**Извличане на код по ICD-10 с помощта на подходи за обработка на естествен език**

Автор: Мартин Попов, ФН: 6MI3400517

**1. Избор на задача (Въведение)**

Медицински специалисти и изследователи често разчитат на ICD-10 кодовете за диагностициране и класифициране на заболявания. Въпреки това, ръчното идентифициране на правилния ICD-10 код въз основа на симптомите на пациента, или описанието на заболяването, може да бъде предизвикателство. Традиционните подходи включват търсения, базирани на ключови думи, или ръчни търсения, които са неефективни и податливи на човешка грешка.

През последните години бяха проучени различни методи за автоматизиране на процеса, като системи, базирани на правила и модели за машинно обучение. Те обаче често изискват обширни етикетирани данни или сложни модели, които може да не са практични в клинични приложения в реалния свят.

Избраният подход има за цел да подобри извличането на ICD-10 кодове чрез използване на техники за обработка на естествен език (NLP) и търсене с Elasticsearch. Фокусира се върху оптимизиране на точността на търсенето и скоростта на извличане, без да разчита на сложни модели за дълбоко обучение, което го прави леко, но ефективно решение.

**2. Избор на корпус и допълнителни данни (Данни)**

Изборът на данни е една от най-сложните стъпки в този проект, понеже има няколко варианта:

**2.1 Използване на ICD-10-CM кодове (клинична модификация)**

Този вариант съдържа ~72 хиляди различни кода. Има dataset с всички заболявания - <https://github.com/k4m1113/ICD-10-CSV/blob/master/codes.csv>, но е твърде подробен и разширен, което доведе до по-лоши резултати. Затова този вариант беше оставен като евентуално бъдещо разширение

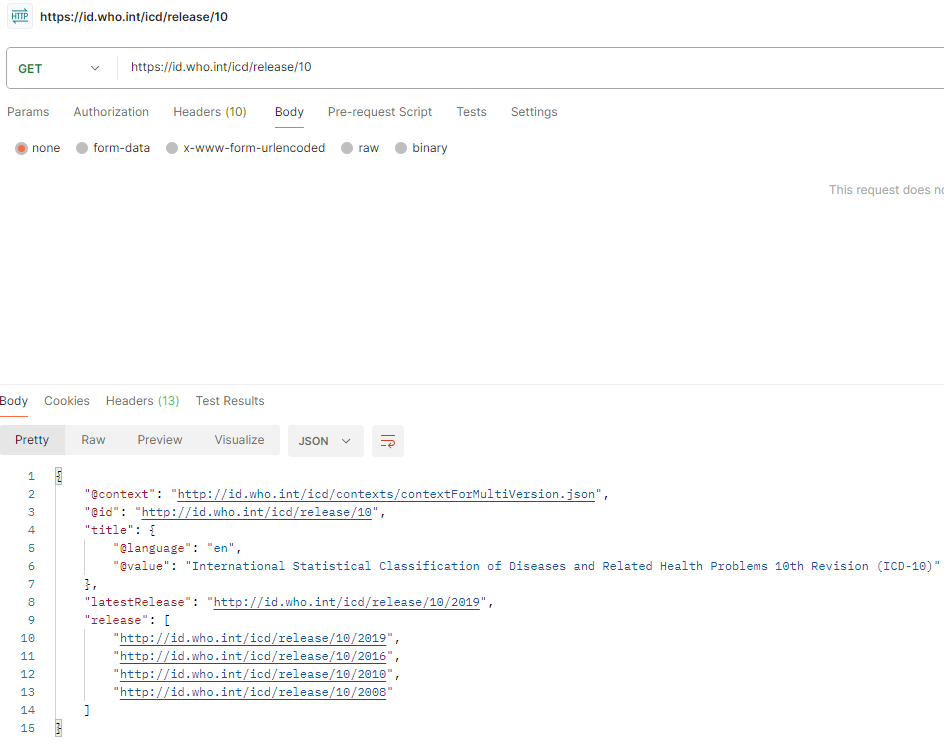
**2.2 Използване на ICD-10 API предоставен от WHO (СЗО)**

Тук използвахме предоставения от СЗО приложно-програмен интерфейс –

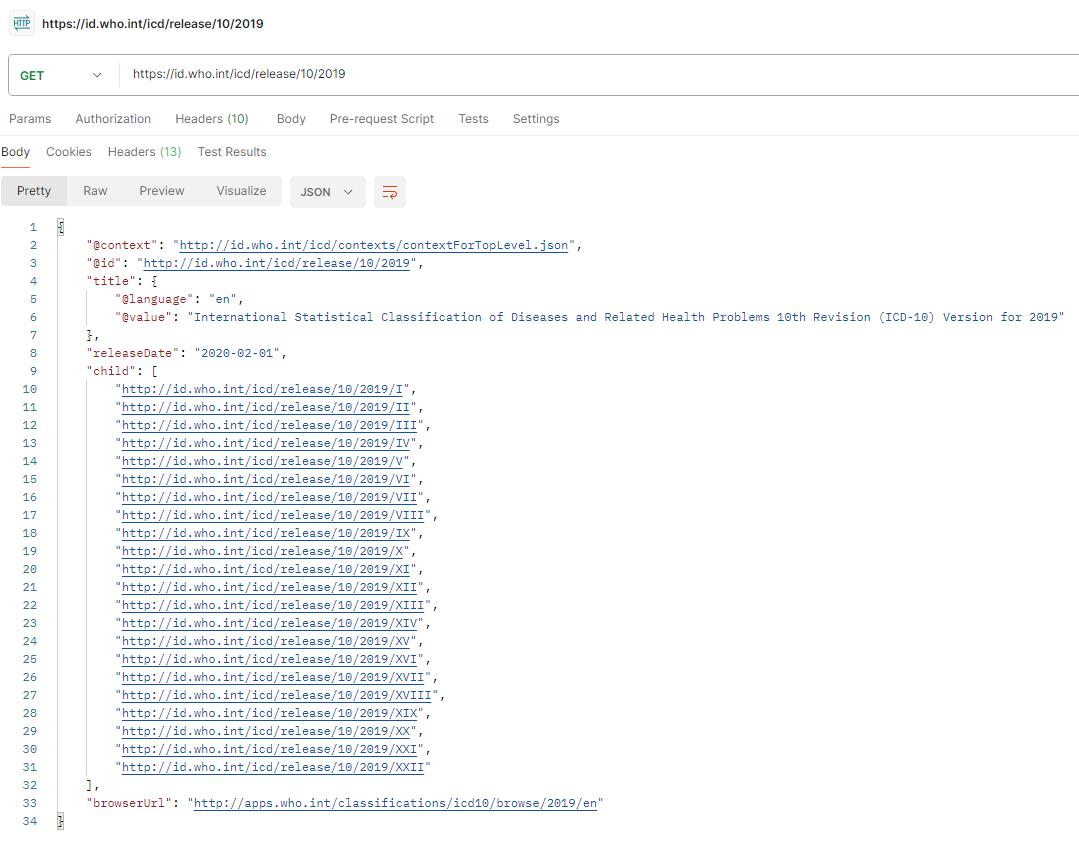
<https://icd.who.int/icdapi>

Събирането на данните протече в следните стъпки:

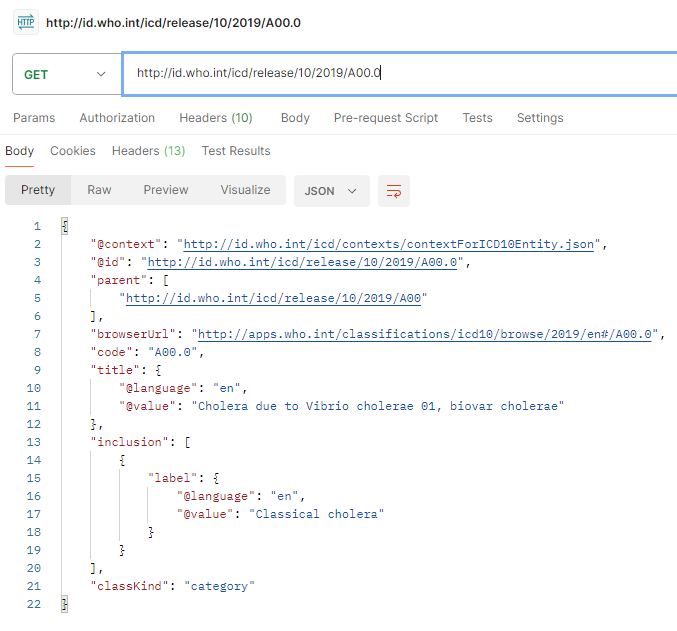
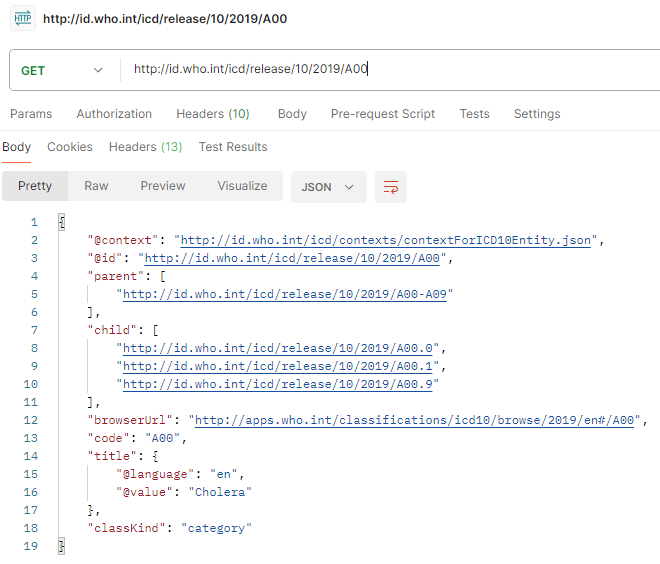
1. Намиране на последния release (2019)



1. Запитване към последния release, запознаване със структурата на резултатите от API-a (всеки резултат има поле “child” – масив от други валидни линкове, към което може да се правят заявки)



1. Заявки към децата докато полето не стане празно или просто го няма



1. Агрегиране на информацията за да получим накрая пълни данни за един код

Имаме 4 вида ресурси – chapter, block, category, subcategory. Всеки има различно име:

Например:

Chapter I - Certain infectious and parasitic diseases

Block A00-A09 - Intestinal infectious diseases

Category A00 – Cholera

Subcategory A00.0 - Cholera due to Vibrio cholerae 01, biovar cholerae

Понеже търсенето е top-down (chapter 🡪 block 🡪 category 🡪 subcategory) за най-ниското ниво нямаме информация в резултата към кой chapter или block принадлежи. Затова, по време на scraping-а агрегираме данните, така че накрая да получим пълна информация за даден ICD10 код.

Набяването на симптомите за различни заболявания се оказа по-сложна задача, понеже доста заболявания имат сходни симптоми, а и никъде особено няма dataset, който директно да свързва заболяване със симптоми. Бяха намерени няколко корпуса с данни, които имаха симптоми за заболявания, но бяха представени в различни видове:

* име на заболяване + симптоми – беше нужно запитване към частта на проекта за дисциплината ИИОЗ, където по описание/име на заболяване вземаме icd10 код
* код на заболяване по Disease Ontology (DOID code) + симптоми – вклюваше създаване на mapping между DOID и ICD10 и навръзване на резултатите

Агрегираните данни излеждат по следния начин:

{

    "icd10code": "A00.0",

    "type": "subcategory",

    "title": "Cholera due to Vibrio cholerae 01, biovar cholerae",

    "chapter": "Certain infectious and parasitic diseases",

    "chapterCode": "I",

    "block": "Intestinal infectious diseases",

    "blockCode": "A00-A09",

    "category": "Cholera",

    "categoryCode": "A00",

    "subcategory": "A00.0",

    "inclusion": [

        "diseases generally recognized as communicable or transmissible",

        "Classical cholera"

    ],

    "exclusion": [

        "carrier or suspected carrier of infectious disease",

        "certain localized infections - see body system-related chapters",

        "infectious and parasitic diseases complicating pregnancy, childbirth and the puerperium [except obstetrical tetanus]",

        "infectious and parasitic diseases specific to the perinatal period [except tetanus neonatorum, whooping cough, congenital syphilis, perinatal gonococcal infection and perinatal human immunodeficiency virus [HIV] disease]",

        "influenza and other acute respiratory infections"

    ],

    "symptoms": [

        "abdominal pain",

        "severe dehydration",

        "fever",

        "vomiting",

        "constipation"

    ]

}

Данните от СЗО имат и полета с това кои неща се включват и кои не се във въпросния ICD10 код, което роля при нагласянето на заявката за търсене (например кое поле с каква тежест да се анализира)

За някои кодове няма подкатегория и затова спира до категория само, те изгелждат по следния начин:

{

    "icd10code": "A33",

    "type": "category",

    "title": "Tetanus neonatorum",

    "chapter": "Certain infectious and parasitic diseases",

    "chapterCode": "I",

    "block": "Other bacterial diseases",

    "blockCode": "A30-A49",

    "category": "Tetanus neonatorum",

    "categoryCode": "A33",

    "subcategory": "",

    "inclusion": [

        "diseases generally recognized as communicable or transmissible"

    ],

    "exclusion": [

        "carrier or suspected carrier of infectious disease",

        "certain localized infections - see body system-related chapters",

        "infectious and parasitic diseases complicating pregnancy, childbirth and the puerperium [except obstetrical tetanus]",

        "infectious and parasitic diseases specific to the perinatal period [except tetanus neonatorum, whooping cough, congenital syphilis, perinatal gonococcal infection and perinatal human immunodeficiency virus [HIV] disease]",

        "influenza and other acute respiratory infections"

    ],

    "symptoms": [

        "poor feeding",

        "irritability",

        "muscle stiffness",

        "spasms",

        "difficulty swallowing"

    ]

}

Някои от използваните datasets:  
<https://www.kaggle.com/datasets/dhivyeshrk/diseases-and-symptoms-dataset/data>

<https://www.kaggle.com/datasets/phoenion/diseases-and-their-symptoms>

<https://www.kaggle.com/datasets/choongqianzheng/disease-and-symptoms-dataset>

Disease ontology datasets със симптоми + mapping:

<https://github.com/dhimmel/hsdn/blob/gh-pages/data/symptoms-DO.tsv>

<https://github.com/DiseaseOntology/HumanDiseaseOntology/blob/main/DOreports/allXREFinDO.tsv>

Този набор от данни служи като основа за индексиране на информация за болестта в Elasticsearch, което позволява бързо и точно запитване.

**3. Избор на подход (Метод)**

Избраният подход е индексиране на данните в Elasticsearch заедно с приложени подходи за обработка на естествен език.

Самата обработка се случва преди подаването на заявката, като се прилагат следните методи:

* **нормализация:** преобразуваме текста в малки букви и премахва ненужните знаци
* **токенизация:** разделяме въведения текст на отделни думи или фрази
* **премахване на стоп думи**
* **лематизация:** намаляваме думите до тяхната основна или речникова форма, за да осигурим последователност

Това ни позволява да получим само думи, които са релевантни в контекста на търсенето.

Пример: Lately, I've been feeling nauseous and have had some stomach cramps. My head is throbbing with pain, and I'm also experiencing shortness of breath and dizziness. I feel exhausted and have trouble sleeping at night.

Резултат: lately feeling nauseous stomach cramp head throbbing pain also experiencing shortness breath dizziness feel exhausted trouble sleeping night

Пример: I have a sore throat, runny nose, sneezing, and mild fever. My body feels a bit tired, but it's not too severe.

Резултат: sore throat runny nose sneezing mild fever body feel bit tired severe

**4. Експерименти върху данните (Експерименти)**

Бяха проведени експерименти, за да се оцени колко добре системата извлича кодове на ICD-10 въз основа на описания на симптомите в реалния свят. Оценката включваше:

* Запитване до системата с различни описания на симптоми на пациенти.
* Измерване на уместността на извлечените ICD-10 кодове.
* Сравняване на резултатите от търсенето преди и след прилагане на NLP техники.

Използваните тестови данни бяха ръчно направени, като броят им е около 10-тина.

**5. Валидиране, анализ на грешки и дискусия (Резултати и дискусия)**

За да потвърдим резултатите, анализираме:

* **Прецизност и припомняне:** оценяване на това колко извлечени кодове по ICD-10 са уместни в сравнение с това колко уместни са пропуснати
* **Анализ на грешката:** проучване на неправилни или по-малко подходящи съвпадения за идентифициране на потенциални подобрения

Потенциални грешки могат да възникнат от двусмислени описания на симптомите, медицински синоними или липсващи връзки между симптоми и заболявания.

Това е може би най-големият проблем при тази задача. Фината настройка на параметрите на Elasticsearch, разширяването на синоними или интегрирането на допълнителни NLP техники може да подобри точността. (например групиране на думите mild fever като един симптом)

**6. Заключение и бъдеща работа**

Проектът демонстрира ефективността на използването на предварителна обработка на естествен език и Elasticsearch за извличане на код на ICD-10. Поради естеството на данните и невъзможност за строго определение на състоянието, той има по-скоро ориентировъчна цел, а не поставяне на диагноза.

**Бъдещи подобрения:**

* **разширение на синоними**
* **интегриране на RAG + LLM модел**
* **разширение с ICD10-CM**

**Код:** [**https://github.com/kachamaka/icd10**](https://github.com/kachamaka/icd10)

**Линк към проекта:** [**https://icd10link.netlify.app/search-by-symptoms**](https://icd10link.netlify.app/search-by-symptoms)