Capstone II Project Proposal :
Categorizing Math Word problems using unsupervised learning

Sukanya Chandramouli

Apr 2018

Background :

DeepMind's **AQuA** dataset consists of approximately 100,000 algebraic/math word problems that has been used to train a program generation model for reasoning,problem solving and rationale generation.In this project we use this extensive dataset to categorize the word problems and map the questions to underlying mathematical concepts.

Dataset

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

Here is an example of a problem object:

{

"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

}

Objective

The goal of this capstone project is to categorize the word problems based on the underlying math concepts. This classification will help in reinforcing the math concepts through video tutorials and also in building the knowledgebase through additional test

problems. I will be using the natural language processing tools and unsupervised learning algorithms to categorize the word problems.

.

Approach

1. The initial analysis will involve building a vector representation (term frequency or tf-idf matrix) using the bag-of-words approach . Tools of the NLTK and GENSIM libraries will be used for text and semantic analysis.
2. The final portion of this project will involve PCA (principal component analysis) as well as clustering and dimensionality reduction algorithms (K-means , agglomerative clustering, naïve-bayes,K-SOM). Depending the various internal metrics we fine tune the model to get the optimal set of clusters for classifying the algebraic word problems

Deliverables

1. Deliverables for this project include the following: 1. The python notebook used for text mining and semantic analysis as well as code for unsupervised learning ; and 2. A report/slide deck that demonstrates the thought process behind the methodology and reports relevant results.