

DataWise – ДатаВайс

**ТЕМА:** Приложение свързано с улесняване запознанството и задълбочаване в сферата на структурите от данни и алгоритми

**1. Автори**

**1.1. Алекс Ивайлов Стефанов**

ЕГН: 0745161445

Адрес: гр.Казанлък, ул. „Добри Кехайов“ №13

Телефон: 0889475177

Email: [rlgalexbgto@gmail.com](mailto:rlgalexbgto@gmail.com)

Училище: ППМГ „Никола Обрешков“

Клас: 11б

**2. Ръководител: Здравка Стефанова Димитрова**

Телефон: 0893422519

Email: [dimitrova@pmgkk.com](mailto:dimitrova@pmgkk.com)

Длъжност: Учител по информатика и информационни технологии

**3. Резюме**

**3.1. Цели:** Проектът "DataWise" е насочен към хора, които се стремят към кариера в програмирането. Той предоставя платформа за усъвършенстване на знанията по структури от данни и алгоритми – умения, които са от критична важност при интервюта в големи технологични компании (FANG: Facebook, Amazon, Netflix, Google).  
**\*Контекст:**  
Интервютата във FANG компаниите често изискват задълбочено познаване на структури от данни и алгоритми. Този проект цели да помогне на кандидатите да се подготвят по най-добрия начин, като им предостави достъп до модерни AI технологии в комбинация с традиционни образователни ресурси.

**3.2. Основни етапи в реализирането на проекта:**

**- Основни дейности:**

**\* Формулиране на идеята**

**\* Изграждане на архитектура**

**\* Конфигуриране на екосистемата**

**\* Изучаване на изкуствения интелект (AI)**

**\* Изграждане архитектурата на AI модела**

**\* Обучение по TypeScript и Front-End технологии**

**\* Обучение по Python**

**\* Изучаване на математическите основи**

**\* Правене на DataSet с тестовите задачи (~10 000 примера)**

**\* Имплементация на AI модела**

**\* Разработка на лого и цялостен дизайн**

**\* Тестване и оптимизация**

**\* Събиране на обратна връзка**

**3.3. Ниво на сложност на проекта**

**Проектът се характеризира с множество слоеве на сложност – както от гледна точка на математическите изчисления, така и от гледна точка на интеграцията на различни технологични екосистеми. Трудността идва от това, че всяко друго приложение ползва библиотеки за разработка на AI продукти или са wrapper-и на съществуващи AI, с “DataWise” искам да покажа, че това не винаги трябва да е случеят, “DataWise.AI” е изцяло създаден от нищото(без библиотеки за AI), на чист Python. Когато всичко ти е дадено на готово е лесно и можеш да се разминеш дори и да не разбираш кода си, когато не ти е дадедно нищо обаче минаваш пред всички трудни стъпки (като програмният език, математиката зад обработването на данните и най-важното събирането на въпросните данни), но в крайна сметка се запознаш с целият процес и наистина разбираш как се работи с „Голям обем от данни“, защото човек трябва да е напълно запознат с това как изобщо да се сдобие с тях и как да ги превърне в нещо, което ще е от полза на хора. По-долу са изброени основните аспекти, като са дадени и подробни примери за математическите предизвикателства:**

* **Математически сложности:**
  + **Линейна алгебра и матрични операции:  
    За изграждането и обучението на AI модела са от съществено значение операции като умножение на матрици, транспониране и изчисляване на инверсии. Например, при предаването на данните през конволюционните слоеве се извършва тензорна операция (тензор-дот продукт), която комбинира информация от няколко измерения и осигурява правилното изчисляване на активациите.**
  + **Функция на активация и нейната диференциируемост:  
    Моделът използва функцията ReLU (Rectified Linear Unit). Тя не само преобразува входните данни, но и трябва да бъде диференцируема, за да може обратното разпространение на грешката (backpropagation) да изчисли производната и. Това включва и сложни математически концепции, като изчисляване на производни в множество точки и справяне с недефинирани стойности при граници.**
  + **Оптимизация и градиентен спуск:  
    Обучението на невронната мрежа става чрез оптимизационни алгоритми като стохастичния градиентен спуск (SGD). Това изисква постоянно изчисляване на градиенти, актуализиране на теглата и настройване на параметри като learning rate. Пример за математическо предизвикателство е адаптивното настройване на learning rate и използването на техники като momentum, за да се избегнат локални минимуми.**
  + **Обратна разпространение (Backpropagation):  
    При обратното разпространение се изчисляват производните на грешката спрямо всички параметри на мрежата. Това включва сложни вериги от диференциране, където се изисква коректно проследяване на градиентите през всяка невронна връзка. Важно е правилното пресмятане на градиентите за всяка операция, включително тези, извършвани в конволюционните слоеве.**
  + **Нормализация и регуларизация:  
    За да се подобри стабилността на обучението, се използват техники като нормализация (batch normalization) и регуларизация (L2-регуларизация). Тези методи добавят допълнителни математически изчисления, които трябва да бъдат интегрирани във функцията за загуба (loss function) и оптимизацията на модела.**
* **Данни (Data set)**
  + **Източник и авторски права:  
    Всички данни, необходими за работата на AI системата, са събрани и напраени лично от мен и са защитени с MIT лиценз.**
  + **Размер и структура:  
    Създадох dataset, съдържащ близо 10 000 примера, които са внимателно подбрани и категоризирани. Тези данни служат за обучение и валидиране на AI модела, осигурявайки богата и разнообразна тренировъчна база.**
  + **Категории:  
    Данните са разделени на следните 10 категории:**
    - **BFS (Обхождане в ширина):**
    - **DFS (Обхождане в дълбочина):**
    - **Two Pointers (Два указателя):**
    - **Dynamic Programming (Динамично програмиране):**
    - **Greedy Algorithm (Жадни алгоритми):**
    - **Backtracking (Обратна проследяемост):**
    - **Binary Search (Двоично търсене):**
    - **Disjoint Set (Несъвместими множества):**
    - **Game Theory (Теория на игрите):**
    - **N/A (Неопределено/Други):**
* **Екосистеми и интеграция:**
  + **Angular (Front-End):  
    Фронтендът е разработен с Angular, който предоставя модерен, интерактивен потребителски интерфейс. Angular комуникира с бекенд услугите чрез RESTful API, като извиква endpoint-и за получаване на резултати от AI модела и ASP.NET-а. Това осигурява динамично обновяване на данните и интерактивност за потребителя.**
  + **Flask (Python):  
    AI моделът и основната логика за обработка на данни са имплементирани с Python и Flask. Тук се извършват сложни математически изчисления, необходими за функционирането на AI модела, като обработка на текстови данни, конволюционни операции и изчисление на грейдиенти.**
  + **ASP.NET:  
    Допълнителни услуги и бизнес логика могат да бъдат реализирани чрез ASP.NET, като се интегрират с останалата част от системата чрез стандартизирани API endpoint-и. Тази част от екосистемата осигурява мащабируемост и стабилност при изпълнение на корпоративни приложения.**
  + **Свързаност на технологичните слоеве:  
    Проектът интегрира различни технологии – Angular, Flask и ASP.NET – които работят синхронно. Angular изпраща заявки към Flask, който обработва данните и изчислява резултатите на базата на сложни математически модели, а ASP.NET допълва системата с допълнителни услуги. Тази мулти-платформена интеграция изисква координация между различните езици и framework-и, което добавя допълнителна степен на сложност.**

**3.4 Логическо и функционално описание на решението**

**- Архитектура**

**Решението е разделено на различни проектни папки, като всяка от тях съдържа специфична функционалност, отговаряща за отделни аспекти на цялостната система. Този модулен подход осигурява гъвкавост, мащабност и лесна поддръжка. Ето структурата на проектните папки:**

* **DataWise.AI:  
  Този компонент съдържа AI модула, разработен на Python, който използва сложни математически изчисления за класификация на текстови задачи. Чрез Flask endpoint-и се обслужват заявките от останалите компоненти на системата.**
* **DataWise.Api:  
  Този проект представлява API слой, базиран на ASP.NET. Той служи като звено за различни операции (endpoint-и) изграждащи масивен и лесно разширим интерфейс.**
* **DataWise.Client:  
  Фронтендът е разработен с Angular и предоставя модерен и интуитивен потребителски интерфейс. Чрез него потребителите могат да взаимодействат с всички endpoint-ове и да получават незабавна обратна връзка за своите решения по задачи, свързани със структури от данни и алгоритми.**
* **DataWise.Core:  
  Този слой осигурява основните услуги и бизнес логика, реализирани на .NET. Той е отговорен за сървърната логика, координацията на процесите и интеграцията между различните компоненти в системата.**
* **DataWise.Data:  
  Слоят за данни се грижи за съхранението, управлението и достъпа до критичните данни на системата. Той предоставя надежден достъп до релационна и нерелационна база данни, необходими за функционирането на цялостната платформа.**
* **DataWise.Common:  
  Този проект съдържа общи константи и преизползваеми помагачи.**

**Тази архитектура осигурява цялостно решение, в което всеки слой играе ключова роля:**

* **Модулността позволява независима разработка и оптимизация на отделните компоненти.**
* **Мащабируемостта е постигната чрез интеграция на различни технологии, като AI моделът, API услугите и фронтенд интерфейсът работят синхронно.**
* **Лесната поддръжка и възможността за бъдещо разширяване се гарантират чрез ясното разделение на отговорностите между проектните папки.**
* **Функционалности**
  + **Локално изпълнение на модела:  
    Моделът може да бъде стартиран локално, което позволява на потребителите да го използват за лични цели. Той обработва входните данни и предоставя автоматична категоризация според предварително обучените категории.**
  + **Уеб интерфейс с категоризация:  
    Потребителите могат да посетят уебсайта, където е интегрирана функционалността за категоризация. Чрез сайта могат да въведат текст, който се обработва от AI модела, и да получат незабавна обратна връзка.**
  + **Knowledge Nexus:  
    Информационна част, насочена към обучение по алгоритми и структури от данни. Този модул помага на потребителите да се подготвят за интервюта или изпити, предоставяйки source код, описание, разлики, сложност, подтипове и друга информация, която лесно може да бъде разширена.**
  + **Data Chartizer  
    Потребителите вкарват голям dataset, който бързо бива обработван, като се вадят колони, по които се избират данни за правене на бърза, персонализирана, вадеща агригирани данни в различни видове диаграма. Използвани са оптимални алгоритми са зареждане и обработване на генерични данни (н. такива от kaggle.com).**

**3.5. Реализация**

**3.5.1 AI Модул – Архитектура на модела**

**AI (изкуственият интелект) е просто алгоритъм, който приема текст, го преобразува в числови вектори (чрез ембединг техника) и след това прилага последователни математически операции – като конволюция, пул и пълно свързване – за да класифицира входните данни.**

### **1. Входни параметри и начална инициализация**

* + - ****vocab\_size:** Брой уникални думи в речника.**
    - ****embedding\_dim:** Размерът на векторното представяне на всяка дума.**
    - ****max\_len:** Максимален брой думи, които се обработват от входното изречение.**
    - ****num\_filters:** Брой конволюционни филтри за всяка група (при всеки размер на филтъра).**
    - ****filter\_sizes:** Списък с размери на филтрите (напр. 2, 3, 4), които определят колко поредни думи се разглеждат едновременно.**
    - ****num\_classes:** Брой категории, в които се класифицира входното изречение.**
    - ****learning\_rate:** Скоростта на обучение, с която се актуализират параметрите.**

**Математическо предизвикателство: Изборът на правилните размери и коефициенти влияе върху стабилността на обучението и възможността за добър generalization.**

### **2. Ембединг слой**

* + - ****Действие:** Преобразува текстови индекси в непрекъснати вектори.**
    - ****Размери:** E ∈ ℝ [vocab\_size×embedding\_dim]**
    - **Всеки входен израз с размер: X ∈ ℝ[N×max\_len] се преобразува в: Xembedded ∈ ℝ[N×max\_len×embedding\_dim] където N е броят на примерите (batch size).**

**Математическо предизвикателство: Представянето на дискретен текст (като индекси) в плътно и смислено пространство – осигуряване на семантична свързаност между подобни думи.**

### **3. Конволюционни слоеве**

**За всеки размер на филтър ƒ (от списъка filter\_sizes):**

* + - ****Инициализация на филтрите:****

**W(ƒ) ∈ ℝ [f×embedding\_dim×num\_filters]**

* + - ****Конволюция:** За всеки "прозорец" от ƒ думи се изчислява:**

**ᴢi,k(ƒ)=ReLU(∑j=1(ƒ)∑d=1(embedding\_dim)Xi+j−1, d⋅Wj, d, k(ƒ))**

**Обяснение на символите:**

* + - **i: Начална позиция на прозореца в изречението (от 1 до max\_len−ƒ+1).**
    - **j: Индекс на думата вътре в прозореца (от 1 до fff).**
    - **d: Измерение на ембединг вектора (от 1 до embedding\_dim).**
    - **k: Индекс на филтъра (от 1 до num\_filters).**
    - **ReLU функция: ReLU(z)=max⁡(0,z).**
    - ****Изходни размери на конволюцията:**  
      За даден ƒ:**

**Z(ƒ) ∈ ℝ [N×(max\_len−f+1)×num\_filters]**

**Математическо предизвикателство: Оптимално извличане на локални характеристики с фиксирани прозорци и справяне с променливата дължина на текстовете.**

### **4. Макс пул (Max Pooling) и Конкатенация**

* + - ****Max Pooling:** За всяка филтърна карта се избира максималната стойност:**

**pk(ƒ)= max⁡1≤i≤max\_len−f+1 zi,k(ƒ)**

**Резултатът за всеки ƒ е:**

**p(ƒ) ∈ ℝ [N×num\_filtersp]**

* + - ****Конкатенация:** Обединяват се пулнатите изходи от всички филтри (ако има mmm различни размера):**

**p = [p(f1);  p(f2);  … ;  p(fm)] ∈ ℝ [N×(num\_filters×m)]**

**Математическо предизвикателство: Съхранение на най-силните характеристики от всяка филтърна карта и правилно обединяване на информацията от различните размери на филтрите.**

### **5. Финален (Напълно свързан) слой**

* + - ****Действие:** Преобразува конкатенирания вектор в изходни логити за класификация.**
    - ****Размери на теглата:****

**Wfc ∈ ℝ [fc\_input\_dim×num\_classes]**

**където:**

**fc\_input\_dim = num\_filters × m**

* + - ****Формула за логити:****

**logits=p⋅Wfc + bfc**

**` Резултатът е:**

**logits ∈ ℝ [N×num\_classes]**

* + - ****Softmax:** Превръща логитите в вероятностно разпределение:**

**ŷi,c = elogitsi,c / ∑c ′= 1num\_classes elogitsi,c′**

**Математическо предизвикателство: Стабилно свързване на конкатенираните характеристики към класове, като се осигури адекватна нормализация чрез softmax.**

### **6. Обратна разпространение (Backpropagation)**

* + - ****Изчисляване на градиенти:**  
      Градиентът на крос-ентропийната загуба спрямо логитите се изчислява като:**

**∂Loss / ∂logits = ŷ − y / N**

**След което градиентите се разпределят през FC слоя, конволюционните слоеве и ембединг слоя.**

* + - ****Актуализация на параметрите:**  
      Всеки параметър θ се актуализира с:**

**θ ← θ − η⋅∂Loss / ∂θ**

**Математическо предизвикателство: Прецизно изчисляване и проследяване на градиентите през нелинейни операции (като ReLU и max pooling), което може да доведе до проблеми с изчезващи или избухващи градиенти.**

* + 1. **Програмни приложения**
* **За разработката и поддръжката на проекта са използвани няколко съвременни инструмента:**
  + **Rider:  
    Отлично IDE за разработка на .NET приложения. Осигурява стабилна поддръжка за C# и ASP.NET, улеснявайки управлението на сложната бизнес логика.**
  + **PyCharm:  
    Популярен IDE за разработка с Python, който е използван за разработка и оптимизация на AI модела и Flask endpoint-ите.**
  + **Visual Studio Code:  
    Лек и гъвкав редактор, използван за работа с Angular и TypeScript, както и за обща редакция на кода. Поддържа множество разширения, които помагат при разработката на уеб приложения.**

**3.5.3 Технологични средства**

* **Проектът използва разнообразни технологични средства, които осигуряват цялостната функционалност и мащабируемост:**
  + **ASP.NET, C#, EF Core:  
    Осигуряват стабилна и мащабируема среда за изграждане на API услугите и бизнес логиката. EF Core позволява лесно управление и достъп до базите данни.**
  + **Angular, TypeScript, Node.js:  
    Създават модерен, интерактивен фронтенд, който комуникира чрез RESTful API с Back-End услугите. Node.js се използва за управлението на зависимостите и стартиране на процесите по разработка.**
  + **Flask, Python:  
    Основата за AI модула, който обработва входните данни, извършва сложни математически изчисления и връща резултатите от класификацията.**
  + **HTML, CSS, Bootstrap:  
    Използват се за оформяне и стилизиране на потребителския интерфейс, осигурявайки отзивчив и модерен дизайн.**

**3.6. Описание на приложението**

**За да се стартира приложението, е необходимо да се изпълнят следните стъпки:**

* **DataWise.AI (Python):  
  Отидете в папката DataWise.AI и изпълнете:**
  + **python app.py – за стартиране на Python endpoint-а, който обслужва заявките към AI модела.**
  + **Ако желаете да обучите собствен модел, използвайте:  
    python train.py**
* **DataWise.Api (.NET):  
  Отидете в папката DataWise.Api и стартирайте .NET API проекта. Обикновено това се извършва чрез командата:**
  + **dotnet run  
    (Уверете се, че имате инсталирана подходящата версия на .NET и че проектът е конфигуриран правилно.)**
* **DataWise.Client (Angular):  
  Отидете в папката DataWise.Client и изпълнете:**
  + **npm start  
    Тази команда стартира Angular клиента, който предоставя потребителския интерфейс на приложението.**

**3.7. Заключение**

**DataWise представлява цялостно и иновативно решение за подготовка на кандидати за технически интервюта, предоставяйки модерни AI технологии и образователни ресурси за усвояване на структурите от данни и алгоритмите. AI моделът, базиран на сложни математически осигурява решение под формата на категория алгоритъм/стратегия, която трябва да се ползва за решаване на задачата.**

**Приложението е насочено към кандидати за програмиране – както студенти, така и млади професионалисти – които се подготвят за интервюта във водещите технологични компании (например FANG). DataWise подпомага потребителите да затвърдят своите знания по структури от данни и алгоритми и да се справят успешно с техническите предизвикателства на интервютата.**

**В заключение, DataWise демонстрира висок технически потенциал и практическа стойност за целевата аудитория. Чрез съчетаването на сложни математически алгоритми и интеграцията на различни технологични екосистеми, проектът осигурява мащабируемост, надеждност и гъвкавост – ключови фактори за успешното обучение и професионално развитие на бъдещите кандидати за технически интервюта.**