졸업프로젝트 보고서

데이터마이닝을 활용한
1) 음식 재료의 신선한 조합 추천과 2) 대학 내 학과 분석

2019년 06월 12일

데이터마이닝 (DMIS) 연구실
고려대학교 정보대학 컴퓨터학과

백진헌 (2016320198)
정소영 (2016320120)

Contents

- ❑ 음식 재료의 신선한 조합 추천
  - ❍KitcheNette: Predicting and Recommending Food Ingredient Pairings using Siamese Neural Networks
  - ❍IJCAI 2019 Accepted
- ❑ 대학 내 학과 분석
  - ❍대학 내 학과 분석을 위한 학과 및 수업 임베딩
  - ❍한국 정보과학회 KCC 2019 Accepted (Oral)

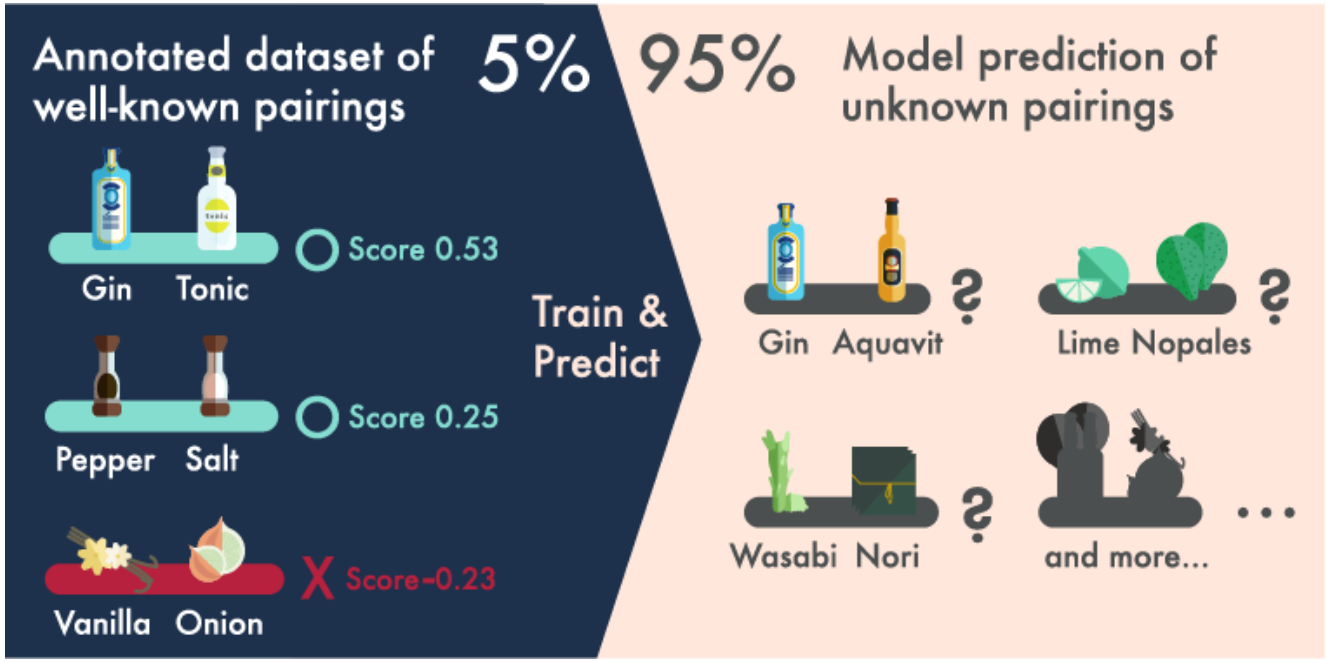
Contents (음식 재료의 신선한 조합 추천)

- ❑ Introduction
- ❑ Related Work (Project)
- ❑ Solution Approach (Main Idea)
- ❑ Experiment / 성능평가
- ❑ Conclusion

1. Introduction

❑ Problem Statement

- ❍ 음식 재료 조합의 경우 경험적으로 얻어진 경우가 많다.
- ❍ 알려지지 않은 음식 재료 조합의 수가 많으므로 새로운 음식 조합을 찾을 수 있다.



1. Introduction

❑Pains and Needs

- ❍ 알려지지 않은 음식 재료 조합을 새롭게 발견할 경우

1) 식품 개발자는 신제품을 개발 가능하고,
2) 식품 소비자는 맛 좋은 다양한 음식을 섭취할 수 있다.

❑Importance

- ❍ 인공지능 기술을 활용한 음식 조합의 다양성을 찾아,

1) 인공지능 학계에는 음식 추천이라는 새로운 분야를 개척하고,
2) 식품 관련 사회에는 인공지능 기술을 접목시킨 음식 재료 조합 선택이 가능하다.

2. Related Work (Project)

- Research on Discovering Food Pairings
  - 음식의 Compound 기반으로 Food Pairing을 예측
  - (Weak) 단순한 Compound 비교 수준에서 그친 연구들이 많음
  - (Weak) 알려진 Compound 개수가 많지 않음

- Research on Recommending Recipes
  - 음식 재료가 나열된 Recipes의 Rating을 예측함
  - (Weak) Large scale의 분석을 하지 못함
  - (Weak) 등장하지 않았던 새로운 재료의 조합을 제안하기는 어려움

3. Solution Approach (Main Idea)

Ingredient Extraction

1M Recipes, 3K Ingredients
[Marin et al., 2019]

Lemon Cupcakes with Blueberry Compote Filling https://www.foodnetwork.com/recipes/lemon-cupcakes-with-blueberry-compote-filling-and-cream-cheese-frosting-recipe-2125195	
Ingredients	1 1/2 cups granulated **sugar**, 1 stick (1/2 cup) unsalted **butter**, 4 **egg whites**, Juice of 2 **lemons**, 2 cups all-purpose **flour**, 1 teaspoon **baking powder**, 1/2 teaspoon **baking soda**, 1/4 teaspoon **salt**, 2/3 cup **buttermilk**, 2 cups **blueberries**, 8 ounces **cream cheese**, 1 tablespoon **vanilla** extract

Ingredient Pairing Score

- PMI Score (co-occurrence probability)
- The more they appear together, the more likely they go well together
- 300K ingredient pairs for training

Ingredient1 [Count]	Ingredient2 [Count]	Co-occurrence	Pairing score
vanilla [51,756]	baking_soda [58,931]	14,657	0.376
	cocoa [6,520]	2,759	0.360
	powdered_sugar [26,729]	6,558	0.314
	nut [9,090]	2,865	0.312
	chocolate_chips [9,172]	2,821	0.307
	onion [191,691]	12	-0.589
	soy_sauce [40,518]	6	-0.483
	salt_and_pepper [46,534]	14	-0.479
	garlic [46,534]	9	-0.477
	pepper [68,984]	26	-0.462

3. Solution Approach (Main Idea)

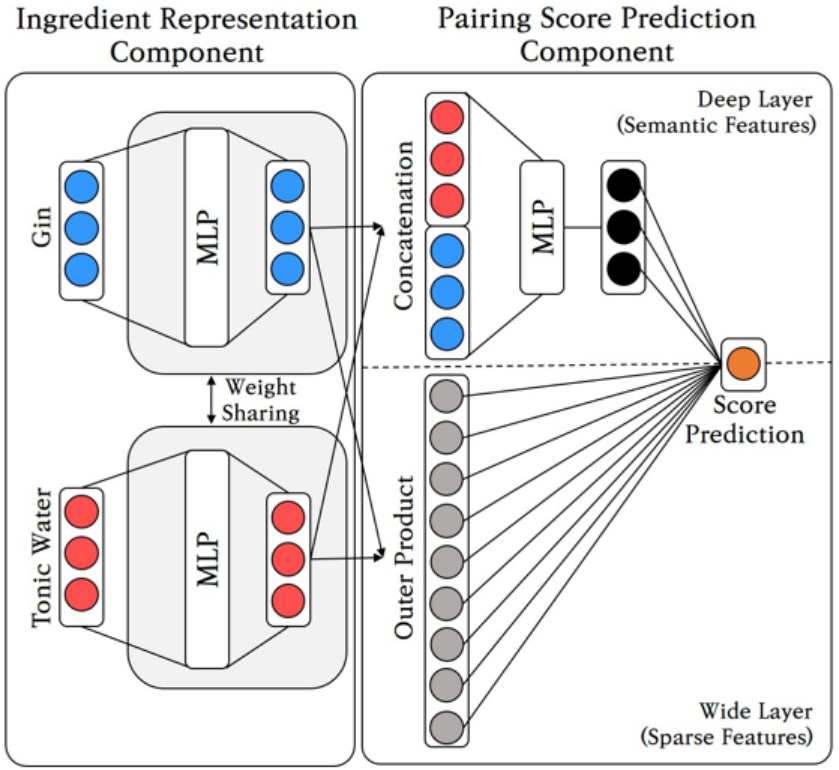
KitcheNette trains 300K pairwise scores of ingredient pairings

Regression Model

Input: ingredient word embedding
Target: a pairwise score of two ingredients

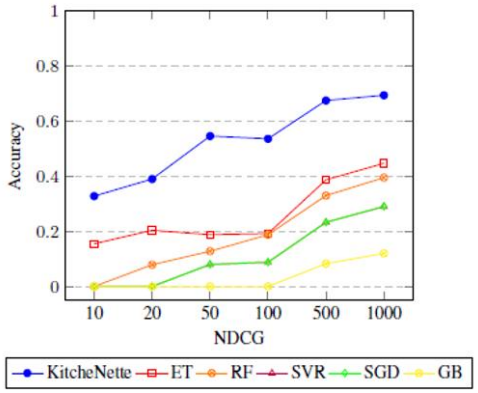
Ingredient Representation Component
Siamese Network (Twin network)

Pairing Score Prediction Component
Wide&Deep Layer [Cheng et al., 2016]
Deep Layer: Concatenation (Semantic Features)
Wide Layer: Outer Product (Sparse Features)

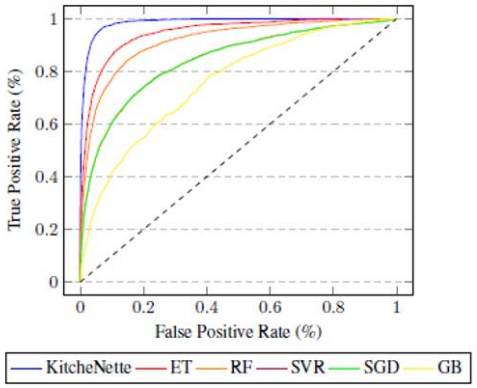


4. Experiment / 성능 평가

Model	Validation					Test				
	RMSE	MSE	MAE	CORR	R2	RMSE	MSE	MAE	CORR	R2
Cosine Similarity	-	-	-	-	-	0.1802	0.0325	0.1328	0.3952	-1.6026
Gradient Boosting	0.1073	0.0115	0.0815	0.3339	0.0773	0.1073	0.0115	0.0815	0.3351	0.0776
SGD	0.0993	0.0099	0.0762	0.4585	0.2102	0.0984	0.0097	0.0759	0.4730	0.2236
Linear SVR	0.0993	0.0099	0.0762	0.4588	0.2105	0.0984	0.0097	0.0759	0.4731	0.2238
Random Forest	0.0802	0.0064	0.0612	0.7015	0.4846	0.0799	0.0064	0.0611	0.7042	0.4885
Extra Tree	0.0742	0.0055	0.0566	0.7664	0.5586	0.0738	0.0054	0.0563	0.7689	0.5637
Siamese Network	0.0726	0.0054	0.0540	0.8223	0.5679	0.0729	0.0054	0.0544	0.8235	0.5662
KitcheNette	**0.0421**	**0.0018**	**0.0320**	**0.9249**	**0.8551**	**0.0417**	**0.0018**	**0.0317**	**0.9266**	**0.8583**



(a) NDCG@K



(b) ROC Curve

Regression Results

- Simple cosine similarity
- 5 traditional ML
- Basic Siamese networks

Ranking & Classification results

- NDCG@K: Ranking performance at top K results
- ROC Curve: Model sensitivity of classification (good or bad pair)

4. Experiment / 성능 평가

Known

Unknown

		champagne	sparkling_wine	prosecco
Case 1	orange_twist	0.33†	0.39*	0.42*
	orange_wedge	0.37*	0.43*	0.45*
	lime_twist	0.34*	0.38*	0.40*
Case 2	elderflower_liqueur	0.34†	0.39*	0.41*
	creme_de_cassis	0.29*	0.33†	0.34*
	lemon_sorbet	0.32*	0.39*	0.42†
Case 3	onion	-0.20†	-0.14*	-0.17*

Case 1 – Similar (foods) & Similar (wines)

For an example known pairing,
orange_twist & champagne

For 8 example unknown pairings,
2 other similar wines (sparkling_wine, prosecco)
& 2 other similar foods (orange_wedge, lime_twist)

5. Conclusion

❑ Summary & Contributions

- ❍ 두 재료 간의 Pairing을 Score화 하는 딥러닝 모델을 제안
- ❍ 5% Pairing 학습을 통해 알려지지 않은 95% Pairing 분석 가능

❑ Future work

- ❍ 두 재료 간의 Co-occurrence Score 예측에만 그쳤지만 Graph 기반의 학습 방법을 통해 재료의 one-to-many 관계를 알아보고자 함
- ❍ 각 재료의 화학적 정보를 함께 임베딩하여 더 깊은 재료 간 추천 결과를 얻음

Contents (대학 내 학과 분석을 위한 학과 및 수업 임베딩)

- ❑ Introduction
- ❑ Related Work (Project)
- ❑ Solution Approach (Main Idea)
- ❑ Experiment / 성능평가
- ❑ Conclusion

1. Introduction

❑ Problem Statement

- ❍ 대학에는 다양한 학과들이 존재하여 학과를 종합적으로 분석하는 것이 어려움

❑Pains and Needs

- ❍ 수험생들이 다양한 학과 정보를 비교 이해하며 탐색하는 것이 어려움
- ❍ 진로 진학 교사들이 모든 학과에 대해 알 수 없어 진학 지도 상담을 하기 어려움

❑Importance

- ❍ 따라서 데이터마이닝을 활용하여 학과 정보를 손쉽게 정리 요약 후 제공해주어,

1) 인공지능 기술을 에듀테크 중 진로 진학 분야에 적용한 새로운 사례를 개척하고,
2) 진로 진학 탐색 및 상담 분야에서 학과 탐색이 용이하게 끔 가능하다.

2. Related Work (Project)

- Research on Classifying Lecture
  - 많은 양의 온라인 강의 수를 분류하는 모델을 제안
  - (Weak) 학과를 구성하는 강의에 대한 분석에 그침
  - (Weak) 단순 강의 분류 뿐 강의에 대한 정보를 종합적으로 분석하지 못함

- Research on Course Map for Describing Relation
  - 학과를 구성하는 강의 정보를 바탕으로 각 강의의 선후 관계를 표현
  - (Weak) 학과와 강의에 대한 종합적 분석을 하지 못하고 단순 선후 관계만 표현 함

3. Solution Approach (Main Idea)

❑서울대 / 고려대 / 연세대 학과 커리큘럼 데이터 (수업 정보) 수집

❍Ex. 고려대 컴퓨터학과 데이터베이스

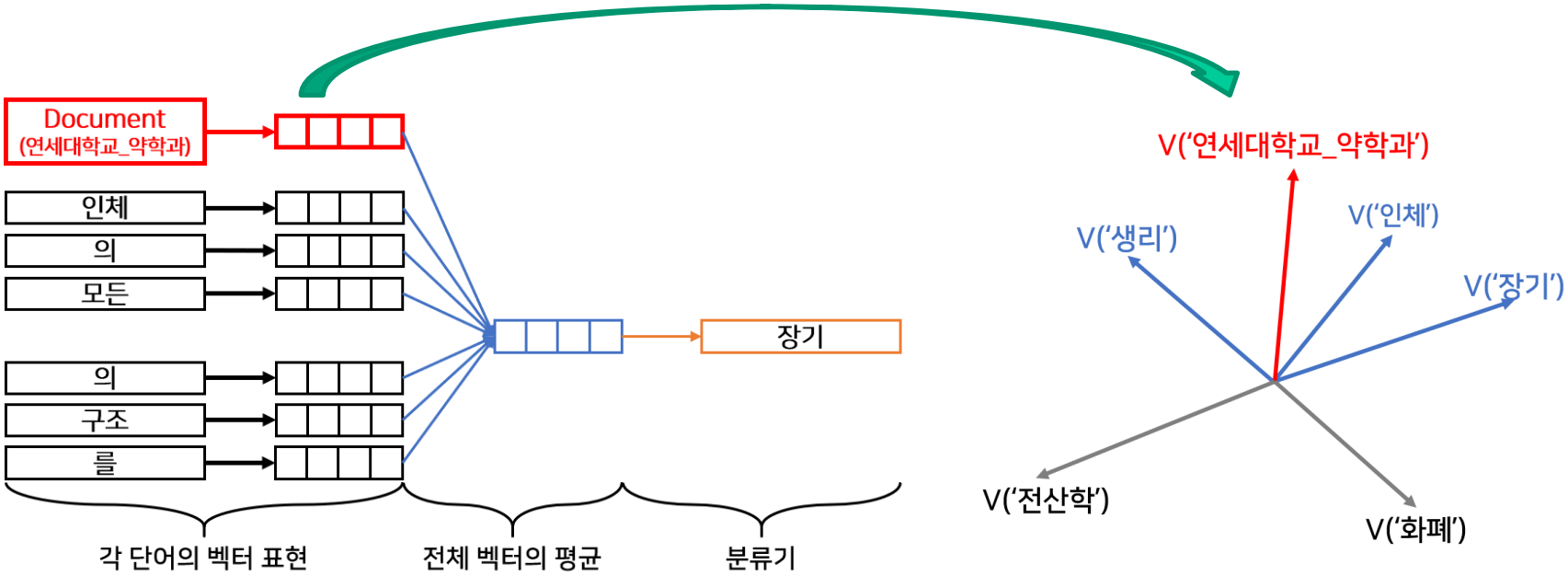
IT 환경의 급속한 변화와 정보화 사회가 점차 성숙해지면서 데이터를 통합 관리하여 유용한 정보를 제공하는 데이터베이스 시스템의 역할을 더욱 중요해지고 있다. 본 강의에서는 이러한 데이터베이스에 대한 기초 개념과 데이터베이스시스템 구축에 필요한 이론 및 기술들을 학습하는 것을 목적으로 한다.
…

	서울대	고려대	연세대	전체
학과 개수	86개	54개	57개	197개
수업 개수	2,973개	2,934개	2,768개	8,675개
학과 평균 수업 개수	34.57개	54.33개	48.56개	44.04개

3. Solution Approach (Main Idea)

❑학과 및 수업 임베딩 방법

❍Ex. 연세대학교 약학과의 학과 표현



3. Solution Approach (Main Idea)

❑학과 표현 방법

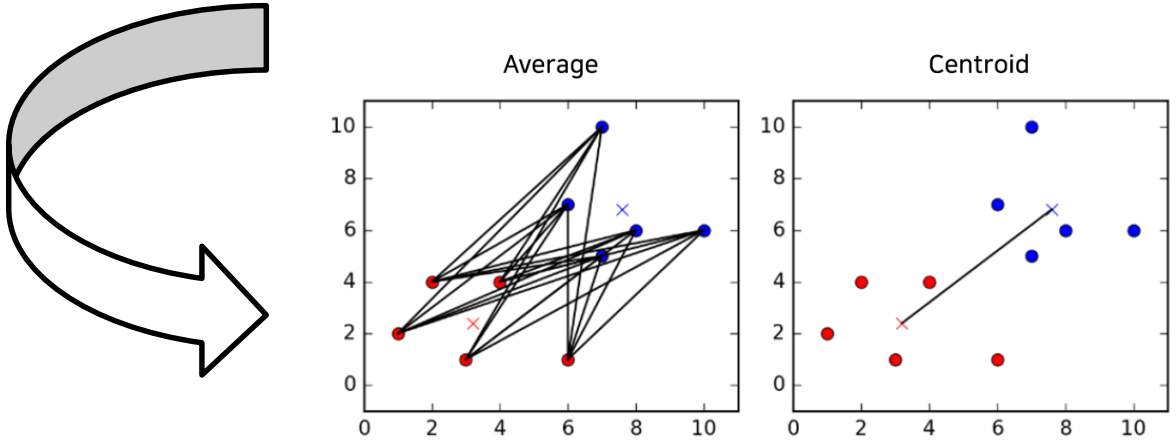
❍학과 임베딩 벡터 활용

✓Doc2Vec 모델 (이전 슬라이드) 결과로 나타난 학과 임베딩 벡터를 학과 표현에 바로 이용

❍수업 임베딩 벡터 활용

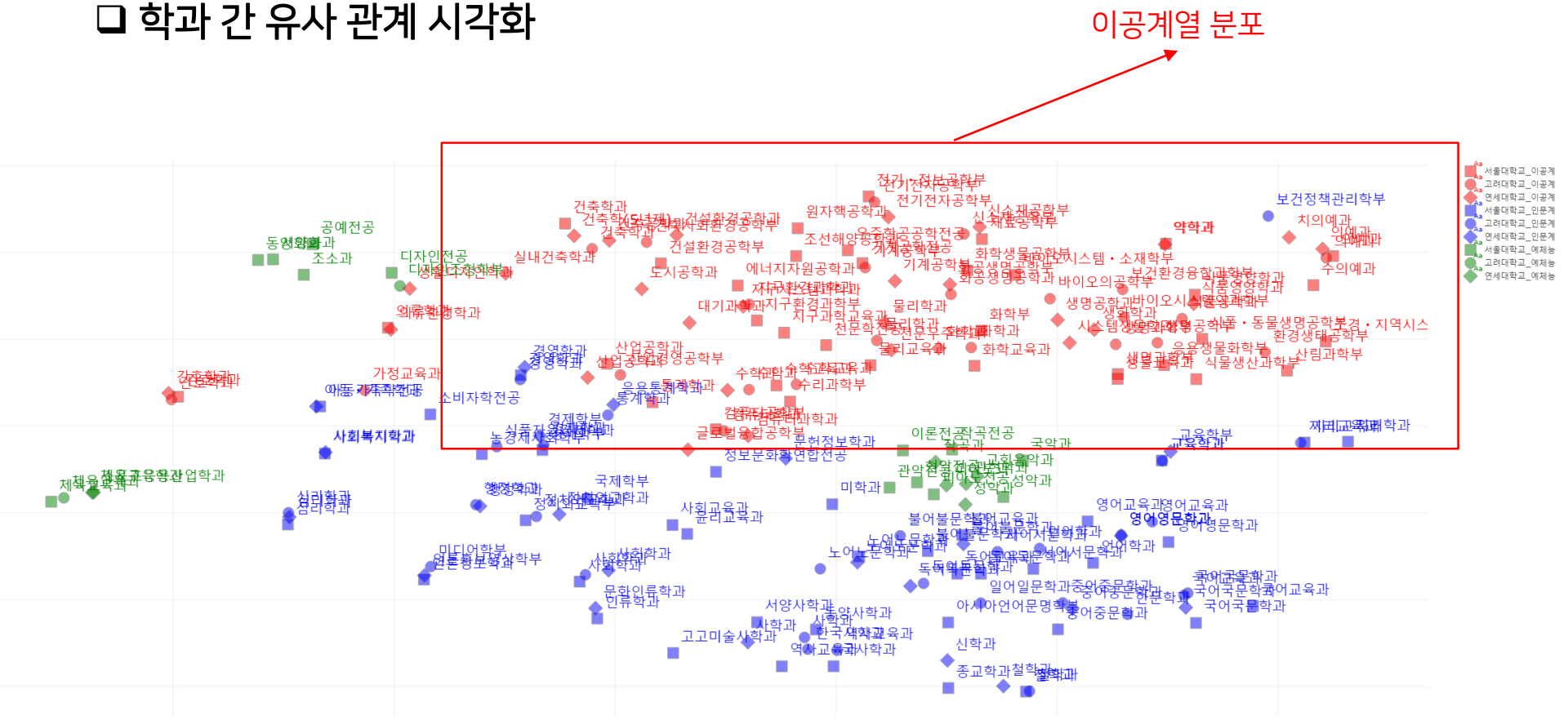
✓Doc2Vec 모델 (이전 슬라이드) 결과로 나타난 수업 임베딩 벡터를 추출

✓한 학과내 수업을 한 클러스터 단위로 보고 클러스터링 방법으로 학과(학과 간 관계) 표현



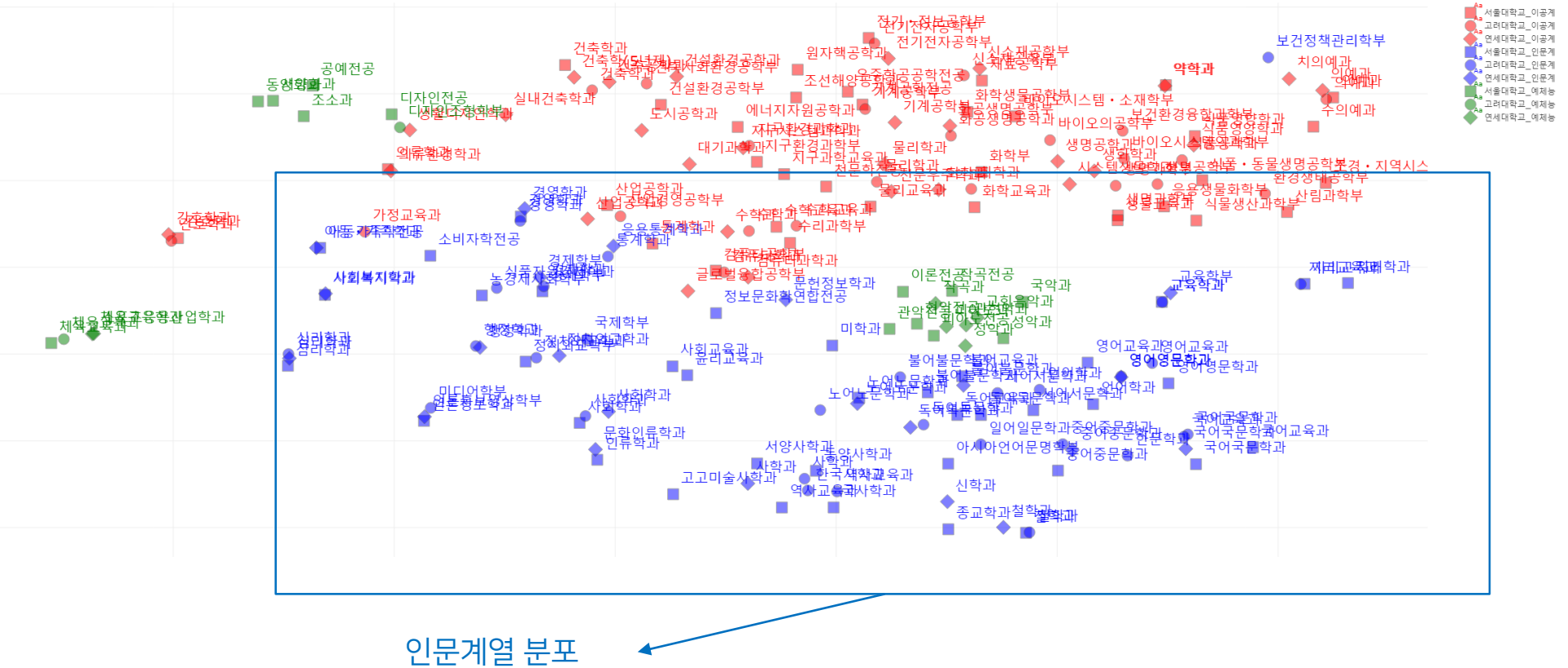
4. Experiment / 성능 평가

❑ 학과 간 유사 관계 시각화



4. Experiment / 성능 평가

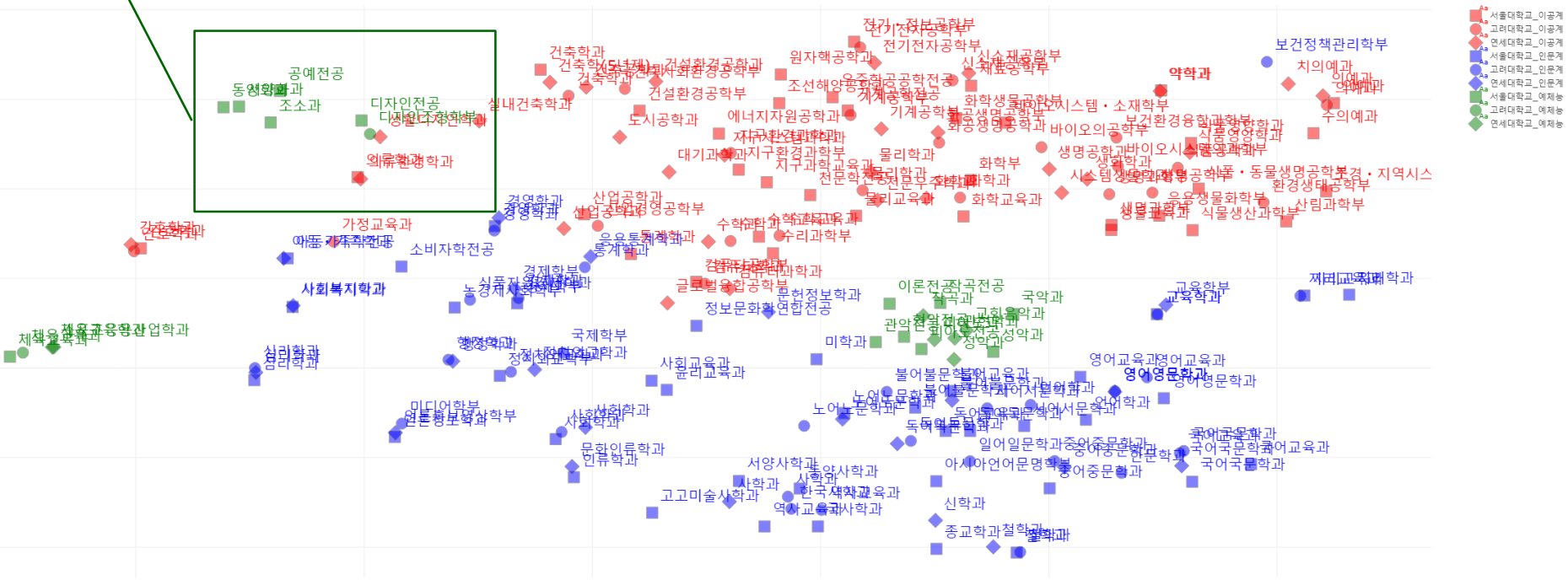
❑ 학과 간 유사 관계 시각화



4. Experiment / 성능 평가

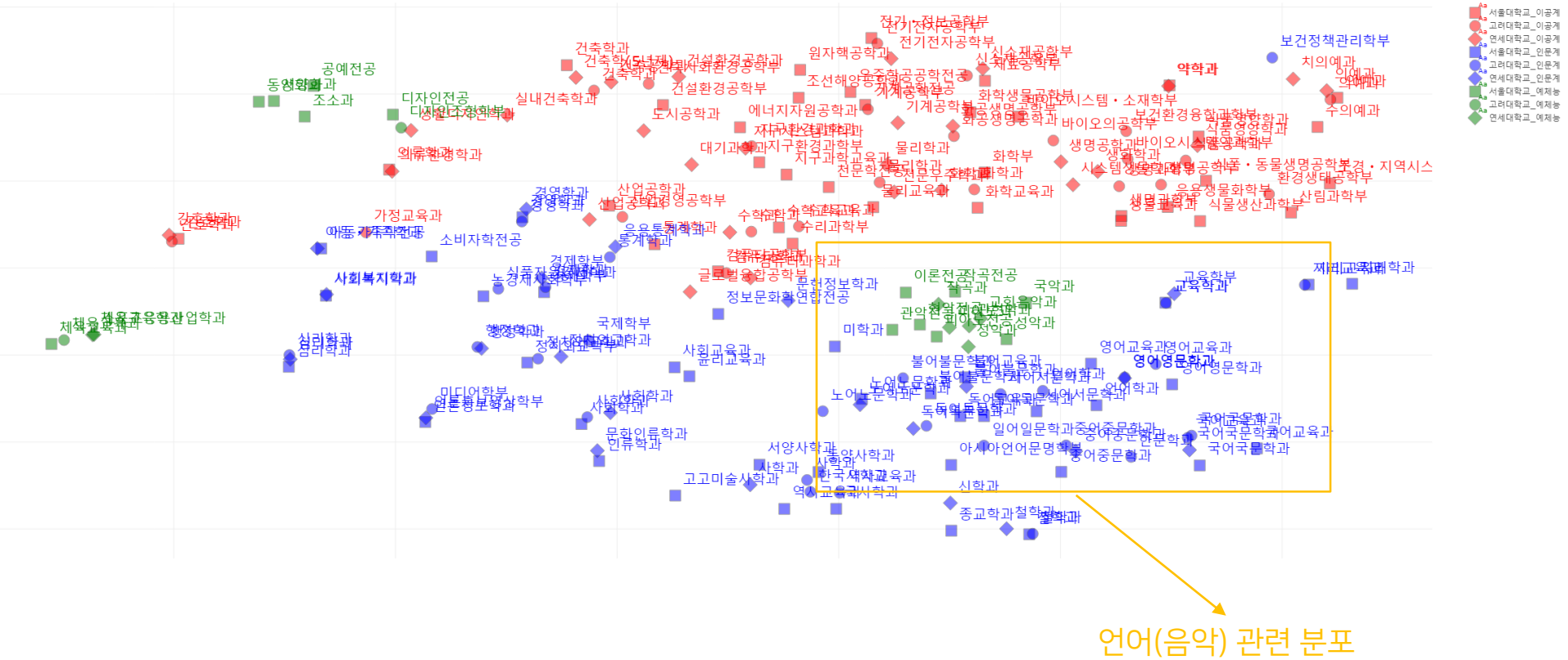
❑ 학과 간 유사 관계 시각화

디자인 관련 분포



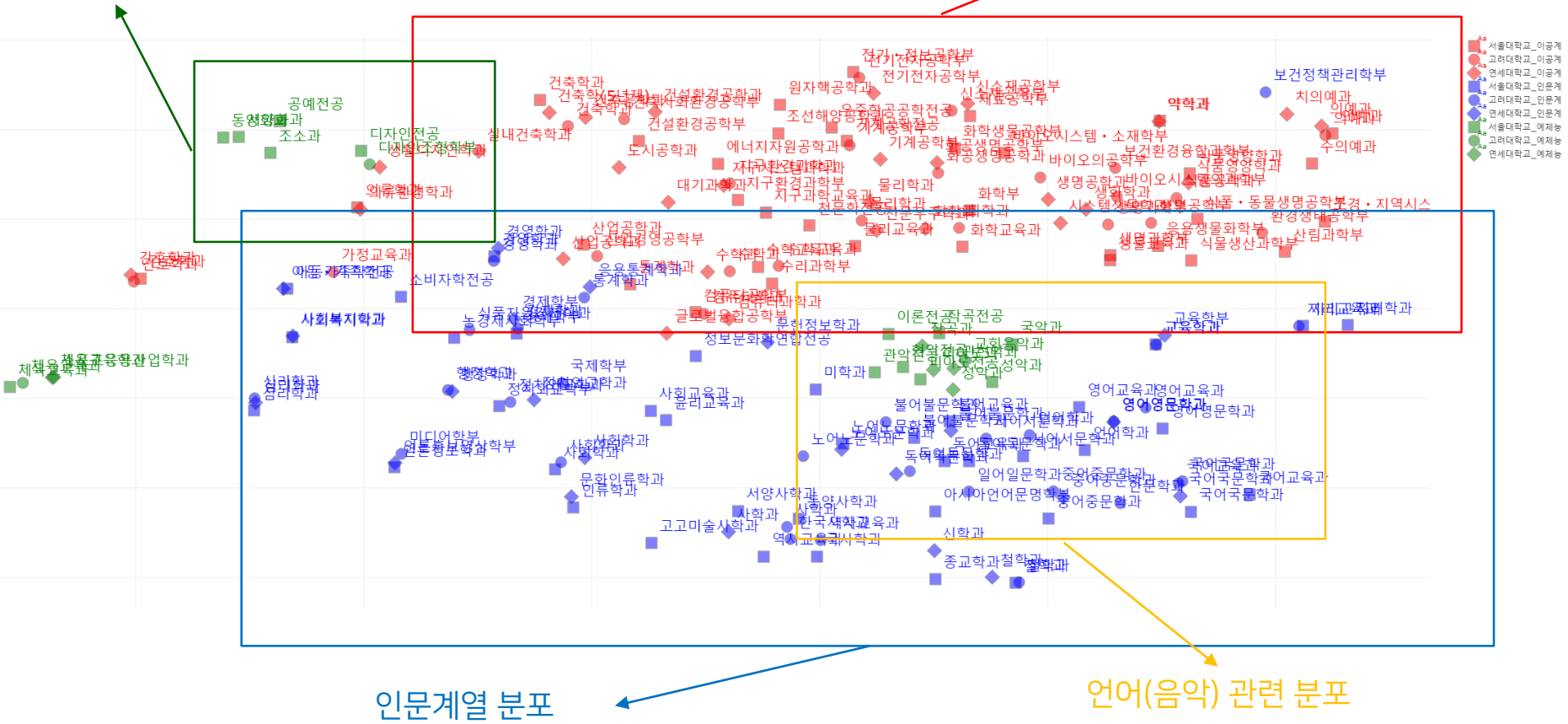
4. Experiment / 성능 평가

❑ 학과 간 유사 관계 시각화



4. Experiment / 성능 평가

❑ 학과 간 유사 관계 시각화



4. Experiment / 성능 평가

❑ 학과 간 유사 관계 도식화

❍학과 임베딩 벡터와 수업 임베딩 벡터 (클러스터링으로 학과 표현)의 비교

✓학과 임베딩 벡터: Doc2Vec 으로 학과 임베딩

✓클러스터(Centroid, Average) 학과 표현: Doc2Vec 으로 수업 임베딩 후 클러스터링으로 학과 표현

	학과 임베딩 벡터	Centroid	Average
1	서울대학교 컴퓨터공학부	서울대학교 컴퓨터공학부	고려대학교 수학과
2	연세대학교 컴퓨터과학과	고려대학교 전기전자공학부	고려대학교 통계학과
3	고려대학교 전기전자공학부	연세대학교 컴퓨터과학과	연세대학교 전기전자공학부
4	서울대학교 전기정부공학부	연세대학교 전기전자공학부	연세대학교 컴퓨터과학과
5	연세대학교 전기전자공학부	서울대학교 전기정보공학부	고려대학교 전기전자공학부

4. Experiment / 성능 평가

❑ 학과 핵심어 추출

❍TF-IDF 가중치 모델을 활용한 각 학교 경제학과 핵심어 추출 및 비교

✓서울대학교, 연세대학교: 시장과 금융 및 거시경제 중시

✓고려대학교: 화폐와 노동 및 미시경제 중시

서울대학교		고려대학교		연세대학교	
금융	시장	화폐	계량	금융	거시경제
기업	제도	노동	노사	금융시장	모형
거시경제	경제사	노동시장	미시	시장	계약

5. Conclusion

- Summary & Contributions
  - 데이터마이닝 기술을 활용한 학과 분석 방법 제안
  - 다량의 학과들을 쉽고 종합적으로 비교 분석 가능

- Future work
  - 임베딩 벡터를 활용하여 학과를 분석하는 다양한 방법 추가 제공
  - 학생 개인화에 맞는 학과 추천 모델 개발