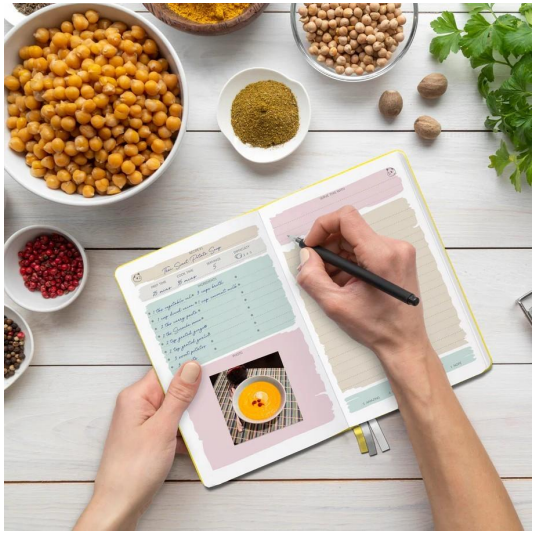
Система за търсене и класиране на рецепти по налични продукти

Курсов проект по извличане на информация и откриване на знания

Изготвил: Екатерина Хамбарлийска
Преподавател: проф. д-р Иван Койчев

Какво прави системата?



- приема списък от налични продукти
- намира рецепти, които ги съдържат частично или изцяло
- подрежда рецептите по степен на релеванстност

Мотивация за избор на тема

Данни

RecipeNLG (Kaggle)

title:

No-Bake Nut Cookies

ingredients:

["1 c. firmly packed brown sugar", "1/2 c. evaporated milk", "1/2 tsp. vanilla", "1/2 c. broken nuts...

directions:

["In a heavy 2-quart saucepan, mix brown sugar, nuts, evaporated milk and butter or margarine.", "St...

link:

www.cookbooks.com/Recipe-Details.aspx?id=44874

source:

Gathered

NER:

["brown sugar", "milk", "vanilla", "nuts", "butter", "bite size shredded rice biscuits"]

Основни компоненти на системата

- Зареждане на данните
- Нормализация и почистване на данните
- Представяне на рецепти като документи
- Индексиране на корпуса
- Търсене по заявка от налични продукти
- Класиране на резултатите
- Оценка на качеството

Почистване и нормализация на данните

Проблем

Списъците със съставки съдържат:

- количества и мерни единици
- описателни думи
- различни форми на една и съща съставка

Цел

- унифицирано представяне на съставките
- по-точно индексиране
- по-стабилно класиране

NER vs ingredients

```
6 baking potatoes  ->  baking potato
1 lb. of extra lean ground beef  ->  extra lean beef
2/3 c. butter or margarine  ->  butter margarine
6 c. milk  ->  milk
1 1/2 c. sugar  ->  sugar
1/2 c. butter  ->  butter
1 egg  ->  egg
1 c. buttermilk  ->  buttermilk
2 c. flour  ->  flour
1/2 tsp. salt  ->  salt
```

Търсене и класиране

- Рецептите се разглеждат като документи
- Използва се **BM25** за изчисляване на релевантност
- Резултатите се подреждат по score
- **TF-IDF** се използва като базов метод за сравнение

Тестване и валидиране

Предизвикателство

- Липсват етикети
- Не е възможна директна човешка оценка за всички заявки

Подход

- Автоматично оценяване чрез self-retrieval
- Заявките се генерират от реални данни в корпуса
- Оценява се качеството на класирането
- Precision@K=|{ релевантни документи сред първите K }| / K
- Recall@K=|{ релевантни документи сред първите K }| / |{ всички релевантни документи }|

📊 Оценяване (self-retrieval)

Evaluation model

TF-IDF

K

5

Брой заявки (sample)

500

Run evaluation

Precision@5

0.1728

Recall@5

0.8640

📊 Оценяване (self-retrieval)

Evaluation model

BM25

K

5

Брой заявки (sample)

500

Run evaluation

Precision@5

0.1848

Recall@5

0.9240

Източници

https://www.kaggle.com/datasets/paultimothymooney/recipenlg/data

https://www.youtube.com/watch?v=ziiF1eFM3_4&t=46s

https://www.itm-conferences.org/articles/itmconf/abs/2022/04/itmconf_icacc2022_02006/itmconf_icacc2022_02006.html