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

정보대학 컴퓨터학과 백진헌, 정소영

◆ Objectives

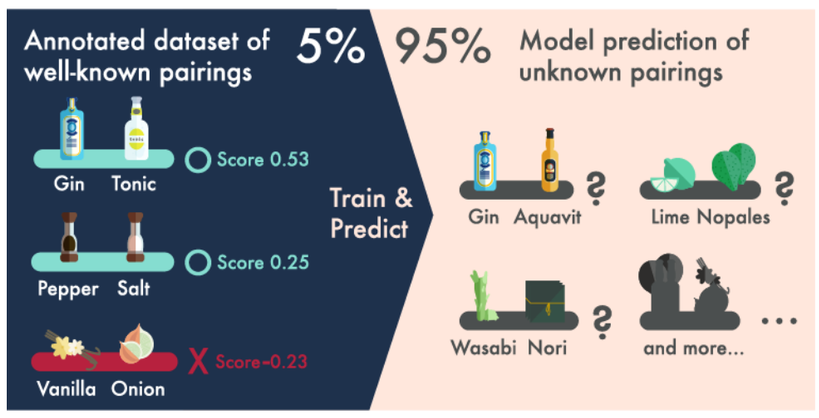
세상에는 데이터마이닝으로 가치를 찾아낼 수 있는 다양한 문제들이 존재한다. 따라서 우리 팀은 존재하는 다양한 문제 중 1) 음식 재료의 신선한 조합 추천이라는 식품-화학 분야와 2) 대학 내 학과 분석이라는 교육 분야의 문제들을 인공지능을 접목한 데이터마이닝을 활용하여 해결하였다.

◆ 음식 재료의 신선한 조합 추천

KitcheNette: Predicting and Recommending Food Ingredient Pairings using Siamese Neural Networks (IJCAI 2019 Accepted)

Introduction

맛있는 음식을 만들기 위해서는 적절한 음식 재료 조합 간의 관계를 찾는 것이 중요하다. 지금까지는 경험적인 방법으로 음식 재료 조합을 획득한 경우가 많았다. 하지만 수많은 재료들 간의 조합을 경험적인 방법으로 만든 다음 직접 맛보는 것에는 시간과 비용이 많이 발생한다. 따라서 우리는 알려진 음식 재료 조합의 수를 바탕으로 기존에 존재하지 않았던 새로운 음식 조합을 찾고자 한다.



위 그림과 같이 알려진 음식 재료 조합이 전체 재료 조합의 5% 밖에 되지 않는다. 따라서 알려진 음식 재료 조합으로 학습 데이터와 평가 데이터를 구축한 다음, 두 재료 조합이 서로 어울리는 재료 조합인지 예측하는 모델을 생성하였다. 그 다음 알려지지 않은 재료 조합인 95%의 조합을 기존 모델을 바탕으로 예측하였다.

Related Work

1. Research on Discovering Food Pairings
해당 연구는 음식의 화합 물질 기반으로 음식 재료 간의 어울림을 판단하였다. 하지만 단순한 화학 물질 비교 수준에서 그쳤으며, 알려진 화학 물질이 많지 않아 음식 재료 간 정확한 비교가 어렵다.
2. Research on Recommending Recipes
해당 연구는 음식 재료가 나열된 레시피의 추천 점수를 예측하였다. 하지만 대규모의 분석을 하지 못했으며, 해당 방식으로는 등장하지 않았던 새로운 재료의 조합을 제안하기는 어렵다.

Our Approach

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

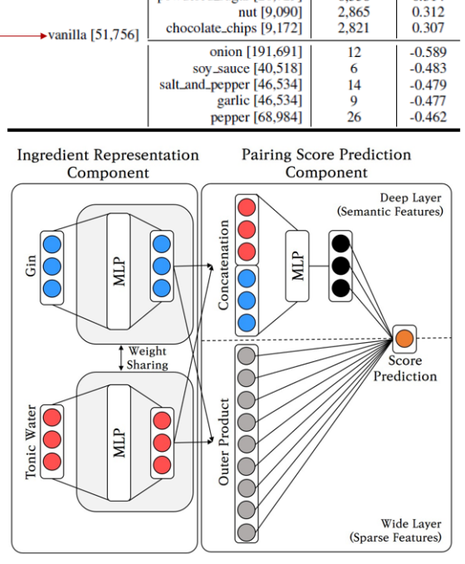
Ingredients: 1 1/2 cups granulated **sugar**, 1 stick (1/2 cup) unsalted **butter**, 4 **egg whites**, Juice of 2 **lemons**, 2 cups all-purpose **flour**, 1 teaspoon **baking powder**, 1/2 teaspoon **baking soda**, 1/4 teaspoon **salt**, 2/3 cup **buttermilk**, 2 cups **blueberries**, 8 ounces **cream cheese**, 1 tablespoon **vanilla** extract

KitcheNette trains 300K pairwise scores of ingredient pairings

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



Regression Model

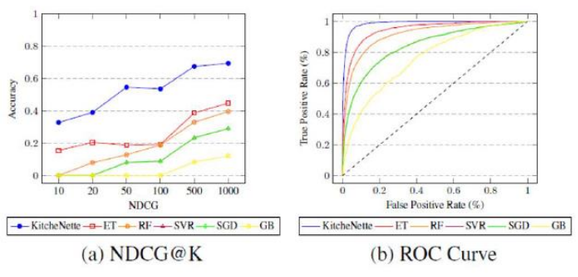
Input: ingredient word embedding
Target: a pairwise score of two ingredients

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)

Experiment

Model	Validation RMSE	MSE	MAE	CORR	R2	Test 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)

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

Known / Unknown

위 결과는 본 모델의 성능을 정량적과 정성적으로 분석한 결과입니다. 위 정량적 분석에서는 우리 모델이 다른 Baseline에 비해 뛰어나다는 것을 알 수 있으며, 정성적 분석에서는 어울리는 음식과 와인의 새로운 조합에 대해 추천해줄 수 있음을 확인하였습니다.

Conclusion

본 연구 결과는 두 재료 간의 Pairing을 점수화 하는 딥러닝 모델을 제안했으며, 알려지지 않은 Pairing을 예측하여 어울리는 새로운 음식 재료 조합을 알 수 있습니다.

◆ 대학 내 학과 분석

대학 내 학과 분석을 위한 학과 및 수업 임베딩 (KCC 2019 Oral Accepted)

Introduction

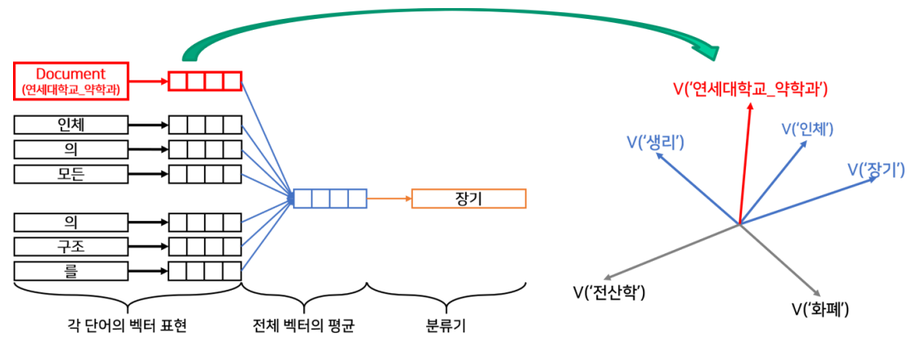
대학에는 다양한 학과들이 존재한다. 하지만 다양한 학과들을 비교 분석하여 종합적으로 분석하기는 어렵다. 따라서 우리는 데이터마이닝을 통해 학과 정보를 정량적으로 표현하여 학과를 분석하는 다양한 방법을 제안했다.

Related Work

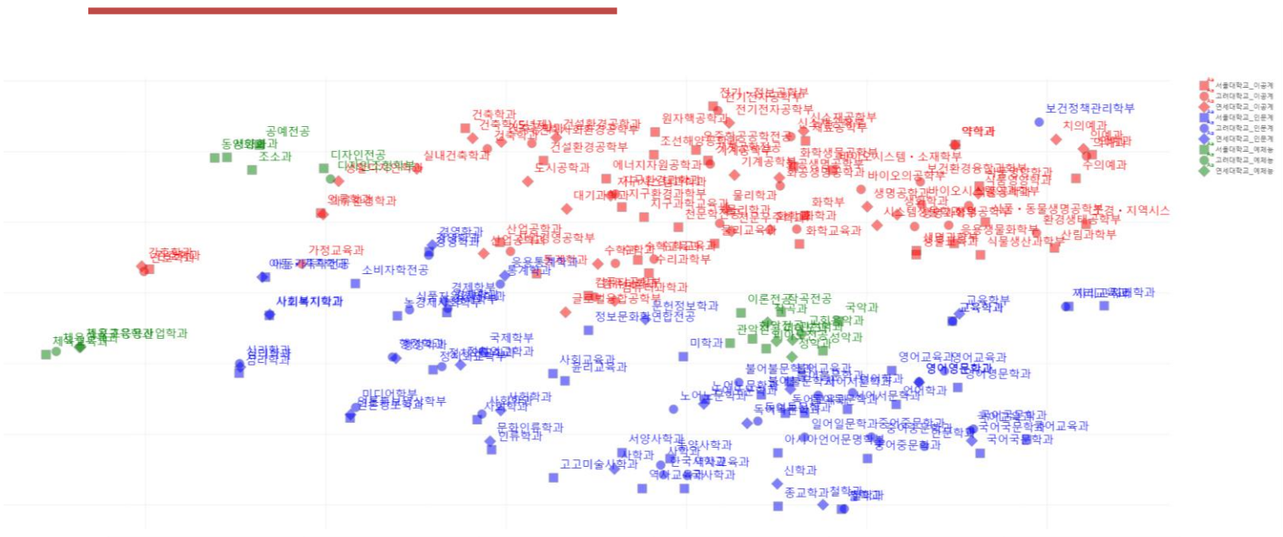
1. Research on Classifying Lecture
많은 양의 온라인 강의 수를 분류하는 모델을 제안했지만 학과를 구성하는 강의에 대한 분석에 그쳤다. 뿐만 아니라 강의 정보를 종합적으로 요약 분석하지 못했다.
2. Research on Couse Map for Describing Relation
학과를 구성하는 강의 정보를 바탕으로 각 강의의 선후 관계를 표현했다. 하지만 학과와 강의에 대한 종합적 분석을 하지 못하고 단순 선후 관계만 표현했다.

Our Approach

학과를 정량적으로 표현하기 위해 학과에서 배우는 커리큘럼 정보인 수업 소개 자료를 바탕으로 각 학과를 표현하였습니다. 본 연구에서 수집한 학교 데이터는 서울대/고려대/연세대 세 학교이며 세 학교의 각 학과는 평균적으로 44.04개의 수업을 가지고 있습니다.
본 연구에서는 학과 및 수업 벡터를 Doc2Vec 모델을 통해 아래와 같이 임베딩 했습니다.



Experiment



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

Conclusion

우리는 본 연구에서 데이터마이닝 기술을 활용한 유사도 기반 학과 분석 방법을 제안하였다. 이를 통해 다양한 학과들을 쉽고 종합적으로 분석 가능하다.