1 Данные

В собранной коллекции имелось 100 пар текстов. Каждая пара состоит из написанного в спокойном состоянии и написанного в фустрированном 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 и была предпринята попытка кластеризации пар по этой разнице.

Пусть F - матрица объект признак для текстов, написанных в фустрированном состоянии, C - матрица объект признак для текстов, написанных в спокойном состоянии, тогда $D = F - C$ - матрица объект-признак для пары текстов. Далее мы будем говорить о нормализованных данных.

2 Кластеризация

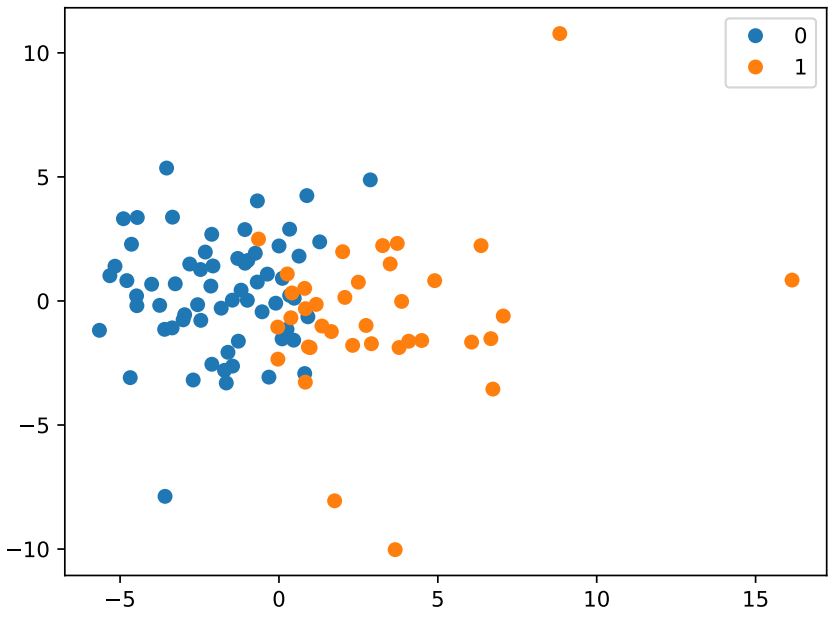

Figure 1: Кластеризация на два класса: визуализация при помощи PCA

Используя различные методы визуализации, в частности, PCA, можно увидеть, что данные представляют собой достаточно непрерывное единое облако точек с небольшими выбросами.

Были опробованы различные методы кластеризации. Было решено остановиться на KMeans с предварительным снижением размерности при помощи метода PCA до 10. Результат кластеризации на два кластера Вы можете видеть на рисунке 2. Как можно видеть, подобная кластеризация просто разрезает вышеупомянутое облако точек приблизительно по центру.

Классы получаются достаточно сбалансированными и при трех кластерах. При большом количестве кластеров получаемое разбиение по большей части просто отсеивает крайние точки.

3 Основные Различия Между Кластерами

Далее выделим главные отличия между двумя кластерами, полученными в прошлом разделе. Сначала определим по каким признакам эти кластеры различаются больше всего. Для этого для каждого признака посчитаем квадрат разности между его средним значением по первому кластеру и по второму. Получим, что приблизительно для 120 признаков эта разница околонулевая. Пусть M - максимальная разница между признаками, тогда возьмем все признаки для которых разница лежит в диапазоне от $[0.7M, M]$. Всего таких признаков получили шесть. Их наименования в порядке возрастания разницы:

- Доля глаголов прошедшего времени, первого лица, единственного числа
- Часть речи: прилагательное
- Часть речи: существительное
- Средняя длина слов (в количестве символов)
- Коэффициент Трейгера
- Часть речи: местоимение-существительное

Посмотрим на распределения для соответствующих признаков. Для каждого признака мы взяли интервал, 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 данные (сделали так, чтобы интеграл по всему интервалу был равен единице). Соответствующие гистограммы Вы можете видеть на 2.

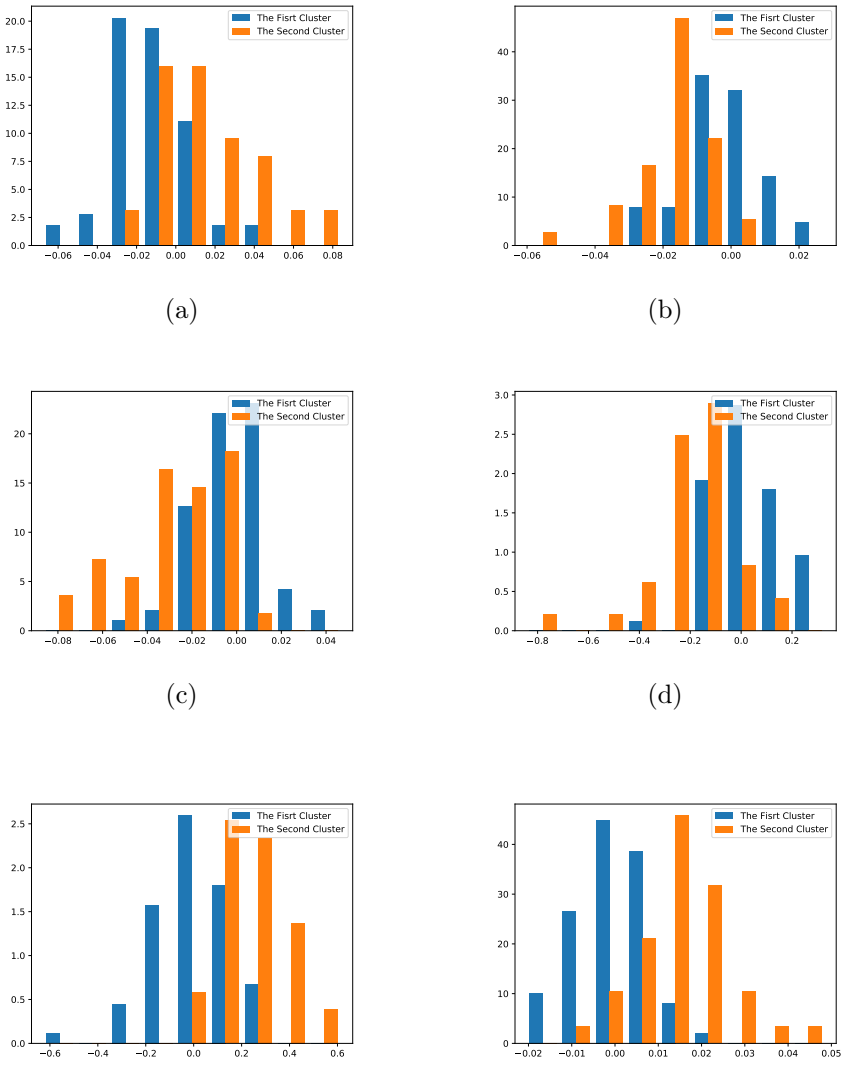

Figure 2: Сравненение распределения для наиболее различающихся признаков: (a) Доля глаголов прошедшего времени, первого лица, единственного числа; (b) Часть речи: прилагательное; (c) Часть речи: существительное; (d) Средняя длина слов (в количестве символов); (e) Коэффициент Трейгера; (f) Часть речи: местоимение-существительное.