Cuisine prediction using recipe-ingredient data

Hasan Kamal, 2015039

Aim: In this project, we aim to design a prediction system that uses recipe-ingredient data to predict the cuisine of a given list of ingredients. We aim to utilize and consequently compare techniques from following domains: Machine Learning, Simple (baseline) heuristics and Network-based heuristics

Dataset: https://www.kaggle.com/c/whats-cooking

Techniques to be evaluated and compared:

1. Machine Learning Domain
    a. Models such as **Neural Networks, SVMs** to be evaluated
    b. Models such as **Neural Networks, SVMs** to be evaluated with Node2Vec **embeddings** as features (where each node corresponds to an ingredient; and the graph is the unipartite projection recipe-ingredient bipartite network)
2. Simple heuristics (these will serve as our baseline models)
    a. Baseline #1: For each ingredient belonging to the given test ingredient list, find the cuisine in which this ingredient is used the most. Among all such cuisines, pick the most frequently occurring cuisine
    b. Baseline #2: Rank the set of training recipes based on number of ingredients common with test ingredient list. Assign weights to recipes based on their rank and add to scores of their corresponding cuisines. Finally, choose the cuisine with the highest score
3. Network-based heuristics
    a. Do **clustering** on ingredient-ingredient network (projection of original bipartite recipe-ingredient network) using hierarchical clustering algorithms. We try to choose clusters at a level such that #clusters = #cuisines in our train set
    We **analyze** if these generated clusters have one-to-one correlation with different cuisines. If yes, we can use these clusters to create a prediction heuristic in which the cuisine having most number of test ingredients (i.e. the cluster having most number of nodes out of a given set of nodes) is picked
    b. Do **clustering** of ingredients using **K-means** on Node2Vec embeddings of ingredient-ingredient network(we set K = # cuisines). We **analyze** if these generated clusters have one-to-one correlation with different cuisines. If yes, we can use these clusters to create a prediction heuristic in which the cuisine having most number of test ingredients (i.e. the cluster having most number of nodes out of a given set of nodes) is picked
    c. Consider the recipe-ingredient bipartite network. For every ingredient in test list, we do: add *w* to scores of cuisines of recipes that are at **distance** of 1 from that ingredient, add *w/2* to scores of cuisines of recipes that are at **distance** of 3 from that ingredient and so on. This is done for each ingredient in the test list and the cuisine with highest score is chosen.