Data 88 - Language Modeling & Text Analysis
Instructor: Professor Adam G. Anderson
Student: Paul Shao

Project Description:
Concerns regarding the effectiveness of elementary and middle school's science education have elevated significantly over the past few years. With a polarizing political landscape and an impasse between federal and state government over the increasing education achievement gap among students with drastically different family and financial backgrounds, an investigation of some of the most commonly asked science exam questions at elementary & middle schools across the U.S. can shed more light on how the current administration (both local and national) navigates the difficult task of providing a impartial, accurate, and modern science education while aligning with the socioeconomic and political interests of the regional voters.

The goal of this project is to visualize and explore the following questions surrounding elementary STEM education in the U.S.:
- What scientific topic(s) do K-12 schools in the U.S. most commonly focus on?
- How does the emphasis on the questions vary through the descriptions of the questions across each scientific subject category?

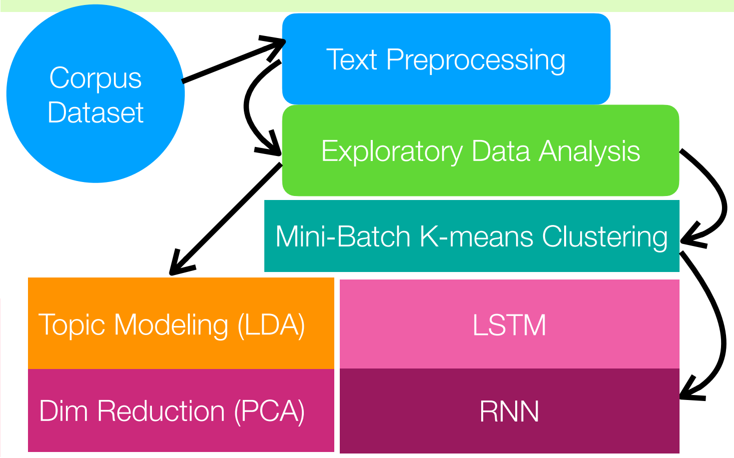
Dataset Used:
To facilitate a broad understanding of how students' understanding of different scientific subjects are assessed across K-12 schools in the U.S., I will incorporate the following dataset in my analysis:
- The Aristo Reasoning Challenge (ARC) corpus, which contains 7787 elementary and middle school standardized exam questions drawn from 12 US states.

Interpretation of Results:
1. Quantitatively assess the distribution of topics/emphasis in K-12 STEM education
2. Evaluate relevance of elementary science education's connection to modern researches and technologies.

Codes: https://github.com/paulshaoyuqiao/k12-stem-exam-topic-modeling
References (with Hyperlinks): NLTK, TensorFlow, Towards Data Science (Visualizing with PyDavis), Understanding LSTMs

Components and Citations:
To allow a thorough analysis, I will break down the investigation into the following segments:
- A combined use of Named Entity Recognition (NER) and Common Entity Recognition (CER) (arXiv:1911.10436 [cs.CL])
- Explanation-based learning based on semi-structure text and constraint patterns (Sebastian Thiem and Peter Jansen, 2019)

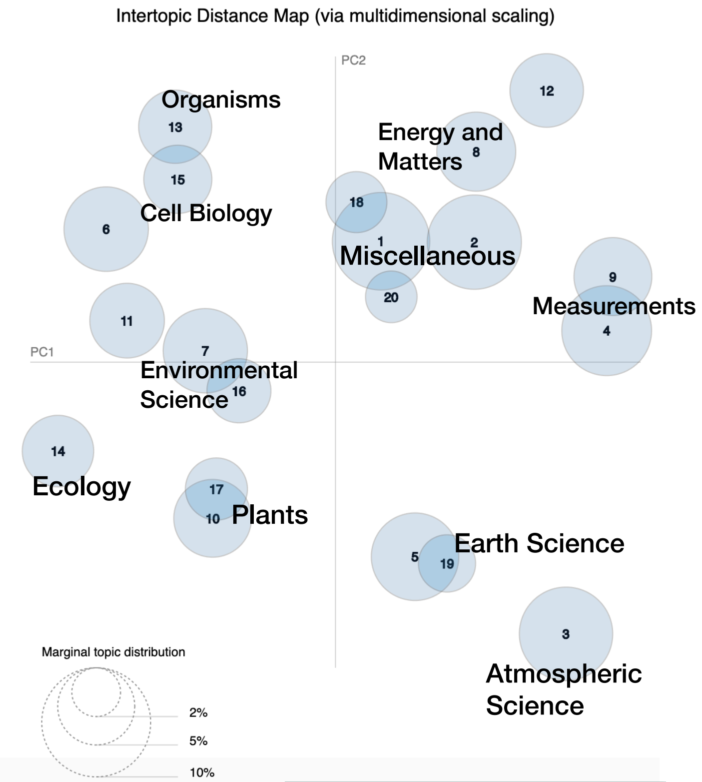
The following workflow (including tools/frameworks) will enable me to complete the analysis:



LDA (Latent Dirichlet Allocation) is a generative statistical model that used unsupervised learning to group texts together by extracting the underlying topics and comparing their levels of similarity and marginal frequencies.

Here, when applying a pre-determined 20 topics, we can see that the topics are further divided into the following main subjects:

Life Science (Organisms), Earth Science (Ecology), Measurement (Experiments), Biology (Cells and Chemistry), Earth Science (Geology), Physics (Energy and Matters), Miscellaneous (Students)



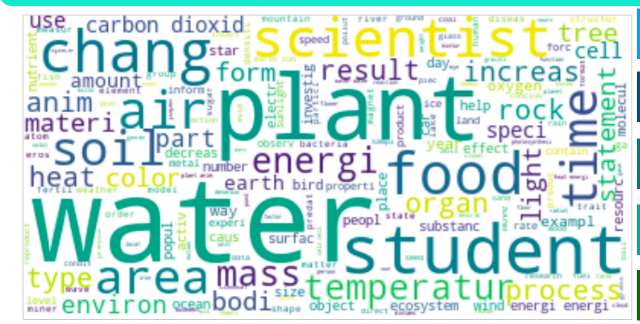
$$\text{Saliency} = f_w \cdot \sum_t P(t|w) \log(P(t|w)/p(t))$$

$$\text{Relevance} = \lambda \cdot P(w|t) + (1-\lambda) \cdot P(w|t)/P(w)$$

Lambda here represents the relevance of a word. A term relevance is determined by their topic-specific probability: the marginal probability the specific term appears within the given topical cluster.

The **Intertopic Distance Map** above represents the similarity between each topical cluster by computing the Jensen-Shannon divergence and applying PCA to project their distances onto a 2D plane for ease of visualization.

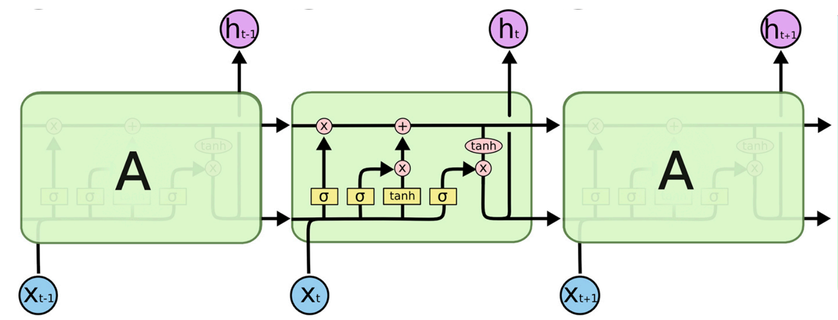
EDA: WordCloud of Word Roots Referenced in ARC Exam Questions



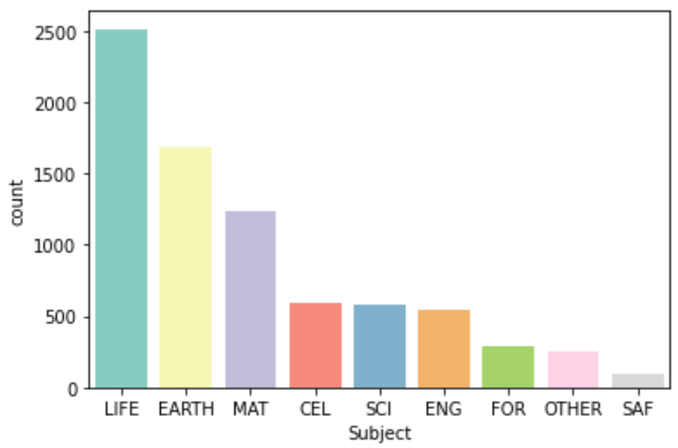
- Tokenization
- Remove Stop Words
- Tagging
- Noun-Filtering
- Port Stemmer

Word cloud is a technique that is used for visualizing the frequency of words in a body of text. The frequency of the word is directly proportional to the size of the text. In order to avoid words with the duplicate roots but different variations affecting the visualization, I also applied the above preprocessing pipeline to each exam question.

As we can see, the word "plant", "food", "soil", "energies", "time" are some of the most frequently used words, suggesting an emphasis on the life sciences, biology, and matters (as in physics).

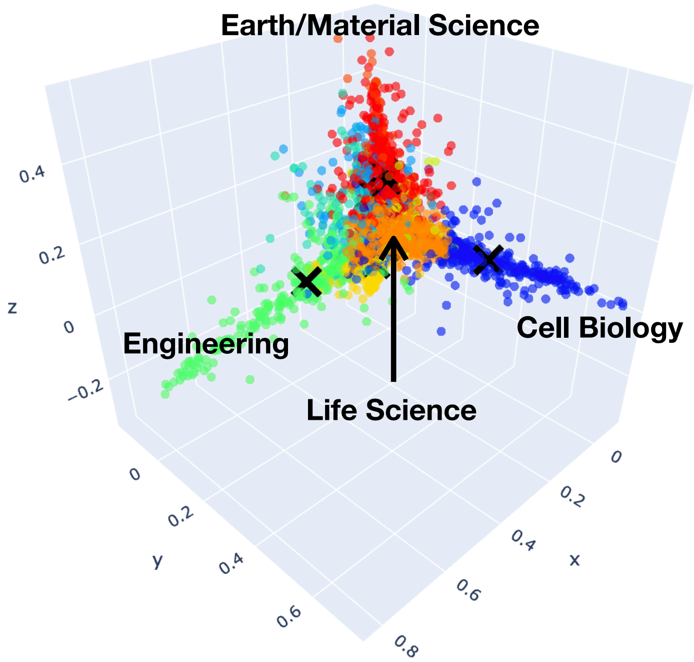


Main Architecture of a Long Short Term Memory (LSTM)-based Recurrent Neural Network (RNN) (for Text Classification)



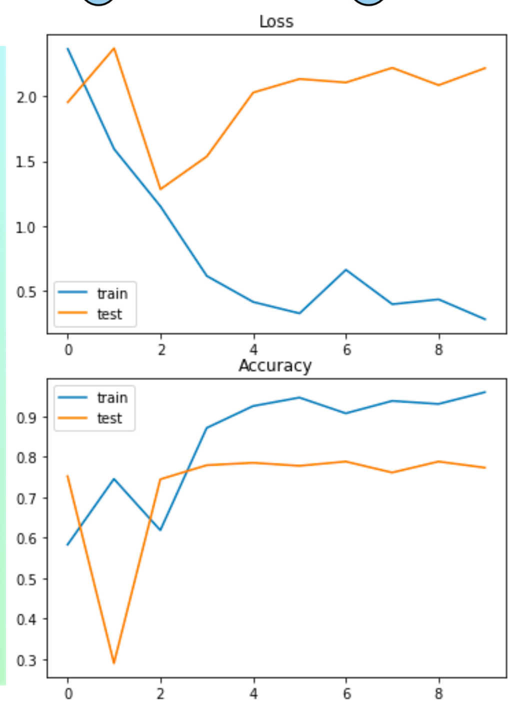
Legends:
LIFE: Life Science
EARTH: Earth Science
MAT: Material Science (Chemistry, Materials)
CEL: Cell Biology
SCI: Generic Science
ENG: Engineering
FOR: Physics (Mechanics)
OTHER: Others
SAF: Experiments & Procedures

Above is a visualization of the distribution of the primary subjects of each question pre-assigned by the ARC. As we can see, Life science, Earth Science, and Material Science (along with a focus on matters and measurements in physics) predominantly represent the majority of the question bank.



K-means Clustering is an unsupervised algorithm that intends to create groups of data points based on the average distance of each point from a set of iteratively improving centroids.

With a pre-determined set of 15 clusters and a principle component analysis (**PCA**) in 3 dimensions, we can see that the topics are broken down along mainly 3 categories: Engineering, Cell Biology, and Earth/Material Science. Toward the center, the topics are concentrated on Forensics and Life Science.



An **LSTM-based RNN** is a state-of-the-art (SOTA) deep learning architecture that performs well in predicting time-series data and working with sequence models.

Here, an LTSM-based RNN is trained on a **TF-IDF** vectorized matrix containing tokens of the ACR Exam questions to predict the underlying subject (one-hot encoded) of the question. Overall a total of 10 training epochs using the Adams optimizer, we can achieve ~76% test accuracy.