Софийски университет „Св. Кл. Охридски”

Факултет по математика и информатика

Курсов Проект

на тема: “Система за търсене и класиране на рецепти по налични продукти”

Студент: Екатерина Милкова Хамбарлийска, ФН: 8MI3400767, спец. ИИОЗ

Курс: 1, Учебна година: 2025/2026

Преподавател: проф.д-р Иван Койчев

==========

Декларация за липса плагиатство:

- Плагиатство е да използваш, идеи, мнение или работа на друг, като претендираш, че са твои. Това е форма на преписване.
- Тази курсова работа е моя, като всички изречения, илюстрации и програми от други хора са изрично цитирани.
- Тази курсова работа или нейна версия не са представени в друг университет или друга учебна институция.
- Разбирам, че ако се установи плагиатство в работата ми ще получа оценка “Слаб”.

12.02.26 г. Подпис на студентите:

Съдържание

1 Увод

Платформите за рецепти съдържат огромен брой кулинарни предложения, което значително затруднява бързото и ефективно намиране на подходящи рецепти спрямо конкретните нужди на потребителите. В реални условия те често разполагат с ограничен набор от продукти и желаят да открият рецепти, които използват наличните съставки изцяло или частично, без да е необходимо закупуването на допълнителни продукти.

Настоящият курсов проект има за цел разработването на система за търсене и класиране на рецепти по налични продукти, базирана на класически методи за Извличане на информация. В рамките на системата рецептите се разглеждат като текстови документи, а съставките – като термини, участващи в процесите на индексиране и търсене.

Основните задачи на проекта са:

- анализ и предварителна обработка на рецептурни данни;
- унифицирано представяне на съставките с цел намаляване на шума в данните;
- представяне на рецептите като текстови документи;
- изграждане на индекс и механизъм за търсене;
- класиране на резултатите по степен на релевантност спрямо подадената заявка;
- сравнение на различни подходи за търсене и класиране;
- автоматична оценка на качеството на получените резултати.

В проекта се извършва сравнителен анализ между базовия модел *TF-IDF* и вероятностния модел *BM25*, като се изследва тяхната ефективност при търсене на рецепти по списък от налични продукти. Поради липсата на предварително зададени етикети за релевантност, качеството на системата се оценява чрез автоматичен подход, базиран на генериране на заявки от реални данни в корпуса.

2 Преглед на търсенето на рецепти

Търсенето на рецепти по съставки може да бъде разглеждано като специализиран случай на извличане на информация, при който рецептите се третират като текстови документи, а съставките – като термини в корпуса. Потребителската заявка представлява списък от налични продукти, който се използва за откриване на най-релевантните рецепти [1].

Класическите системи за извличане на информация използват обърнат индекс за ефективно търсене и прилагат модели за претегляне на термини като TF-IDF(Term Frequency – Inverse Document Frequency) и вероятностния модел BM25(Best Matching 25), които позволяват извличане на резултатите, подредени според тяхната релевантност [1].

Специфично предизвикателство при търсенето на рецепти е обработката на съставките, които често съдържат количества, мерни единици и описателни думи, както и различни форми на една и съща съставка. Това налага прилагането на предварителна обработка и унифициране на термините с цел по-точно индексиране и търсене.

Качеството на търсенето се оценява чрез стандартни метрики за извличане на информация като Precision и Recall, които позволяват сравнение между различни модели и подходи за класиране на резултатите [1].

3 Проектиране

3.1 Използвани данни

В проекта е използван корпусът RecipeNLG от Kaggle, който съдържа 1 312 871 рецепти, събрани от различни кулинарни уеб източници. Корпусът предоставя богата и разнообразна база от данни, подходяща за изследване на задачи, свързани с информационно извличане и препоръчване на рецепти.За всяка рецепта в корпуса са налични следните основни атрибути:

- заглавие на рецептата;
- списък от съставки в свободен текст;
- списък от съставки, извлечени чрез Named Entity Recognition (NER);
- линк към оригиналния източник на рецептата.

Полето със сурови съставки представлява списък от текстови низове, които често съдържат допълнителна информация, несъществена за задачата по търсене. В тези низове присъстват количества, мерни единици, описателни думи и пояснения, както и различни варианти на изписване на една и съща съставка. Това води до значителна нееднородност в данните и затруднява директното им използване за индексиране и търсене.

Корпусът предоставя и предварително извлечени NER съставки, които представляват именовани обекти от тип *ingredient*. Тези съставки са по-структурирани и в по-голяма степен отговарят на семантичните единици, необходими за задачата. Въпреки това, дори при NER данните се наблюдават вариации в изписването и остатъчен шум, което налага прилагането на допълнителна нормализация.

Тези особености налагат разработването на механизъм за предварителна обработка и нормализация, който да осигури унифицирано представяне на съставките. Такова представяне е ключово за коректното индексиране на рецептите и за стабилното функциониране на механизмите за търсене и класиране.

При проектирането на системата е взета предвид и възможността за бъдещо разширение към автоматично извличане на рецепти от уеб източници. В подобен сценарий данните биха били още по-нееднородни и шумни, поради което наличието на стабилна логика за почистване и нормализация на съставките е от съществено значение. Поради тази причина механизмът за предварителна обработка е проектиран като независим модул, приложим както към корпуса RecipeNLG, така и към потенциално извлечени в бъдеще рецепти.

3.2 Архитектура на системата

Системата е проектирана с модулна архитектура, която позволява ясно разделение на отговорностите между отделните компоненти и улеснява бъдещо разширяване и експериментиране с различни методи за търсене и класиране. Всеки модул изпълнява конкретна функция и комуникира с останалите чрез ясно дефинирани входове и изходи.

В основата на архитектурата стои модулът за предварителна обработка на данните, който отговаря за почистването и унифицирането на съставките. Този модул приема като вход както сурови текстови списъци със съставки, така и предварително извлечени NER съставки, и ги преобразува до унифицирано представяне, подходящо за индексиране. По този начин се осигурява независимост от конкретния източник на данни и възможност за бъдещо включване на нови източници.

Следващият основен компонент е модулът за избор на източник на съставки, който определя дали за дадена рецепта да се използват NER съставките или нормализираните сурови съставки. Когато NER информация е налична, тя се предпочита, тъй като предоставя по-структурирано представяне. В противен случай системата използва резултатите от нормализацията на суровите съставки. Тази логика позволява гъвкаво използване на наличните данни и по-устойчиво поведение при непълни или шумни записи.

Получените нормализирани съставки се използват от модула за представяне на документи, който разглежда всяка рецепта като документ, описан чрез множество от термини. Това представяне е съвместимо с класическите модели за информационно извличане и позволява директно индексиране без допълнителна токенизация или лингвистична обработка.

Върху представените документи се изграждат два индекса – BM25 и TF-IDF, реализирани в отделни модули. Те отговарят за изчисляване на релевантността между заявката и документите и връщат класиран списък от рецепти. Използването на два различни модела позволява сравнителен анализ на тяхното поведение и качество.

Модулът за изграждане на заявки приема като вход списък от продукти, въведени от потребителя, и прилага същата логика за нормализация, използвана при документите. Това гарантира консистентност между представянето на заявките и документите и намалява риска от несъвпадения поради различно изписване.

Над функционалните модули е изграден интерактивен потребителски интерфейс, реализиран със Streamlit. Интерфейсът позволява избор на модел за търсене, въвеждане на налични продукти, визуализация на резултатите и стартиране на експерименти за оценяване на качеството на търсенето.

Отделно от основния поток на търсене е реализиран модул за оценяване, който използва подхода self-retrieval за автоматична оценка на качеството на системата. Този модул е логически независим от търсещата функционалност и може да бъде използван както офлайн, така и директно през потребителския интерфейс.

Избраната архитектура осигурява яснота, разширяемост и възможност за експериментиране с различни компоненти, без да се нарушава цялостната структура на системата.

3.3 Представяне на знанията

В разработената система знанията се представят чрез унифицирано терминно описание на рецептите и заявките. Всяка рецепта се разглежда като документ, описан чрез списък от нормализирани съставки, които играят ролята на термини в класическия модел за

информационно извличане. По този начин пространството на документите и заявките е общо и съпоставимо.

Съставките се третират като независими термини, без да се отчита редът им или допълнителна синтактична структура. Това опростено представяне е достатъчно за целите на търсенето по налични продукти и позволява директното използване на модели като BM25 и TF-IDF. Заявките на потребителя се изграждат по същия начин – чрез нормализиране на въведените продукти и представянето им като множество от термини.

Този подход осигурява консистентност между документите и заявките и минимизира влиянието на шум и вариации в изписването на съставките.

4 Реализация, тестване/експерименти

4.1 Използвани технологии, платформи и библиотеки

Проектът е реализиран на програмния език Python, поради широката му поддръжка на библиотеки за обработка на данни и информационно извличане. За реализацията са използвани следните основни технологии и библиотеки:

- pandas – за зареждане и обработка на данните от корпуса RecipeNLG;
- rank-bm25 – за реализация на BM25 модел;
- scikit-learn – за изграждане на TF-IDF индекс;
- streamlit – за създаване на интерактивен потребителски интерфейс;
- pickle – за сериализация и повторна употреба на изградените индекси и документи.

Изборът на тези технологии е мотивиран от тяхната стабилност, разпространеност и подходящост за прототипиране и експериментиране.

4.2 Реализация

4.2.1 Предварителна обработка и нормализация на съставките

Нормализацията на съставките е реализирана чрез подход, който обработва всеки елемент от списъка със съставки поотделно. Процесът включва премахване на количества, мерни единици, описателни думи и стоп-думи, почистване на пунктуация и свеждане на думите до единствено число. Тази обработка позволява извличането на канонични имена на съставките, подходящи за индексиране.

На Фигура 1 са показани примерни резултати от прилагането на нормализацията върху реални входни данни. Сложните низове със съдържание на количества и мерки се преобразуват до унифицирани представяния, което значително намалява шума в данните.

Следва да се отбележи, че въпреки значителното намаляване на шума в данните, при правило-базирания подход в някои случаи се запазват описателни прилагателни, например във фрази като „extra lean beef“. Това поведение е нежелателно, но е прието като ограничение на настоящата реализация и може да бъде адресирано в бъдеща версия чрез по-задълбочена семантична обработка.

```
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
```

Фигура 1: Примерни резултати от нормализация на съставки – премахване на количества, мерни единици и описателни думи.

4.2.2 Изграждане на индекс и търсене

След приключване на предварителната обработка и нормализация, всяка рецепта се представя като документ, описан чрез списък от канонични термини (съставки). Върху това представяне се изграждат два независими индекса – TF-IDF и BM25.

TF-IDF моделът използва класическата схема за претегляне на термини, базирана на честотата им в документа и обратната честота в корпуса. BM25 представлява вероятностен модел,който допълнително отчита дължината на документа, което е особено важно при рецепти с различен брой съставки.

При търсене потребителската заявка се обработва чрез същата логика за нормализация, използвана при документите, което гарантира консистентност между заявките и корпуса. Резултатът от търсенето е класиран списък от рецепти, подредени според изчислената релевантност.

4.3 Провеждане на експерименти

4.3.1 Начин на оценяване

За оценяване на качеството на разработената система за търсене на рецепти по налични продукти са използвани стандартни метрики от областта на информационното извличане – Precision и Recall. Тези метрики са широко приети при оценка на търсещи системи, тъй като измерват различни аспекти от поведението на модела и позволяват обективно сравнение между различни подходи за класиране на резултатите [2].

Precision@K (прецизност) измерва каква част от първите K върнати резултата са релевантни спрямо подадената заявка. В контекста на търсене на рецепти по съставки, висока стойност на Precision@K означава, че сред първите резултати преобладават рецепти, които реално съответстват на наличните продукти.

$$\text{Precision@K} = |\{\text{релевантни документи сред първите } K\}| \ / \ K$$

Recall@K (чувствителност) измерва каква част от всички релевантни документи в корпуса се намират сред първите K върнати резултата. Тази метрика отразява способността на системата да „намери“ релевантните документи и е особено подходяща при сценарии, в които е важно релевантният резултат да се появи в горната част на класирането.

Recall@K=|{ релевантни документи сред първите K }| / |{ всички релевантни документи }|

Поради липсата на ръчно анотирани данни за релевантност, в проекта е използван автоматичен подход за оценяване, известен като self-retrieval.[3] При този подход за всяка рецепта нейните собствени нормализирани съставки се използват като заявка, а релевантен документ се счита самата рецепта. В този контекст за всяка заявка съществува точно един релевантен документ, което прави Recall@K основната метрика за оценка.

4.3.2 Експерименти

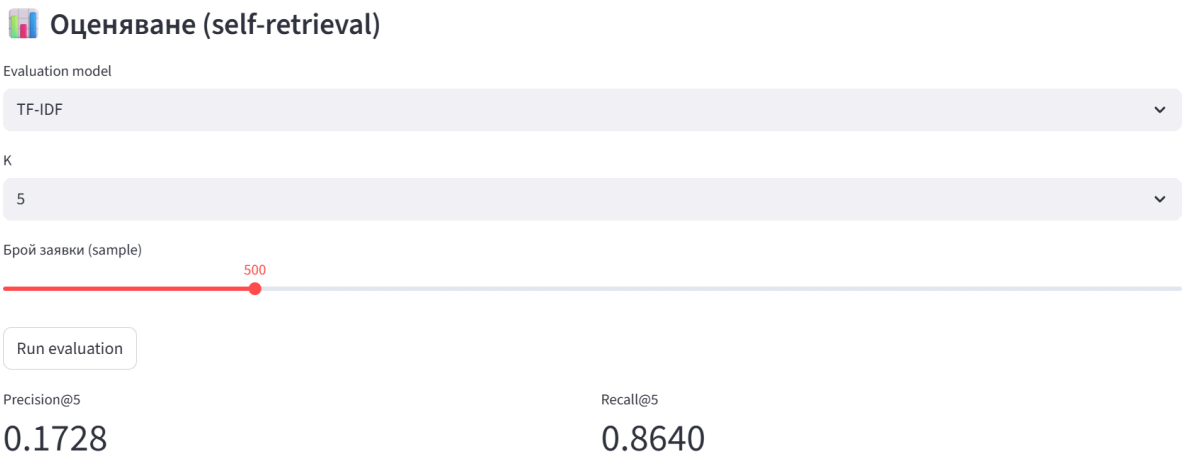
Получените експериментални резултати показват, че моделът BM25 постига по-високи стойности на Recall@5 в сравнение с TF-IDF, което означава, че по-често успява да класира оригиналната рецепта сред първите резултати. Това може да се види на Фигура 2 и Фигура 3 Ниските стойности на Precision@5 са очаквани и произтичат от естеството на self-retrieval оценяването, при което сред върнатите K резултата има само един релевантен документ. Въпреки това Precision и Recall заедно дават ясна представа за поведението и ефективността на използваните модели [2].

В рамките на експериментите беше анализирано влиянието на параметъра b на модела BM25, който контролира степента на нормализация по дължина на документите. Този параметър е от особено значение при рецептурни данни, тъй като броят на съставките в различните рецепти варира значително.

Проведените експерименти показват, че по-високи стойности на параметъра b

водят до по-добро качество на търсенето. Най-добри резултати бяха постигнати при стойност b = 0.95, при която се наблюдава максимална стойност на Recall@K, както и по-добър баланс между Precision@K и Recall@K в сравнение с по-ниски стойности на параметъра.

Това поведение може да бъде обяснено с факта,че при по-високи стойности на \(b\) моделът налага по-силна нормализация по дължина на документа, което ограничава влиянието на рецепти с голям брой съставки. По този начин се предпочитат по-компактни рецепти, които по-често отговарят на сценария за търсене по налични продукти. На Фигура 3 и Фигура 4 са показани резултатите от оценяването при различни стойности на параметъра, като се вижда превъзходството на настройката b = 0.95 пред b=0.4. Не е много голяма разлика, но съществува.



Фигура 2: Резултати от оценката за TF-IDF

Фигура 3: Резултати от оценката за BM25 за b= 0.4

Фигура 4: Резултати от оценката за BM25 за b= 0.95

5 Заключение

В курсовия проект беше разработена система за търсене и класиране на рецепти по налични продукти, базирана на класически методи за извличане на информация. Рецептите се разглеждат като текстови документи, а съставките – като термини, преминаващи през предварителна обработка и нормализация с цел намаляване на шума в данните.

Експерименталната оценка показа, че вероятностният модел BM25 се представя по-добре от базовия TF-IDF подход при търсене по съставки. Анализът на параметрите на модела показа, че силната нормализация по дължина на документите е особено подходяща за рецептурни данни и води до по-качествено класиране на резултатите.

Получените резултати потвърждават приложимостта на класическите модели за извличане на информация в контекста на търсене на рецепти по налични продукти и очертават възможности за бъдещо развитие чрез допълнително подобряване на класирането и използване на по-богати семантични представяния.

6 Използвана литература

[1] Записки към курса „Извличане на информация“, воден от проф. д-р И. Койчев

Факултет по математика и информатика,

Софийски университет „Св. Климент Охридски“, 2024.

[2] Evaluation measures (information retrieval) — Wikipedia article on standard metrics in information retrieval, including Precision and Recall.

[3] Self-Retrieval: End-to-End Information Retrieval with One Large Language Model —

OpenReview article on self-retrieval approaches in information retrieval.

https://openreview.net/forum?id=H3at5y8VFW