Dish’s Cuisine Exploration

Machine Learning Project

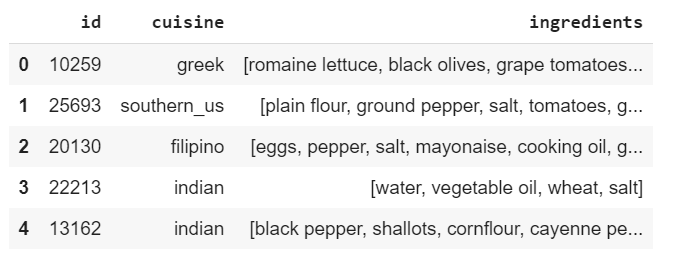
สิทธิชัย ศรีเดช 6110412011

ธันชกร สุวรรณเจียรมณี 6110412012

วัฒนธรรมทางด้านอาหารแต่ละประเทศมีลักษณะที่แตกต่างกันแต่บางประเทศก็มีลักษณะคล้ายกันจากวัฒนธรรมหรือภูมิประเทศที่คล้ายคลึงกัน ในปัจจุบันได้เกิดการผสมผสานหรือฟิวชั่นอาหารประเทศต่าง ๆ ทำให้เกิดเมนูอาหารใหม่ๆ ซึ่งการผสมผสานอาหารให้ลงตัวต้องอาศัยวัตถุดิบที่คล้ายคลึงกัน เพื่อให้ได้เมนูใหม่ที่ลงตัวและได้รสชาติที่อร่อย ฉะนั้นในการศึกษาครั้งนี้จึงสนใจที่จะสำรวจวัตถุดิบอาหารของชาติต่าง ๆ ทั่วโลก เพื่อสำรวจและหาความคล้ายคลึงกันของวัตถุดิบรวมถึงประเทศที่มีอาหารคล้ายคลึงกันและหา insight จากข้อมูลชุดนี้เพื่อนำไปประยุกต์ใช้ โดยข้อมูลที่ใช้คือมาจากเว็บ Kaggle.com ซึ่งข้อมูลประกอบด้วยประเภทอาหารแบ่งตามประเทศและรายการส่วนผสมของแต่ละประเทศ โดยข้อมูลเก็บในรูปแบบ JSON มีขั้นตอนวิธีดังนี้

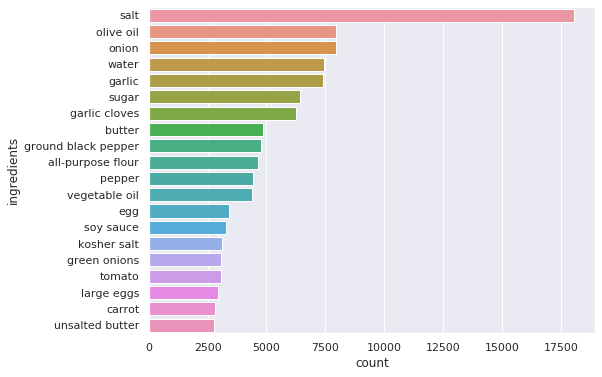
**1. import dataset และ Data preparation**

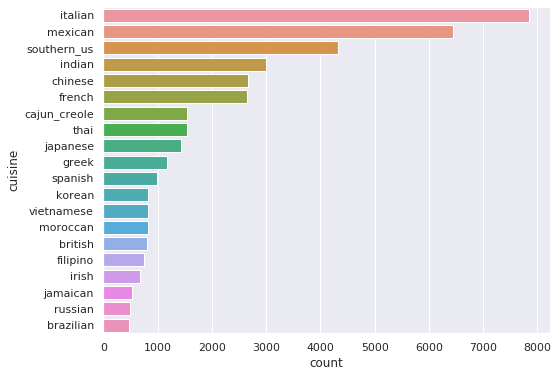
ในส่วนนี้จะทำการ import ข้อมูลเข้าแล้วทำการ clean ข้อมูลเพื่อให้สามารถนำไปวิเคราะห์ได้



**2. Exploratory data analysis**

ในส่วนแรกจะทำการนับจำนวนประเทศและจำนวนวัตถุดิบทั้งหมด เพื่อดูว่ามีจำนวนข้อมูลแต่ละประเภท ซึ่งได้ผลลัพธ์ดังนี้



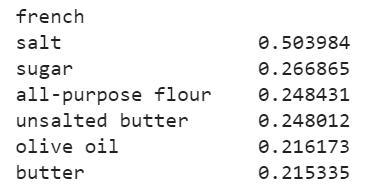
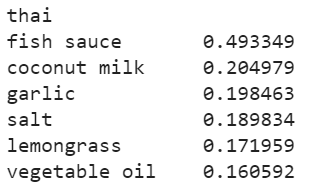
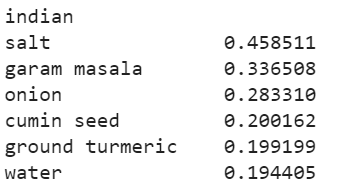


**3. Method**

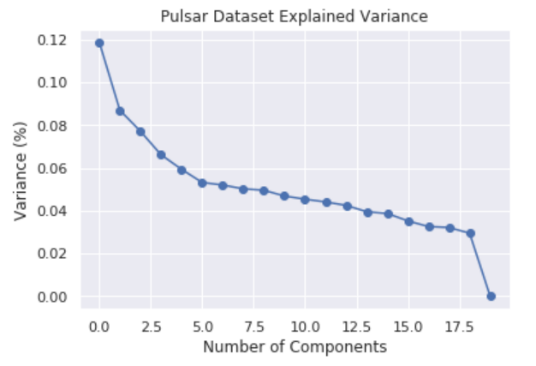
ในส่วนนี้จะทำการแบ่งวิเคราะห์ออกเป็น 2 ส่วนคือ ส่วนที่ 1 นับจำนวนวัตถุดิบที่พบมากที่สุดในแต่ละประเทศโดยใช้ method TF-IDF และแบ่งกลุ่มตามประเทศที่มีวัตถุดิบที่คล้ายคลึงกัน และส่วนที่ 2 หาความคลายคลึงของวัตถุดิบที่เกิดขึ้นร่วมกันหรือไม่เกิดขึ้นร่วมกัน โดยใช้ method Word2vec จากนั้นทำการนำกลุ่มของประเทศจากการแบ่งกลุ่มในส่วนที่ 1 มาหาวัตถุดิบที่ใช้ร่วมกันบ่อย ๆ หรือ ไม่เคยเกิดขึ้นร่วมกัน มาปรับใช้ในการสร้างเมนูใหม่ๆ มีขั้นตอนวิธีดังนี้

**ส่วนที่ 1 วัตถุดิบที่เป็นตัวแทนของแต่ละประเทศและแบ่งกลุ่มประเทศตามที่มีวัตถุดิบที่คล้ายคลึงกัน**

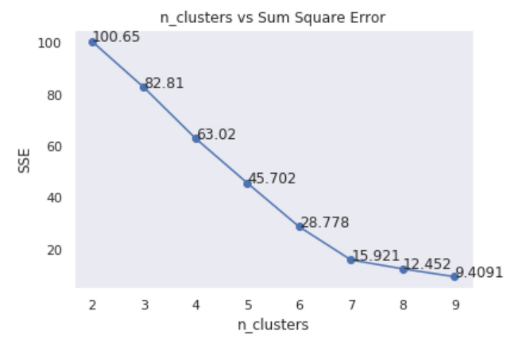
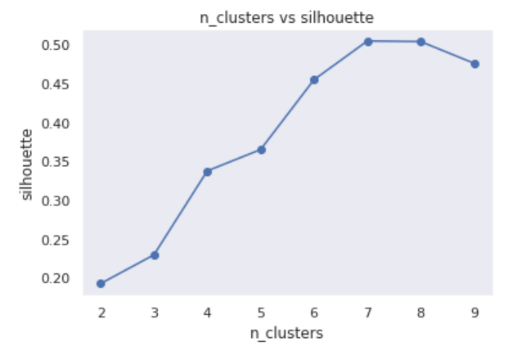
1. หาวัตถุดิบที่เป็นตัวแทนของแต่ละประเทศโดยการใช้ method TF-IDF ในส่วนนี้ได้นำผลลัพธ์มาแสดงเพียง 3 ประเทศ ดังนี้



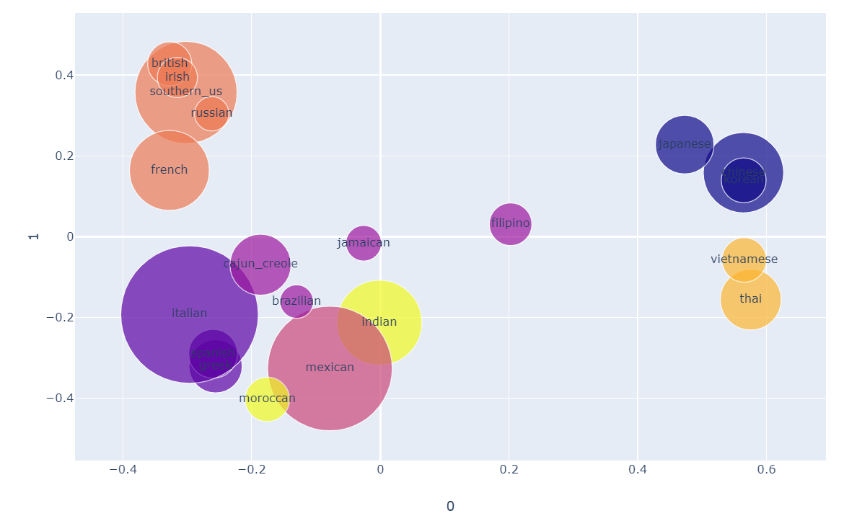
2. แบ่งกลุ่มประเทศตามวัตถุดิบ ในขั้นตอนนี้ใช้ feature เป็นวัตถุดิบในเบื้องต้นจะทำการลดมิติของข้อมูลโดยใช้ PCA เพื่อลดจำนวน feature ซึ่งจำนวนวัตถุดิบมีจำนวนที่มากเกินไปและบางวัตถุดิบมีความสัมพันธ์กัน และได้จำนวนกลุ่มที่เหมาะสมประมาณ 6 กลุ่มสำหรับการทำ PCA ผลลัพธ์ที่ได้เป็นดังรูป

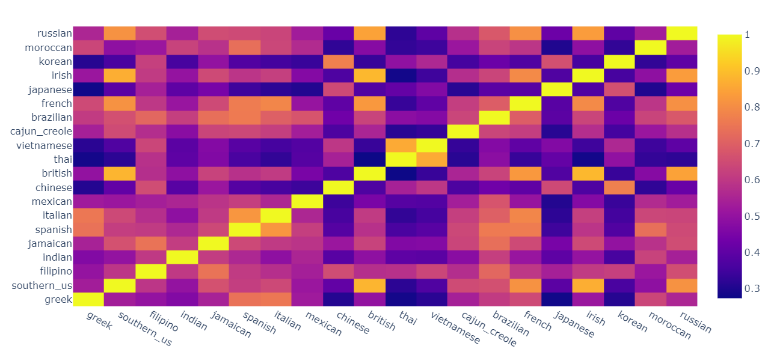
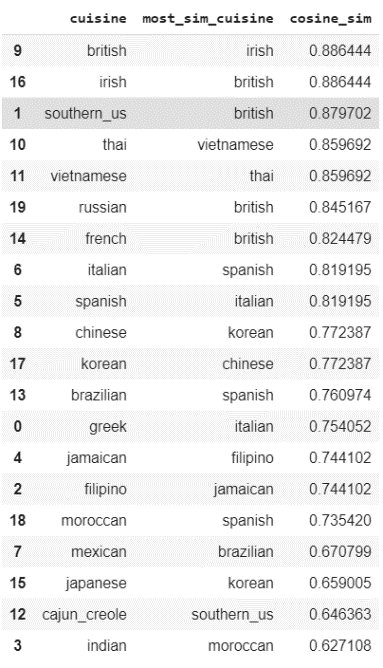


จากนั้นทำการเลือก k ที่เหมาะสมที่สุดในการแบ่งกลุ่มจากข้อมูลชุดนี้ โดยใช้ elbow method ในการเลือก k ที่เหมาะสม จากรูปด้านล่างพบว่า K = 7 เป็นจำนวนที่เหมาะสมที่สุด



จากนั้นทำการแบ่งกลุ่มข้อมูลโดยใช้ K-mean method



 จากนั้นหาความคล้ายคลึงของวัตถุดิบจาก Cosine similarity ผลลัพธ์ที่ได้สามารถบ่งบอกว่าถ้าจะทำอาหารฟิวชั่นระหว่างชาติไหนสามารถดูจาก similarity ได้ กรณีที่ similarity มาก แสดงว่ามีโอกาสมากที่วัตถุดิบเหล่านั้นจะสามารถสน้างเมนูใหม่ๆ ได้ลงตัว กรณี similarity น้อย แต่ไม่ได้หมายความว่าจะเป็นไปไม่ได้

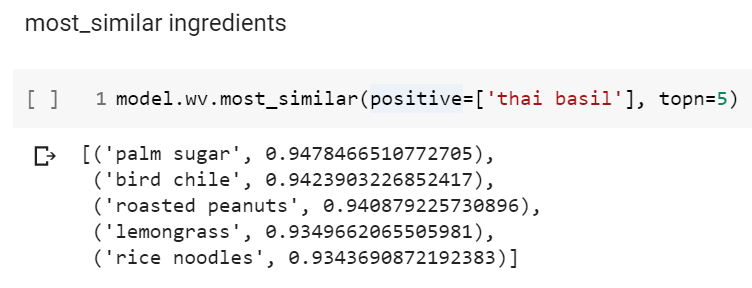
จากผลลัพธ์ที่ได้ จะเห็นว่ากลุ่มต่าง ๆ ได้นำเสนอลักษณะของอาหารและวัตถุดิบที่คล้ายคลึงกัน รวมทั้งบ่งบอกถึงวัฒนธรรมหรือภูมิประเทศที่คล้ายคลึงกันอีกด้วย ผลลัพธ์ที่ได้สามารถนำไปประยุกต์ใช้เพื่อสร้างเมนูใหม่ๆ จากวัตถุดิบที่คล้ายคลึงกันได้ ทำให้สามารถใช้วัตุดิบที่มีอยู่ในประเทศนั้น ๆ ได้อย่างคุ้มค่า ทำให้ร้านอาหารมีเมนูใหม่ๆ สามารถเพิ่มยอดขายของร้านได้และสามารถลดต้นทุนการสั่งวัตถุดิบเนื่องจากใช้วัตถุดิบเดียวกันหรือคล้ายกัน

**ส่วนที่ 2 หาความคลายคลึงของวัตถุดิบ** ขั้นตอนนี้จะใช้ word2vec ซึ่งเป็นการแสดงคำให้อยู่ในรูปของ vector เพื่อนำเสนอความคล้ายคลึงของคำเพื่อแสดงถึงความหมายคำ โดยกำหนดให้วัตถุดิบเป็น feature จากผลลัพธ์ทำให้ได้ space ใหม่ 100 dim ได้ผลลัพธ์ดังนี้

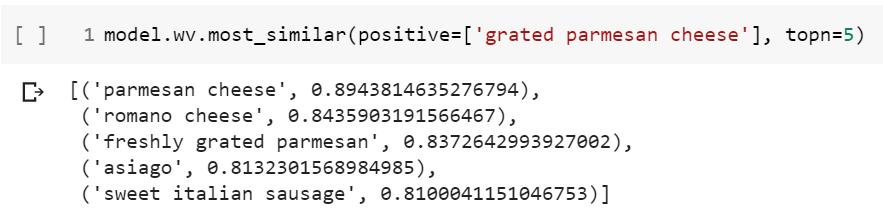
1. วัตถุดิบที่ไม่เข้าพวกมากที่สุด

2. วัตถุดิบที่เกิดขึ้นร่วมกันบ่อยกับวัตถุดิบนั้น ๆ

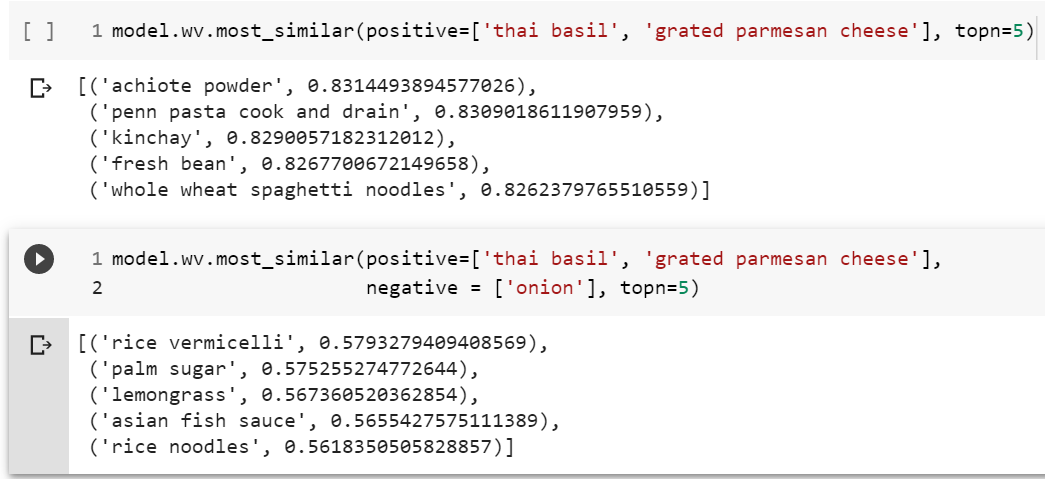
# เลือกวัตถุดิบจากอาหาร thai



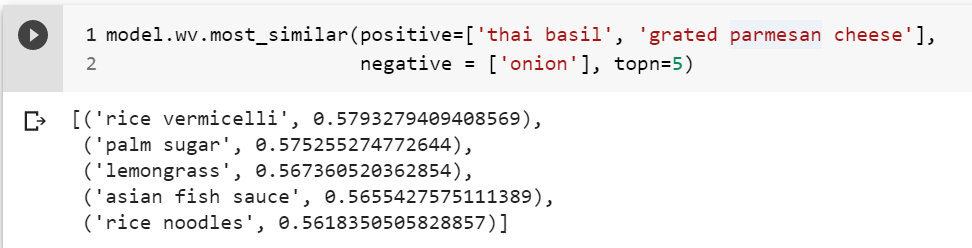
# เลือกวัตถุดิบจากอาหาร italian



3. หาวัตถุดิบที่เป็นตัวเชื่อมระหว่างวัตถุดิบสองอย่าง ตัวอย่างเช่น 'thai basil' และ 'grated parmesan cheese' ผลลัพธ์ที่ได้คือ



4. หาวัตถุดิบที่เป็นตัวเชื่อมระหว่างวัตถุดิบสองอย่างแต่ไม่อยากได้วัตถุดิบของชาติอื่น ๆ เป็นส่วนผสม เช่น ต้องการผสม 'thai basil' และ 'grated parmesan cheese' แต่ไม่อยากให้อาหารออกแนว spanish



**การประยุกต์ใช้**