Амбулаторните листове се пишат докато се лекува пациента – при посещение при лекар.

Има голям брои документи от здравната система, повечето електронни писани от медицински персонал. Съдържат диагнози, оплаквания, лечение, резултати от лечението, рискови фактори, фамилна история. В тях се съдържа лична информация. Обикновено по-голямата част от документите не се ползва.

От големия брои данни може да се добият данни.

За да се направи текст mining на клинични документи е трудно понеже има много понятия на областта както и различни здравни дисциплини използващи различни понятия, нужни са знания за domain-a

Може да се използва за :

* Откриване и прогнозиране на инфекции
* Намиране на страничните ефекти на лекарства
* Откриване на ранни симптоми на рак и други заболявания
* Препоръчване на лекарства
* Изследване на резултатите от лечението с дадено лекрство

Съдържат правописни грешки, абревиатури, отрицания, съкращения на думи

rule-based and machine learning-based methods such as unsupervised and supervised methods

Active learning is described that is used for selecting appropriate training material for the machine learning algorithms.

Данните се увеличават експоненциално

//Да сложа пример за анамнеза

According to Pakhomov et al. (2005) there are 30% non-word tokens, abbreviations, acronyms, misspellings, wrongly used grammar etc. in clinical text, which is a good indication of the noisiness of clinical text.

Generally patient records are written by highly skilled physicians and nurses using domain specific terms. For example, patient record text is very domain specific depending on which medical discipline it is written in. Each discipline or domain within medicine uses its own set of terms that can be incomprehensible by other disciplines.

Отрицанията са често използвани за да може да се отхвърли някое заболяване

A Bulgarian clinical corpus containing several hundred thousand patient records from the specialties general practice, endocrinology, metabolic disorders cardiology, ophthalmology, gastroenterology, pneumology and physical therapy used for text mining and big data analytics (Boytcheva et al. 2015). • Another Bulgarian clinical corpus containing 500,000 pseudonymised outpatient records on diabetes (Boytcheva et al. 2017b).

Patient record text is different from standard text. Patient records contain plenty of misspellings (up to 10%) and domain specific abbreviations (up to 10%) and acronyms (up to 5%). Patient records also contain incomplete sentences, often the subject or patient is missing in the sentence. In the assessment field there are many negations (up to 10%) since the physician tries to exclude symptoms while reasoning to find the disorder of the patient. In addition to the negations, the content contains vague or uncertain expressions (up to 12%) regarding the factuality of the findings and disorders. Patient record text is written by different professions and also varies between different medical specialties. Discharge summaries are often wellwritten and structured, since they are written for a broader audience than the personnel at the clinical unit. Most available clinical corpora for research are in English; however, there are some corpora in other languages available for research.

Стигнала съм до 6-та глава