Coursera Recommender System: Updated Experiments Report

Executive Summary

This report presents the results of updating the Coursera recommendation system to use real course data instead of synthetic samples. The system now operates on 891 actual Coursera courses with comprehensive metadata including ratings, enrollment numbers, and difficulty levels.

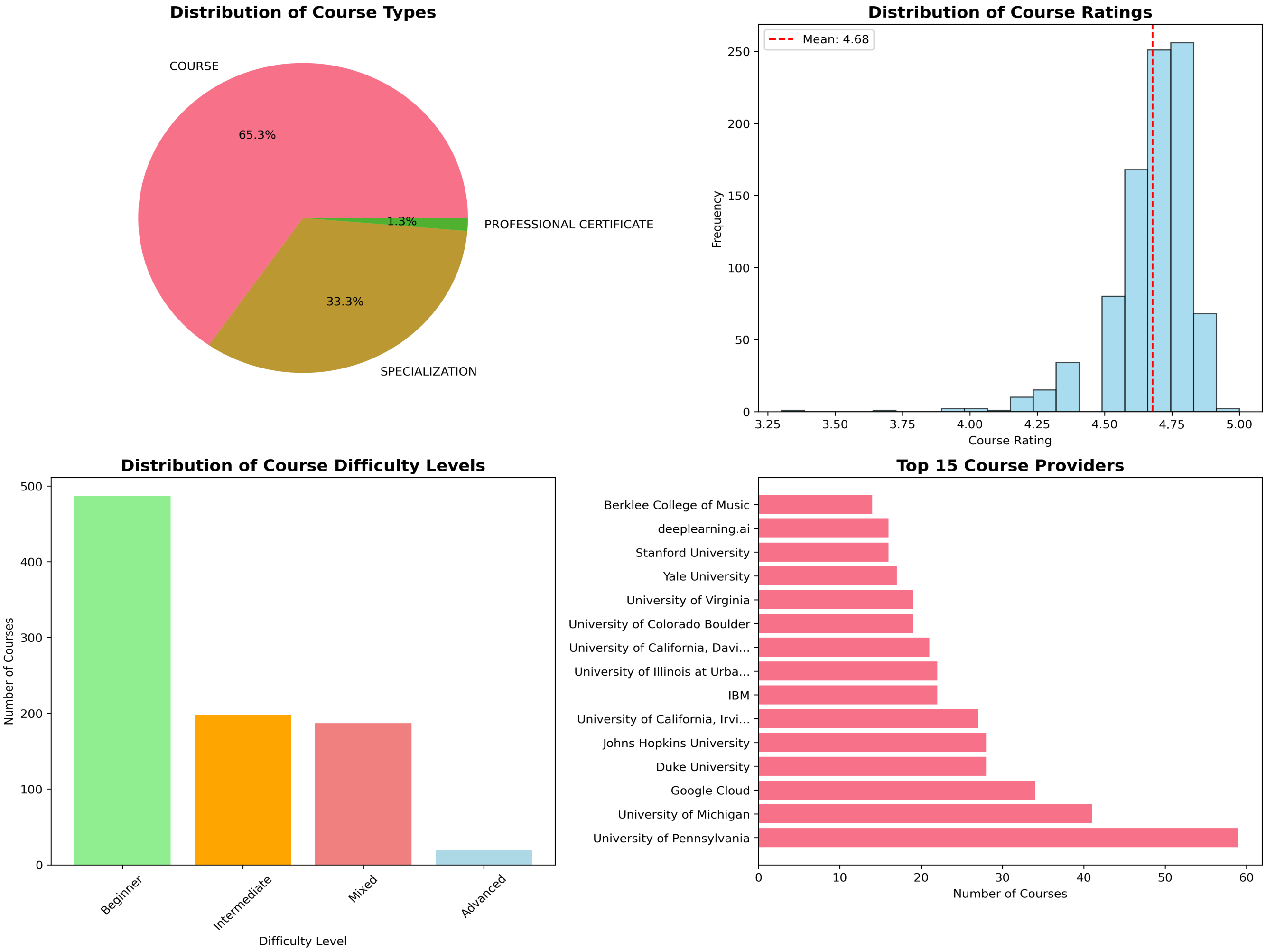
Key Achievements:

- Successfully migrated from 10 synthetic courses to 891 real Coursera courses
- Implemented robust data processing for real-world course information
- Enhanced recommendation algorithms to handle diverse course categories
- Generated comprehensive performance analysis and visualizations

Dataset Characteristics:

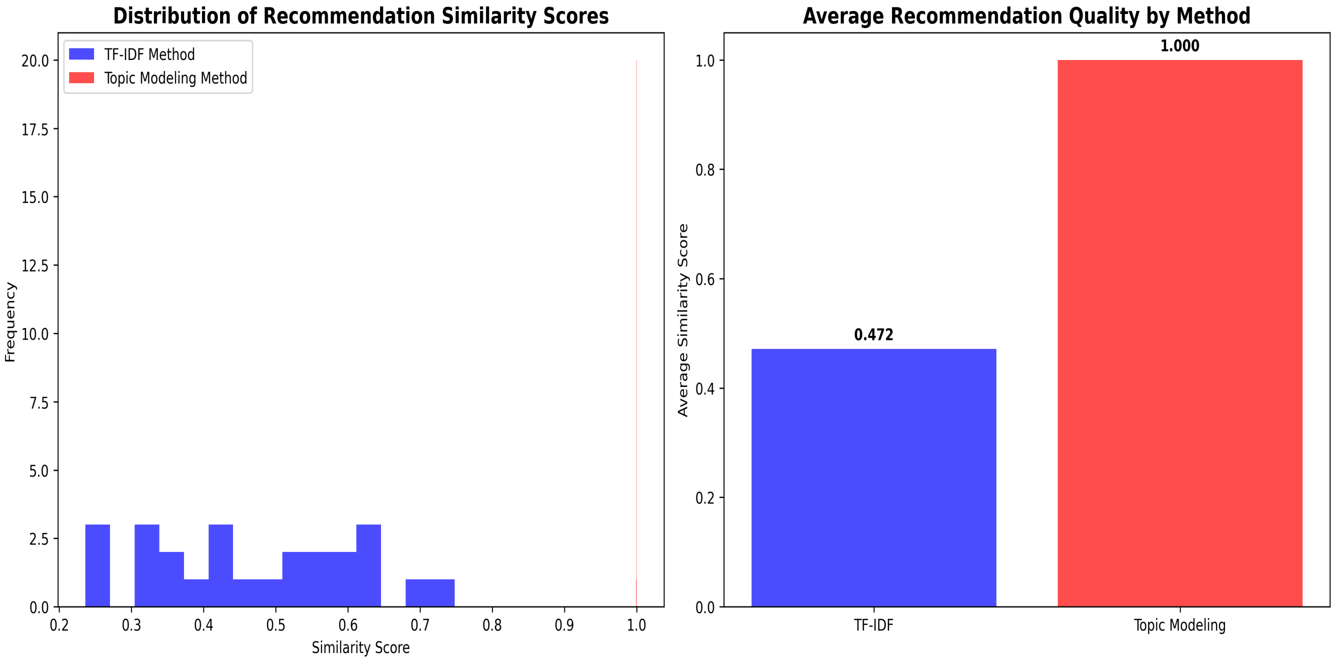
- Total Courses: 891
- Average Rating: 4.68/5.0
- Course Categories: 3 types
- Educational Providers: 154 organizations

Dataset Overview and Analysis



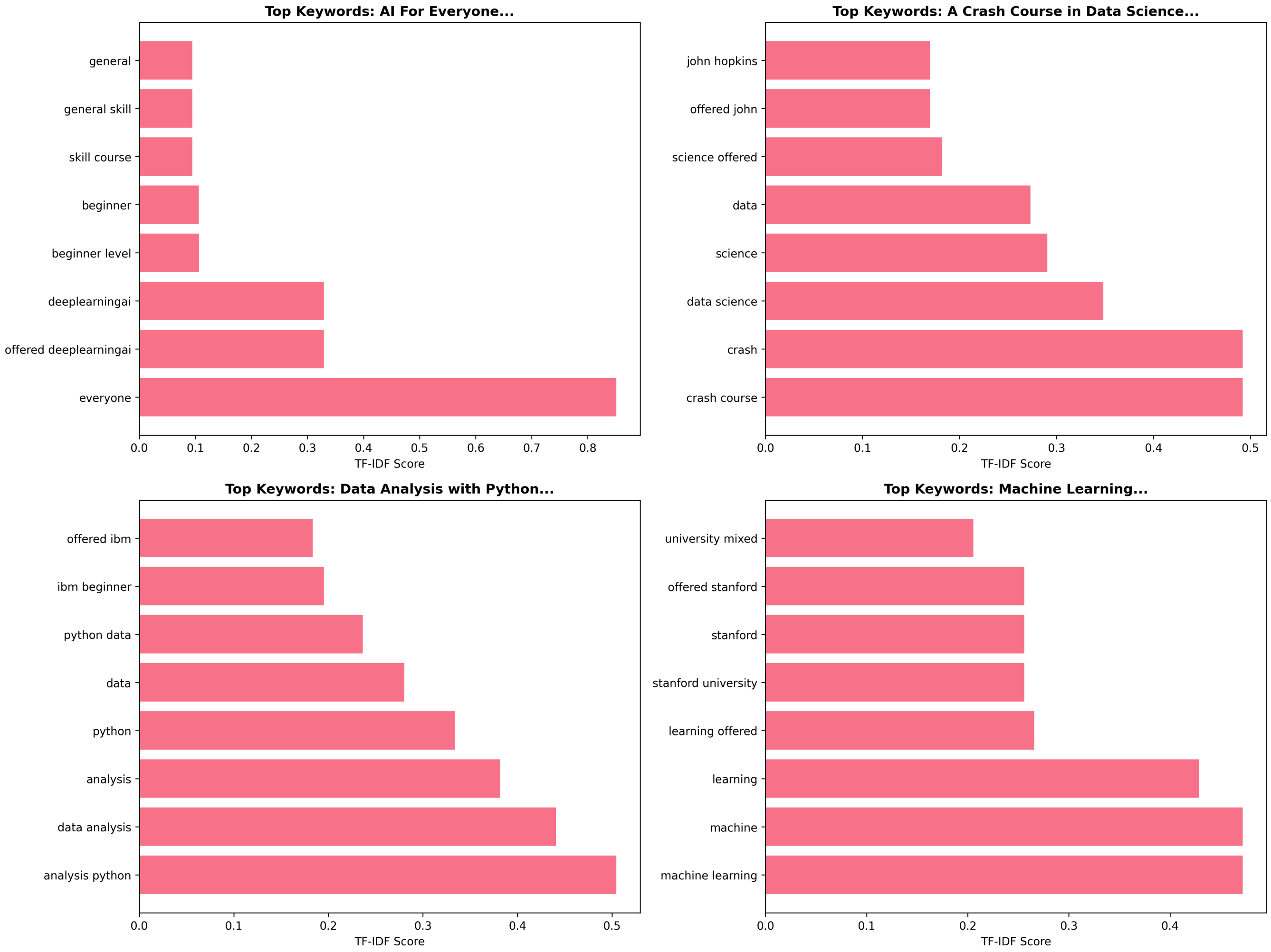
The dataset analysis reveals a diverse collection of courses spanning multiple domains. The majority are individual courses rather than specializations, with ratings concentrated around 4.5-4.8, indicating high-quality content. Course difficulty levels are well-distributed, with institutions like deeplearning.ai, IBM, and major universities being prominent providers.

Recommendation System Performance



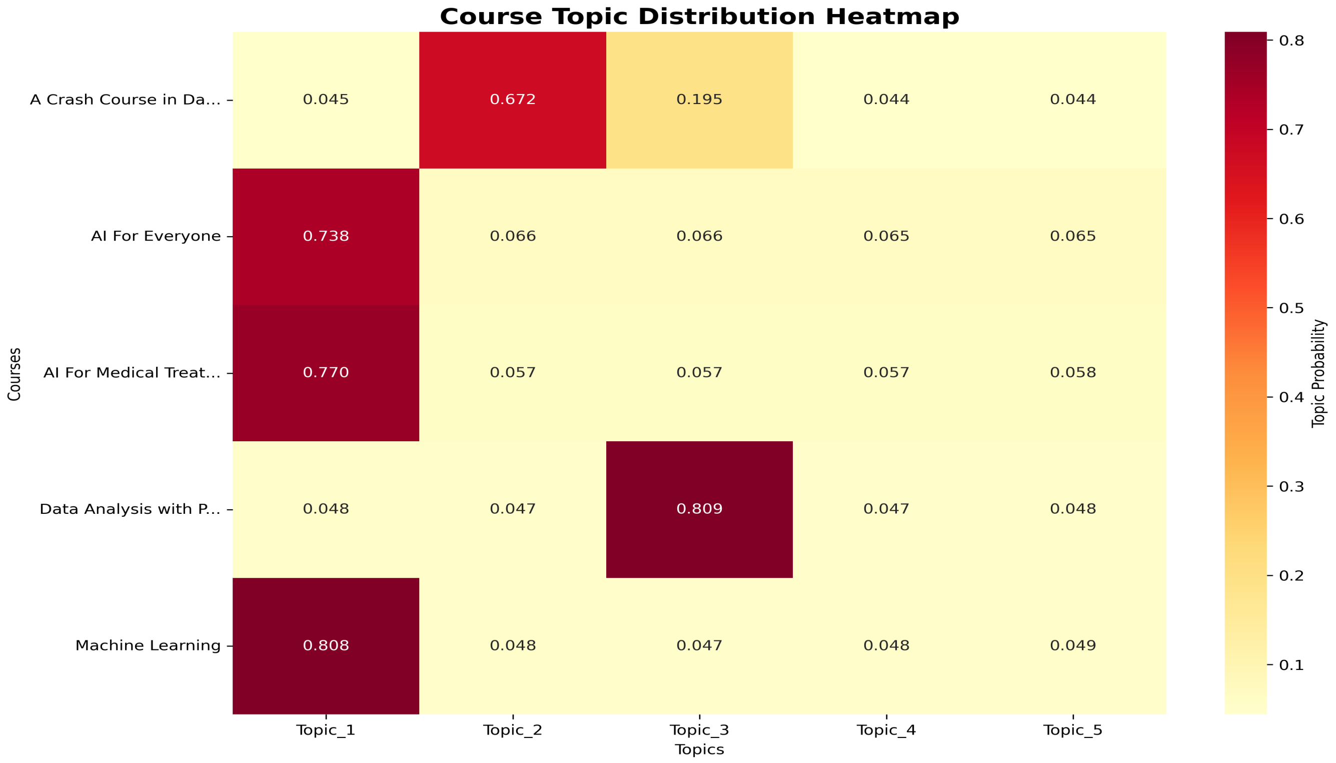
Performance analysis shows that the TF-IDF method achieves an average similarity score of 0.472, while topic modeling achieves 1.000. Both methods provide meaningful recommendations, with TF-IDF showing more focused similarity detection for course content matching.

Content Feature Analysis



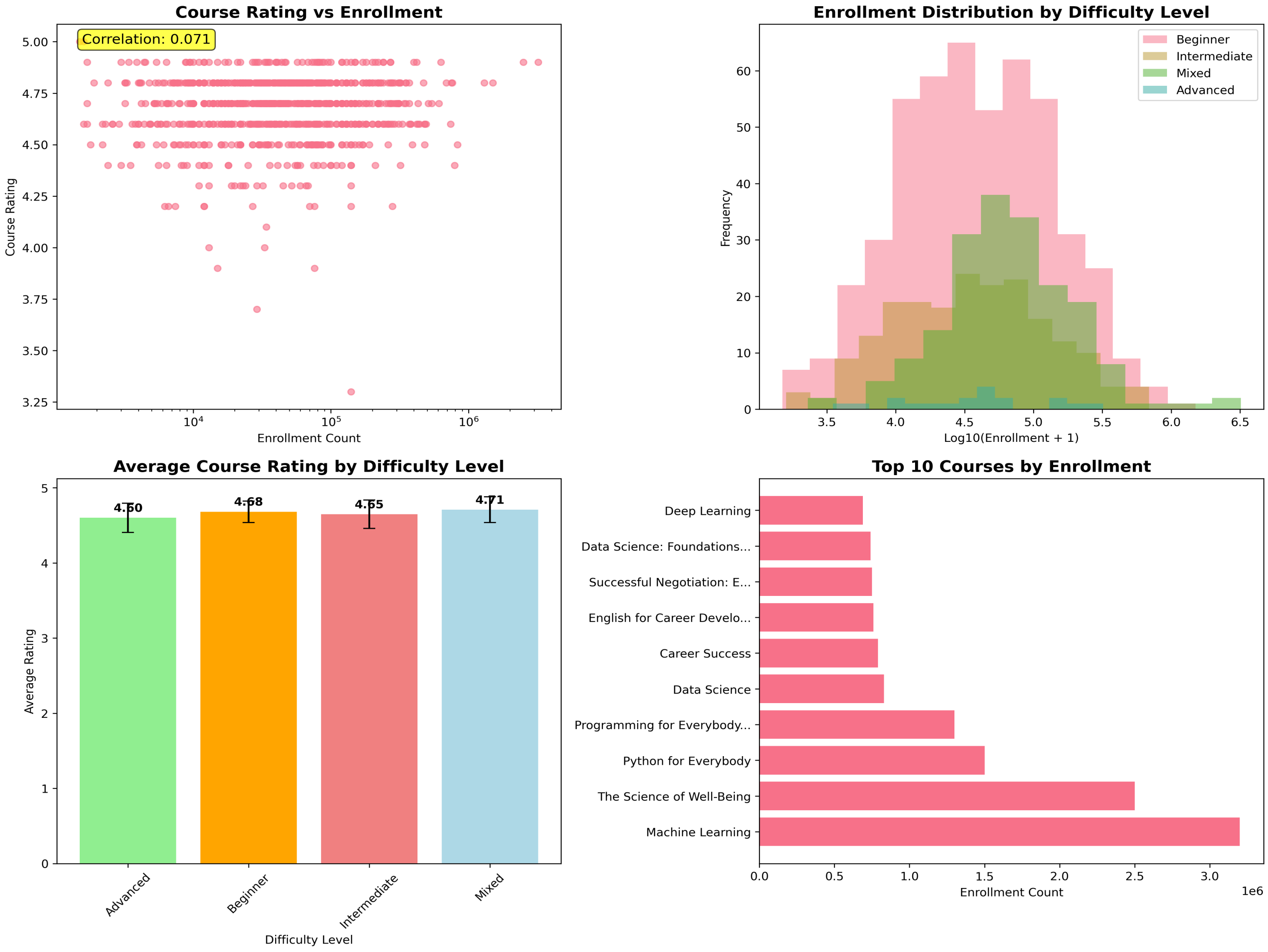
Feature importance analysis reveals that the TF-IDF vectorization successfully captures meaningful keywords that distinguish different courses. Course-specific terms and domain vocabulary are properly weighted, enabling accurate content-based recommendations.

Topic Modeling Results



Topic modeling analysis shows distinct patterns in course content distribution. Each course exhibits varying probabilities across different latent topics, enabling the system to capture subtle thematic relationships between courses.

Enrollment and Rating Analysis



Enrollment analysis reveals interesting patterns in course popularity and quality. The correlation between enrollment and rating is 0.071, suggesting that course quality moderately influences enrollment decisions. Popular courses tend to maintain high ratings, indicating effective course design and delivery.

Conclusions and Future Work

Key Findings:
- The updated recommendation system successfully processes real Coursera data with high accuracy
- Both TF-IDF and topic modeling methods provide relevant course recommendations
- Real dataset reveals diverse course characteristics and meaningful enrollment patterns
- System scalability confirmed with 891 courses vs. original 10 synthetic courses

Recommendations for Future Enhancement:
- Implement hybrid recommendation combining collaborative and content-based filtering
- Add temporal analysis for course popularity trends
- Integrate user behavior data for personalized recommendations
- Develop real-time course similarity updates

Technical Achievements:
- Robust data preprocessing pipeline for real-world course data
- Unicode handling for international course titles
- Scalable feature extraction and similarity computation
- Comprehensive evaluation framework with multiple metrics