**Кейс**

1. **Участники команды: Долгушин Михаил, Исмакова Даяна**
2. Общее описание
3. Токенизировать на отдельные слова в сообщении, с частью речи, с подписью имя ли это собственное, bool.
4. Используя **словарь мед. Симтомов**, найти есть ли они в токенизированных сообщениях, отметить.
5. Используя **словарь мед. Названий**, найти содержатся ли они в разбитых на токены сообщениях, отметить. (токен Кардиомагнил, в нем содержится слово кардио, к примеру, явно отсылающее к медицине. Ну или сразу мед. Препаратов распространенные названия в этом словаре будут).
6. **Word2vec** на слова, “Лекарство”, “Препарат”, “Вызывать”, “Симптомы” из них тоже составить словарь и проставить метрику сообщениям.
7. Токенизировать сообщения с ненулевыми значениями по параметрам в предыдущих пунктов на биграммы, ..., вплоть до пентаграммы, проставить метрику tf(bool).
8. Пример признаков токена... документа.

|  |  |  |  |
| --- | --- | --- | --- |
| Док | Есть симптом | Есть мед названия | Word2vec словарь |
| 1 | 0 | 1 | 1 |
| 2 | 1 | 0 | 0 |
| 3 | 1 | 1 | 1 |

|  |  |  |  |
| --- | --- | --- | --- |
| Док\Биграммы | Биграммы 1 | Биграммы 2 | 3 |
| Док 1 | 1 | 1 | 1 |
| Док 2 | 0 | 1 | 0 |
| Док 3 | 1 | 1 | 0 |

...

|  |  |  |  |
| --- | --- | --- | --- |
| Док\Пентаграммы | Пентаграммы 1 | Пентаграммы 2 | Пентаграммы 3 |
| Док 1 | 0 | 0 | 1 |
| Док 2 | 0 | 1 | 0 |
| Док 3 | 0 | 1 | 1 |

1. Вручную разметили обучающую выборку текстов и тестовую (3000 текстов в сумме, обучающих 2500, тестовых 500), отобранные при анализе словарями на ADR 0 или 1.

Обучаем классификатор наивного байеса (NLTK.NaiveBayes...) для векторной модели биграммной, затем триграммной и тд. Вплоть до пентаграммной.

Алгоритм наивного Байеса:

(Классы C: ADR или unADR)

Для каждого уникального токена **w** из «тренировочного» набора вычисляется столько значений P(w, c), сколько классов **c** нас интересует.

Далее, в «неизвестном» тексте каждому слову назначаются вычисленные вероятности и вычисляется величина:

P(c|d)=P(c) P(w1|c) P(w2|c) … P(wn|c)

где P(c|d) – вероятность принадлежности документа **d** классу **с**,

n – количество уникальных слов в документе **d**.

Получим 4 классификатора. Результат - если среднее вероятностное значение данных классификаторов > 50, тогда у нас класс ADR.

В сообщениях, которые таки ADR, если верить классификаторам, находим симптомы по словарям симптомов и имена собственные лекарств для вывода результата. И составляем векторную модель: лекарство, тип побочного эффекта, Доля случаев от общего числа упоминаний о приеме лекарства. Это и будет результат.

1. Общая архитектура приложения

