# Tema 3

Text mining și sisteme de recomandare
Mineritul datelor și analiza datelor (MDAD)

2023 - v1.0
Deadline: 02.06.2023 (23:59)

## Descriere generală

În această temă veți explora unelte pentru indexarea și căutarea de texte folosind baze de date vectoriale (i.e. Weaviate) dar și implementarea unui sistem de recomandare pentru rețete alimentare.

## Setul de date

Veți folosi setul de date Food.com Recipes and Interactions, ce conține peste 170K rețete și 700K review-uri date acestora, pe care le puteți găsi în fișierele `RAW_recipes.csv` și `RAW_interactions.csv` respectiv. Mai multe informații despre conținutul fiecărui fișier puteți găsiți pe *Kaggle* în secțiunea *Data Card*.

## Căutare bazată pe texte

Folosind Weaviate, pe care îl veți instala și configura, veți stoca descrierile rețetelor din setul de date. Utilizați (doar) câmpurile *ID* și *Description* din `RAW_recipes.csv` pentru a popula o tabelă de date vectoriale.
Pe baza datelor vectoriale astfel stocate realizați o scurtă analiză calitativă plecând de la 10 query-uri (cât mai reprezentative) propuse de voi pentru a găsi, folosind *Similarity Search (nearText)*, cele mai relevante descrieri de rețete. Câteva exemple de query-uri ar fi:

- *Autumn inspired food*
- *Vibrant colors and fresh fish cooked in a mediterranean style*
- *Low fat Christmas dishes with no added sugar*

**Atenție!** NU folosiți query-urile de mai sus în analiza voastră! Folosiți, evident, limba engleză!

## Sisteme de recomandare

Plecând de la setul de date propus antrenați un sistem de recomandare folosind `interactions_train.csv` pe care îl veți evalua apoi folosind `interactions_test.csv`

Veți pleca de la codul descris în Notebook practic - Sisteme de recomandare (II) și veți construi un sistem de recomandare pentru rețetele din setul de date propus.

**Atenție!** Deoarece ID-urile rețetelor și ale utilizatorilor nu sunt secvențiale (e.g., pentru rețete încep de la 33 și se termină la 538k), va fi necesară crearea unei mapări între ID-ul acestora și un index.

# Livrabile

Predarea temei constă în a încărca pe Moodle o **arhivă cu codul** implementat de voi pentru rezolvarea temei împreună cu **raportul tehnic** care va conține:

- O scurtă analiză (tip EDA) a setului de date
- Cele 10 query-uri propuse (minim 5, maxim 30 de cuvinte per query)
- Rezultatele (complete) obținute pentru cele 10 query-uri
- Analiza calitativă asupra acestor rezultate (raportat la query-urile propuse)
- Detalii de rulare (timp, număr de pași, etc.) a antrenării sistemului de recomandare, precum și rezultatele obținute (i.e. RMSE)