Analyzing Food Similarities Using NLP

Jong-Yol Lee
Department Of Fine Arts
University of San Carlos
Cebu-city, Philippines
23103132@usc.edu.ph

***Abstract*—By performing cosine similarity analysis and clustering analysis using simple TF-IDF and NMF(Non-Negative Matric Factorization) among natural language processing algorithms, it can obtain valid similarity analysis results even with a simple algorithm.**

***Index Terms*—TF-IDF, Cosine similarity, NMF, Clustering**

I. Introduction

This document analyzes ten food descriptions to identify similarities between foods using a natural language processing algorithm and confirms its effectiveness.

II. Collection of food descriptions

In order to find similar foods using a natural language processing algorithm and compare results between algorithms, descriptions of five Filipino foods and five Korean foods are collected from the wiki site.

III. Analyzing Similarities and clustering

This document creates a TF-IDF matrix, uses cosine similarity and NMF to find similar foods among 10 foods, and compares the results of the two algorithms.

A. Cosine similarity

Similar types of food can be identified using the cosine similarity of the TF-IDF matrix.

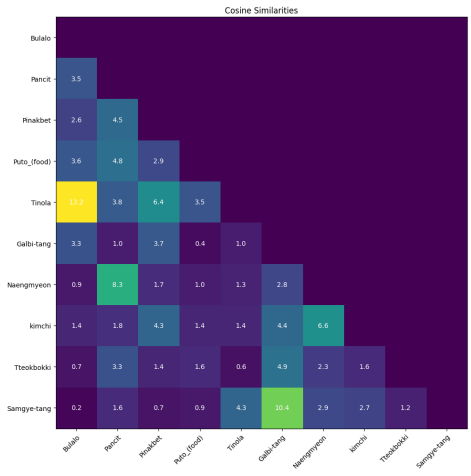

Fig. 1. Cosine similarity.

The top three pairs with high similarity are (Tinola, Bulalo), (Galbi-tang, Samgye-tang), and (Pancit, Naengmyeon). The first pair, Tinola and Bulalo, have something in common: they are Filipino soups made by boiling vegetables and meat. The second pair, Galbi-tang and Samgye-tang, have something in common: a Korean soup made by boiling vegetables and meat. The last pair, Pancit and Naengmyeon, have differences in that they are Filipino and Korean food, respectively, but they have one thing in common: noodle food.

B. NMF(Non-Negative Matric Factorization)

Clusters in the TF-IDF matrix can be classified using the NMF algorithm. In this document, it was classified into four clusters, and the results are shown in the table below.

TABLE I
Cuisine Group

Group No.	***Cuisine***
1	Pancit, Naengmyeon, Kimchi, Tteokbokki
2	Bulalo, Pinakbet, Tinola
3	Galbi-tang, Samgye-tang
4	Puto

C. Comparison of Cosine Similarity and NMF results

The top three pairs of foods with high cosine similarity were classified into the same cluster in NMF cluster analysis. Puso, which was uniquely classified into cluster 4, showed a low cosine similarity value compared to other foods.