1 Данные

В собранной коллекции имелось 100 пар текстов. Каждая пара состоит из написанного в спокойном состоянии и написанного в фрустрированном состоянии одним и тем же человеком текстов. Для каждого текста были вычислены одинаковые признаки (всего признаков 198), и для каждой пары была вычислена разница между текстами.

Перед нами ставится задача кластеризировать эту разницу, т.е. мы хотим сгруппировать людей по схожести шаблона изменения состояния со спокойного на фустрированное. Более того, добавим ограничение, что нас интересует качественная схожесть, а именно, мы будем считать, что если есть два человека, у которых некоторый признак при переходе состояния увеличивается, то значение этого признака не изменяется качественно. Это обуславливает далее описываемую предобработку данных.

Во-первых, все признаки были отмасштабированы по максимальному абсолютному значению, что привело к тому, что все значения лежали на отрезке от -1 до 1, причем знак сохранился. Во-вторых, ко всем данным был применен модифицированный сигнум:

$$\text{sign}_\epsilon(x) = \begin{cases} -1, \text{if } x < -\epsilon, \\ 0, \text{if } |x| \leq \epsilon, \\ 1, \text{if } x > \epsilon. \end{cases},$$

где ϵ - параметр. В наших экспериментах был выбран параметр, равный 0.2. Данный сигнум лучше обычного, поскольку он также показывает знак, но при этом принебрегает близкими к нулю значениями.

2 Метод Кластеризации. Количество Кластеров.

Перед основной кластеризацией был использован DBSCAN для поиска шумов и метрика Score Function для оценки качества того, насколько отделение шумов улучшает качество кластеризации по сравнению с качеством на одном кластере. Было получено, что DBSCAN не обнаружил таких шумов, удаление которых приводило бы к улучшению качества.

Зафиксируем следующую метрику качества кластеризации:

$$M(C) = \text{SF}(C) + \exp\left(-K\frac{\text{MaxSize}(C) - \text{MinSize}(C)}{N}\right),$$

где SF - метрика для оценки качества кластеризации Score Function, MaxSize, MinSize - размеры наибольшего и наименьшего кластеров, N - количество кластеризируемых объектов, K - количество кластера. Первое слагаемое этой метрики отвечает за качество, второе же слагаемое штрафует за несбалансированность классов, поскольку данная выборка имеет не достаточно большой размер, чтобы правильно в последствие оценить распределение значений признаков в маленьких кластерах.

Мы использовали различные известные алгоритмы кластеризации, при этом мы также варьировали количество кластеров в диапазоне от 1 до 10. Результаты Вы можете видеть на рис.1.

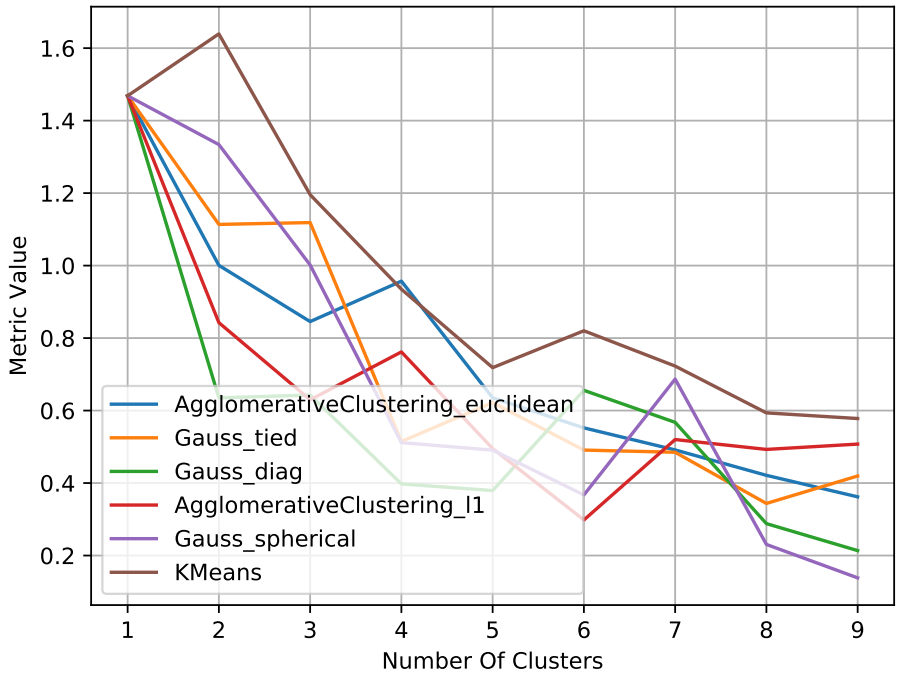


Figure 1: ....

Из результатов данного эксперимента следует, что наилучшим методом кластеризации является KMeans, причем оптимальным количеством кластеров является два. Далее мы будем использовать именно метод KMeans. Однако кроме кластеризации на два кластера, мы также рассмотрим

и кластеризацию на 4, как на наибольшее количество кластеров, для которого значение метрики упало менее, чем в два раза.

Крому того, что нам нужно кластеризовать, нам следует выделить наиболее различающиеся признаки у кластеров. Мы это делаем тривиальным образом: высчитать среднее значение всех признаков для каждого кластера, получить разницу между средними для всех пар кластеров, оценить информативность, как максимальную по всем парам разницу для каждого признака, отранжировать их по информативности и выбрать все, у которых информативность больше 70% от максимальной.

3 Кластеризация на два кластера

Наиболее различающиеся признаки получаются следующие (записаны в порядке возрастания):

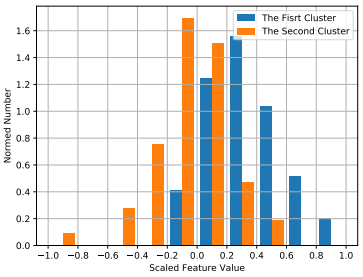
- Коэффициент Трейгера
- Коэффициент опредмеченности действия (кол. глаголов / кол. существительных)

Посмотрим на распределения значений для соответствующих признаков, причем мы будем смотеть на отмасштабированные значения. Для каждого признака мы взяли интервал [-1,1], в котором лежат значения этого признака, разбили его на 10 частей, посчитали для каждого кластера количество объектов со значением этого признака в соответствующей части и нормализовали полученные распределения (сделали так, чтобы интеграл по всему интервалу был равен единице). Соответствующие гистограммы Вы можете видеть на 2.

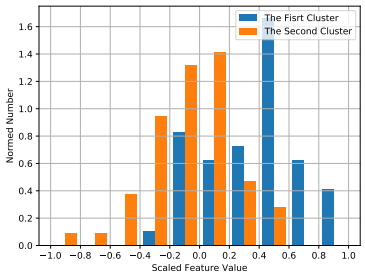
4 Кластеризация на четыре кластера

Наиболее различающиеся признаки получаются следующие (записаны в порядке возрастания):

- Доля глаголов 1 лица
- Сем. связь: QNT



(a)

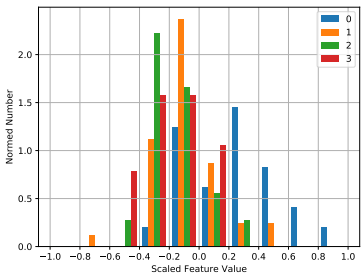


(b)

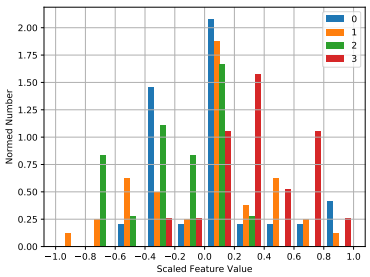
Figure 2: Сравненение распределения для наиболее различающихся признаков: (a) Коэффициент Трейгера; (b) Коэффициент опредмеченности действия (кол. глаголов / кол. существительных).

- Сем. роль: адресат
- Число знаков пунктуации / Число слов
- Средняя длина слов (в количестве символов)
- Доля глаголов прошедшего времени
- Коэффициент Трейгера
- Словарь: Лексика положительной рациональной оценки и ментальных действий

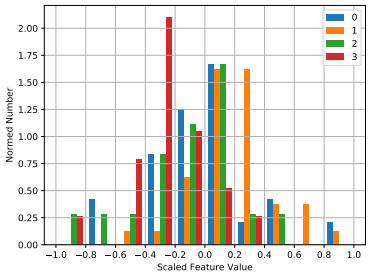
Посмотрим на распределения значений для соответствующих признаков, причем мы будем смотеть на отмасштабированные значения. Для каждого признака мы взяли интервал [-1,1], в котором лежат значения этого признака, разбили его на 10 частей, посчитали для каждого кластера количество объектов со значением этого признака в соответствующей части и нормализовали полученные распределения (сделали так, чтобы интеграл по всему интервалу был равен единице). Соответствующие гистограммы Вы можете видеть на 3 и 4.



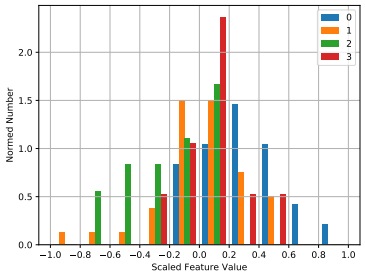
(a)



(b)

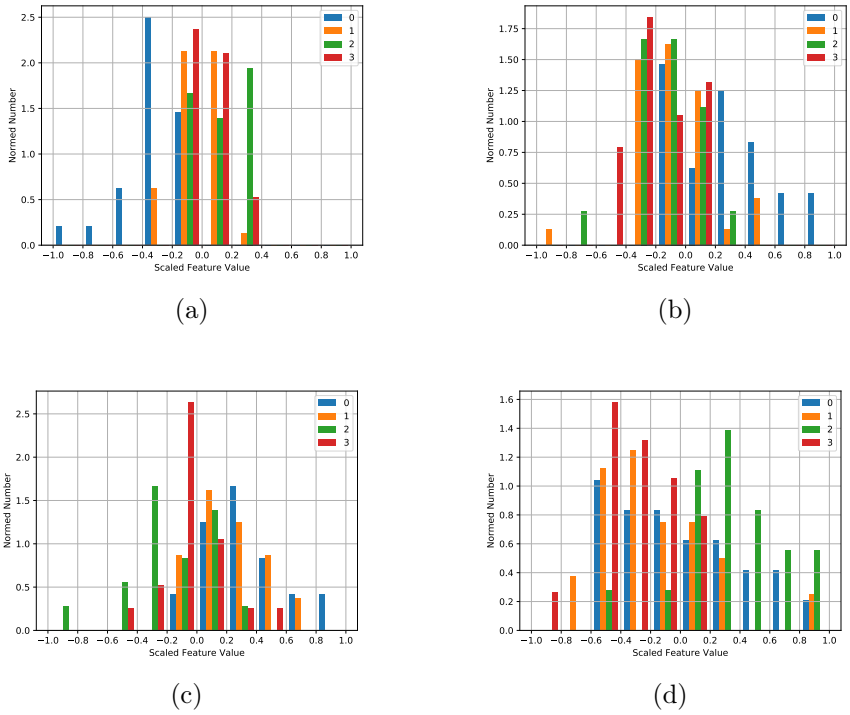


(c)



(d)

Figure 3: Сравненение распределения для наиболее различающихся первых четырех признаков: (a) Доля глаголов 1 лица; (b) Сем. связь: QNT; (c) Сем. роль: адресат; (d) Число знаков пунктуации / Число слов.



(a) (b)

(c) (d)

Figure 4: Сравненение распределения для наиболее различающихся последних четырех признаков: (a) Средняя длина слов (в количестве символов); (b) Доля глаголов прошедшего времени: QNT; (c) Коэффициент Трейгера; (d) Словарь: Лексика положительной рациональной оценки и ментальных действий.