Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра автоматизации систем вычислительных комплексов

ИССЛЕДОВАТЕЛЬСКИЙ ПРОЕКТ

Кластеризация пользователей социальной сети

Выполнил:
студент 521 группы
Боднарюк Василий

Москва, 2018

Содержание

Введение

В современном мире социальное взаимодействие через Интернет приобретает повсеместный характер. Наиболее ярко подобное взаимодействие иллюстрируется на примере социальных сетей. Пользователи социальных сетей, особенно подростки, склонны раскрывать большое количество личной информации. Помимо прямо указанных личных данных и интересов, к информации, характеризующей пользователя, можно отнести *лайки* (одобрение, проявляемое пользователем в виде публичной отметки) определенных постов и медиаконтента, а также *репосты* (размещение чьего-либо поста на своей странице).

На основе информации, оставляемой подростками, можно производить дробление данной социальной группы. Целью такого дробления может являться *рекламный таргетинг*, заключающийся в демонстрации рекламы по интересам пользователей.

Существует возможность производить таргетинг продукта по словам, используемым пользователем на его странице. К сожалению, таргетинг по конкретным словам не всегда эффективен, поскольку слов, характеризующих продукт, может быть очень большое количество, и упомянуть их все – часто невыполнимая задача. В то же время слов, характеризующих интересы пользователя, как правило, не так много, и пересечение с выявленными для продукта словами оказывается пустым.

В связи с вышесказанным разбиение пользователей на группы по интересам – актуальная задача. Предполагается, что удобства таргетинга количество групп должно находиться в интервале от 5 до 15. Такое разбиение можно производить, используя алгоритмы кластеризации.

Работа посвящена исследованию возможности кластеризации подростков, являющихся пользователями социальных сетей. Задача работы в следующем: на основе информации о встречаемых на страницах пользователей словах разбить пользователей на кластеры. Для новых пользователей необходимо иметь возможность относить их к какому-либо из ранее выделенных кластеров. Кластеризация является задачей обучения без учителя.

К подзадачам относятся определение числа кластеров, на которые следует разбивать множество пользователей, и обучение модели с целью предсказания кластера по новым данным.

Программный код в поддержку проекта размещен по адресу https://github.com/Vasilesk/sphere/tree/master/extra/snsdata_clustering

Исходные данные

В качестве исходных данных выступает информация о частоте встречаемости 36 слов на 30000 страницах подростков в социальной сети Facebook. Данные собраны в 2006-м году Бреттом Ланцем (Brett Lantz) в рамках социологического исследования по изучению подростков, которое проводилось в Университете Нотр-Дам, США. Данные используются в книге Ланца “Machine Learning with R” [1], а также в учебном курсе “Statistical Learning with R” [2] профессора California State University, East Bay, США Эрика Сьюса (Eric Suess). В обоих случаях с их помощью иллюстрируется возможность проведения кластеризации с помощью алгоритма k средних.

Формат исходных данных [3] следующий: для каждого из 30000 объектов имеется описание, представляющее собой вектор длины 36, каждый элемент вектора отвечает соответственно одному из слов basketball, football, soccer, softball, volleyball, swimming, cheerleading, baseball, tennis, sports, cute, sex, sexy, hot, kissed, dance, band, marching, music, rock, god, church, jesus, bible, hair, dress, blonde, mall, shopping, clothes, hollister, abercrombie, die, death, drunk, drugs. Элемент вектора – атрибут объекта – целое число, обозначающее частоту встречаемости слова на странице. Пример описания первых десяти объектов исходных данных приведен на рисунке 1.

	0	1	2	3	4	5	6	7	8	9
basketball	0	0	0	0	0	0	0	0	0	0
football	0	1	1	0	0	0	0	0	0	0
soccer	0	0	0	0	0	0	0	0	0	0
softball	0	0	0	0	0	0	0	1	0	0
volleyball	0	0	0	0	0	0	0	0	0	0
swimming	0	0	0	0	0	0	0	0	0	0
cheerleading	0	0	0	0	0	0	0	0	0	0
baseball	0	0	0	0	0	0	0	0	0	0
tennis	0	0	0	0	0	0	0	0	0	0
sports	0	0	0	0	0	0	0	0	0	0
cute	0	1	0	1	0	0	0	0	0	1
sex	0	0	0	0	1	1	0	2	0	0
sexy	0	0	0	0	0	0	0	1	0	0
hot	0	0	0	0	0	0	0	0	0	1
kissed	0	0	0	0	5	0	0	0	0	0
dance	1	0	0	0	1	0	0	0	0	0
band	0	0	2	0	1	0	1	0	0	0
marching	0	0	0	0	0	1	1	0	0	0

	0	1	2	3	4	5	6	7	8	9
music	0	2	1	0	3	2	0	1	0	1
rock	0	2	0	1	0	0	0	1	0	1
god	0	1	0	0	1	0	0	0	0	6
church	0	0	0	0	0	0	0	0	0	0
jesus	0	0	0	0	0	0	0	0	0	2
bible	0	0	0	0	0	0	0	0	0	0
hair	0	6	0	0	1	0	0	0	0	1
dress	0	4	0	0	0	1	0	0	0	0
blonde	0	0	0	0	0	0	0	0	0	0
mall	0	1	0	0	0	0	2	0	0	0
shopping	0	0	0	0	2	1	0	0	0	1
clothes	0	0	0	0	0	0	0	0	0	0
hollister	0	0	0	0	0	0	2	0	0	0
abercrombie	0	0	0	0	0	0	0	0	0	0
die	0	0	0	0	0	0	0	0	0	0
death	0	0	1	0	0	0	0	0	0	0
drunk	0	0	0	0	1	1	0	0	0	0
drugs	0	0	0	0	1	0	0	0	0	0

Рисунок 1. Описание первых десяти объектов исходных данных

Предполагается, что данные являются репрезентативными, поскольку собраны в рамках научного исследования. Помимо этого, как отмечено выше, их используют в своих курсах специалисты отрасли анализа данных.

Сумма частот встречаемости слов, упорядоченная по возрастанию, приводится в таблице 1. На рисунке 2 приведено количество слов, чья частота попадает в определенный интервал.

Рисунок 3 отображает частоту встречаемости слова *baseball* и значение функции вероятности распределения Пуассона с параметром лямбда, равным выборочному среднему частоты встречаемости *baseball*.

Вышеназванные таблица и рисунки иллюстрируют зависимости, согласующиеся с эмпирическими представлениями о предметной области, например, очевидно, что музыка более популярна среди подростков, чем Библия.

слово: частота	**слово: частота**
bible: 640	volleyball: 4294
marching: 1218	clothes: 4455
abercrombie: 1535	softball: 4836
drugs: 1813	die: 5523
hollister: 2096	sex: 6282
tennis: 2620	soccer: 6683
drunk: 2639	rock: 7300
blonde: 2968	church: 7445
kissed: 3096	football: 7569
baseball: 3148	mall: 7721
cheerleading: 3199	basketball: 8020
dress: 3329	band: 8988
jesus: 3362	cute: 9686
death: 3427	shopping: 10590
hot: 3798	hair: 12677
swimming: 4032	dance: 12755
sports: 4199	god: 13959
sexy: 4236	music: 22135

Таблица 1. Сумма частот встречаемости слов в выборке

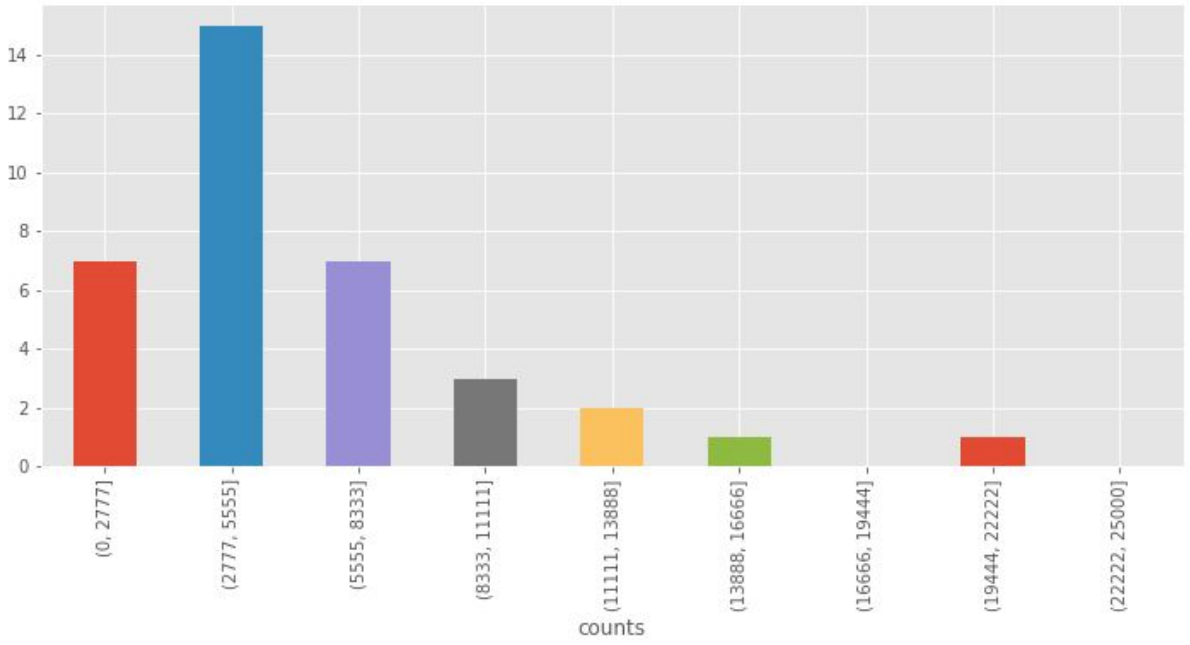

Рисунок 2. Количество слов, суммарная частота встречаемости которых попадает в соответствующий интервал

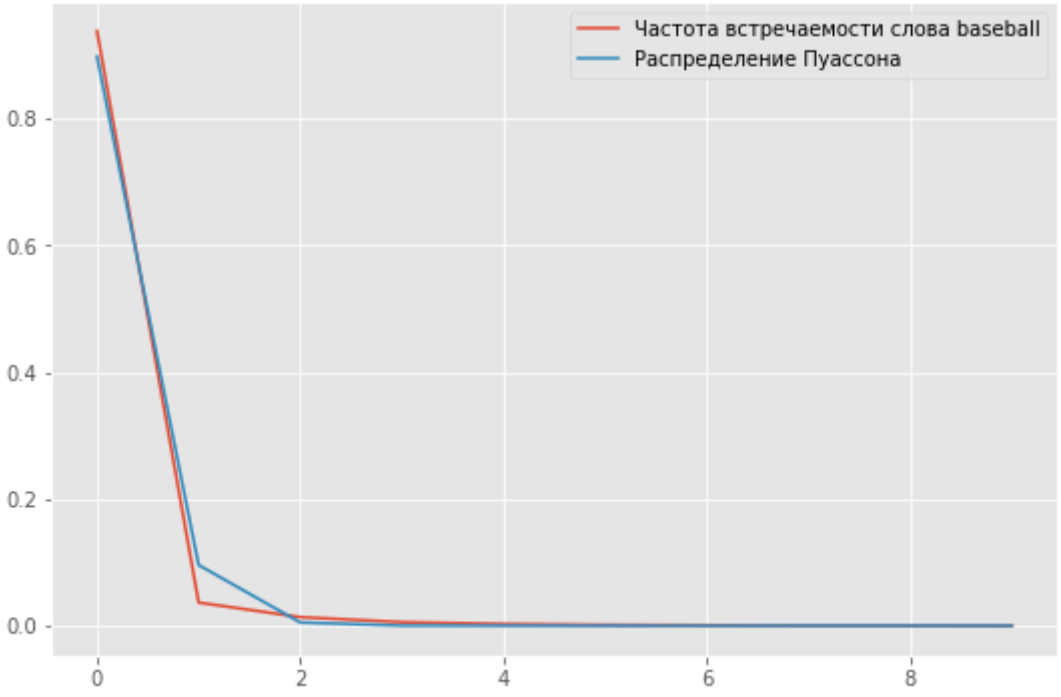

Рисунок 3. Частота встречаемости слова baseball и значение функции вероятности распределения Пуассона

Формальная постановка задачи

Формальная постановка задачи кластеризации взята из источника [4] и выглядит следующим образом.

Пусть X – множество объектов, Y – множество номеров кластеров. Задана функция расстояния между объектами $\rho(x,x')$. Имеется конечная обучающая выборка объектов $X^m = \{x_1,\dots,x_m\} \subset X$.

Требуется разбить выборку на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из объектов, близких по метрике ρ, а объекты разных кластеров существенно отличались. При этом каждому объекту $x_i \in X^m$ приписывается номер кластера $y_i \in Y^k$.

Алгоритм кластеризации – это функция $X \to Y$, которая любому объекту $x \in X$ ставит в соответствие номер кластера $y \in Y$. Множество Y заранее неизвестно, и ставится задача определить оптимальное число кластеров, руководствуясь ограничениями содержательной постановки задачи (от 5 до 15 кластеров) и критерием качества кластеризации.

Критерием качества кластеризации выступает значение silhouette [5], которое рассчитывается как сумма по объектам $x \in X$ значений

$$(b - a) / \max(a, b),$$

где a – среднее расстояние между рассматриваемым объектом и объектами его кластера,

b – среднее расстояние между рассматриваемым объектом и объектами *ближайшего* к нему кластера.

Ближайший кластер – это кластер, которому рассматриваемый объект не принадлежит, и среднее расстояние от рассматриваемого объекта до объектов которого по подобным кластерам минимально.

Предобработка исходных данных

Предобработка исходных данных включает в себя несколько этапов, среди которых следует выделить заполнение отсутствующих значений атрибутов объектов, выявление и удаление аномалий, нормализацию значений признаков [6]. Исходя из содержательной постановки задачи, а именно того факта, что признаки – это слова, помимо нормализации исследуется вариант перехода к мере TF-IDF [7].

В исходных данных значения атрибутов не содержат пробелов, поэтому нет необходимости решать проблему заполнения отсутствующих значений.

Выявление аномалий производилось с помощью анализа диаграмм частот встречаемости слов в выборке. Примеры диаграмм без аномалий и с аномалией приведены на рисунках 4 и 5 соответственно. Рисунок 6 отображает аномальное значение, которое не удается увидеть на рисунке 5 из-за масштаба.

Подобным образом были выявлены 8 аномальных объектов. Они были удалены из исходной выборки и далее не рассматриваются.

Проблема соразмерности признаков является актуальной в контексте решаемой задачи (см. рис. 2). Данная проблема решалась двумя способами.

Первый вариант решения предполагал нормализацию признаков. Нормализация конкретного признака производится следующим образом: из значения признака объекта вычитается среднее значение по всем объектам выборки, производится деление на стандартное отклонение.

Во втором варианте каждое описание рассматривается как документ, а значение признака – как частота встречаемости слова в документе. Производится переход к мере TF-IDF.

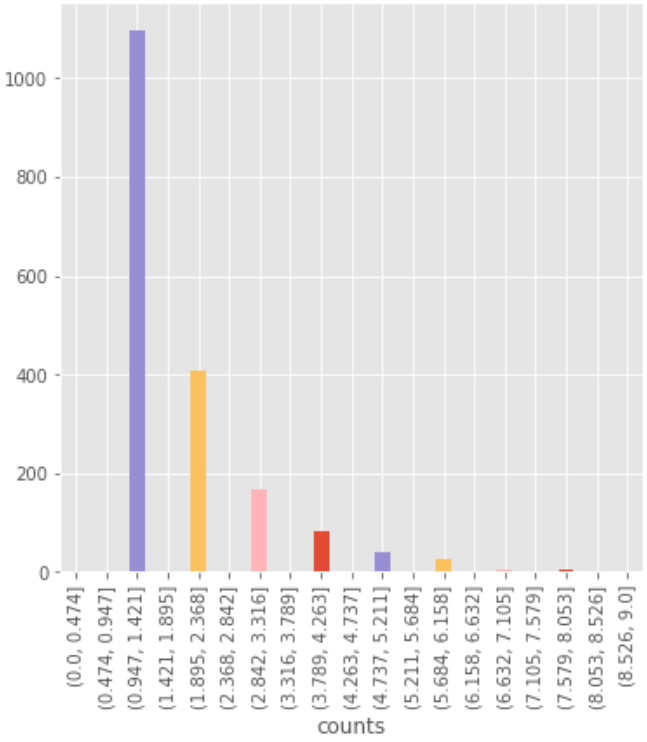

Рисунок 4. Частоты встречаемости определенного количества слов. Аномалии отсутствуют

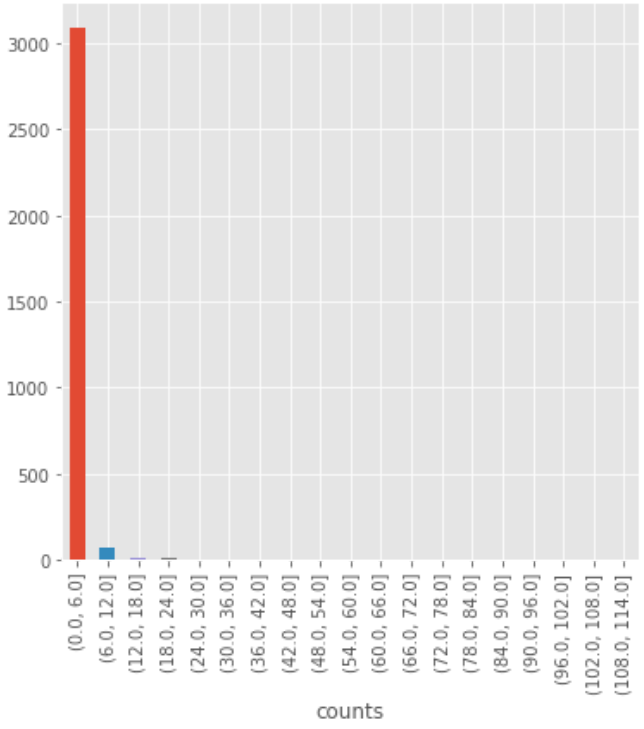

Рисунок 5. Частоты встречаемости определенного количества слов.

Аномалия порождает несбалансированный график

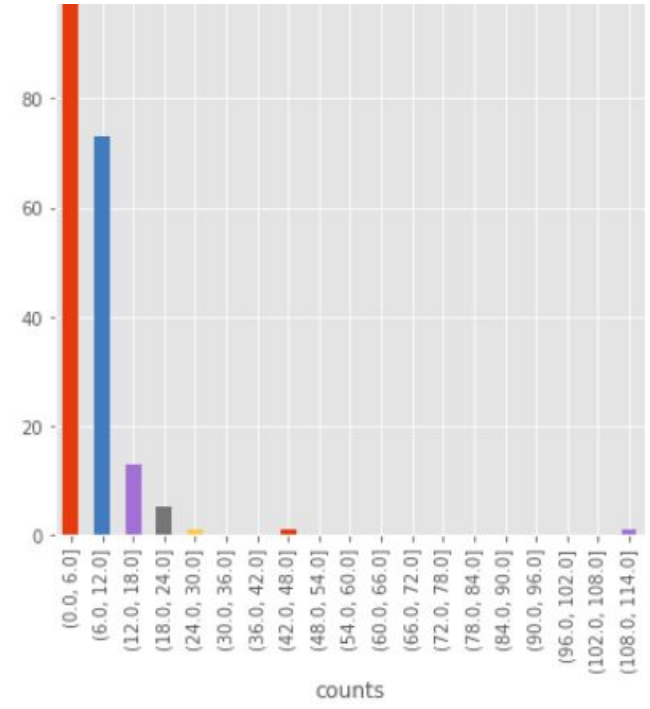

Рисунок 6. Аномальное значение при масштабировании графика

Решение задачи кластеризации

Решение задачи кластеризации производится для двух вариантов предобработанных данных – нормализованный вариант и вариант TF-IDF.

Исследуются два алгоритма кластеризации – алгоритм k средних [8] и EM-алгоритм разделения смеси Гауссовских распределений [9].

Первым этапом решения задачи кластеризации является определение числа кластеров k. Качество кластеризации на варианте TF-IDF оказалось хуже, в связи с чем выбор числа кластеров производится на основе работы алгоритмов на нормализованных данных (рисунки 7, 8).

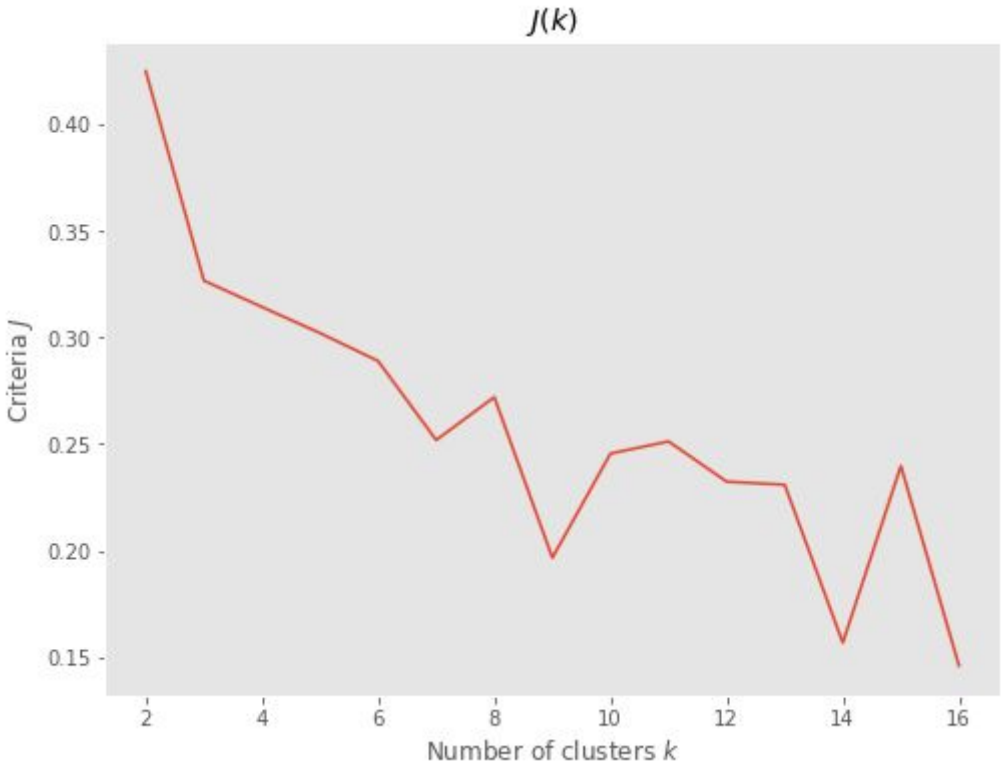

Рисунок 7. Алгоритм k средних.

Значение критерия качества в зависимости от количества кластеров

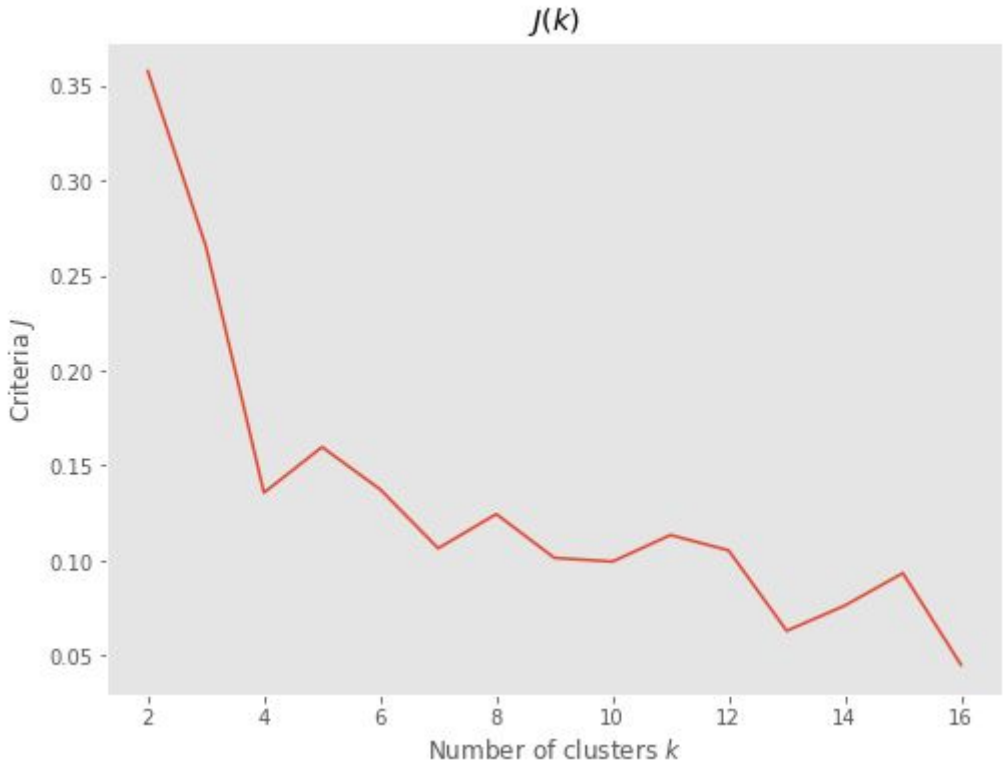

Рисунок 8. EM-алгоритм.

Значение критерия качества в зависимости от количества кластеров

Рисунки 7 и 8 иллюстрируют наличие локального максимума значения критерия качества при k=8. Таким образом, принято решение производить кластеризацию на 8 кластеров.

Результаты и их сравнение

Результаты кластеризации для числа кластеров, равного 8, выглядят следующим образом. Топ слов рассчитан как относительная доля упоминаний слова по всей выборке, попавшего в кластер, поделенная на размер кластера.

Нормализованные данные, алгоритм k средних

Время обучения: 6.86 секунд

Значение метрики качества silhouette: 0.2886142256155253

Размеры кластеров:

	cluster	size
0	5	19532
1	1	4663
2	4	2429
3	3	965
4	2	844
5	6	609
6	0	529
7	7	421

Топ слов в кластерах:

Кластер 0 : bible jesus god church death

Кластер 1 : dress hot cute mall shopping

Кластер 2 : abercrombie hollister cheerleading clothes shopping

Кластер 3 : kissed drugs sex blonde drunk

Кластер 4 : softball baseball volleyball basketball sports

Кластер 5 : music die death soccer god

Кластер 6 : marching band music rock dress

Кластер 7 : tennis sports church soccer blonde

Нормализованные данные, EM-алгоритм

Время обучения: 24.2 секунды

Значение метрики качества silhouette: 0.14209979055475316

Размеры кластеров:

	cluster	size
0	3	17439
1	1	5484
2	7	2455
3	4	2063
4	6	1208
5	0	603
6	5	500
7	2	240

Топ слов в кластерах:

Кластер 0 : abercrombie hollister kissed drugs tennis

Кластер 1 : baseball softball cheerleading volleyball tennis

Кластер 2 : bible marching drugs blonde kissed

Кластер 3 : music cute dance shopping sexy

Кластер 4 : hollister abercrombie dress cheerleading mall

Кластер 5 : bible jesus marching god church

Кластер 6 : blonde drunk drugs sex kissed

Кластер 7 : marching drugs band tennis kissed

Данные TF-IDF, алгоритм k средних

Время обучения: 5.39 секунд

Значение метрики качества silhouette: 0.12478214144072033

Размеры кластеров:

	cluster	size
0	4	12753
1	5	3225
2	1	2747
3	2	2626
4	0	2580
5	3	2216
6	6	2069
7	7	1776

Топ слов в кластерах:

Кластер 0 : god bible jesus church death

Кластер 1 : cute mall sexy shopping hot

Кластер 2 : dance dress shopping hot abercrombie

Кластер 3 : basketball football sports baseball volleyball

Кластер 4 : sex kissed cheerleading softball drugs

Кластер 5 : music rock death drugs shopping

Кластер 6 : marching band music rock death

Кластер 7 : soccer sports swimming football tennis

Данные TF-IDF, EM-алгоритм

Время обучения: 28 секунд

Значение метрики качества silhouette: 0.027161106877676245

Размеры кластеров:

	cluster	size
0	2	12520
1	1	6360
2	3	4293
3	0	2599
4	7	1811
5	5	899
6	6	822
7	4	688

Топ слов в кластерах:

Кластер 0 : hollister abercrombie blonde tennis dress

Кластер 1 : baseball basketball volleyball soccer cheerleading

Кластер 2 : music sexy die death cute

Кластер 3 : marching dress drunk jesus tennis

Кластер 4 : marching kissed blonde tennis baseball

Кластер 5 : bible marching jesus volleyball tennis

Кластер 6 : volleyball bible hollister abercrombie blonde

Кластер 7 : kissed drugs drunk blonde sex

Сравнение результатов

Лучшее значение выбранной метрики качества демонстрирует алгоритм k средних на нормализованных данных. Топ слов, полученный для этого случая, выглядит наиболее удачным с точки зрения эмпирических данных, поскольку слова по кластерам имеют отношение к какой-то сфере интересов подростков. Интерпретировать кластеры можно следующим образом

Кластер 0 (bible jesus god church death)

Подростки, имеющие склонность к религии

Кластер 1 : dress hot cute mall shopping

Любители гламура, которые следят за собой

Кластер 2 : abercrombie hollister cheerleading clothes shopping

Опытные шопоголики, разбирающиеся в брендах одежды

Кластер 3 : kissed drugs sex blonde drunk

Любители несколько маргинальной стороны подростковой жизни

Кластер 4 : softball baseball volleyball basketball sports

Спортсмены, специализирующиеся на играх с мячом

Кластер 5 : music die death soccer god

Самый крупный кластер, тяжело интерпретировать

Кластер 6 : marching band music rock dress

Любители музыки и, вероятно, музыкальных фестивалей

Кластер 7 : tennis sports church soccer blonde

Спортсмены иных дисциплин

Данные в нормализованном виде кластеризовались лучше как с точки зрения критерия качества, так и с точки зрения изучения топа слов.

Однако TF-IDF данные по результатам кластеризации не имеют столь ярко выраженного перекоса в пользу одного определенного кластера. Например, на них алгоритм k средних дает результаты, схожие с результатами кластеризации на нормализованных данных, однако кластер, интерпретируемый как “любители маргинальной стороны подростковой жизни” не выделятся в случае TF-IDF.

Следует отметить, что алгоритм k средних имеет преимущество в скорости работы по сравнению с EM-алгоритмом.

Также важным замечанием является то, что во всех рассмотренных вариантах кластеризации присутствует кластер большого размера, что говорит о сложностях, возникающих при попытке разбить страницы подростков на группы, используя лишь рассмотренные признаки.

Заключение

В процессе исследования были рассмотрены алгоритм k средних и EM-алгоритм разделения смеси Гауссовских распределений в задаче кластеризации страниц подростков-пользователей социальной сети.

Данные на вход алгоритмов были предобработаны, алгоритмы запущены на двух вариантах предобработанных данных.

Выбран критерий качества кластеризации, после получения результатов работы алгоритмов очевидно, что критерий качества согласуется с эмпирическими представлениями о кластеризации в предметной области.

Согласно критерию качества и ограничениям предметной области выбрано оптимальное значение количества кластеров.

Кластеризация произведена в четырех вариантах, для каждого варианта оценены качество и время работы.

Результаты наиболее удачной кластеризации интерпретированы на основе эмпирических знаний.

Таким образом, в результате проделанной работы обучены модели, способные предсказывать принадлежность объекта тому или иному интерпретированному кластеру. Согласно критерию качества, из моделей выбрана лучшая.

Список литературы

1. Brett Lantz “Machine Learning with R”, Packt Publishing, ISBN-13: 978-1782162148 (http://books.tarsoit.com/Machine%20Learning%20with%20R%20-%20Second%20Edition.pdf)
2. Eric Suess “Statistical Learning with R” (http://www.sci.csueastbay.edu/~esuess/stat6620/)
3. Данные 30000 пользователей-подростков социальной сети Facebook https://github.com/Vasilesk/sphere/blob/master/extra/snsdata_clustering/snsdata.csv)
4. Формальная постановка задачи кластеризации, понятие алгоритма (http://www.machinelearning.ru/wiki/index.php?title=%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F)
5. Clustering - Silhouette Coefficient (http://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient)
6. Tasks to prepare data for enhanced machine learning (https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/prepare-data)
7. TF-IDF (http://www.tfidf.com/)
8. Clustering - K-means (http://scikit-learn.org/stable/modules/clustering.html#k-means)
9. The Expectation Maximization Algorithm (http://www.cs.cmu.edu/~dgovinda/pdf/recog/EM_algorithm-1.pdf)