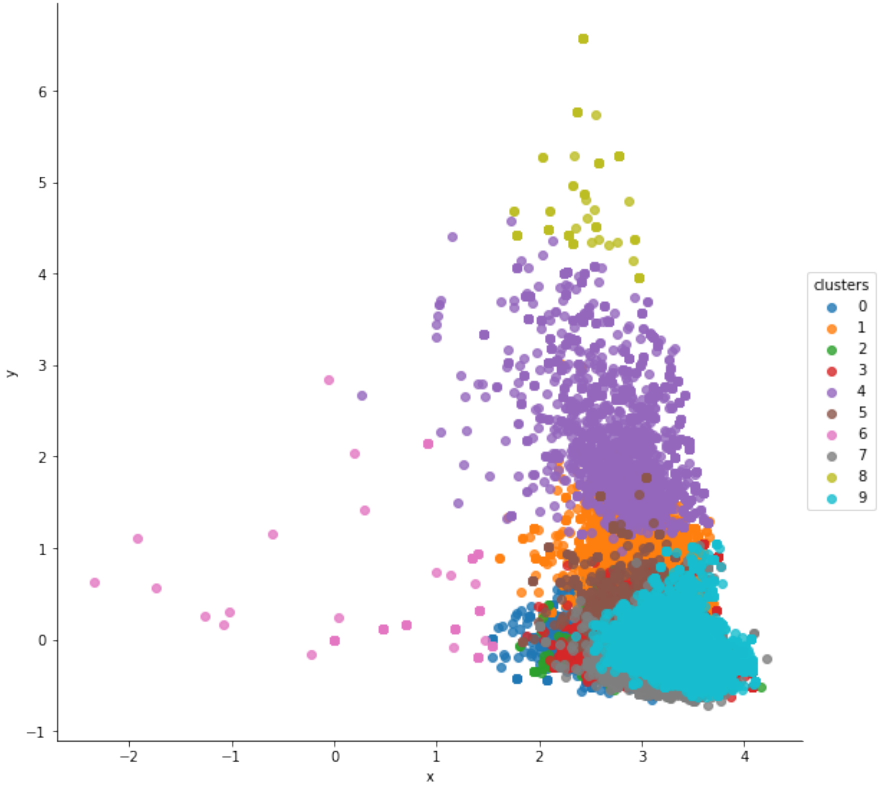
| Probability and Ratio | $k = solid$ | $k = gas$ | $k = water$ | $k = fashion$ |
|---|---|---|---|---|
| $P(k\|ice)$ | $1.9 \times 10^{-4}$ | $6.6 \times 10^{-5}$ | $3.0 \times 10^{-3}$ | $1.7 \times 10^{-5}$ |
| $P(k\|steam)$ | $2.2 \times 10^{-5}$ | $7.8 \times 10^{-4}$ | $2.2 \times 10^{-3}$ | $1.8 \times 10^{-5}$ |
| $P(k\|ice)/P(k\|steam)$ | 8.9 | $8.5 \times 10^{-2}$ | 1.36 | 0.96 |

# Clustering
## *Kmeans with GloVe Vectors*

- Clustering with Kmeans , using Cosine similarity measure as the scoring function
- #Clusters set to 10
- We get the cluster breakup which is more balanced the previous model



# Clustering
## *Kmeans with GloVe vectors*



Further analysis of the Cluster Map (with dimensionality reduction with SVD) shows good separation between clusters

# Clustering

## *Kmeans with GloVe vectors*

Analyzing the cluster/questions we see a better performance than TFIDF based clustering

| |
|---|
| 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? |

# Summary

- Using NLP and vector space model, we can build a ML model to categorize the mathematical word problems
- We get better performance with word embeddings / GloVe vectorization
- We can apply deep learning / CNN model for Topic Identification
- Model performance can be improved with Deep learning methods using multilayer networks