САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Направление: 01.03.02 «Прикладная математика, фундаментальная информатика и программирование»

Кафедра технологии программирования

**ОТЧЕТ О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ**

**Тема задания: Методы анализа пользовательской активности в социальной сети Reddit**

**Выполнил (а):** Камаев Кирилл Владимирович, группа 19-Б08-пу

**Руководитель научно-**

**исследовательской работы:** Блеканов Иван Станиславович, кандидат технических наук, доцент кафедры технологии программирования, заведующий кафедрой технологии программирования

Санкт-Петербург

2021

[**Введение**](#_ph81v9sy9zyq) **3**

[Актуальность](#_fvg37g6biw6c) 4

[Цель курсовой работы](#_tyr8pjen7mdm) 5

[Задачи](#_rrq7awaerhyl) 5

[**Анализ и построение социальных графов**](#_azpi3s31gly) **5**

[1.1 Этапы построения социального графа](#_m8gabdp4mroa) 6

[1.2 Сбор данных](#_y3zmkkvkelwt) 6

[1.3 Выявление наиболее влиятельных пользователей (простой способ)](#_jeijmrguim07) 7

[1.4 Построение связей между сабреддитами, как способ выявление лидеров мнений](#_5bzmm2tpimya) 8

[1.5 Промежуточный итог](#_wvv1e3w80c3g) 11

[**Анализ постов и комментариев пользователей**](#_n5ux6gnrwwtk) **11**

[2.1 Сбор данных и обработка](#_3atabnbwko3q) 12

[2.2 Построение облака слов](#_uguqef28c1tu) 13

[2.3 Промежуточный итог](#_is1iq3erm4zx) 15

[**Заключение**](#_qfu6vhtdflsq) **16**

[Результаты работы](#_u961km4yvff0) 16

[**Список используемых источников**](#_yhfxqgasfas2) **16**

# Введение

По мере того как социальные сети и пользовательский контент, который их заполняет, продолжают расти в распространенности, размерах и влиянии, понимание того, как пользователи взаимодействуют и создают этот контент, становится все более важным. Понимание этой динамики сообщества может оказаться полезным для измерения доверия к контенту, предоставления групповых рекомендаций на основе ролей или оценки стабильности и роста группы.

Reddit был основан в июне 2005 года Стивом Хаффманом и Алексис Оганян и финансируется за счет как рекламы, так и платной подписки на Reddit Gold. Reddit стал проектом с открытым исходным кодом, размещенным на GitHub 18 июня 2008 года. В настоящее время Reddit нанимает 10 сотрудников на полную ставку. Сайт написан на Python с использованием веб-фреймворка Pylons. Трафик Reddit продолжает расти в последние годы, приближаясь к 100-му наиболее посещаемому сайту в Интернете во всем мире и в топ-50 в США (Alexa). На главной странице Reddit отображаются избранные пользователем или выбранные по умолчанию сабреддиты. Сабреддит (*subreddit*) - это портал сообщества, на котором размещаются отдельные материалы. Сами сабреддиты также создаются пользователями и охватывают широкий круг тем. При просмотре сабреддита отображается список представленных материалов, упорядоченных по дате отправки, популярности, спорности или общему наивысшему баллу за период времени. У каждого поста есть текстовый заголовок, оценка, основанная на голосовании пользователей за и против, либо ссылка на веб-страницу, либо текстовое описание, а также раздел с цепочкой комментариев для обсуждения пользователями Reddit. В разделе комментариев пользователи могут обсуждать отправленные материалы и голосовать за другие комментарии, позволяя фильтровать популярные комментарии вверх. [1, c. 1]

Reddit - это социальный новостной сайт, основанный на пользовательском контенте. Контент представляет собой либо ссылку на другую веб-страницу, либо текстовый блок, сопровождаемый заголовком. За каждое сообщение пользователи сайта голосуют, повышая или понижая его рейтинг. Пользователи также могут обсуждать представленные материалы в разделе комментариев, прикрепленном к каждому материалу. Reddit - это богатый данными сайт со сложной сетевой структурой. Каждый сабреддит функционально независим друг от друга, но иногда имеет общие материалы и даже происходит перекрестное размещение. Пользователи сайта определяют содержание и рейтинг содержания, а также масштабные обсуждения каждой статьи. Структура Reddit наиболее похожа на социальный новостной сайт Digg, разделяя почти все функции, за исключением видимости действий пользователя. [2, c. 1]

Анализ пользовательской активности социальной сети включает в себя: [3, c. 1-3]

1. Анализ и построение социальных графов
   1. Выявление лидеров мнений (далее инфлюенсеров)
   2. Построение связей между сабреддитами для выявление инфлюенсеров
2. Анализ постов и комментариев пользователей
   1. Построение облака слов
   2. Применение анализа для максимизации рейтинга поста

Также существуют множество других методов анализа социальной сети, способствующие выявлению структурных связей для применения их в таких отраслях как реклама и кибербезопасность. [8, c. 1-5]

## Актуальность

За последние два десятилетия новостные онлайн-сайты медленно, но неуклонно продвигались в направлении повышения вовлеченности пользователей посредством комментирования новостных статей и простого обмена статьями на сайтах социальных сетей, таких как Facebook и Twitter. Однако окончательное решение о том, какие статьи получают наибольшую известность благодаря размещению в верхней части главной страницы, по-прежнему остается исключительно за редакторами новостного сайта. Сайты социальных новостей устраняют этот последний остаток экспертного контроля, позволяя самим пользователям голосовать за то, какие посты заслуживают наибольшего внимания, и даже отправлять посты на эти сайты.

Сейчас социальные сети являются неотъемлемой частью нашей жизни, что означает огромное влияние на информацию, которую мы получаем и распространяем. Таким образом, появление необходимости в анализе пользовательской активности социальных сетей связано с повышением эффективности продвижение продукта компании и кибербезопасностью. Для этого было изобретено множество методов, в частности построение социального графа, рассматриваемого в данной работе.

## Цель курсовой работы

Ознакомиться с некоторыми методами анализа социальных сетей и оценить их эффективность на практике.

## Задачи

1. Собрать большой набор данных с Reddit. В данной работе я сосредоточусь на популярных разделах Reddit, чтобы найти больше активности пользователей и, следовательно, больше интересных данных.
2. Анализ Reddit на высоком уровне. Начиная с сабредетов и продолжая рассматривать сообщения и комментарии.
3. Определить роль каждого из методов в анализе и их эффективность.

# 

# 

# 

# 

# Анализ и построение социальных графов

Социальный граф (англ. Social graph) — это граф, узлы которого представлены социальными объектами, такими как пользовательские профили с различными атрибутами (например: имя, день рождения, родной город и т. д.), сообщества, медиа-контент и т. д., а ребра — социальными связями между ними.

Неявный социальный граф (англ. Implicit social graph) — это такой граф, который можно сформировать на основе взаимодействий пользователя со своими «друзьями» и группами «друзей» в социальной сети. В этом графе в отличие от обычного социального графа нет явного указания «друзей», то есть нет явных социальных связей.

С помощью социальных графов решают такие задачи, как: идентификация пользователей; социальный поиск; генерация рекомендаций по выбору «друзей», медиа-контента, новостей; выявление «реальных» связей или сбор открытой информации для моделирования графа.

Обработка данных социальных графов связана с рядом проблем, как например различия социальных сетей, закрытость социальных данных.

# 

## 1.1 Этапы построения социального графа

В данной работе я буду строить граф на примере сабреддита r/wallstreetbets. Шаги построения графа следующие [6]:

1. Определите активных пользователей субредита, узнав сабреддиты, в которых пользователь прокомментировал не менее 10 различных материалов за последние 6 месяцев.
2. Просуммировать количество ссылок между двумя сабреддитами; это будем весовым коэффициентом.
3. Отфильтровать полученный набор данных, удалив петли и обратные края. (например, если у нас есть askreddit → pics, удалить pics → askreddit). Кроме того, необходимо сохранять только ребра с не менее чем 200 активными пользователями, чтобы результирующий набор данных имел приемлемый размер для этого анализа.

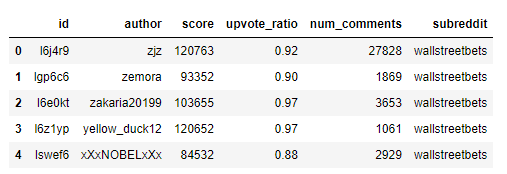
## 

## 1.2 Сбор данных

Для построения графа связей требуется множество данных о пользователях, и, в частности, о их постах. Для решения данной задачи используется библиотека PRAW.

PRAW - это оболочка Python для API Reddit, которая позволяет извлекать данные из сабреддитов. API Reddit выдает примерно один запрос в секунду, что кажется довольно разумным для небольших проектов.

Для данной работы я выбрал сабреддит r/wallstreetbets, как один из наиболее активных и быстроразвивающихся. Для построения графа были выбраны 250 (данное число обусловлено тем, что при больших значениях графики становятся практически нечитаемыми, а при меньших значениях выборка нерепрезентативна) наиболее популярных постов и их данные (автор, количество оценок и комментариев):



## 1.3 Выявление наиболее влиятельных пользователей (простой способ)

Для выявлениях наиболее влиятельных людей (инфлюенсеров) построим гистограмму распределения пользователей по количеству постов.[7,с.3]

Как можно заметить по гистограмме наиболее влиятельный пользователь - DeppFuckingValue, причем с большим отрывом от остальных. Если увеличить выборку до 1000 наиболее популярных постов распределение становится более репрезентативным:

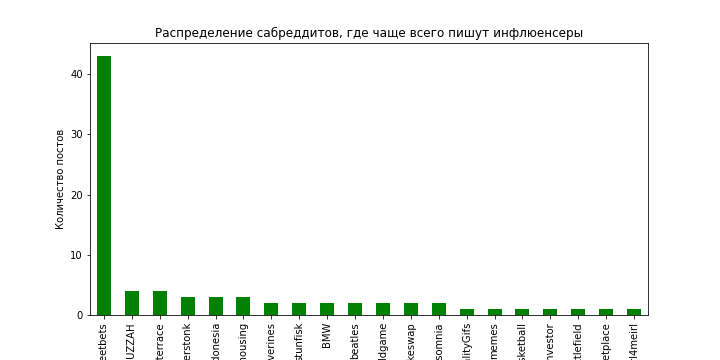
И уже он становится не самым активным пользователем. Поэтому данный метод не является наиболее эффективным, т.к требует больших данных. Но данное решение поможет дальше построить граф связи сабреддитов в качестве альтернативы. Таким образом, данный метод можно использовать для быстрого анализа, ввиду его быстрого подсчёта.

## 1.4 Построение связей между сабреддитами, как способ выявление лидеров мнений

Пользователи подписываются на сабреддиты, которые сосредоточены вокруг различных интересующих их тем или сообществ. Однако эти сабреддиты не организованы каким-либо систематическим образом, и пользователи Reddit обычно узнают о новых сабреддитах друг от друга, что означает, что из одного сабреддита они находят другой новый сабреддит. И один сабреддит влияет на другой сабреддит прямым или косвенным образом.

Мы можем определить наиболее влиятельных пользователей, определив важные узлы с помощью сетевого анализа. Для этого можно использовать библиотеку NetworkX для создания графиков и просмотра того, как пользователи влияют на другие сабреддиты, и что общего у всех сабреддитов. [7]

Для этого можно взять 10 самых популярных постов пользователей, которые мы получили из предыдущих шагов, и посмотреть на распределение количества постов, которые они чаще всего пишут на сабреддитах:



Очевидно, что наибольшее количество постов будет на изначальном сабреддите r/wallstreetbets, на котором мы искали пользователей. Далее идут уже другие, на которых чаще всего инфлюенсеры пишут или ссылаются.

Для понимание связи сабреддитов необходимо рассмотреть граф их связей. [6] Из построения понятно, что r/wallstreetbets будем иметь максимальную степень вершины (а именно 7), в том время как среднее значение по всем равно 1.84. Общее количество вершин - 49, а рёбер - 49, согласно данной таблице ниже:

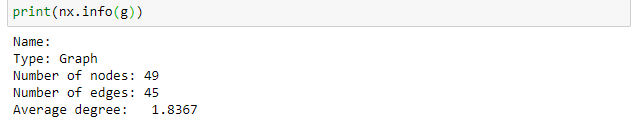
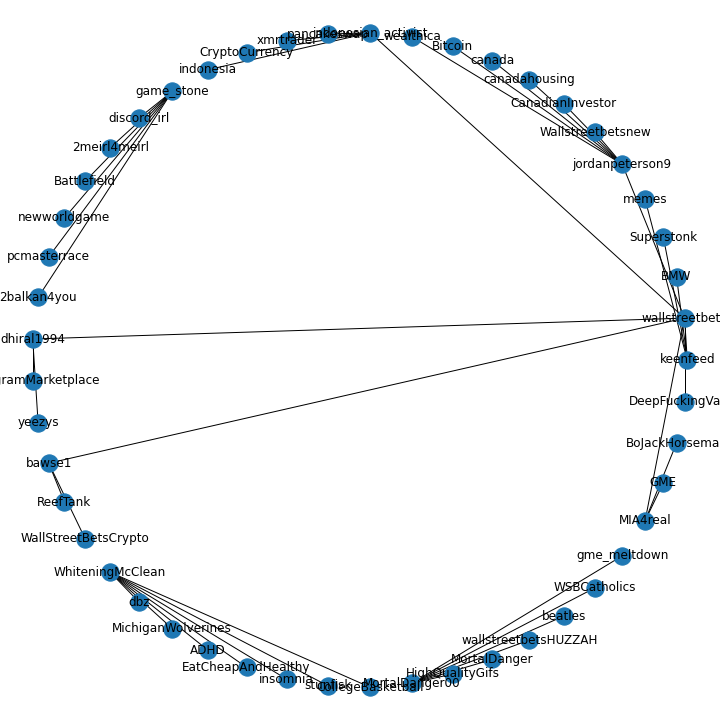
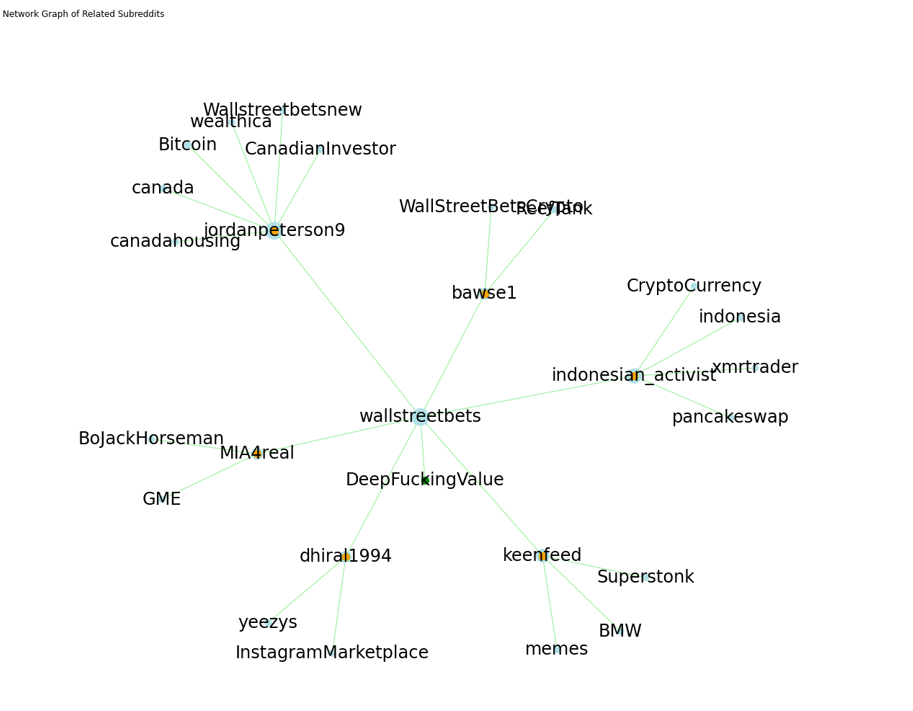


График ниже показывает как сабреддиты связаны друг с другом.



И наконец мы можем построить граф взаимодействия сабреддитов. Узлы влияния отображаются зеленым цветом. Между тем, сабреддиты отображаются светло-голубым цветом и имеют размер в соответствии с соответствующими степенями вершин, которые есть у каждой из них. Инфлюенсеры, у которых связи не только с r/wallstreetbets, выделены оранжевым цветом:



Из приведенных выше графов и анализа можно сказать, что 'DeepFuckingValue', 'indonesian\_activist', 'MIA4real', 'keenfeed' являются одними из самых влиятельных пользователей, которые оказывают влияние на r/wallstreetbets. Существует множество других сабреддитов, которые связаны друг с другом пользователями. Сабреддиты, расположенные ближе друг к другу на этой диаграмме, с большей вероятностью имеют пользователей, которые комментируют их оба. Также гистограмма в нашем анализе показала, насколько часто пользователи r/wallstreetbets комментируют посты в других субреддитах. Пользователи, которые читают /r/wallstreetbets, также читают GME, Bitcoin, CryptoCurrency, memes, superstonk и т.д.

Наша выборка из 250 лучших постов в сабреддите может иметь нереалистичные результаты, так как данные не являются последовательными и в них может быть перекос, поэтому одни и те же инфлюенсеры могут появляться слишком часто.

## 1.5 Промежуточный итог

Для выявление лидеров мнений было проанализировано 250 постов с сабреддита r/wallstreetbets. Этого недостаточно для полноценного анализа и выявления лидеров мнений, так как выборка получается нерепрезентативной. Но хватает, чтобы рассмотреть работу методов на их примере. Первый метод показал, что пользователь с ником DeepFuckingValue является наиболее влиятельным, второй метод это подтвердил, так как на графе он был ближе всего. Так получилось, потому что этот пользователь является администратором, поэтому в целом этот результат можно считать успешным в частности. Например, если мы хотим купить рекламу только на одном сабреддите, то будет логично обратиться к администратору. Но обычно требуется, чтобы реклама была более обширной, поэтому стоит воспользоваться методом построения графа социальных связей для выявления настоящих лидеров мнений.

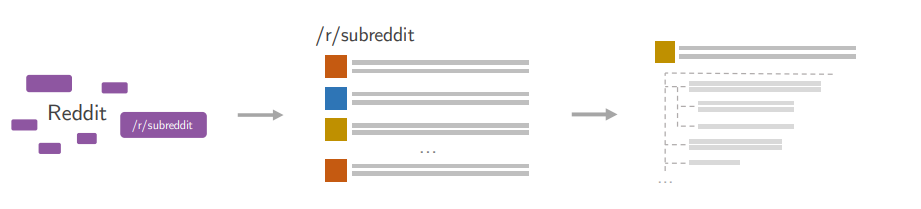
# Анализ постов и комментариев пользователей

В Reddit пользователи могут отправлять контент в виде текстов, изображений и ссылок на внешние ресурсы. Отправленный контент (также называемый просто постами) может быть прочитан другими пользователями и обсужден в комментариях. Пользователи могут подписаться на несколько сабреддитов, чтобы получать последние опубликованные материалы на своих первых страницах. Важной особенностью Reddit является голосование, которое представляет собой механизм, влияющий на видимость и рейтинг как постов, так и комментариев. На самом деле, пользователям разрешено голосовать за или против сообщения других пользователей, чтобы каждая публикация имела оценку. Это показатель, основанный на разнице между количеством положительных и отрицательных голосов, и он существенно влияет на порядок показа

постов и комментариев пользователям. Однако точное количество голосов за и

против не отображается публично.

Таким образом, мы можем проанализировать текст лучших постов, чтобы использовать эти данные для максимизации голосов “за” уже на своих постов. Например, для этого можно использовать похожую лексику. То же самое касается и комментариев, поэтому можно считать, что посты и комментарии в данном контексте это одно и то же, поскольку обладают похожей структурой “заголовок-тело-оценка”:



## 2.1 Сбор данных и обработка

Чтобы делать вывод о взаимосвязи использовании конкретных слов и рейтинга, необходимо собрать множество данных. К счастью, мне удалось найти датасет на 4 миллиона записей (2 миллиона комментариев с положительным рейтинг и столько же с негативным). В общем сумме датасет весит около 2 гигабайт. Этого должно хватить для полноценного анализа. [4]

Далее необходимо произвести обработку каждой записи по следующей схеме:

1. Разбить каждую запись на отдельные слова, удалив знаки препинания, преобразовав это в вектор слов для удобства дальнейшей обработки.
2. Привести все слова к нижнему регистру, потому что, например, для алгоритма слова Reddit и reddit будут разными в силу того, что заглавная и строчная буквы имеют разную кодировку.
3. Произвести лемматизацию каждого слова. Лемматизация — это алгоритмический процесс нахождения леммы слова в зависимости от его значения. Лемматизация обычно относится к морфологическому анализу слов, целью которого является удаление флективных окончаний. Это помогает в возвращении базовой или словарной формы слова, которое известно как лемма.
4. Удаление стоп слов. Стоп слова - это шумовые слова, не несущие смысла при алгоритмической обработки. К ним можно отнести предлоги, суффиксы, причастия, междометия, цифры, частицы и т. п.

Также, вместо лемматизации на 3 шаге можно использовать стемминг. Стемминг - это своего рода нормализация слов, которая заключается в том, чтобы обрезать слово до корня. Но выбор пал на лемматизацию, потому что она является более мощной операцией, которая учитывается морфологический анализ слов, то есть использует более глубокие лингвистические знания. Следовательно у нас появляется больше возможностей для анализа.

## 2.2 Построение облака слов

Облака слов или облака тегов - это графические представления частоты слов, которые выделяют слова, чаще встречающиеся в исходном тексте. Чем крупнее слово в визуализации, тем чаще оно встречается в тексте. Этот тип визуализации может помочь специалистам по оценке в проведении исследовательского текстового анализа путем выявления слов, которые часто встречаются в наборе интервью, документов или других текстов. Он также может быть использован для передачи наиболее важных моментов или тем на этапе подготовки отчета.

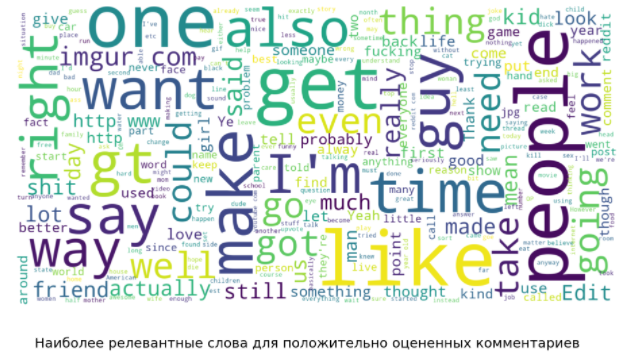
Облака слов не будут точно отражать содержание текста, если для одной и той же идеи используются немного разные слова (например, "большой", "огромный", "гигантский"). Они также не обеспечивают контекст, поэтому значение отдельных слов может быть утеряно. Из-за этих недостатков облака слов лучше всего подходят для исследовательского качественного анализа. И поэтому данный метод требует тщательной подготовки наших данных (алгоритм подготовки описан выше).

Алгоритм построение облака слов заключается в нескольких этапах:

1. Посчитать слова
2. Убрать нерелевантные слова
3. Отсортировать их по частоте встречаемости в тексте в убывающем порядке
4. Взять первые N слов
5. Визуализировать их на двумерном графике, размер шрифта слова будет пропорционально зависеть от его частоты в тексте, как и его расположение (ближе к центру будут наиболее частотные слова).

Для построение облака слов я буду использовать библиотеку wordcloud, а также nltk для фильтрации слов, которые привносят мало смысла для контекста. Во всех построениях я буду полагать параметр N равным 500. [5]

Для начала рассмотрим комментарии, рейтинг которых больше 200, и построим для них облако слов.

Как можно заметить, нет какого-то особенного слова для того, чтобы получить больше положительных оценок. Больше всего выделяются слова с положительной семантикой (“want”, “like” и т.п.), а остальные имеют нейтральный окрас. Из этого следует, что стоит использовать положительно-нейтрально окрашенную лексику. Тогда существует вероятность, что пост понравится людям больше. Данная гипотеза подтверждается рассмотрением отрицательно оцененных комментариев.

Повторим процедуру для постов уже с рейтингом меньше -200:



Для данного облака слов характерны слова с негативным окрасом. В частности, это слова направленные на выражение расизма, гомофобии, сексизма, сарказма и прочего социально неприемлемого контента. Однозначно можно сказать, что Reddit - это площадка, на которой общество способно само модерировать контент и снижать в выдаче комментарии, которые могут кого-то оскорбить.

## 2.3 Промежуточный итог

Построение облака слов помогает лучше сориентироваться в новом контенте, понять какие слова больше всего использует сообщество и возможно определить ключевые слова, определяющие его трендовое настроение. В частности, я интерпретировал облака слов в контексте анализа пользовательской активности для того, что определить какие слова могут помочь косвенно максимизировать вероятность получение высокого. К таким косвенным фактором также можно отнести время, когда происходит размещение, и рейтинг, который получает пост в первые минуты, так как сообщество склонно лавинообразно оценивать, повторяя друг за другом.

# Заключение

## Результаты работы

В данной курсовой работе мне пришлось собрать множество данных о пользователях социальной сети Reddit. Было проанализировано множество записей о них. Первый метод, который я выделил для рассмотрения, - построение социального графа пользователей. Данный метод позволяет нам определить важных для сообщества пользователей, которые определяют весь трендовый контент (также называемые инфлюенсеры или лидеры мнений), а также сабреддиты, на которые они способны перенаправлять траффик. Это является очень важной частью определения социальной иерархии пользователей, в частности их социальных ролей (создатель контента и потребитель). В реальности данный анализ применяют маркетологи, чтобы определить влиятельных людей для максимизации эффективности продвижения продукта. Далее по иерархии конструкции сайта я рассматривал комментарии пользователей, а именно как определенные слова могут влиять на конечный рейтинг комментария. Для такого анализа я выбрал метод построения облака слов, который позволяет определить наиболее частотные слова в предложениях. В результате оказалось, такой подход является косвенным, но всё-ещё актуальным для анализа. В комментариях с положительной оценкой авторы чаще всего использовали положительно окрашенную лексику, а в комментариях с отрицательной оценке чаще всего оказывались слова с резко негативным оценочным окрасом. Таким образом, можно сделать вывод об общем настроении сообщества. Reddit является дружелюбным сообществом, в противоположность, например, таким токсичным сообществам как 2ch. [8] Такой подход позволяет выбирать аудиторию, подходящую данному продукту. Также, его можно использовать для выявления кибербуллинга и экстремизма в сообществе в целом, что нельзя сделать, если анализировать конкретного человека.

Резюмируя всё вышенаписанное, можно сказать, что анализ социальных сетей является мощным инструментом, включающим себя множество интерпретируемых методов.

# Список литературы

1. About Reddit [wikipedia] URL: ]<https://ru.wikipedia.org/wiki/Reddit>
2. Troy Steinbauer, Information and social analysis of Reddit [stanford.edu] URL: <http://snap.stanford.edu/class/cs224w-2011/proj/tbower_Finalwriteup_v1.pdf>
3. Jasneek Chugh, “How influencers on Reddit form a network of related subreddits?” [medium.com] URL:<https://medium.com/web-mining-is688-spring-2021/how-influencers-on-reddit-form-a-network-of-related-subreddits-30884134c137>
4. Evan Hallmark, Данные для второго метода. URL: <https://www.kaggle.com/ehallmar/reddit-comment-score-prediction/code>
5. Sachin Thukral, “Analyzing behavioral trends in community driven discussion platforms like Reddit” [arxiv.org]. URL: <https://arxiv.org/pdf/1809.07087.pdf>
6. Max Woolf, “How to Create a Network Graph Visualization of Reddit Subreddits” [minimaxir.com]. URL: <https://minimaxir.com/2016/05/reddit-graph/>
7. Cody Buntain, “Identifying Social Roles in reddit Using Network Structure” [sci-hub.ru]. URL: <https://sci-hub.ru/10.1145/2567948.2579231>
8. Смирнова О.С., Петров А.И., Бабийчук Г.А., “Основные методы анализа, используемые при исследовании социальных сетей” [cyberleninka.ru]. URL: <https://cyberleninka.ru/article/n/osnovnye-metody-analiza-ispolzuemye-pri-issledovanii-sotsialnyh-setey/viewer>

Код для практического использования методов:

1. Построение социальных графов. URL: <https://colab.research.google.com/drive/1FZxiwiGWAzrb9HyUK0rBaQvOd7m3R4bw?usp=sharing>
2. Построение облака слов. URL: <https://colab.research.google.com/drive/1Jz3Kntyvu47s1wnhW93gNEwYrVhOIo8f?usp=sharing>