Автоматическое выделение симптомов заболеваний из текстов жалоб пациентов.

Обзор литературы

http://www.frccsc.ru/sites/default/files/docs/events/med/Deviatkin.pdf?386

- По тексту сгенерировать большое количество различных вариантов терминов.
- Осуществить нежесткое сравнение сгенерированных вариантов с терминами из медицинского тезауруса, оценить их сходство.
- Отранжировать варианты по оценке сходства
- Выбрать варианты, наиболее похожие на термины в тезаурусе.

Медицинский тезаурус:

UMLS Метатезаурус + UMLS Семантическая сеть

- Метатезаурус сопоставляет концепты различных других медицинских тезаурусов с единым кодом (уникальным идентификатором концепта, УИК)
- Объединяет в себе MeSH, SNOMED-CT, ICD-10, и др.
- Единственный ресурс на русском – MeSHRUS ~ 27 тыс. Концептов; 85 тыс. Терминов (доступ затруднен)
- Семантическая сеть сопоставляет каждому УИК семантический тип: болезнь, симптом, микроорганизм, хим. вещество, и др.
- Этот ресурс использовался для извлечения из текстов упоминаний болезней, симптомов и частей тела

Обзор литературы

http://www.aclweb.org/anthology/W14-3416

Использование системы паттернов:

Annotating the first MEDLINE corpus of Abstracts with HPO provided us with a corpus of 10,000 annotated sentences. The 13,477 annotated units were replaced by a keyword –SYMPTOM– in order to facilitate the discovery of patterns. Results produced 988 patterns, among which 326 contained the keyword symptom. Based on these patterns, several remarks can already be made:

Several annotated signs or symptoms are regularly associated with a third term, which can be another sign or symptom: {symptom}{symptom}{and}{stress};

Обзор литературы

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5944189/pdf/hir-24-148.pdf

Извлечение на основе правил.

Обзор корпусов

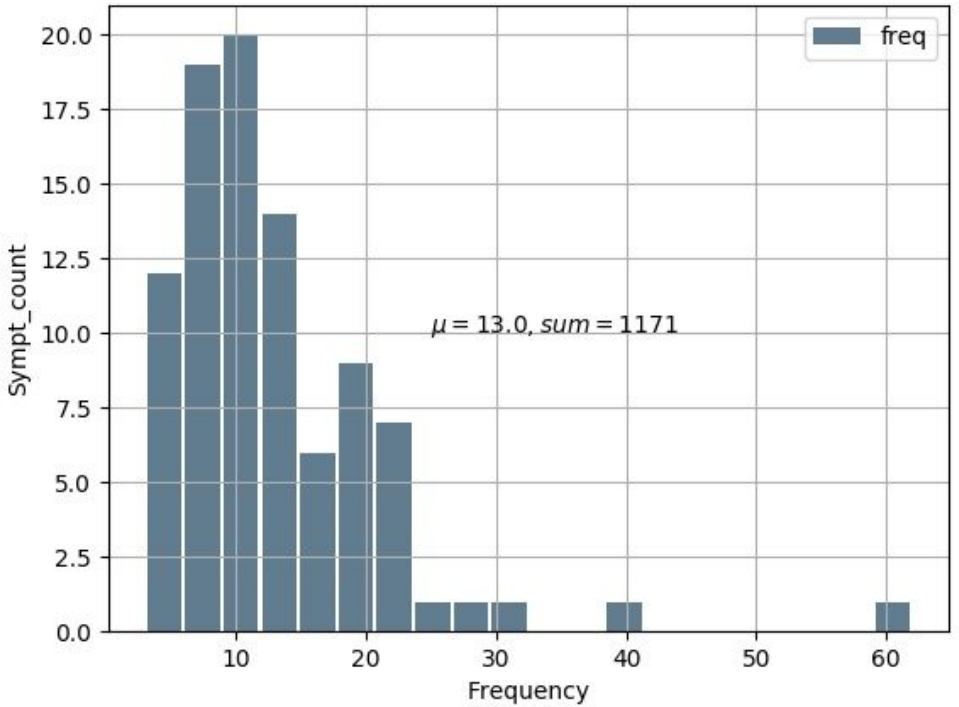
Удалось получить доступ к следующему корпусу:
http://nlp.isa.ru/index.php/component/portal/?view=corpusclinical
http://nlp.isa.ru/corpora/clinical/guidelines_v.1.6.1.pdf

Однако симптомы в данном корпусе составлены врачами. Таким образом их сложно использовать для парсинга описаний состояний пациентов.

Частоты симптомов

Гистограмма распределения симптомов по их частоте среди описаний.

В данных есть дисбаланс классов, если мы будем решать эту задачу через классификацию.



Частоты симптомов

Примеры 5 наиболее редких и частотных симптомов.

sympt	freq
невозможность движения в руках	3
боль в спине во время беременности	4
возраст от 50 до 60 лет	4
односторонняя боль	4
температура 38 градусов и больше	24
насморк	27
боль в груди	30
повышение температуры	39
головная боль	62

Условные группы симптомов

В представленных данных был файл, в котором симптомы сгруппированы по условным болезням. Если бы мы решали задачу классификации, то мы могли бы обучить классификатор типов болезни и классификаторы симптомов к каждому типу, повысив тем самым точность решения. Однако к одному описанию могут относится симптомы, которые принадлежат разным условным болезням. Скорее всего, это нормальная ситуация.

Частотность условных болезней:

Условная болезнь	частота
Сыпь и пятна на коже	85
Болевые ощущения в спине	82
Болевые ощущения в грудной клетке	72
Головная боль у взрослых	62
Заложенность и выделения из носа	59
sum	360 из 446 описаний

Условные группы симптомов

Остальные 86 - это пересечения различных условных болезней в одном описании. Пара примеров:

Условная болезнь	частота
Головная боль у взрослых&Заложенность и выделения из носа	14
Болевые ощущения в грудной клетке&Головная боль у взрослых	3
Болевые ощущения в спине&Сыпь и пятна на коже	3

Нормализация текстов

1. Исправление опечаток YandexSpeller (https://pypi.org/project/pyaspeller/). Для работы необходим доступ к сети.

2. Токенизация с помощью WordPunctTokenizer из nltk.

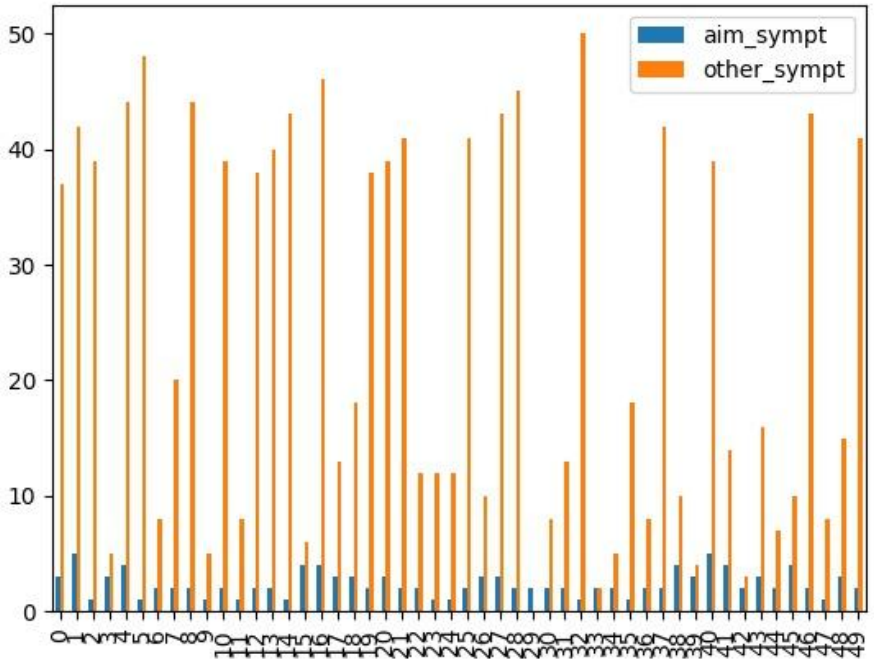
3. Лемматизация с помощью pymorphy2.

Данные: описания состояний пациентов

Попробуем пересечь нормализованные тексты состояний пациентов с нормализованными целевыми симптомами и остальными. Некий “брутфорс”.

Будем пересекать по униграммам, `2_ngrams`, `3_ngrams`, `4_ngrams`. Если у симптома и описание состояния есть хотя бы `1` общий элемент - считаем, что они пересекаются.

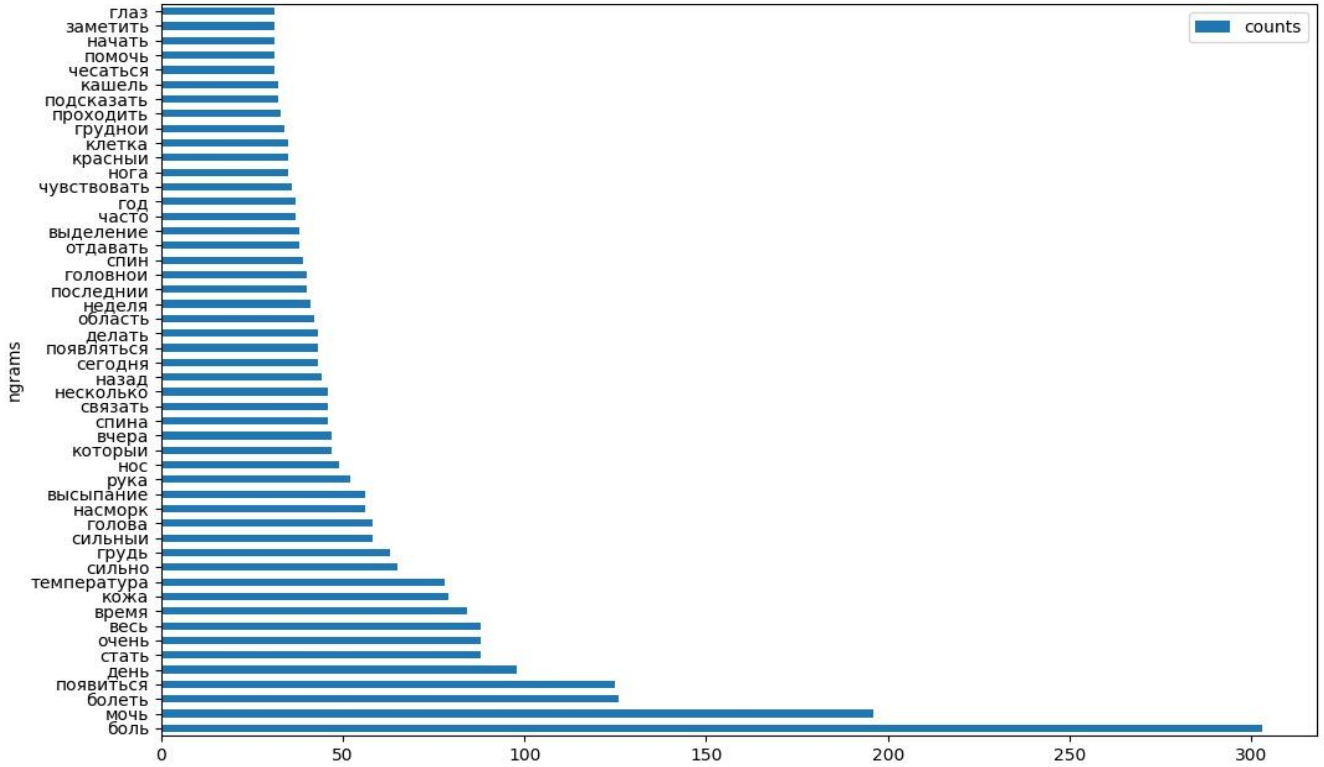
Возьмем `50` рандомных описаний. Пересечение по токенам. Мы видим, что каждое из описаний имеет большое количество пересечений с нецелевыми симптомами.



Данные: описание состояния пациентов

По всей видимости, это связано с высокой частотой некоторых токенов, которые присутствуют в описаниях и в большом количестве симптомов.

Попробуем повысить “точность” с помощью повышения порядка N-грамм.

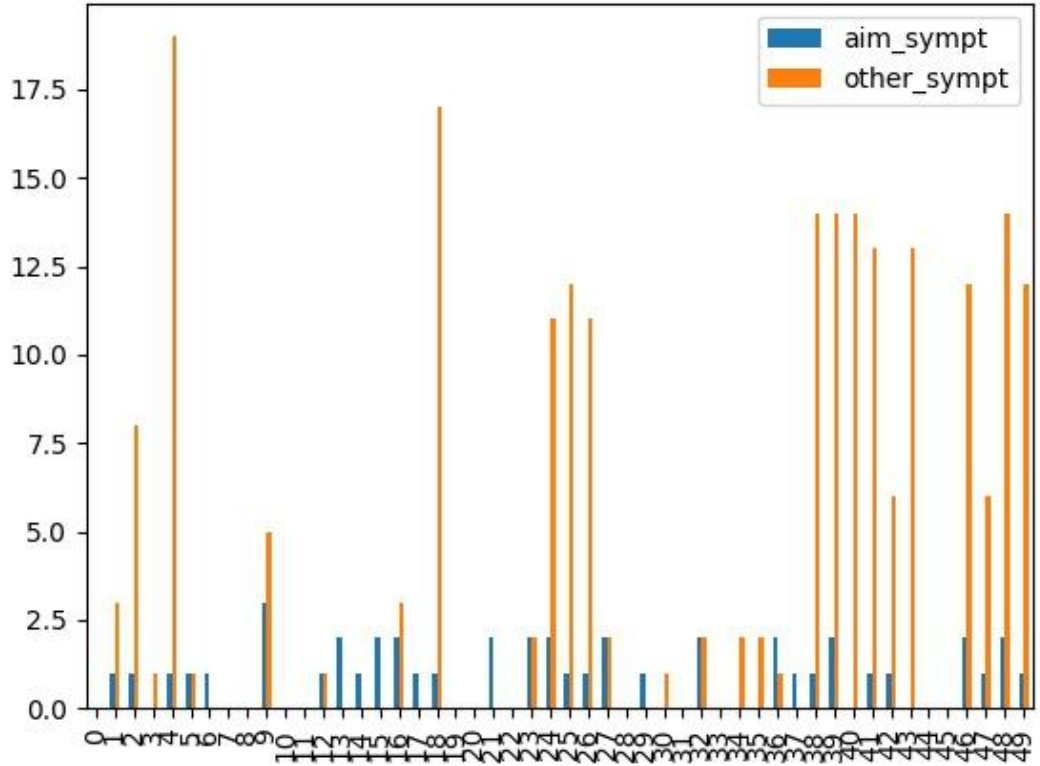


Данные: описание состояния пациентов

Возьмем 50 рандомных описаний, пересечения по 2_ngrams.

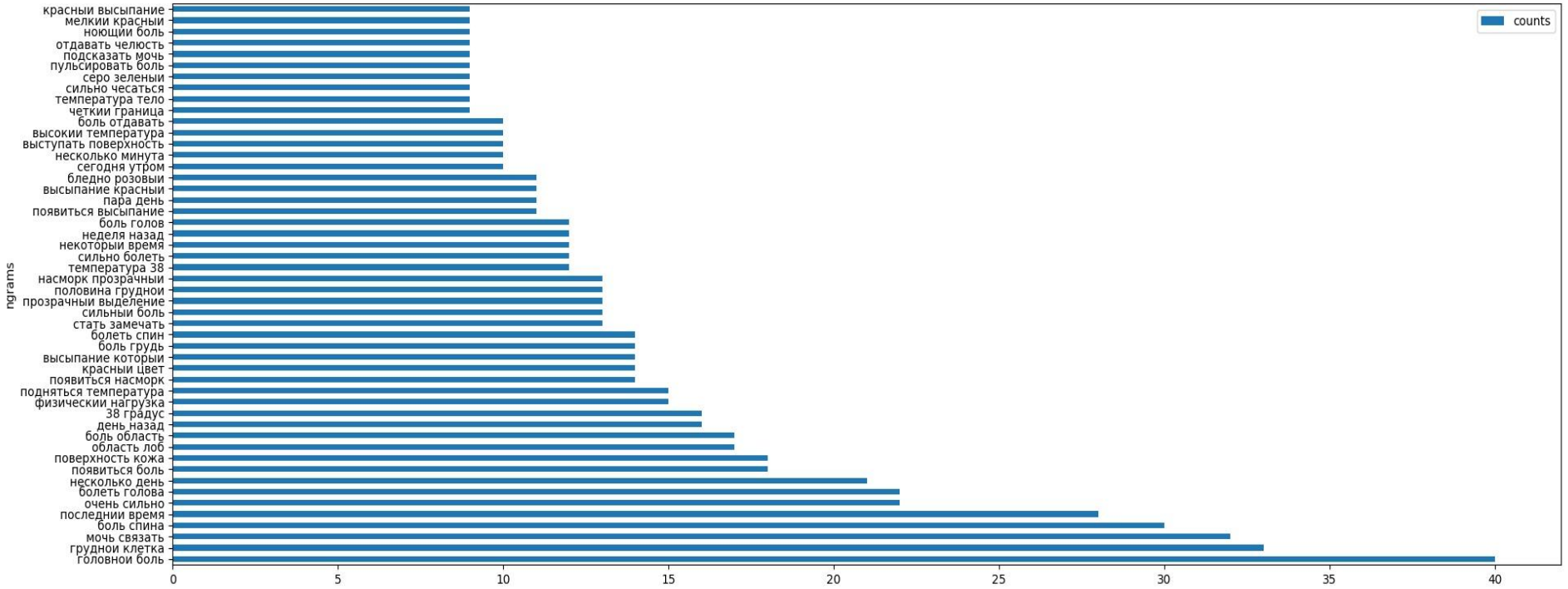
Количество пересечений с нецелевыми симптомами уменьшилось.

Появились описания без пересечений с симптомами.



Данные: описание состояния пациентов

Частоты `2_ngrams`.



Данные: описание состояния пациентов

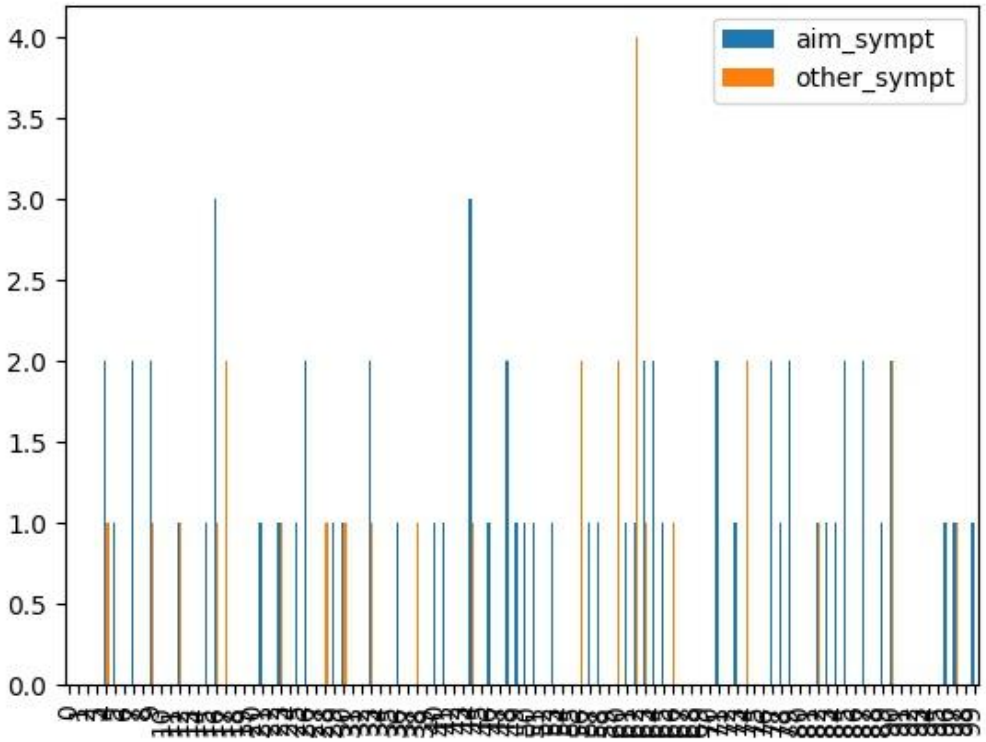
Возьмем 100 рандомных описаний, пересечения по 3_ngrams.

Количество пересечений с нецелевыми симптомами уменьшилось.

Появились описания без пересечений с симптомами.

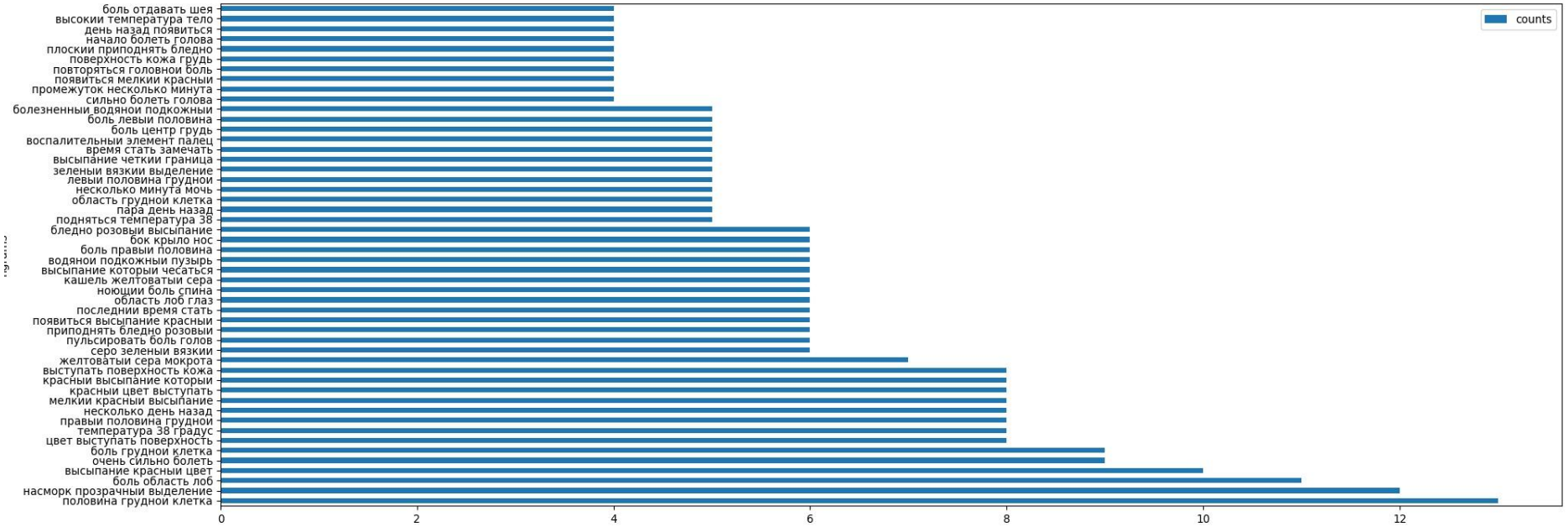
По всей видимости существуют некоторые последовательности токенов, которые почти целиком есть в симптомах и описаниях.

Учитывая распределение частот 3_ngrams, возможно, можно говорить о том, что некоторые симптомы вводились на основании больших частей описаний.



Данные: описание состояния пациентов

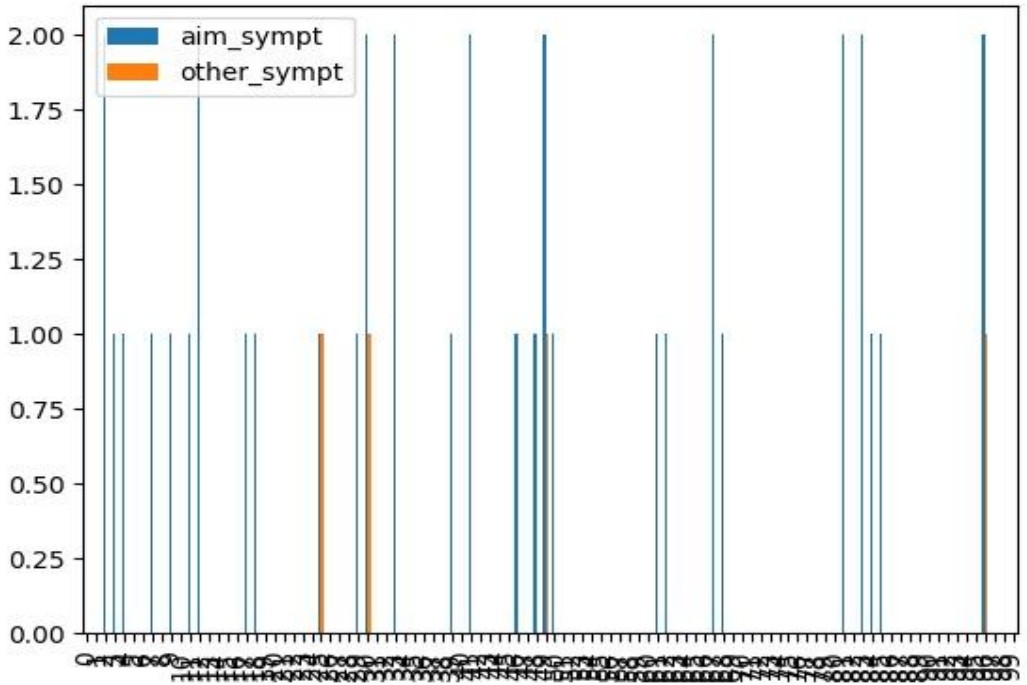
Частоты `3_ngrams`.



Данные: описание состояния пациентов

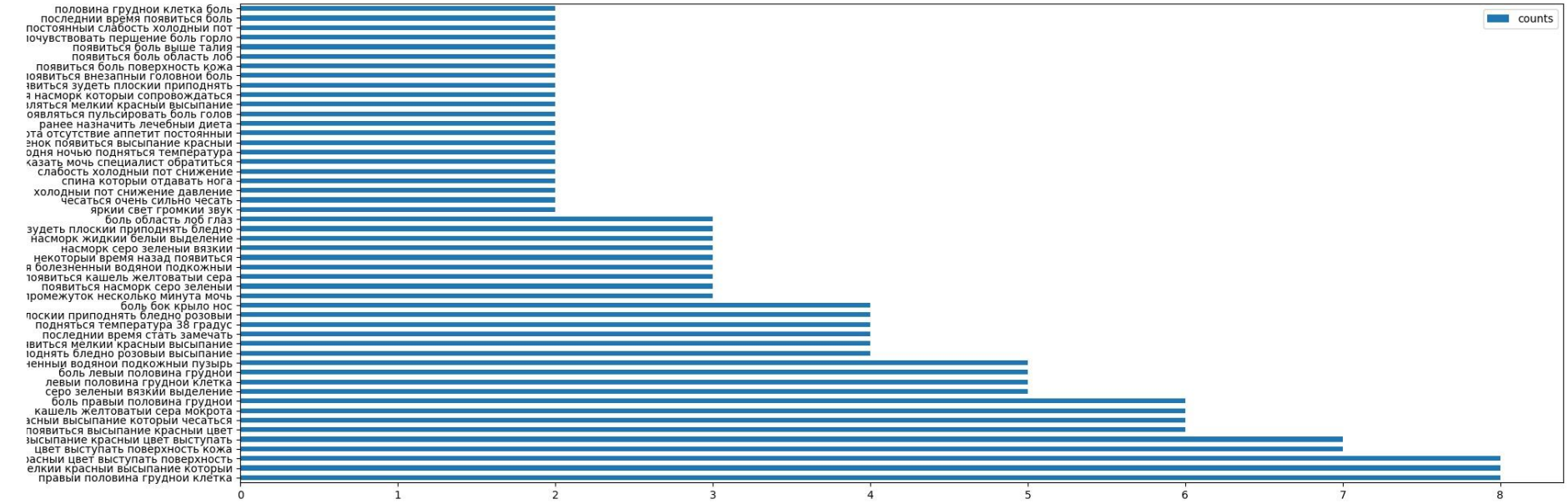
Возьмем 100 рандомных описаний, пересечения по 4_ngrams.

Данные закономерности о пересечениях нграмм порядка 3 и 4 можно будет использовать для создания признаков для моделей классификации или элементов rule-based подхода.



Данные: описание состояния пациентов

Частоты `4_ngrams`.



Реализованные подходы: rule-based

Первоначально попробуем rule-based подход для того, чтобы посмотреть данные более детально, найти возможные признаки для машинного обучения, выявить противоречия и ошибки в разметке.

Этапы выделения симптомов.

1. Составление соответствующего конфига.
2. Нормализация входного текста описания состояния пациента.
3. Генерация ngrams от 2 до максимальной длины симптома из исходного текста описания.
4. Проверка: если среди сгенерированных ngram такие, в которых есть слова, из указанных в файле конфига ключей для keywords симптома. То есть, исходя из части конфига для симптома "боль в груди", данная проверка будет положительна, есть будет биграмма, у которой лемма первого слова *грудь,* а лемма второго *болеть*.
5. Проверка: если среди сгенерированных ngram такие, в которых есть слова, из указанных в файле конфига ключей для delimiters симптома. Данная проверка будет положительна, есть будет ngram, в которой есть леммы: *центр, грудь, болеть*.
6. Есть п.4 == True, а п.5 == False - тогда записываем симптом.

Реализованные подходы: rule-based

Пример конфига для выделения симптома *боль в груди:*

```
"symptoms": [
        "боль в груди"
    ],
 "keywords": {
   "object": ["грудь", "сердце", "клетка", "грудной"],
   "feel": ["боль", "болеть", "болевой", "покалывание", "гореть", "уронить", "посинелый", "заколоть",
     "давить"],
   "place": [],
   "operators": []
 },
 "delimiters": {
   "object": ["центр", "грудь", "клетка", "грудной", "половина", "сторона"],
   "feel": ["боль", "болеть", "приблизительно"],
   "place": ["левый", "центр", "правый", "кожа"],
   "operators": []
 }
```

Реализованные подходы: rule-based

Результаты

1. В качестве метрики качества использовали `Jaccard similarity coefficient`.
2. Реализовали правила для `15` симптомов.
3. Значение метрики `0.234636` на всех данных.
4. Выявили большое количество ошибок в разметке.

Особенности использования симптомов

Особенности использования симптомов.

В разметке целевых классов наблюдаются некие нюансы, которые приведут к снижению качества используемых подходов.

1. Использование синонимичных симптомов. Синонимичные симптомы не всегда используются вместе.
2. Лишний симптом.
3. Пропуск симптома.

Особенности использования симптомов

Особенности использования групп симптомов.

Можно выделить группы синонимичных симптомов. Например: *"боль в грудной клетке", "боль сконцентрирована в области грудины", "боль в груди"*. Иногда синонимичные симптомы не всегда используется все вместе при похожем описании.

Примеры:

"Около двух недель назад начало **болеть в области грудины**. Началось после сильного волнения, ночью сводило сердце как периодами, потянет и отпустит как нерв. Что это может быть?", [**"боль сконцентрирована в области грудины", "боль в грудной клетке"**]],

"Хожу в спортзал около двух лет. Недавно начало болеть в **области грудной клетки**. То ли она сама, а то ли за ней. Особенно часто, когда делааю жим лёжа или качаю пресс. Подскажите, что мне предпринять?",
[**"боль в грудной клетке", "боль сконцентрирована в области грудины", "боль в груди"**]],

"Вчера на вылазке был с друзьями и очень устал, был практически истощен и к вечеру, когда лег отдыхать, начало **давить в груди.** Я только месяц назад сдавал кучу анализов все в норме. Что со мной?",
["боль в положении лежа", "**боль в грудной клетке**", "**боль сконцентрирована в области грудины", "боль в груд**и"]]

Особенности использования симптомов

Особенности использования групп симптомов.

Можно выделить группы синонимичных симптомов. Например: *"боль в грудной клетке", "боль сконцентрирована в области грудины", "боль в груди"*. Иногда синонимичные симптомы не всегда используется все вместе при похожем описании.

2019-01-04 18:51:07,194 element: DataObject(description='Решил бросить курить и заняться спортом с женой. После таких занятий возникают **болевые ощущения в грудной клетке**, боль также отдает в челюсть. Жена мне в таких случаях дает нитроглицерин - помогает. Может это быть связанно с изменением образа жизни?', symptoms=['**боль в груди**', 'боль отдает в челюсть', 'боль при значительной физической нагрузке'])

Особенности использования симптомов

Особенности использования групп симптомов.

Можно выделить группы синонимичных симптомов. Например: *"повышение температуры", "температура 38 градусов и больше", "высокая температура тела", “температура тела выше 38 градусов”*. Ситуация аналогична предыдущей группе симптомов.

2019-01-03 20:21:55,222 element: DataObject(description='Вчера пришла с работы, почувствовала какое-то недомогание, решила померить температуру, на градуснике было - **38,5 градусов**. Еще на работе очень часто чихала.', symptoms=['**температура 38 градусов и больше**', 'чихание', '**повышение температуры', 'температура тела выше 38 градусов**'])

2019-01-03 20:21:55,359 element: DataObject(description='Сильные головные боли. Температура **до 40 градусов**, рвота и отсутствие аппетита. Постоянная слабость, холодный пот и снижение давления.', symptoms=['головная боль', **'высокая температура тела', 'повышение температуры**', '**температура 38 градусов и больше**', 'рвота'])

2019-01-03 20:21:55,480 element: DataObject(description='Повышение температуры тела **до 38 градусов**, ближе к вечеру наступилпа лихорадка.', symptoms=['**температура 38 градусов и больше**', 'лихорадка', **'повышение температуры**'])

Особенности использования симптомов

Особенности использования групп симптомов.
Можно выделить группы синонимичных симптомов. Например: *"повышение температуры", "температура 38 градусов и больше", "высокая температура тела", "температура тела выше 38 градусов"*. Так как набор лексики схожий, данные группы мы будем путать между собой.

2019-01-03 20:31:00,814 element: DataObject(description='Бывает, болит голова, сопровождается **повышенной температурой тела**. Бывает это не часто, а когда болит, закладывает нос. Возможно, это воспалительный процесс.', symptoms=['заложен нос', '**высокая температура тела**', '**повышение температуры**'])

2019-01-03 20:54:08,525 element: DataObject(description='У мужа третий день **температура 39,3** держится, спит постоянно. Когда просыпается, говорит, что видит всё мутно. Не ест ничего, а если съест, то его рвет сразу же. От гриппа таблетки пили, не помогает. ТЧо это ещё может быть?', symptoms=['рвота', 'нарушение зрения', '**высокая температура тела**'])

2019-01-03 20:31:01,112 element: DataObject(description='Вечером **поднялась высокая температура тела**, закружилась голова, боль в голове такой силы, что отдает и пульсирует в висках.', symptoms=[**'высокая температура тела**', 'пульсирует в висках', 'головная боль', '**повышение температуры**'])

2019-01-03 20:31:01,255 element: DataObject(description='Во время просмотра телевизора появилась внезапная головная боль. Решила померить **температуру она оказалась очень высокой**. До этого жаловалась только на заложеность носа. Скажите, чем мне можно помочь?', symptoms=['головная боль', 'боль внезапная', '**высокая температура тела'**, 'заложен нос'])

Особенности использования симптомов

Особенности использования групп симптомов.

Некоторые синонимы близкие по смыслу, имеющие иерархические отношения, как в примере ниже, не всегда используются однозначно.

ГРУДЬ + МЕСТО vs ПРОСТО ГРУДЬ

2019-01-05 22:10:24,901 element: DataObject(description='Вчера я почувствовала **боль в левой половине грудной клетки.** Самочувствие было плохое, поэтому я измерила температуру - она была выше 38 градусов. Как боль в груди и температура связаны между собой?', symptoms=['температура тела выше 38 градусов', **'боль в левой половине грудной клетки', 'боль в груди**', 'температура 38 градусов и больше'])

2019-01-05 22:23:38,222 element: DataObject(description='Болит **в правой половине грудной клетки**. Боль жгучая, ноющая, но это боль не острая, не колющая. Часто возникает чувство отдышки.', symptoms=['**боль в правой половине грудной клетки**', 'одышка', 'жгучая боль', 'ноющая боль'])

Особенности использования симптомов

Особенности использования групп симптомов.

В данном случае не всегда ясно на основании чего используются те или иные симптомы.

НАСМОРК vs НАСМОРК + ВЫДЕЛЕНИЯ vs ЗАЛОЖЕН НОС

2019-01-06 00:53:45,775 element: DataObject(description='На работе включили кондиционер, а к вечеру появился **насморк**. **Выделения прозрачные**. Как вылечить?', symptoms=['**насморк', 'насморк с прозрачными выделениями'**])

2019-01-06 00:53:43,470 element: DataObject(description='На днях гуляли с друзьями, было довольно прохладно и я вполне могла заболеть. На следующий день ожидания оправдались: появился **насморк с жидкими белыми выделениями**. Может посоветуете что-нибудь?', symptoms=['**насморк с жидкими белыми выделениями**']) *насморк*?

2019-01-06 01:22:42,710 element: DataObject(description='Головная боль ощущается в височной части. Болевые ощущения сильнее в утреннее время из-за того, что всю ночь Был **заложен нос**.', symptoms=['**заложен нос**', 'головная боль']) *насморк*?

2019-01-06 01:22:45,254 element: DataObject(description='Вчера длительное время пробыла на улице, попала под дождь, на утро появился **насморк**, а сейчас полностью перестал дышать нос, почти не различаю запахи, капли для носа помогают но не долго, как быть?', symptoms=['**насморк', 'заложен нос**'])

Особенности использования симптомов

Особенности использования групп симптомов.

В данном случае не всегда ясно на основании чего используются те или иные симптомы.

СЛАБОСТЬ vs НЕДОМОГАНИЕ

2019-01-06 17:58:54,676 element: DataObject(description='Я занимаюсь спортом, и последнее время испытываю **недомогание**. Еще появились боли в спине, в основном болит между лопатками. Что мне сделать, чтобы снять боли?', symptoms=['**недомогание**', 'боль преимущественно между лопатками'])

2019-01-06 17:59:00,048 element: DataObject(description='Сегодня проснулась утром и почувствовала сильную боль в спине, **недомогание**. После долгого сидения или лежания, чувствую онемение в ногах, боль сопровождается недомоганием. Когда лежу, становится чуть легче.', symptoms=['боль сопровождается недомоганием', '**слабость**', 'боль после долгого лежания', 'боль после долгого сидения', '**недомогание**', 'онемение в ногах'])

Особенности использования симптомов

Пропуск симптома:

2019-01-04 08:26:44,333 element: DataObject(description='Вчера упал с велосипеда на руль, **удар пришелся на грудь**. Сначала была сильная жгучая боль, которая со временем утихла. Сегодня утром боль не прекратилась, дополнительно стало отдавать в шею. Почему стала болеть шея, может это перелом ребер?', symptoms=['боль отдает в шею', 'жгучая боль'])

2019-01-06 00:21:42,733 element: DataObject(description='Я просидел дома перед компьтером 36 часов без перерыва. Когда работа закончилась я вышел на улицу. От яркого света у меня внезапно **заболела голова**. Может у меня мигрень?', symptoms=['боль внезапная']) *есть симптом мигрень?*

2019-01-06 19:36:41,693 element: DataObject(description='Ребенку 7 лет, 3 дня сходил в школу, на 4 день у него озноб начался, температура небольшая поднялась, сопли потекли, чихает постоянно. Горло болит, красное. Каждую осень такое происходит, может у него алдлергия на что-нибудь или в школе его заразхили чем-то?', symptoms=['озноб']) *только озноб?*

Особенности использования симптомов

Лишний симптом:

2019-01-06 02:02:33,938 element: DataObject(description='Я шел по улице, переписывался с заказчиком по работе. Не заметил перед собой столб, ударился головой. Прошло несколько часов, у меня головокружение, тошнота. Не могу уснуть. Подскажите, что делать??', symptoms=['головная боль', 'тошнота', **'рвота'**, 'плохой сон'])

2019-01-06 19:47:04,773 element: DataObject(description='У сына 4-х лет затяжной насморк, выделения белые и прозрачные, 3 недели не можем справиться. Промывание носа не помогает. Общее состояние нормальное. Как вылечить?', symptoms=['насморк с жидкими белыми выделениями', 'насморк с прозрачными выделениями', 'насморк', **'чихание'**])

2019-01-06 22:23:23,567 element: DataObject(description='У моей дочери 12 лет после прогулки на лыжах сильно болит голова, поднялась температура тела. На следующй день появился неприятный насморк с прозрачными выделениями. Есть ли методы для быстрого лечения? ', symptoms=['температура 38 градусов и больше', 'насморк с прозрачными выделениями', **'озноб'**, 'головная боль', 'повышение температуры'])

Реализованные подходы: classification

Разбивка данных: так как данных мало, сделаем 10 фолдов. В обучающем множестве 402 описания, в тестовом 44.

Признаки: тексты после нормализации. Мешок слов + tfidf.

```
cv_word = CountVectorizer(ngram_range=(1, 5), analyzer='word', stop_words=stopWords)
tf_idf_word = TfidfVectorizer(ngram_range=(1, 5), analyzer='word')
```

Размер векторов признаков: 68466 признака.

Классы: 92 класса, так как объект может иметь несколько классов, то будем использовать MultiLabelBinarizer и OneVsRestClassifier из sklearn.

Реализованные подходы: classification

LogisticRegression с подбором гиперпараметра C с помощью пакета bayes_opt для каждого фолда.

	fold_#	precision	recall	f1	jaccard
0	1	0.622899	0.504348	0.53604	0.443704
1	2	0.634085	0.535088	0.550344	0.466111
2	3	0.755195	0.718182	0.724412	0.66963
3	4	0.595455	0.463636	0.504004	0.396481
4	5	0.683471	0.570248	0.592731	0.54537
5	6	0.676522	0.591304	0.607708	0.537831
6	7	0.607975	0.508065	0.529215	0.453274
7	8	0.680435	0.582609	0.602029	0.524892
8	9	0.6403	0.579365	0.592753	0.56553
9	10	0.727135	0.644628	0.667674	0.646212

CV score f1 : Mean - 0.5906910 | Std - 0.0667856 | Min - 0.5040043 | Max - 0.7244123

CV score jaccard : Mean - 0.5249035 | Std - 0.0879210 | Min - 0.3964815 | Max - 0.6696296

Реализованные подходы: classification

GradientBoostingClassifier с подбором гиперпараметров learning_rate, n_estimators с помощью пакета bayes_opt для каждого фолда.

	fold_#	precision	recall	f1	jaccard
0	1	0.648986	0.53913	0.560248	0.453333
1	2	0.678363	0.561404	0.581871	0.465185
2	3	0.719697	0.681818	0.684652	0.615185
3	4	0.731818	0.509091	0.569268	0.481852
4	5	0.646281	0.570248	0.58267	0.539136
5	6	0.693623	0.66087	0.663307	0.56672
6	7	0.611022	0.548387	0.561277	0.493579
7	8	0.651087	0.530435	0.560663	0.476732
8	9	0.594907	0.52381	0.536508	0.478788
9	10	0.786777	0.694215	0.713784	0.627652

CV score f1 : Mean - 0.6014248 | Std - 0.0617632 | Min - 0.5365079 | Max - 0.7137845
CV score jaccard : Mean - 0.5198161 | Std - 0.0635703 | Min - 0.4533333 | Max - 0.6276515