**PDF Raporu: Yemek Tarifi Eşleştirme Analizi 2102131013 Yaren İnce**

**1. Giriş**

**Ödevin Amacı**  
Bu ödevin amacı, bir yemek tarifi veri setinde yer alan malzemelere dayalı olarak tarifler arasında benzerlik analizi yapmaktır. TF-IDF ve Word2Vec modellerini kullanarak, verilen bir tarifin malzemelerine en benzer tarifleri bulmayı ve modellerin performansını anlamsal ve Jaccard benzerlik ölçütleriyle değerlendirmeyi hedefledik.

**Veri Seti**  
Veri seti, "recipe\_final (1).csv" dosyasından yüklenmiştir ve 48,735 tarif içermektedir. Her tarif; recipe\_id, recipe\_name, ingredients\_list, aver\_rate, review\_nums, calories, fat, carbohydrates, protein, cholesterol, sodium, fiber ve image\_url gibi sütunlara sahiptir. Analiz, yalnızca ingredients\_list sütununa odaklanmıştır.

**2. Yöntem**

**Benzerlik Hesaplama**

* **Ön İşleme:** Malzeme listeleri tokenize edildi, stop words çıkarıldı, kelimeler lemmatize ve stem edilerek iki ayrı corpus (lemmatized ve stemmed) oluşturuldu.
* **TF-IDF:** Malzemelerin metin temsilleri TfidfVectorizer ile vektörleştirildi. Cosine benzerliği kullanılarak giriş tarifine en benzer 5 tarif bulundu.
* **Word2Vec:** Lemmatized ve stemmed corpuslar için CBOW ve Skip-gram modelleri, farklı pencere boyutları (2, 4) ve vektör boyutları (100, 300) ile eğitildi. Ortalama vektörler alınarak cosine benzerliği hesaplandı.
* **Jaccard Benzerliği:** Modellerin önerdiği tarif kümeleri arasında örtüşme oranı hesaplandı.

**Kullanılan Modeller ve Teknikler**

* **TF-IDF:** tfidf\_lemmatized, tfidf\_stemmed
* **Word2Vec:** 16 model: lemmatized/stemmed, CBOW/Skip-gram, window=2/4, vector\_size=100/300 kombinasyonları
* **Kütüphaneler:** pandas, numpy, nltk, sklearn, gensim

**3. Sonuçlar ve Değerlendirme**

**Her Model için İlk 5 Benzer Metin**  
Örnek giriş: ['pork belly', 'smoked paprika', 'kosher salt', 'ground black pepper'] (Homemade Bacon)

* **tfidf\_lemmatized:**
  1. Homemade Bacon
  2. Crispy Pork Belly
  3. Filippino Lechon Kawali
  4. Caramelized Pork Belly (Thit Kho)
  5. Split Pea Soup with Pork Belly
* **tfidf\_stemmed:**
  1. Homemade Bacon
  2. Crispy Pork Belly
  3. Filippino Lechon Kawali
  4. Caramelized Pork Belly (Thit Kho)
  5. Split Pea Soup with Pork Belly
* **lemmatized\_model\_skipgram\_window2\_dim300 (En Başarılı):**
  1. Homemade Bacon
  2. Crispy Pork Belly
  3. Hot Dog Sausages
  4. How to "Dry-Brine" Pork Chops
  5. Pork Roast with the World's Best Rub

**Benzerlik Skorları ve Tablolar**

* **Anlamsal Değerlendirme Tablosu**  
  | Model | Ortalama Puan | |------------------------------------------|---------------| | tfidf\_lemmatized | 3.20 | | tfidf\_stemmed | 2.20 | | lemmatized\_model\_cbow\_window2\_dim100 | 2.80 | | lemmatized\_model\_skipgram\_window2\_dim100 | 3.20 | | lemmatized\_model\_cbow\_window4\_dim100 | 2.20 | | lemmatized\_model\_skipgram\_window4\_dim100 | 2.80 | | lemmatized\_model\_cbow\_window2\_dim300 | 3.20 | | lemmatized\_model\_skipgram\_window2\_dim300 | 3.80 | | lemmatized\_model\_cbow\_window4\_dim300 | 2.20 | | lemmatized\_model\_skipgram\_window4\_dim300 | 3.20 | | stemmed\_model\_cbow\_window2\_dim100 | 1.80 | | stemmed\_model\_skipgram\_window2\_dim100 | 2.20 | | stemmed\_model\_cbow\_window4\_dim100 | 1.40 | | stemmed\_model\_skipgram\_window4\_dim100 | 1.80 | | stemmed\_model\_cbow\_window2\_dim300 | 2.20 | | stemmed\_model\_skipgram\_window2\_dim300 | 2.80 | | stemmed\_model\_cbow\_window4\_dim300 | 1.40 | | stemmed\_model\_skipgram\_window4\_dim300 | 2.20 |
* **Jaccard Benzerlik Matrisi (Örnek)**  
  | Model | tfidf\_lemmatized | tfidf\_stemmed | lemmatized\_model\_skipgram\_window2\_dim300 | |------------------------------------------|------------------|---------------|------------------------------------------| | tfidf\_lemmatized | 1.00 | 1.00 | 0.25 | | tfidf\_stemmed | 1.00 | 1.00 | 0.25 | | lemmatized\_model\_skipgram\_window2\_dim300 | 0.25 | 0.25 | 1.00 |

**Hangi Model(ler) Daha Başarılıydı?**

* En yüksek anlamsal puanı (3.80) alan **lemmatized\_model\_skipgram\_window2\_dim300** en başarılı modeldir. Skip-gram modelleri, kelimelerin bağlamsal ilişkilerini daha iyi yakaladı.
* **TF-IDF vs. Word2Vec:** TF-IDF, yüzeysel kelime eşleşmelerine odaklanırken (ortalama 3.20 ve 2.20), Word2Vec (özellikle Skip-gram) anlamsal ilişkileri daha iyi modelledi.
* **Model Yapılandırmalarının Etkisi:** Vector\_size=300, daha fazla bilgi saklayarak performansı artırdı. Window=2, daha lokal bağlamları yakalayarak daha iyi sonuçlar verdi. CBOW, Skip-gram’e göre daha az anlamlı sonuçlar üretti.

**4. Sonuç ve Öneriler**

**Genel Çıkarımlar**

* Skip-gram tabanlı Word2Vec modelleri, özellikle lemmatized verilerle, en anlamlı benzerlik sonuçlarını üretti.
* TF-IDF, hızlı ve basit eşleşmeler için etkili, ancak bağlamsal derinlik eksik.
* Jaccard matrisi, lemmatized ve stemmed modeller arasında düşük örtüşme gösterdi; Skip-gram modelleri daha tutarlı.

**Öneriler**

* **Hangi Model?** Lemmatized Skip-gram (window=2, vector\_size=300), anlamsal olarak zengin sonuçlar için idealdir. TF-IDF, hızlı ve basit eşleşmeler için uygundur.
* **Görev Türleri:** Word2Vec, karmaşık tarif önerileri ve bağlamsal analiz için; TF-IDF, temel malzeme eşleşmeleri için tercih edilebilir.