# CSE-6242 Team 143

# Deep Skill Stack: A Tool for Creating an Online Curriculum using Machine Learning

Hanlin Chen
Rachel Small
Aubrey Green
Sataporn Woraislpchai
Steven Luu
Maoz Zisman

## Summary

**Deep Skill Stack offers multiple interfaces for users to discover and visualize online courses, allowing them to build their own course curriculum.**

**By combining publicly available data of online courses and creating word embedding for each course, Deep Skill Stack generated a rich similarity visualization among courses, which allows the generation of recommended courses and curriculum per user's input.**

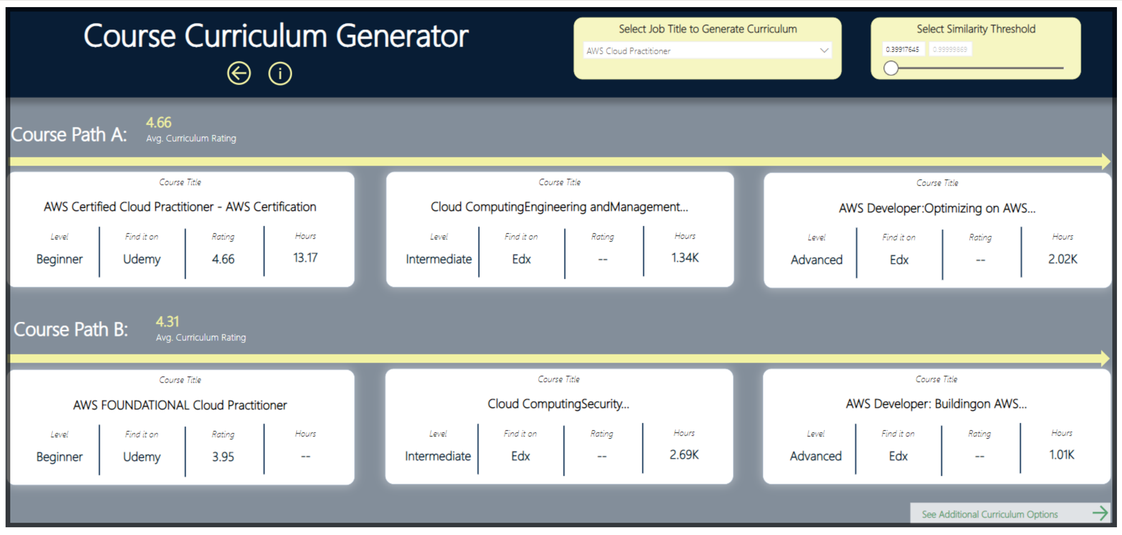
## Our Approach

### Algorithm

- Trained word2vec model on our dataset to generate embeddings and sentence embeddings for each course.
- Generate sentence embeddings for queries and compute the cosine similarity against our dataset to find most similar courses.
- For Course Plan:
  - Return 2 most dissimilar courses among 5 most similar courses for each difficulty (beginner, intermediate, advanced).
- For Graph: Use T-SNE dimensionality reduction to generate (x,y) coordinate pairs to visualize top courses per query.

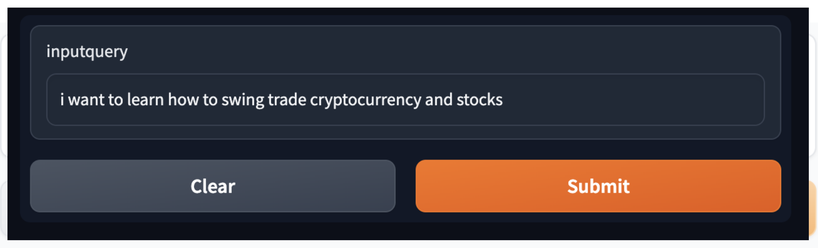
### Deployment

- Generated output datasets for each query using Google Colab and provided to PowerBI.
- Created course plan with top 2 results from each diffifculty level.
- Created graph in PowerBI to plot T-SNE results and provide tooltips for each course.

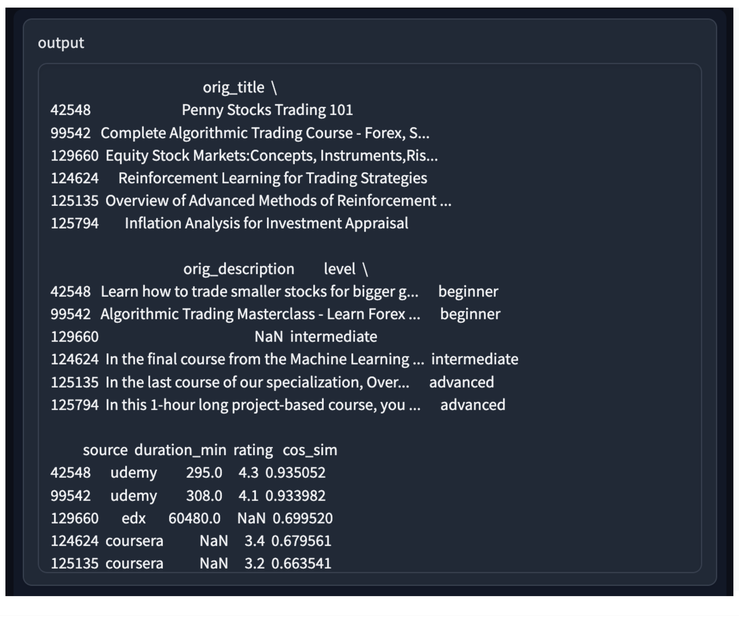



- The Course Curriculum Generator returns the courses most similar to the selected job title at varying levels of difficulty.
- This allows the user to receive several options of course paths, the average rating, and the duration of each course.


- The Course Explorer Tool shows the 40 courses most similar to each job title in addition to their related proximity to one another.
- This allows for the user to browse similar courses and explore related, or non-related job titles.




- The web API hosted on grad.io allows the user to submit queries directly to our model to return a DataFrame or json object containing the course recommendations.



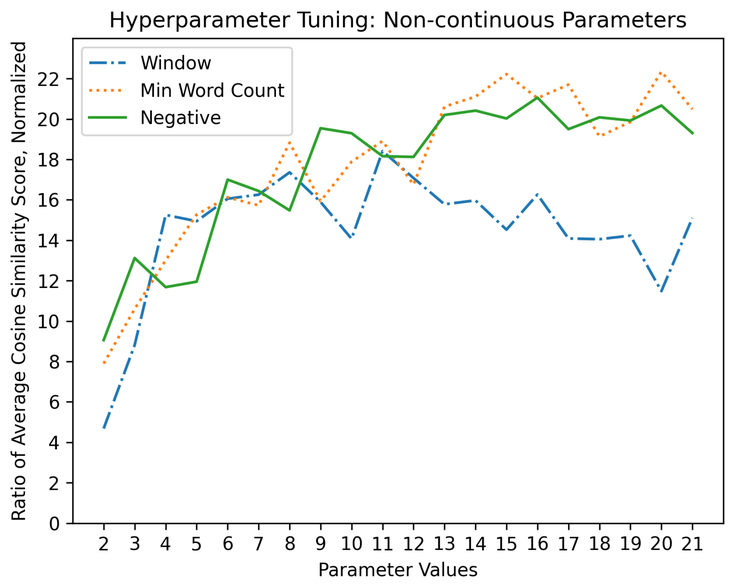

UDACITY coursera SKILLSHARE. udemy edX

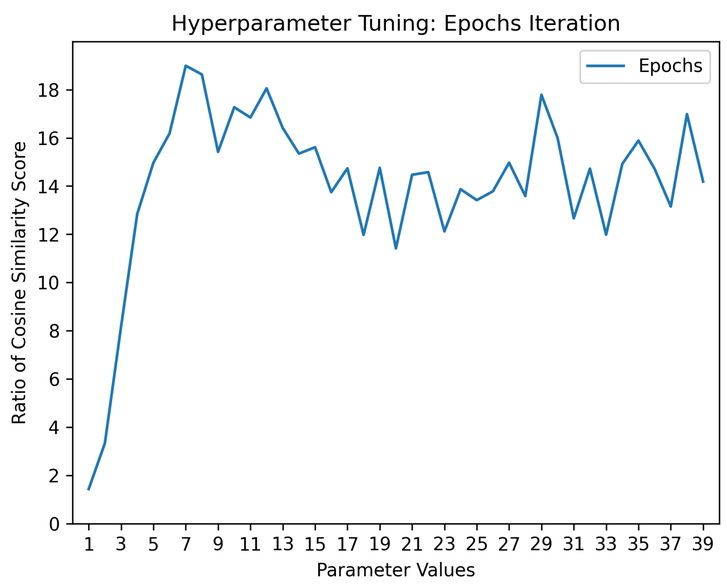
| Source | Courses | Percent of Total |
|---|---|---|
| Udacity | 267 | 0.20% |
| Skillshare | 1,080 | 0.82% |
| Edx | 2,365 | 1.80% |
| Coursera | 3,424 | 2.61% |
| Udemy | 123,915 | 94.55% |

**Average number of words in title & description = 23**

Source: Data set was provided by https://www.kaggle.com/datasets/hossainqh/udemy-courses?select=Course_info.csv, https://www.kaggle.com/datasets/patrickgendotti/udacity-course-catalog, https://www.kaggle.com/datasets/imuhammad/edx-courses, https://www.kaggle.com/datasets/siddharthm1698/coursera-course-dataset, https://www.kaggle.com/datasets/mahmoudahmed6/skillshare-top-1000-course

## Experiments & Evaluation




**The default word2vec parameter performed well by the output of the T-SNE. We try to improve the cosine similarity score by tuning the 4 hyperparameters in this experiment. Parameters that deviated from the default model (n=5) did not produce significant improvement in the cosine similarity score.**