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.

Первый этап предобработки. Все признаки были отмасшатиборованы таким образом: отрицательные и положительные значения масштабировались на максимальное абсолютные значения среди всех отрицательных и среди всех положительных соответственно. Данная предобработка обусловлена тем, что некоторые признаки при переходе текстов из спокойного в фустрированное состояние имеют не одинаковые пределы для увеличения и для уменьшения. Так, например, средняя длина предложений не может стать отрицательной, хотя увеличиться может до сколь угодно больших значений.

Второй этап предобработки. 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. Данный сигнум лучше обычного, поскольку он также показывает знак, но при этом принебрегает близкими к нулю значениями.

В наших экспериментах был выбран параметр, равный 0.05, т.е. мы считали, что у признака нет существенного изменения, если он изменился менее чем на 5% от максимального изменения в эту сторону.

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. DBSCAN не обнаружил шумов, поэтому далее мы будем предполагать, что наши данные не содержат выбросов.

Зафиксируем следующую метрику качества кластеризации:

$$M(C) = \mathrm{SF}(C)\dot{\min}\left(1, \frac{\mathrm{MinSize}}{10}\right)$$

где SF - метрика для оценки качества кластеризации Score Function, MinSize - размер наименьшего кластера. Первый множитель этой метрики отвечает за качество, второй же штрафует, если в кластере меньше 10 объектов. Такие кластеры мы не сможем качественно проанализировать ввиду их небольшого размера и логичней всего было бы интерпретировать их как выбросы, однако это противоречит предположению сделанному на основе результатов DBSCAN.

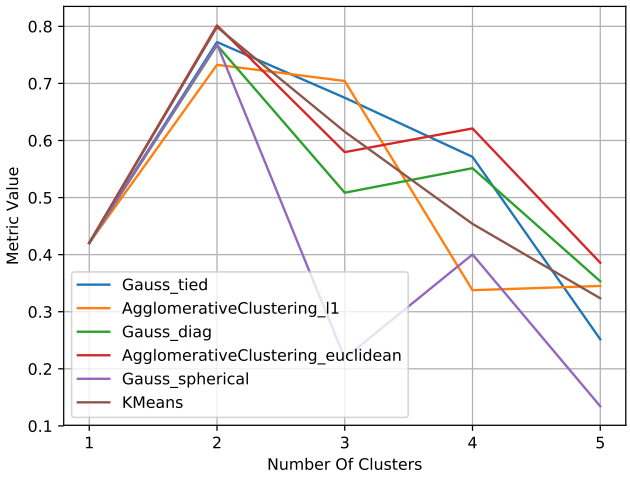


Figure 1: Результаты подбора оптимального метода и оптимального количества кластеров

Мы использовали различные известные алгоритмы кластеризации,

при этом мы также варьировали количество кластеров в диапазоне от 1 до 5. Результаты Вы можете видеть на рис.1.

Из результатов данного эксперимента следует, что наилучшим методом кластеризации является KMeans, причем оптимальным количеством кластеров является два.

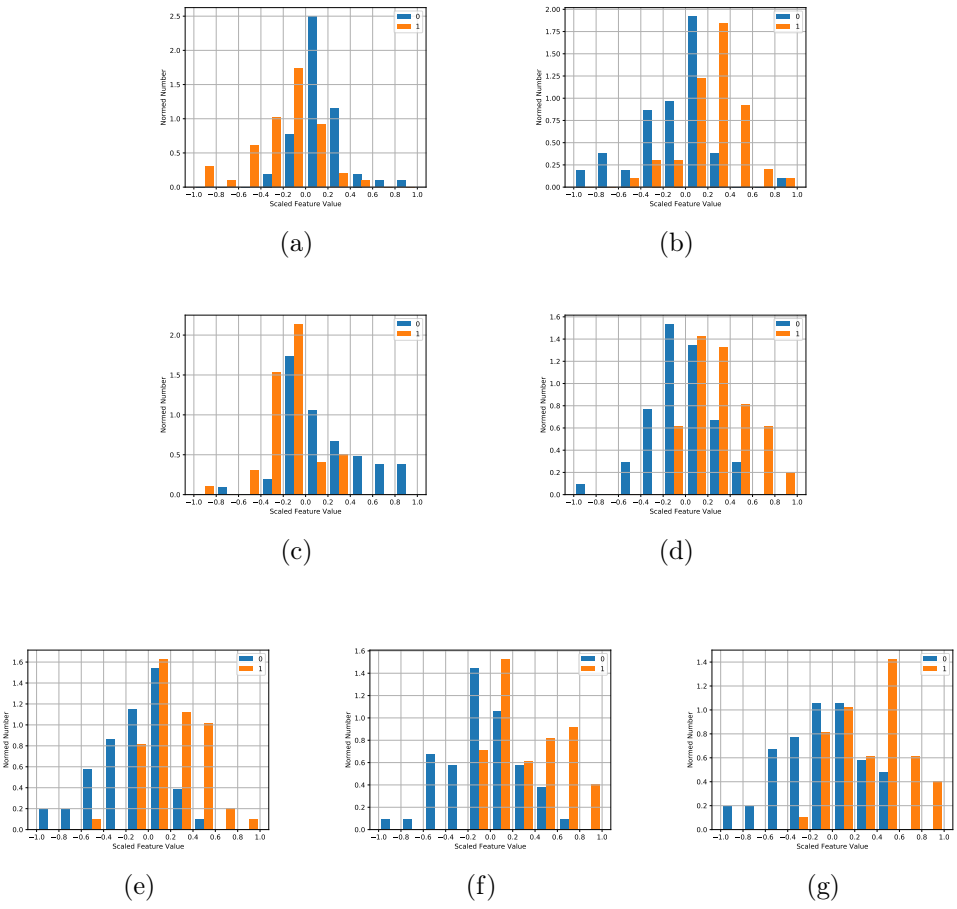
3 Кластеризация на два кластера

Выделим наиболее различающиеся признаки у двух кластеров. Мы это сделаем следующим образом: высчитать среднее значение всех признаков для каждого кластера, получить разницу между ними и выбрать все, у которых информативность больше 75% от максимальной.

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

- Коэффициент логической связности
- Часть речи: местоимение-существительное
- Средняя длина слов (в количестве символов)
- Коэффициент Трейгера
- Число знаков пунктуации / Число слов
- Часть речи: глагол
- Коэффициент опредмеченности действия (кол. глаголов / кол. существительных)

Посмотрим на распределения значений для соответствующих признаков, причем мы будем смотеть на отмасштабированные значения, т.е. значения, которые получены после первого этапа предобработки. 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 (интеграл по всему интервалу равен единице). Соответствующие гистограммы Вы можете видеть на 2.



(a) (b)

(c) (d)

(e) (f) (g)

Figure 2: Сравненение распределения для наиболее различающихся признаков (перечислены в порядке возрастания): (a) Коэффициент логической связности; (b) Часть речи: местоимение-существительное; (c) Средняя длина слов (в количестве символов); (d) Коэффициент Трейгера; (e) Число знаков пунктуации / Число слов; (f) Часть речи: глагол; (g) Коэффициент опредмеченности действия (кол. глаголов / кол. существительных).