

Recipy

Motor de búsqueda de recetas usando grafos

Luis Ibarra¹ y Marcos Valdivié¹

¹Facultad de Matemática y Computación, Universidad de La Habana

Introducción

El procesamiento de redes complejas es un campo en constante evolución que permite la representación y análisis de sistemas complejos a través de la creación de grafos. En este trabajo, se explorará el uso de un conjunto de recetas para la creación de grafos que permitan extraer información relevante sobre las mismas. A través del análisis de estas redes, se espera obtener una comprensión más profunda de las interacciones y relaciones entre los elementos del sistema.

Se plantea la construcción de distintos sistemas que permitan realizar distintos análisis sobre el dataset Food.com, a partir de los cuales se definirán y evaluarán un conjunto de consultas para la extracción de información relevante. Por último, se propone la elaboración de una herramienta informática que permita acceder de forma sencilla a la información extraída.

Datos

Se analizaron tres conjuntos de datos cada uno provisto con diferentes elementos y dimensiones (Tabla 1):

- Cocina al Minuto
- Food.com [majumder2019generating]
- RecipeNLG [bien-etal-2020-recipenlg]

Cocina al Minuto

Conjunto de datos creado a partir del libro Cocina al Minuto. Este presenta pocas muestras y no posee los pasos para la confección de la receta, por lo que no es posible algunos de los análisis.

RecipeNLG

Conjunto de datos más grandes de los analizados. El tamaño de este conjunto afecta el rendimiento del motor de búsqueda en tiempo y espacio, necesitando más recursos de los disponibles para poder ejecutar la aplicación.

Food.com

Este conjunto de datos es el más completo con respecto a la variedad de datos que ofrece. Tiene un tamaño mediano y es el utilizado en la implementación final del motor de búsqueda ya que se obtiene un balance tamaño e información.

Grafos

Para la conformación de la base de datos para el motor de búsqueda se contruyen diferentes grafos que contienen información relacionada a la interacción entre los componentes que de una manera u otra conforman las recetas.

Grafo Bipartito Receta-Ingrediente

Se define como un grafo no dirigido en el cual sus nodos representan recetas o ingredientes y sus aristas representan la participación de un ingrediente en la confección de la receta. Dado que el grafo es bipartito, no se permiten las aristas entre ingredientes ni entre recetas.

En este grafo, el grado (*degree*) del nodo representa la cantidad de recetas que necesitan el ingrediente dado, y de manera recíproca también representa la cantidad de ingredientes utilizados para confeccionar una determinada receta.

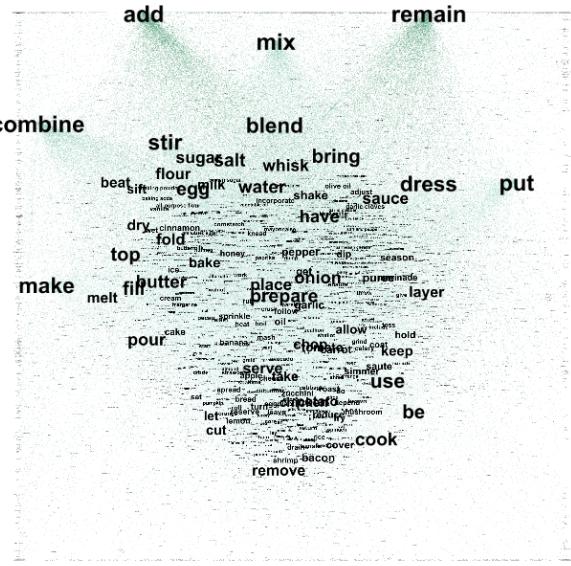
Grafo Bipartito Ingrediente-Acción

Se define como un grafo bipartito y no dirigido, formado por dos grupos: los ingredientes y el conjunto de acciones que se emplean sobre esos ingredientes.

Para la construcción se tomó el conjunto de instrucciones que posee cada receta del dataset *Food.com*, se extrajo la lista de tokens para cada instrucción y se clasificaron los mismos utilizando el módulo *tokenize* de *nltk* y el método *pos_tag_sents* de esta misma biblioteca. Dada la lista de tokens descrita anteriormente, se filtraron los verbos y los sustantivos, haciendo coincidir estos últimos con la lista de ingredientes usados en la receta, y se les asignó la acción correspondiente, formando de esta forma pares <acción, ingrediente> que definen las aristas del grafo. Por último, se le asignó peso a las aristas

Table 1: Atributos de los conjuntos de datos.

Atributo	Cocina al Minuto	Food.com	RecipeNLG
Cantidad de Recetas	555	230,186	1,312,871
Cantidad de Ingredientes	214	14,927	170,204
Cantidad de Pasos	-	2,248,564	9,709,075
Cantidad de Comentarios	-	1,132,367	-

**Figure 1:** Grafo Ingrediente-Acción (8223 nodos y 55547 aristas).

de acuerdo al índice *Intersection Over Union* (IOU) de la cantidad de pares a los que pertenece cada acción y cada ingrediente (fig. 1).

Dicho de otra forma,

$$w(X, Y) = \frac{|P_{<X,Y>}|}{|P_X \cup P_Y|}$$

Donde $w(X, Y)$ representa el peso de la arista entre la acción X y el ingrediente Y , $P_{<X,Y>}$ representa el conjunto de los pares $< X, Y >$ y P_X y P_Y representan los pares en los que se encuentra X e Y respectivamente.

Grafo Receta-Receta

Se define como un grafo no dirigido en el cual sus nodos representan recetas y sus aristas están ponderadas con una métrica de similitud entre recetas. Para este tipo de grafo se confeccionaron dos versiones basadas en diferentes métricas:

1. Similitud de Jaccard entre los conjuntos de ingredientes utilizados por ambas recetas (Figuras 12 y 15).
2. Similitud vectorial entre representación semántica de recetas (Figura 16).

Similitud de Jaccard Para la confección de este grafo las aristas son ponderadas mediante la similitud de Jaccard (Ecuación 1).

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

1. Se representaron las recetas como el conjunto de ingredientes que se necesitan para ser preparadas.
2. Se calcula la similitud de Jaccard para cada par de nodos obteniendo el peso de la arista correspondiente.

Similitud semántica de recetas Para la confección de este grafo se realizaron un conjunto de pasos para la vectorización semántica de las recetas.

1. Se representaron las recetas como una lista de las instrucciones de su preparación.
2. Cada instrucción fue codificada por el modelo *Universal Sentence Encoder* [cer2018universal] el cual devuelve un vector de dimensión 512 por cada oración procesada, obteniendo una representación final para la receta con dimensión (*CantInstrucciones*, 512).
3. Se procedió a disminuir la dimensión hasta un vector de dimensión 256. Esto se logra mediante un entrenamiento no supervisado con modelo encoder-decoder (Figura 2).
4. Para el peso de las aristas se utiliza la similitud de exponencial inversa de la distancia entre vectores (Ecuación 2)

$$invExpSim(x, y) = e^{-||x-y||} \quad (2)$$

Grafo Ingrediente-Ingrediente

Se define como un grafo no dirigido en el cual sus nodos representan ingredientes y sus aristas están ponderadas con una métrica de similitud entre ingredientes. Para este tipo de grafo se confeccionaron tres versiones basadas en diferentes métricas:

1. Similitud de Jaccard entre los conjuntos de ingredientes utilizados por ambos ingredientes (Figuras 10 13).

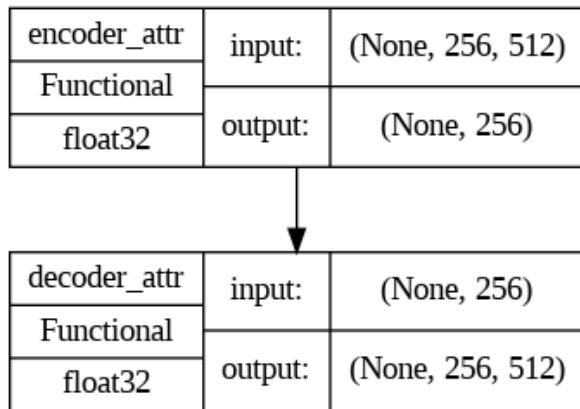


Figure 2: Arquitectura encoder-decoder para el vectorización de recetas.

2. Similitud vectorial entre representación sintáctica de ingredientes mediante vectorización por frecuencia de términos.
3. Similitud por *Pointwise mutual information* (PMI) definida en **teng2012recipe** ((Figuras 11 14)).

Similitud de Jaccard La confección de este grafo se realiza de manera similar a su contraparte de recetas.

1. Los ingredientes se representan por el conjunto de recetas en los que se usan.
2. Se calcula la similitud de Jaccard (Ecuación 1) entre ingredientes para ponderar las aristas entre estos.

Este grafo provee información acerca de cómo están asociados los ingredientes con las recetas. Una valor de similitud cercano a 1 indica que los ingredientes pertenecientes a la arista se comparten en la mayoría de las recetas en que son utilizados. En caso de estar cercano a 0, indica que no coinciden en casi ninguna receta dichos ingredientes.

Similitud vectorial entre representación sintáctica de ingredientes En la confección de este grafo se realiza una transformación del nombre de los ingredientes a vectores mediante un conteo de las palabras presentes en su nombre.

1. Cada ingrediente se representa por su nombre.
2. Los nombres son vectorizados a una tabla de términos de frecuencia normalizada.
3. Las aristas entre los ingredientes son ponderadas por la similitud de coseno entre las representaciones vectoriales de cada par de ingredientes.

Una razón por la cual construir este grafo es para dar información acerca de las diferentes variantes de ingredientes. Estas distintas variantes, por lo general poseen nombres compuestos con la clase principal en su nombre, por ejemplo:

- Queso
- Queso Parmesano
- Queso Ravioli
- Queso Gouda

Similitud por PMI En la confección de este grafo se calcula el PMI [teng2012recipe] (Ecuación 3) entre cada par de ingredientes:

$$PMI(a, b) = \log \frac{p(a, b)}{p(a)p(b)}, \quad (3)$$

donde

$$p(a) = \frac{\# \text{ de recetas que contienen } a}{\# \text{ de recetas}},$$

$$p(a, b) = \frac{\# \text{ de recetas que contienen } a \text{ y } b}{\# \text{ de recetas}}.$$

1. Cada ingrediente es representado por el conjunto de recetas en el cual es utilizado.
2. Se calcula el PMI entre cada par de ingredientes y se pondra la arista con este valor.

En este grafo, los pesos pueden llegar a ser negativos o positivos. En este caso nos interesa los pesos positivos, ya que estos pesos indican que la probabilidad de que coincidan dos ingredientes al mismo tiempo es mayor que la probabilidad que coincidan de forma independiente, por lo tanto su ocurrencia está condicionada a ser más probable dado que tengo un ingrediente.

$$PMI(a, b) = \log \frac{p(a, b)}{p(a)p(b)} > 0$$

$$\frac{p(a, b)}{p(a)p(b)} > 1$$

$$p(a, b) > p(a)p(b)$$

$$p(a|b)p(b) > p(a)p(b) \quad p(b|a)p(a) > p(a)p(b)$$

$$p(a|b) > p(a) \quad p(b|a) > p(b)$$

Ranking

El ranking se lleva a cabo para tomar la información más relevante de los datos de acuerdo a un criterio. Se emplea para dos funcionalidades, disminuir la cantidad de los nodos de los grafos y recuperar los elementos más similares a una consulta dada de los datos.

Ranking de Nodos

Dada la cantidad de nodos presentes en los grafos, en algunas situaciones existe la necesidad de reducir esta cantidad. Para esto se toman diferentes estrategias en dependencia del tipo de grafo con el que se trabaje:

1. Ranking de recetas por valoraciones de usuarios.
2. Ranking de ingredientes por ocurrencias acumuladas.

Ranking de recetas por valoraciones de usuarios

Dado un conjunto de valoraciones a recetas dadas por usuarios se procede a definir una medida de importancia con estas. Para esto se realizan los siguientes pasos:

1. Se extraen para cada receta la cantidad de valoraciones y la media de estas.
2. Dichos valores son normalizados entre 0 y 1 al dividirlos por la cantidad máxima de ambas métricas por receta.
3. Se calcula la métrica F1 entre ambos valores para obtener la importancia final de la receta (Ecuación 4).

$$ImpReceta(r, Val) = F1(A(r, Val), M(r, Val)), \quad (4)$$

donde

$$F1(a, b) = 2 \frac{a\Delta b}{a + b},$$

$$A(r, Val) = \frac{|[r \sqsubseteq Val]|}{\max\{|[p \sqsubseteq Val]| \mid p \in R\}},$$

$$M(r, Val) = \frac{\text{mean}(|[r \sqsubseteq Val]|)}{\max\{\text{mean}(|[p \sqsubseteq Val]|) \mid p \in R\}},$$

Val es la lista de valoraciones dadas por los usuarios, la expresión $[r \sqsubseteq Val]$ denota la lista de valores de las valoraciones dadas a la receta r y R es el conjunto de recetas.

Esta métrica previene que recetas con pocas y muy altas valoraciones tengan más importancia que recetas con muchas valoraciones pero más bajas. De manera que las recetas más importantes tengan que tener al mismo tiempo una gran cantidad de valoraciones y que estas sean buenas (Figura 3).

Este algoritmo fue utilizado para reducir el número de recetas de Food.com para la construcción de los grafos tipo Receta-Receta. La reducción se hizo a 5000 nodos dado que a partir de esta cantidad la relevancia de las recetas es muy baja (Figura 18)

Ranking de ingredientes por ocurrencias acumuladas Dado un conjunto de ingredientes se toman los que participan en la mayor cantidad de enlaces entre recetas. Para realizar el ranking se realizan los siguientes pasos:

1. Ordenar por grado de mayor a menor los nodos ingredientes del grafo bipartito de ingredientes-recetas.
2. Seleccionar los ingredientes, por este orden, hasta que la suma de los grados de los nodos seleccionados represente un porcentaje fijo de la suma total de todos los grados del grafo.

Este algoritmo fue utilizado para reducir el número de ingredientes de Food.com para la construcción de los grafos tipo Ingrediente-Ingrediente. Los ingredientes resultantes contiene un 90% de las ocurrencias con solo 521 lo que representa solo el 3.5% del total (Figura 17)

Ranking de Resultados a Consultas

Para la extracción de ingredientes y recetas dadas una consulta se emplean dos mecanismos para hacer el cálculo de la relevancia.

- Distancia de Levenshtein: Se calcula la distancia de Levenshtein entre la consulta y el nombre de la entidad a extraer.
- Similitud Vectorial: Se calcula la similitud de coseno entre la vectorización de la consulta y los vectores de la entidad. Las entidades son vectorizadas mediante una tabla TF-IDF.
- Ranking de recetas por ingredientes que las usen.

Ranking de recetas por ingredientes que las usen

Dado un conjunto de ingredientes se quiere encontrar las recetas que los usen, para esto se toman dos variantes. En primer lugar se ordena por la cantidad de ingredientes que tienen las recetas en común con los ingredientes requeridos. Esta aproximación lleva a que en casos de que la intersección sea completa se ordene por nombre lo que no necesariamente es deseable. El otro método es vectorizar las recetas por sus ingredientes mediante TF-IDF y realizar el ranking con la similitud de coseno entre la vectorización TF-IDF de los ingredientes consultados.

Motor de Búsqueda

Para la confección del motor de búsqueda de recetas se conformó un paquete de Python, **recipy**, el cual contiene las funciones necesarias para la creación y consulta de grafos. Para la visualización de los

resultados se creó una aplicación usando **streamlit** la cual permite al usuario interactuar con la API.

En este el usuario puede hacer distintas consultas a los distintos grafos para recuperar la información deseada.

Métodos de Búsqueda

Para los algoritmos funcionen es conveniente tener el nodo exacto sobre el cual se realiza la consulta. Para la obtención de dicho nodo a partir de una consulta de texto se emplean los métodos explicados en la sección Ranking de Resultados a Consultas para la obtención de un ranking de posibles candidatos. Con este resultado el usuario puede seleccionar el más adecuado y con esta información realizar la consulta a un nodo concreto.

Consultas

La aplicación provee al usuario de un conjunto de consultas predefinidas.

Buscar recetas por nombre Dado una consulta devuelve las primeras mejores 100 recetas de acuerdo al criterio de búsqueda seleccionado (Figura 4). Esta consulta se realiza sobre el grafo bipartito de receta-ingrediente, filtrando los nodos recetas.

Buscar ingredientes por nombre y ver recetas que los usan Dado una consulta devuelve los primeros mejores 100 ingredientes de acuerdo al criterio de búsqueda seleccionado, estos ingredientes luego pueden ser añadidos a una canasta de la cual se seleccionaran los ingredientes para hacer la búsqueda de recetas con estos, devolviendo los mejores 100 resultados de acuerdo al segundo criterio seleccionado (Figura 5). Ambas partes de la consulta se realizan sobre el grafo bipartito de receta-ingrediente, filtrando los nodos ingredientes en la primera y calculando las métricas correspondientes en la segunda.

Buscar ingrediente que mezcle bien con otro Dado una consulta devuelve los primeros mejores 100 ingredientes de acuerdo al criterio de búsqueda seleccionado, estos ingredientes luego pueden ser seleccionados, devolviendo los mejores 100 ingredientes con los cuales se puede complementar (Figura 6). Esta consulta es realizada sobre el grafo PMI de ingredientes.

Buscar recetas por similitud entre estas La similitud entre recetas se viene dada de dos formas:

- Por coincidencia de ingredientes: En este caso se devuelven las que tengan mayores coincidencias con la receta dada. Calculando dichas recetas por el grafo bipartito de ingrediente-recetas (Figura 7).

- Por similitud semántica: En este caso se devuelven las recetas ordenadas por pesos adyacentes a la receta dada en el grafo de similitud semántica (Figura 8).

Reemplazo de ingredientes Dado un ingrediente, es posible hacer una lista ordenada de aquellos otros por los que puede ser sustituido. Esto se logra tomando la partición del grafo *ingrediente-ingrediente* a la cual pertenece el ingrediente dado para obtener una lista de posibles reemplazos. Luego, se ordenan estos de acuerdo al índice de Jaccard de la cantidad de acciones comunes entre ambos ingredientes, utilizando el grafo de <acción, ingrediente>. De esta forma se obtiene una medida, para cada elemento de la lista obtenida, de cuan seguro es usar dicho ingrediente como reemplazo del original.

Se implementó una herramienta usando streamlit mediante la cual el usuario inserta un ingrediente en un cuadro de texto y se le muestra una lista con los elementos por los cuales puede ser sustituido (Figura 9).

Conclusiones

En este trabajo se estudió e implementó una herramienta informática para la extracción de información relevante del dataset de recetas Food.com a partir del uso de diversos algoritmos de Inteligencia Artificial, Sistemas de Recuperación de Información y Análisis de Redes Complejas. En particular, se construyeron diversos sistemas complejos a partir de las relaciones existentes entre los ingredientes, las recetas, los pasos de preparación y los reviews de usuarios existentes en el dataset mencionado, a partir de los cuales fue posible realizar consultas como: Qué ingredientes puede sustituir un ingrediente dado?, Qué recetas puedo elaborar dada esta lista de ingredientes?, Qué recetas son similares a una receta dada?, entre otros.

Anexos

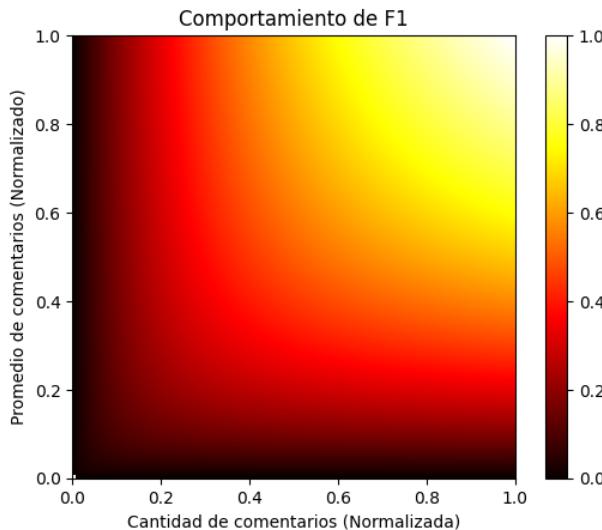


Figure 3: Variación del valor de $F1$ en dependencia de los argumentos.

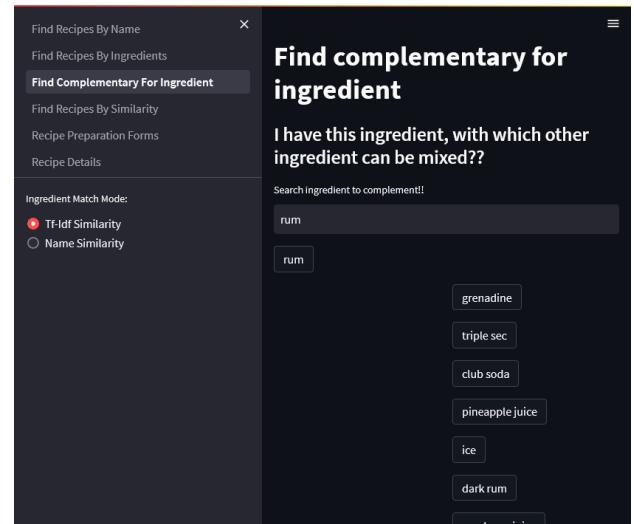


Figure 6: Búsqueda de ingredientes para complementar un ingrediente dado.

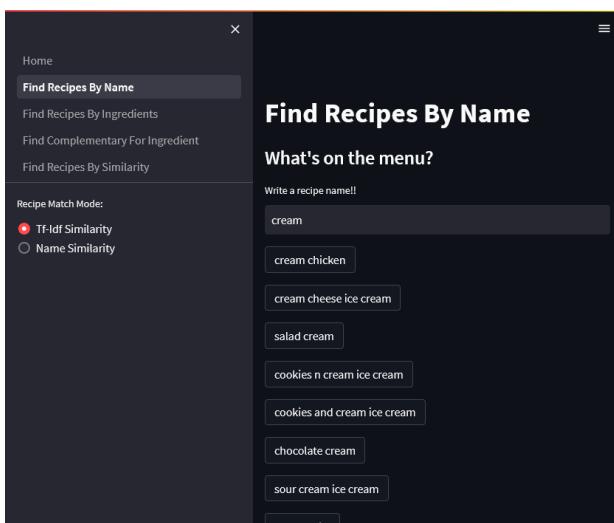


Figure 4: Búsqueda de receta por nombre con ranking TF-IDF.

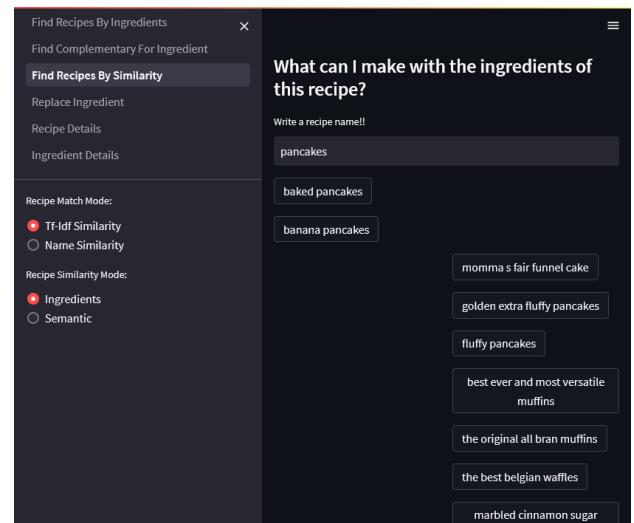


Figure 7: Búsqueda de ingredientes para complementar un ingrediente dado.

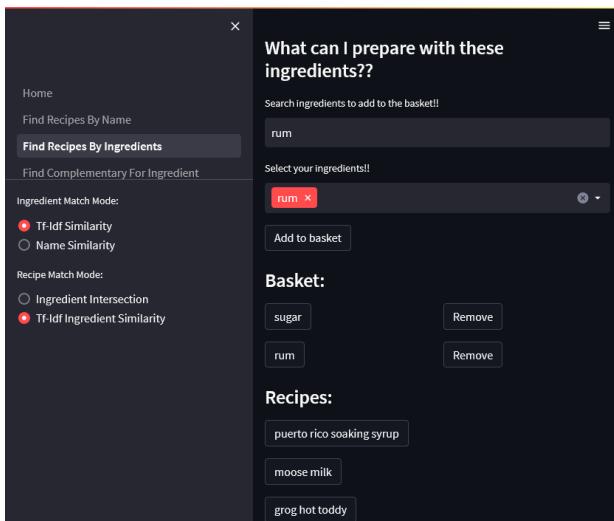


Figure 5: Búsqueda de ingredientes por nombre con ranking TF-IDF y de recetas para realizar con estos.

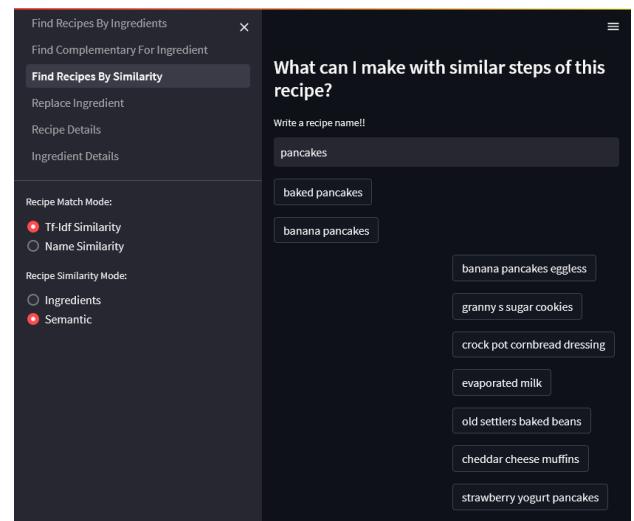


Figure 8: Búsqueda de ingredientes para complementar un ingrediente dado.

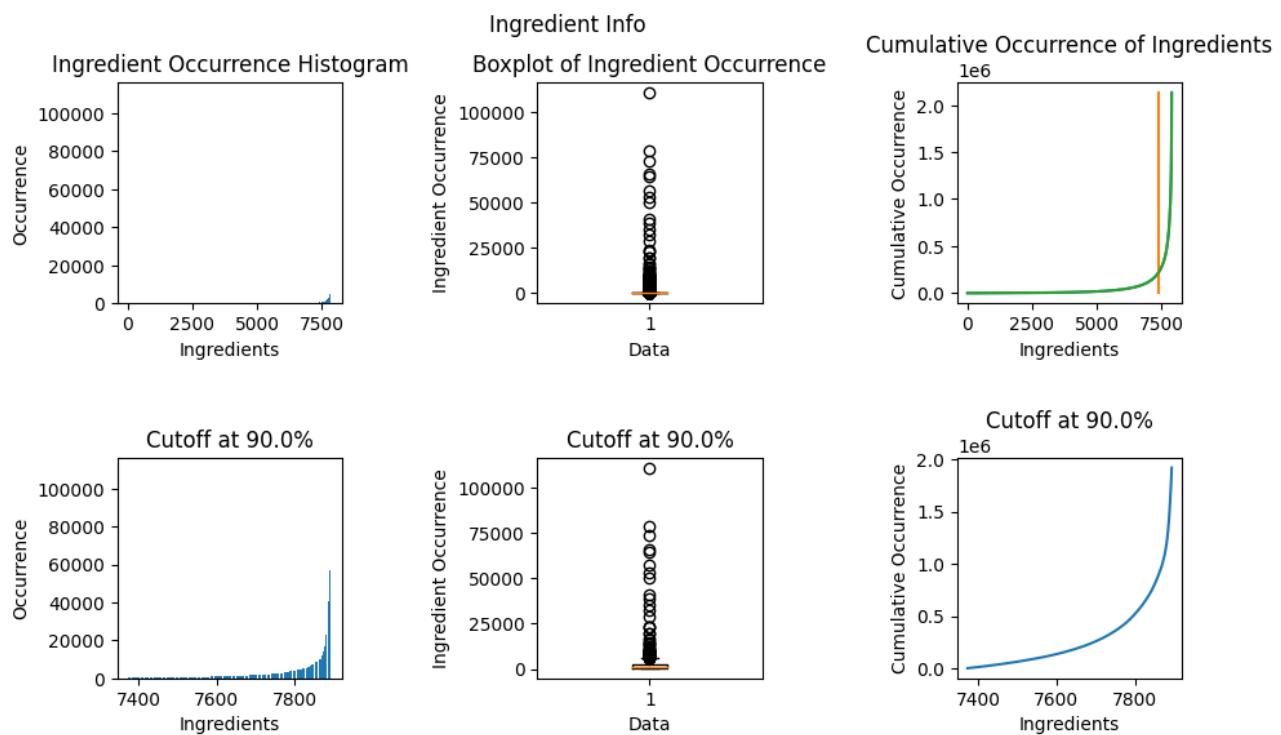


Figure 17: Reporte de importancia de los ingredientes dado la cantidad de veces que son usados en las recetas.

Replace Ingredient

How can I replace this ingredient?

Select your ingredient!

salt

Here is a list of some possible replacements for salt:

- pepper
- onion
- paprika
- vinegar
- shallot

Figure 9: Página de reemplazo de ingredientes.

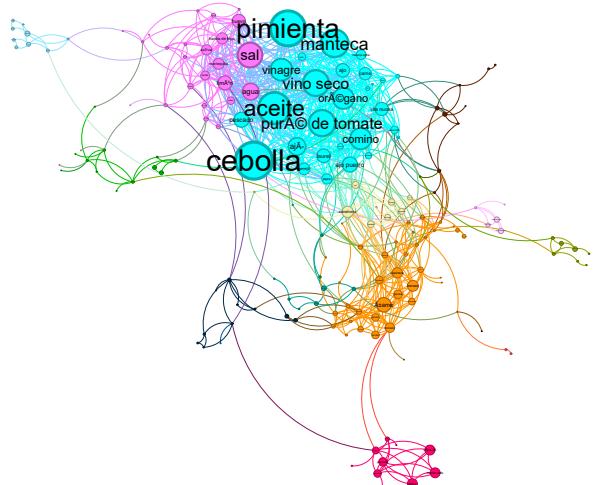


Figure 10: Cocina al Minuto Grafo Ingrediente-Ingrediente ponderado con Similitud de Jaccard entre Recetas.

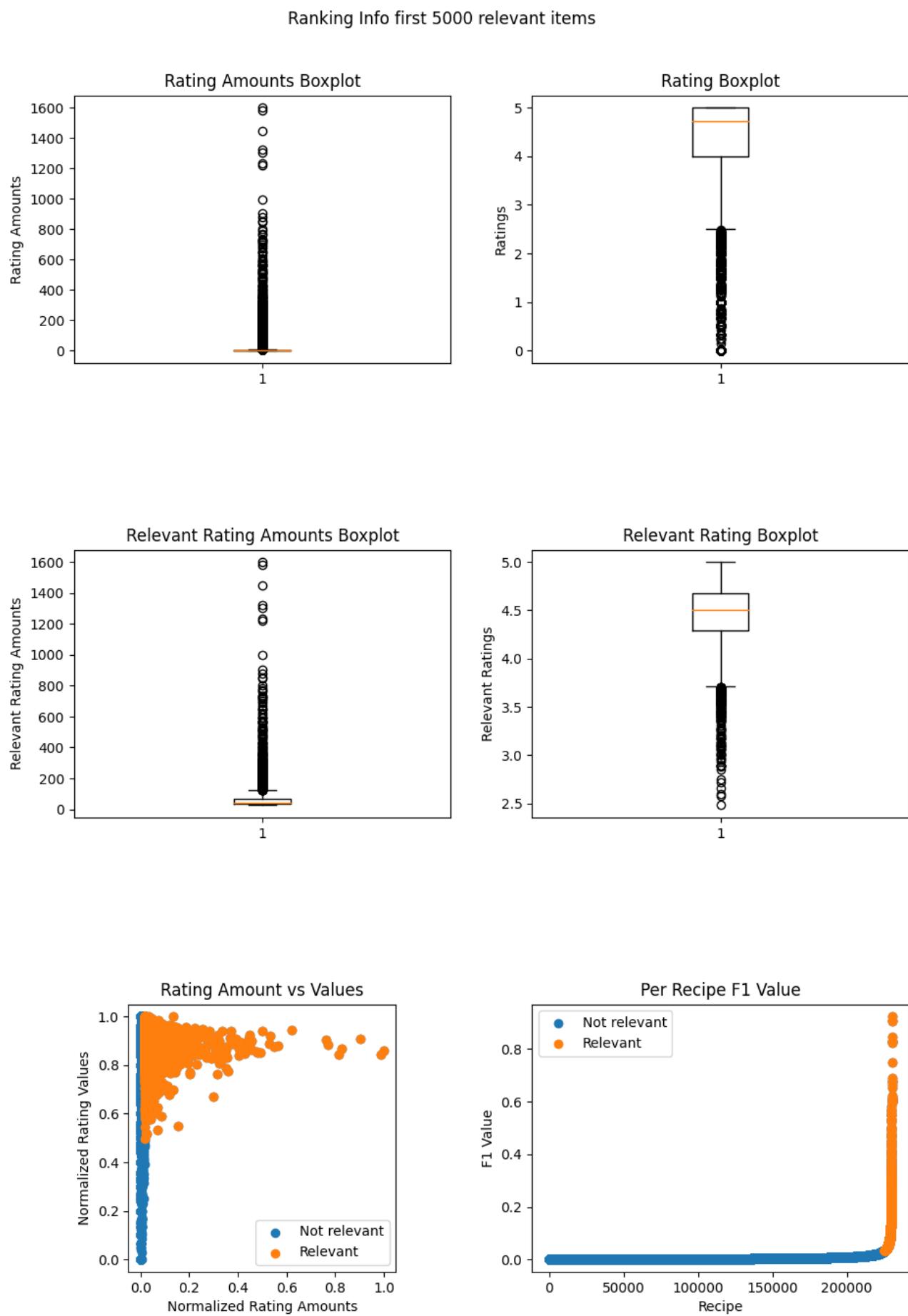


Figure 18: Reporte de importancia de las recetas dada la métrica F1.

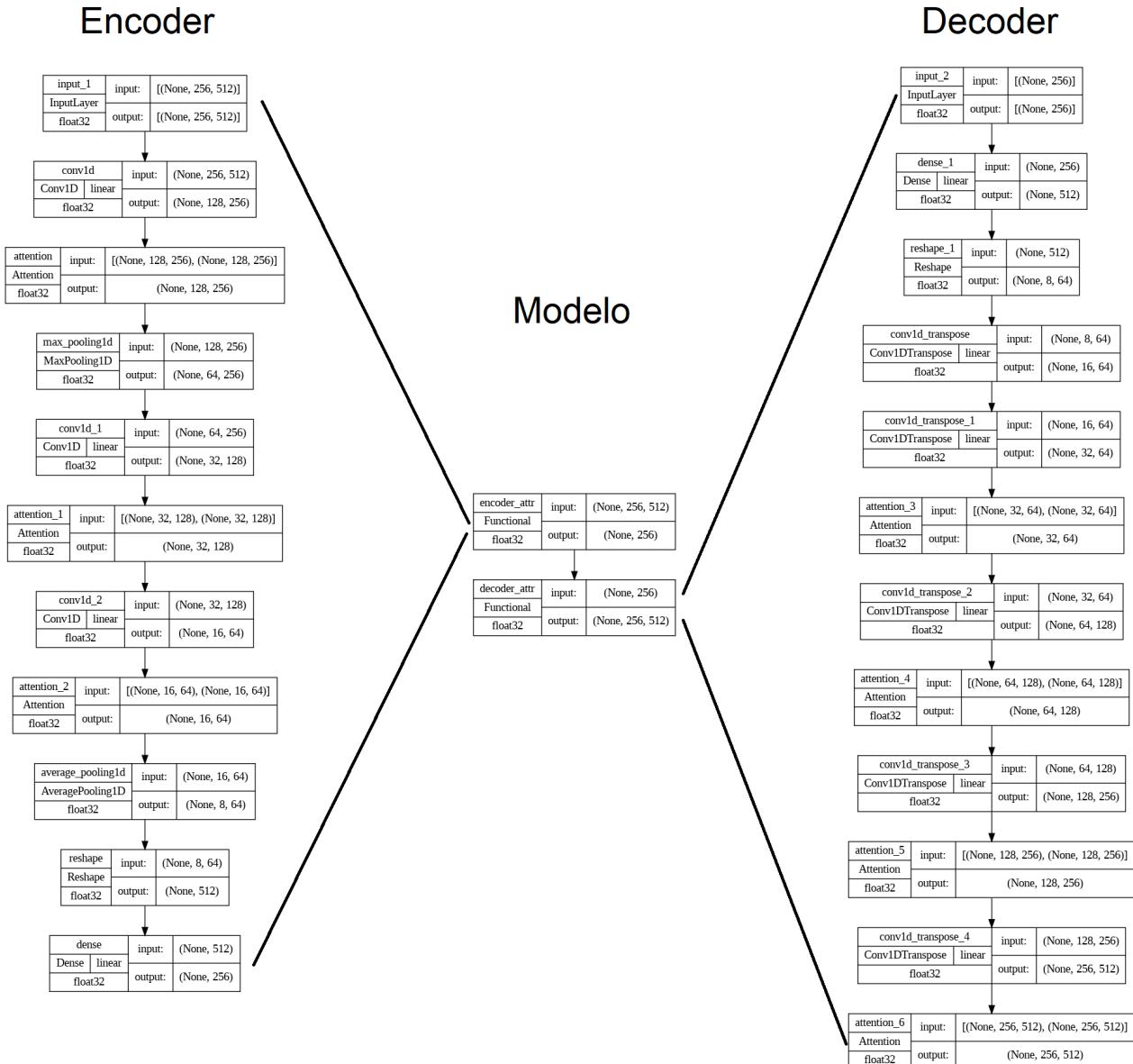


Figure 19: Arquitectura encoder-decoder para el vectorización de recetas.

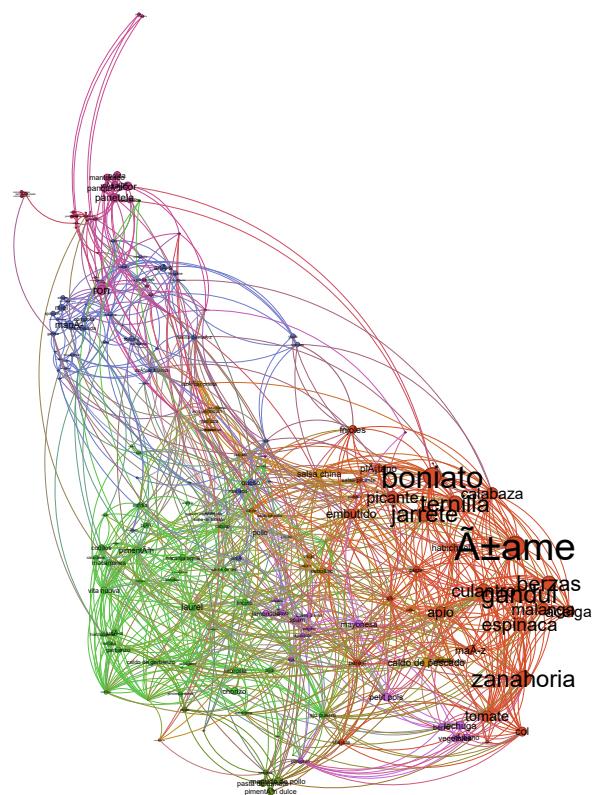


Figure 11: Cocina al Minuto Grafo Ingrediente-Ingrediente ponderado con Métrica de PMI entre Recetas.

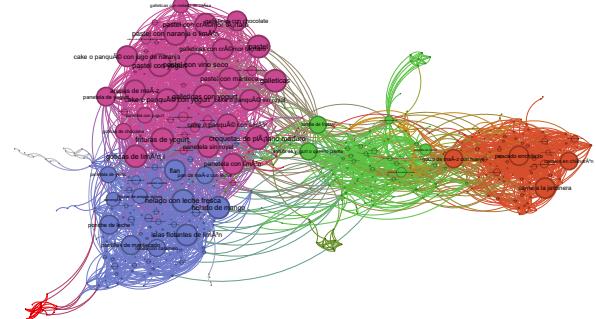


Figure 12: Cocina al Minuto Grafo Receta-Receta ponderado con Similitud de Jaccard entre Ingredientes.

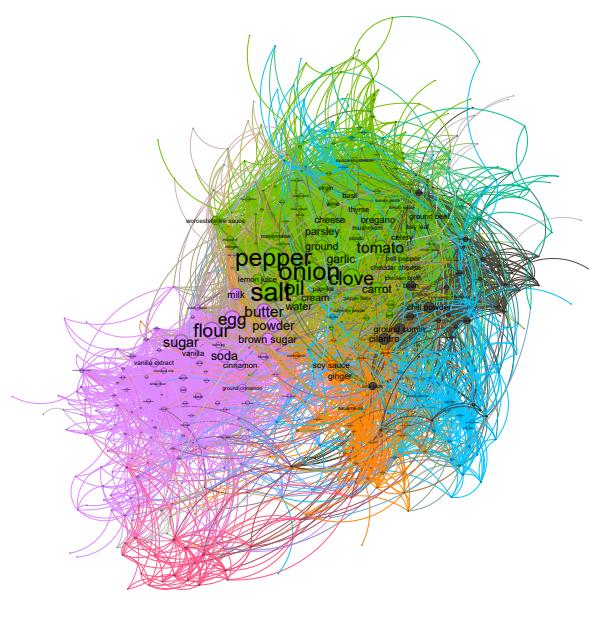


Figure 13: Food.com Grafo Ingrediente-Ingrediente ponderado con Similitud de Jaccard entre Recetas.

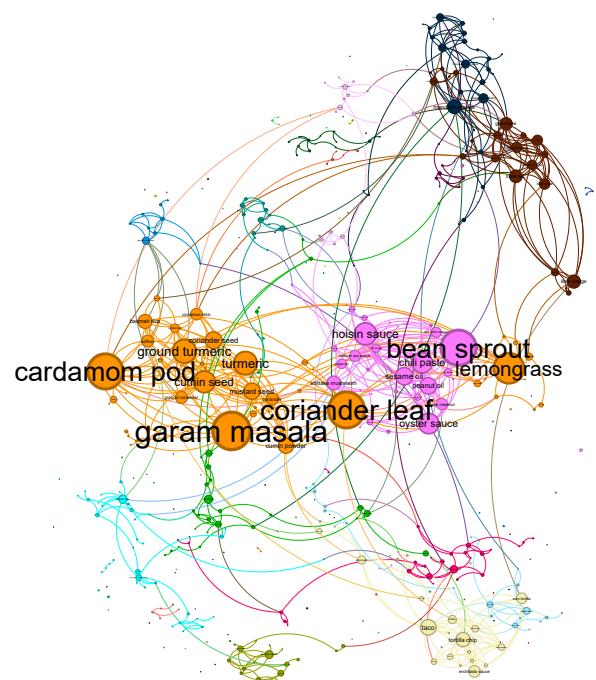


Figure 14: Food.com Grafo Ingrediente-Ingrediente ponderado con Métrica de PMI entre Recetas.

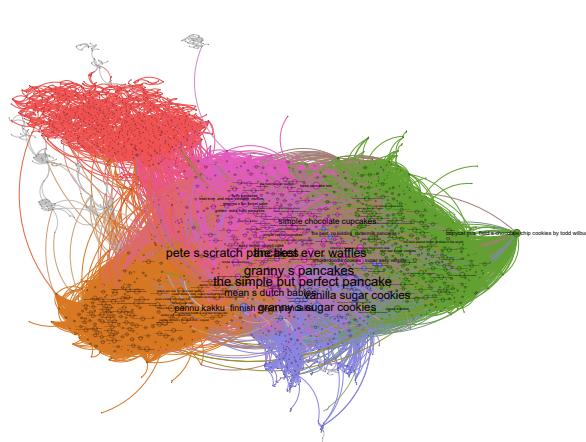


Figure 15: Food.com Grafo Receta-Receta ponderado con Similitud de Jaccard entre Ingredientes.

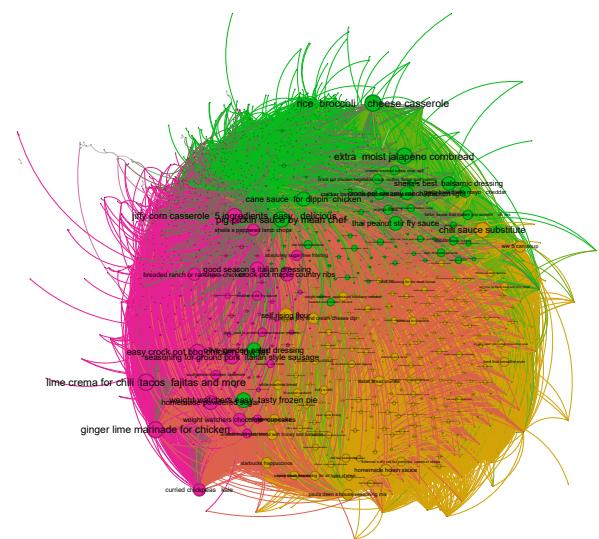


Figure 16: Food.com Grafo Receta-Receta ponderado con Similitud Semántica entre Recetas.