Proyecto Final

Métodos Analíticos: Ingredientes

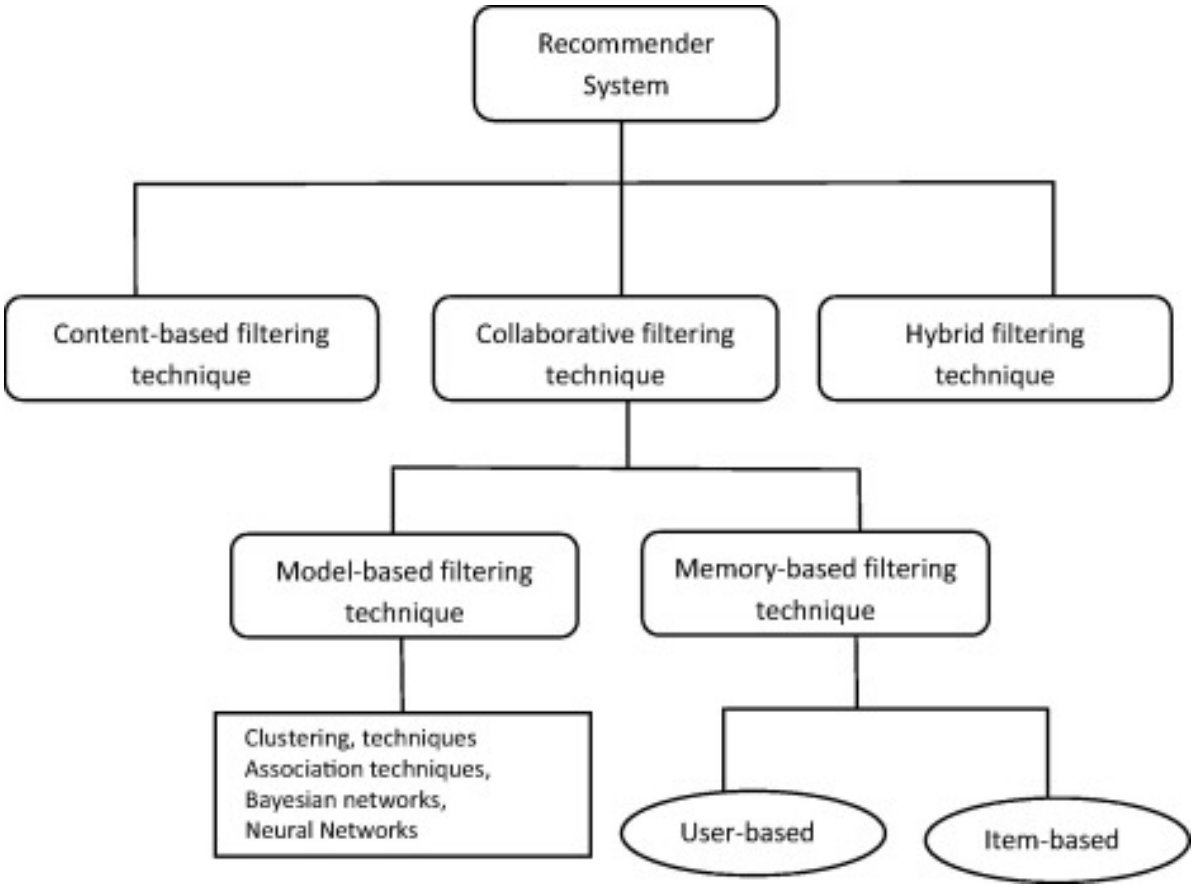
Jorge Altamirano
Uriel Miranda
Rodrigo Cedeño

Introducción

Objetivo: Desarrollamos un sistema de recomendación en donde se identifican distintos ingredientes para poder clasificarlos en 20 tipos de cocina.

Los ingredientes son capturados por usuarios en Yummly.com por lo que son *ruidosos*: incluyen errores ortográficos y tipográficos, marcas (algunas recetas parecen ser patrocinadas).

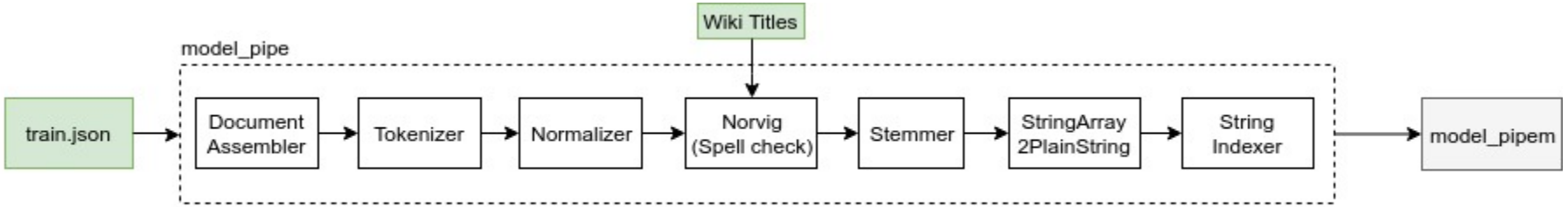
Sistemas de recomendación

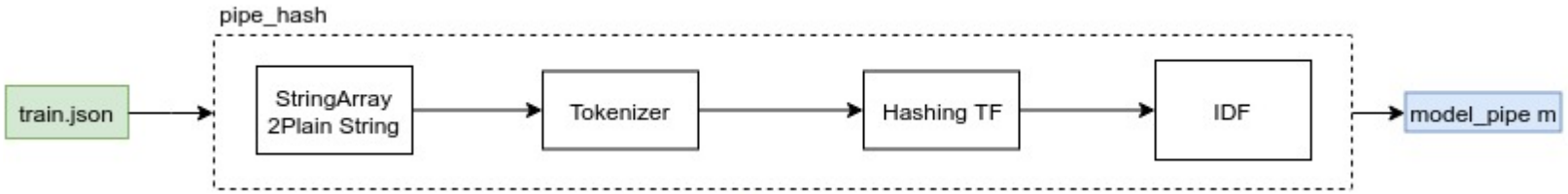


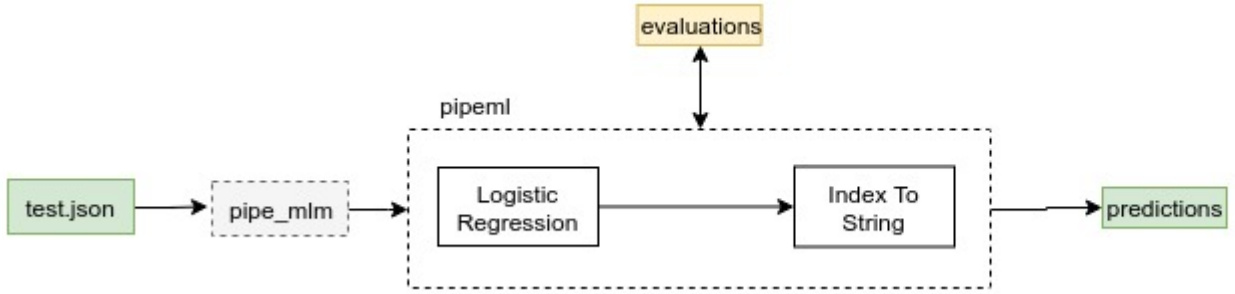
Implementación

- El dataset fue obtenido de la página de Kaggle [2], y los datos provienen de Yummly [3]
- Ayudan a las personas a encontrar qué comer basado en preferencias personales.
- Modelos desarrollados en PySpark: tanto lenguaje natural (NLP) como Machine Learning.
- Análisis de datos en R

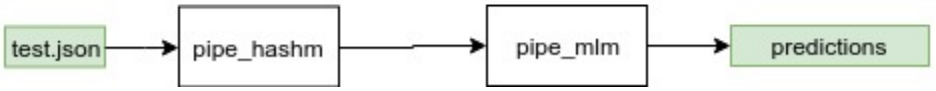
Diagramas



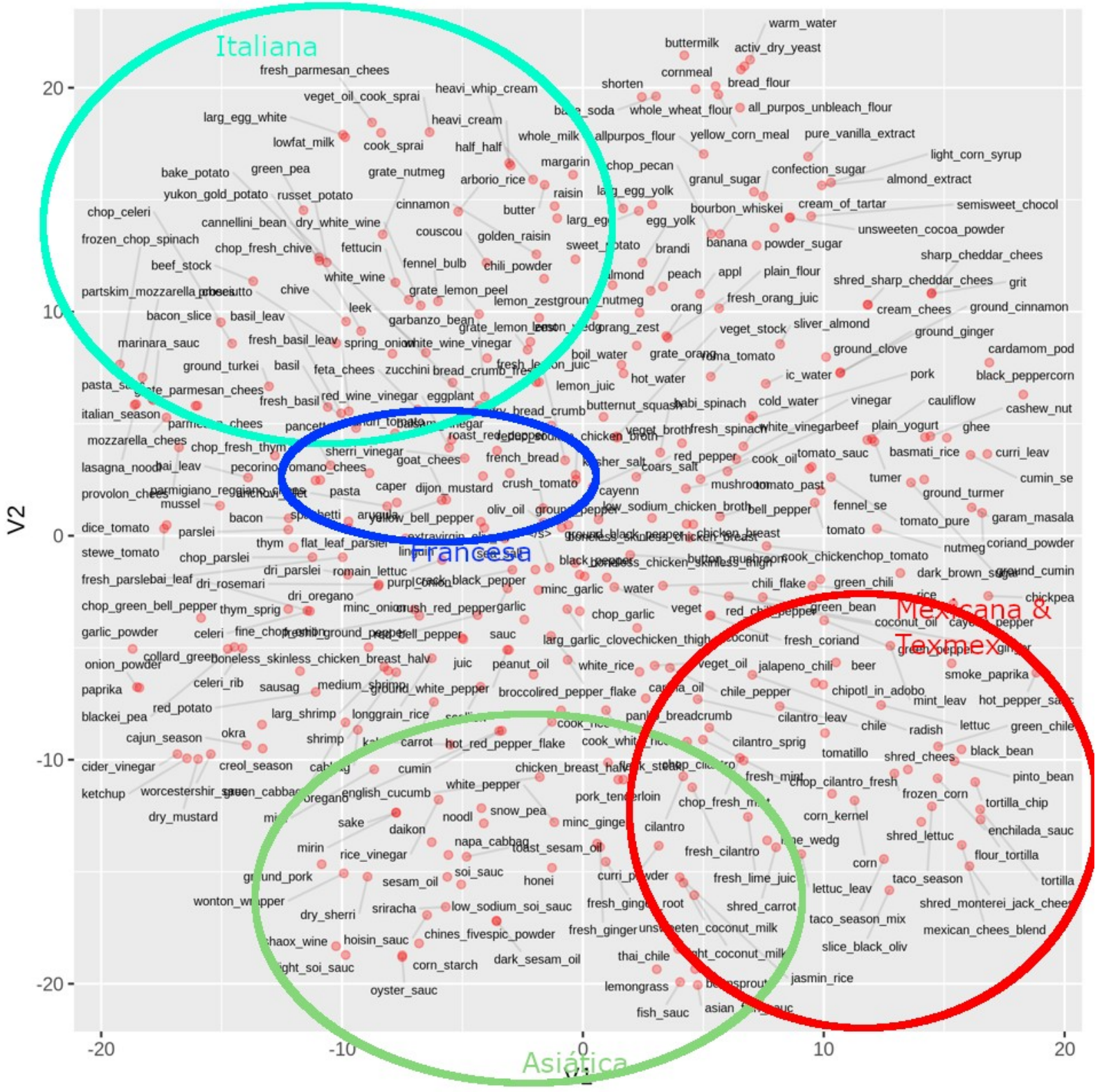




Prueba Kaggle



Visualización Word2Vec



Conclusiones

- Mejoramos el performance de nuestro ML basándonos en técnicas NLP de la clase para mejorar un F1 Score de: 0.60 a 0.68
- Realizamos una técnica más limpia: de utilizar muchísimo Regexp a generalizar basándonos en el paquete NLP: tenemos más robustés en nuestra clasificación para ingredientes desconocidos o que no aparecieron en el entrenamiento.
- Reinterpretamos la proximidad coseno con Word2Vec.
- Utilizamos las herramientas aprendidas en clase:
  - Hashing (TDIDF)
  - Tokenizing
  - Stemming
  - Spell Checker: alimentado por los títulos de Wikipedia en sus secciones Español, Inglés, Italiano, Alemán, Vietnamita.

Bibliografía

- D Jurafsky, J H Martin. (2018). 4: Language Modeling with N-Grams. Speech and Language Processing(250). Stanford, CA: Prentice Hall.
- John Snow Labs - SparkNLP: https://nlp.johnsnowlabs.com/
- Apache Spark: https://spark.apache.org
- Notas del Curso
- Bibliografía completa en nuestro GitHub: philwebsurfer/metodos-analiticos-2018