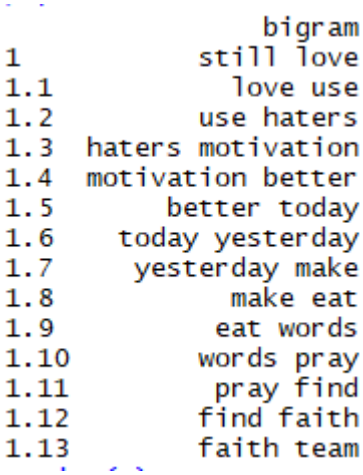
Исходная коллекция данных представляет собой данные о пользователях социальной сети Twitter, которые включают в себя информацию и описание профилей, а также одну из последних публикаций и информацию о ней для каждого из пользователей. Полнотекстовые данные, а именно поле text, в котором содержится текст публикации, считается содержимым документа, а прочие собранные сведения – метаданными.

Часть 1. Представление документов

Был разработан скрипт на языке R, представляющий исходную текстовую коллекцию в вероятностной, векторной и графовой моделях.

Вероятностная модель: исходные текстовые данные представляются в виде множества всевозможных би-грамм слов каждого текста. Было решено использовать би-граммы, так как сообщения в Twitter являются небольшими текстами ввиду ограниченной размерности.



	bigram
1	still love
1.1	love use
1.2	use haters
1.3	haters motivation
1.4	motivation better
1.5	better today
1.6	today yesterday
1.7	yesterday make
1.8	make eat
1.9	eat words
1.10	words pray
1.11	pray find
1.12	find faith
1.13	faith team

Рисунок 1 – Пример документа, представленного в вероятностной модели

Для примера документа, представленного в вероятностной модели, используется следующий изначальный текст: «`still love use haters motivation better today yesterday make eat words pray find faith team`».

Векторная модель: документы представляются в виде вектора, длина которого равна размеру словаря корпуса документов. На основе этого создается матрица терминов-документов, в которой в качестве веса используется мера TF-IDF (Term Frequency Inverse Document Frequency). Таким образом весь корпус текстов представляется в виде одного документа,

содержащим матрицу терминов-документов. Часть данной матрицы представлена на рисунке 2.

Docs	Terms feel	hope	kiddo	yesterday	your
1	0.000000	0.000000	0.000000	0.5982827	0.000000
10	0.000000	0.000000	0.000000	0.0000000	0.000000
2	1.002168	1.077341	2.169772	0.0000000	1.455539
3	0.000000	0.000000	0.000000	0.0000000	0.000000
4	0.000000	0.000000	0.000000	0.0000000	0.000000
5	0.000000	0.000000	0.000000	0.0000000	0.000000
6	0.000000	0.000000	0.000000	0.0000000	0.000000
7	0.000000	0.000000	0.000000	0.0000000	0.000000
8	0.000000	0.000000	0.000000	0.0000000	0.000000
9	0.000000	0.000000	0.000000	0.0000000	0.000000

Рисунок 2 – Часть терм-документной матрицы для корпуса текстов

Графовая модель: текст представляется в виде графа, где вершинами являются термины текста, а ребра – это связи между ними. Связи определяются в зависимости от того, стоят ли данные термины рядом друг с другом в тексте или нет. Если связей несколько, то увеличивается вес ребра. Граф задается матрицей.

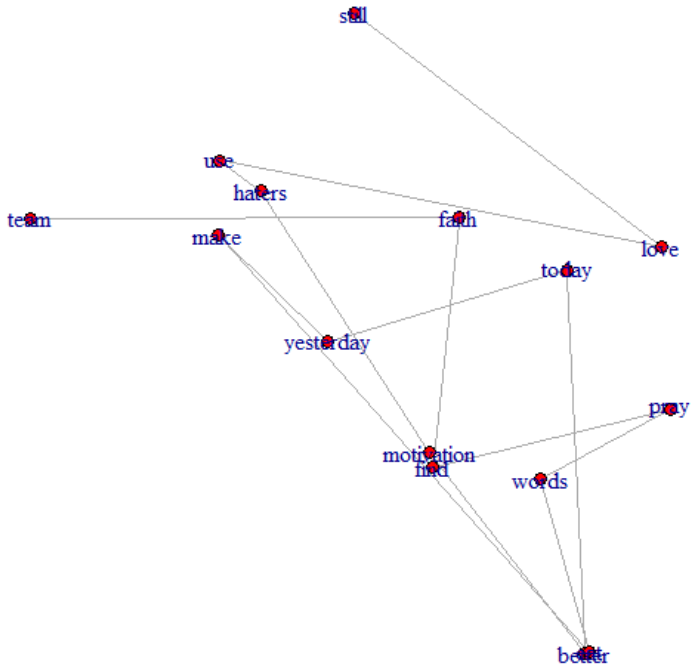


Рисунок 3 – Пример документа представленного в графовой модели

Часть 2. Кластеризация документов

Цель кластеризации: разделение корпуса текстов на кластеры в зависимости от их схожести (по тематике и структуре).

Было решено использовать алгоритм кластеризации DBSCAN с заранее неизвестным количеством кластеров. Для этого построенные графовые модели текстов были преобразованы в терм-документную матрицу, где в качества терминов выступают связанные между собой в графе слова (н-р, «still love», «love you» и т.д.).

Этот алгоритм кластеризации, основан на плотности — если дан набор точек в некотором пространстве, алгоритм группирует вместе точки, которые тесно расположены, помечая как шум точки, которые находятся одиноко в областях с малой плотностью (ближайшие соседи которых лежат далеко).

Алгоритм DBSCAN:

1. Находим точки в ε окрестности каждой точки и выделяем основные точки с более чем minPts соседями.
2. Находим связные компоненты основных точек на графе соседей, игнорируя все неосновные точки.
3. Назначаем каждую неосновную ближайшему кластеру, если кластер является ε - соседним, в противном случае считаем точку шумом.

В результате была проведена кластеризация всех исходных документов коллекции. Но сначала была проведена дополнительная очистка данных:

- удалены специальные символы и окончания слов (н-р, ‘ve, ‘s и т.д);
- удалены знаки пунктуации;
- удалены все цифры;
- удалены стоп-слова английского языка;
- все буквы переведены в нижний регистр;
- удалены лишние пробелы между терминами;
- произведен стемминг всех терминов.

Терм-документная матрица была преобразована к весам TF-IDF меры. Также для проведения корректной кластеризации были удалены все документы, в которых нет терминов (пары слов), это документы коллекции, которые изначально состояли из одного слова или преобразовались в такие из-

за предобработки. Были удалены все разряженные термины коллекции (порог разреженности равен 0,999).

Для проведения кластеризации функции dbscan необходимо подать на вход матрицу расстояний между объектами, а также значения коэффициентов eps = 0,3 и minPts = 10. Матрица расстояний была вычислена с использованием косинусного расстояния и евклидова.

При просмотре результатов, кластеризация на основе матрицы косинусного расстояния показала лучшие результаты, было выделено больше кластеров, а также меньшее количество шумовых точек.

Для оценки качества кластеризации было решено использовать внутреннюю метрику оценки качества – силуэт, которая показывает насколько объект похож на свой кластер по сравнению с другими кластерами. Значение данной метрики лежит между -1 и 1, чем ближе данная оценка к 1, тем качество разбиения лучше.

Среднее значение ширины силуэта для кластеризованных данных равно 0,97. Это говорит о хорошем качестве полученного разбиения текстов коллекции алгоритмом кластеризации dbscan.

Часть 3. Классификация документов

Целью классификации является разделение корпуса текста по трем классам: негативные, нейтральные и позитивные тексты. Так как коллекция документов состоит из твиттов, то тексты небольшие и зачастую достаточно сложно конкретно определить их тематику, зато эмоциональная составляющая сразу прослеживается.

Таким образом изначальная коллекция из 16597 документов была классифицирована следующим образом: используется готовый словарь эмоционально окрашенных терминов, оценивается настроение пары слов (графовая модель), на его основе каждому тексту присваивается определенная метрика. Если значение метрики < 1, то текст негативный, если равен 0, то нейтральный, при метрике > 1 текст имеет позитивную окрашенность.

Была выполнена предобработка данных, также, как и в части кластеризации.

Таблица 1 – Результаты классификации

Эмоциональная окраска	Количество документов
Негативная	4335
Нейтральная	4836
Позитивная	7426

Для оценивания качества классификации вручную были просмотрены порядка 1000 документов и их эмоциональная окраска согласно проведенной классификации, почти все результаты были корректными.

Также для оценки качества классификации были посчитаны ошибки 1-го (опровержение верной гипотезы) и 2-го (принятие ложной гипотезы) рода:

- ошибки 1-го рода: $\frac{5}{100} = 0{,}05$;
- ошибки 2-го рода: $\frac{6}{100} = 0{,}06$.