**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 кодове на ICD-10 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)**

Наборът от данни, използван за този проект, включва:

* **Набори от данни за симптоми от Kaggle**: Съдържа описания на симптомите, използвани за обработка на въвеждане от потребителя.
* **Данни за ICD-10 от API на СЗО**: Предоставя официални кодове на заболявания и описания, използвани за индексиране в Elasticsearch.

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)**

**а. Основни лечения на компютърна лингвистика.**

За да подобрим точността на търсенето, прилагаме следните техники за предварителна обработка:

* **Нормализация:** преобразуваме текста в малки букви и премахва ненужните знаци
* **Токенизация:** разделяме въведения текст на отделни думи или фрази
* **Лематизация:** намаляваме думите до тяхната основна или речникова форма, за да осигурим последователност

**b. Допълнителни алгоритми**

Не са използвани допълнителни алгоритми (като класификатори за машинно обучение или модели за дълбоко обучение). Вместо това проектът разчита на възможностите за търсене на Elasticsearch, комбинирани с NLP-базирана обработка на текст.

**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 НЛП техники.

**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. Фината настройка на параметрите на Elasticsearch, разширяването на синоними или интегрирането на допълнителни NLP техники може да подобри точността.

**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. Настоящият подход предоставя лек, ефективен метод за търсене на медицински кодове въз основа на описания на симптомите.

**Бъдещи подобрения:**

* **разширение на синоними**
* **размито съвпадение** (търсения, базирани на сходство)
* **интегриране на RAG модел**

**Линк към проекта:** [**https://github.com/kachamaka/icd10**](https://github.com/kachamaka/icd10)