Федеральное государственное образовательное бюджетное учреждение высшего образования

**«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»**

Кафедра анализа данных и машинного обучения

Факультета информационных технологий и анализа больших данных

Дисциплина «Машинное обучение в семантическом и сетевом анализе»

Направление подготовки «Прикладная математика и информатика»

Профиль «Прикладное машинное обучение»

**Курсовая работа на тему:**

«Построение онтологических эмбеддингов графов знаний»

Выполнил:

студент группы ПМ21-2 Зайцев А.В.  ~~(~~\_по\_д\_пи\_сь\_)

Руководитель:

к.э.н., доцент

Иванюк В.А.

(подпись)

**Москва 2024**

**СОДЕРЖАНИЕ**

Оглавление

[ВВЕДЕНИЕ 3](#_Toc164129276)

[I. Теоретические основы 3](#_Toc164129277)

[1. Онтологические эмбеддинги: определение, принцип работы 3](#_Toc164129278)

[2. Графы знаний: основные понятия, виды, применение 4](#_Toc164129279)

[3. Применение онтологических эмбеддингов графов знаний 5](#_Toc164129280)

[II. Создание графа знаний 5](#_Toc164129281)

[1. Выбор набора данных для построения графа 5](#_Toc164129282)

[2. Предварительный анализ и очистка данных 6](#_Toc164129283)

[3. Преобразование данных 7](#_Toc164129284)

[4. Формирование графа знаний и векторное представление. 9](#_Toc164129285)

[III. Построение онтологических эмбеддингов графа знаний 13](#_Toc164129286)

[1. Настройка параметров модели и обучение. 13](#_Toc164129287)

[2. Оценка качества модели. 14](#_Toc164129288)

[3. Представление результатов обучения в наглядном виде и анализ полученных выводов. 15](#_Toc164129289)

[ЗАКЛЮЧЕНИЕ 19](#_Toc164129290)

[СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ 20](#_Toc164129291)

# ВВЕДЕНИЕ

В современном информационном обществе огромное количество данных генерируется и распространяется каждую секунду. От поисковых запросов и социальных медиа до биомедицинских и научных исследований, данные охватывают различные области знаний и информационные источники. Однако эффективное использование этого огромного объема информации представляет собой сложную задачу из-за ее разнообразия, неструктурированности и часто неоднозначности.

В этом контексте графы знаний и онтологии играют ключевую роль в организации и представлении знаний. Они позволяют моделировать связи и отношения между сущностями в информационном пространстве, создавая более структурированный и понятный формат для анализа и использования данных. Однако для эффективного использования графов знаний необходимо обеспечить их эмбеддинги – компактные векторные представления графов, которые сохраняют семантическую информацию о сущностях и их взаимосвязях.

В данной курсовой работе рассматривается тема построения онтологических эмбеддингов графов знаний, которая представляет собой актуальное исследовательское направление в области машинного обучения и искусственного интеллекта. Целью работы является изучение существующих методов и подходов к созданию эмбеддингов для графов знаний.

# 

# Теоретические основы

## 1. Онтологические эмбеддинги: определение, принцип работы

Онтологические эмбеддинги - это мощный инструмент для преобразования сложных структур знаний, таких как графы знаний или онтологии, в непрерывные векторные пространства. В основе этого подхода лежит идея представления сущностей и их отношений в виде векторов фиксированной размерности. Эти вектора сохраняют семантическую информацию о связях между сущностями и позволяют выявлять и анализировать сходства и различия между ними.

Принцип работы онтологических эмбеддингов заключается в том, чтобы каждой сущности и отношению в графе знаний сопоставить векторное представление в многомерном пространстве. Эти представления строятся таким образом, чтобы схожие сущности имели близкие векторы, а семантически связанные сущности - близкие расположения в пространстве. Таким образом, даже сложные связи и паттерны между сущностями могут быть захвачены и сохранены в виде распределенных представлений.

Онтологические эмбеддинги находят применение во многих областях, включая информационный поиск, семантический анализ, обработку естественного языка и машинное обучение. Они позволяют решать разнообразные задачи, такие как классификация, кластеризация, сопоставление сущностей, предсказание отношений и многое другое, используя мощные методы и алгоритмы работы с векторными данными.

Таким образом, онтологические эмбеддинги представляют собой эффективный инструмент для анализа и работы с комплексными структурами знаний, открывая новые возможности для исследования и применения в различных областях науки и технологии.

## 2. Графы знаний: основные понятия, виды, применение

Граф знаний - это структура данных, которая позволяет организовать и представить знания в виде сущностей (узлов) и их взаимосвязей (ребер). Он играет важную роль в организации и анализе информации в различных областях, таких как информационные технологии, наука о данных, биомедицинская информатика и многие другие.

Основные понятия в графах знаний включают узлы (сущности), которые могут представлять собой объекты, концепции, события или факты, и ребра (отношения), которые определяют связи между этими сущностями. Кроме того, в графах знаний могут также присутствовать атрибуты и метаданные, которые расширяют информацию о сущностях и их отношениях.

Графы знаний могут быть направленными, где отношения имеют определенное направление, или ненаправленными, где отношения симметричны. Они также могут быть взвешенными, если ребрам присваиваются числовые значения или атрибуты, и невзвешенными, если ребра не имеют дополнительной информации.

Применение графов знаний охватывает широкий спектр областей. Они используются для создания баз данных знаний, где знания организованы и хранятся в виде графа для более эффективного доступа и анализа. Графы знаний также применяются в информационном поиске, где они помогают находить связи между сущностями и предоставлять более релевантные результаты. Кроме того, они используются в социальных сетях для моделирования связей между пользователями, а также для рекомендации контента.

Таким образом, графы знаний играют важную роль в представлении, анализе и использовании знаний в различных областях, и их применение продолжает расширяться и развиваться.

## 3. Применение онтологических эмбеддингов графов знаний

Онтологические эмбеддинги графов знаний представляют собой мощный инструмент, который находит широкое применение в различных областях исследований и промышленных приложений. Их способность преобразовывать сложные структуры знаний в векторные представления позволяет эффективно обрабатывать и анализировать данные, а также использовать их в методах машинного обучения и глубокого обучения.

В области информационного поиска онтологические эмбеддинги графов знаний позволяют улучшить качество поисковых запросов и рекомендаций путем учета семантических связей между терминами и сущностями. Это помогает пользователям быстрее находить нужную информацию и делать более точные выводы.

В анализе социальных сетей онтологические эмбеддинги могут использоваться для выявления сообществ, идентификации влиятельных узлов и прогнозирования динамики взаимодействия между участниками сети. Это позволяет лучше понимать структуру социальных связей и влияние различных факторов на формирование сообществ и распространение информации.

В биомедицинской информатике онтологические эмбеддинги могут быть использованы для анализа медицинских данных, выявления сходств между болезнями и лекарствами, а также для прогнозирования эффективности лечения и развития заболеваний. Это помогает улучшить диагностику и лечение пациентов, а также исследовать новые методы лечения и профилактики заболеваний.

Применение онтологических эмбеддингов графов знаний также включает поиск семантических аналогий, кластеризацию и классификацию сущностей в графах знаний. Это позволяет автоматизировать процессы анализа данных и делать более точные прогнозы на основе имеющихся знаний.

# Создание графа знаний

## Выбор набора данных для построения графа

Набор данных, используемый в курсовой работе, взят с платформы Kaggle1.

Этот набор данных содержит текстовые данные, извлеченные из сообщений с платформы Reddit, относящихся к различным религиям. Каждая строка в наборе данных представляет сообщение и включает в себя два основных столбца: «Document», содержащий текст сообщения, и «Subreddit», определяющий связанный религиозный текст.

Столбцы:

Документ: в этом столбце содержится текстовое содержание сообщений Reddit, связанных с различными религиозными дискуссиями, опытом и запросами.

Subreddit: в столбце указывается религиозная принадлежность или тема, связанная с каждым сообщением.

Датасет "Religion" является хорошим выбором для создания графа знаний по следующим причинам:

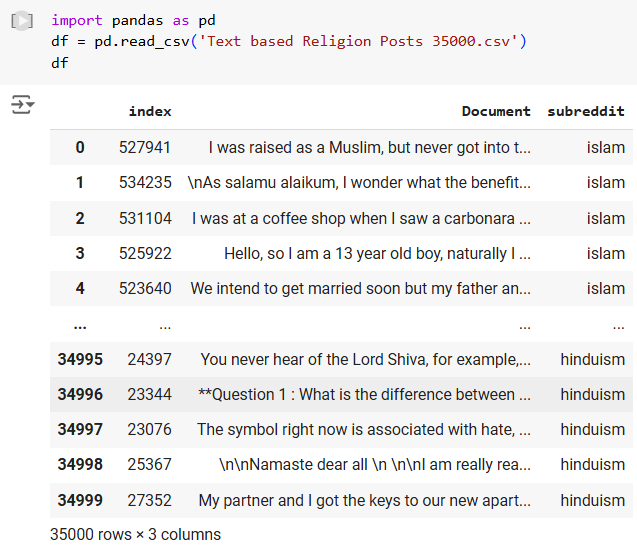
1. Разнообразие религиозных тем: Датасет содержит сообщения из различных религиозных сообществ, что позволяет создать граф, охватывающий широкий спектр религиозных убеждений и тематик. Это позволяет более полно охватить разнообразие мировоззрений и пониманий религиозных вопросов.

2. Текстовые данные: Содержание сообщений представлено в текстовом формате, что обеспечивает большой объем информации для анализа и обработки.

3. Соотнесение с соответствующими религиозными сообществами: Каждое сообщение в датасете помечено соответствующим подразделом (subreddit), что позволяет легко идентифицировать его связь с определенной религиозной темой или сообществом. Это облегчает построение графа, где узлы представляют собой сообщества или темы, а ребра - связи между ними на основе семантической или тематической близости.

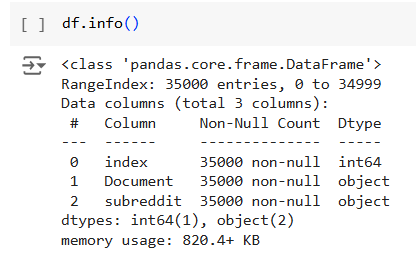
## Предварительный анализ и очистка данных

Представление данных в датасете:



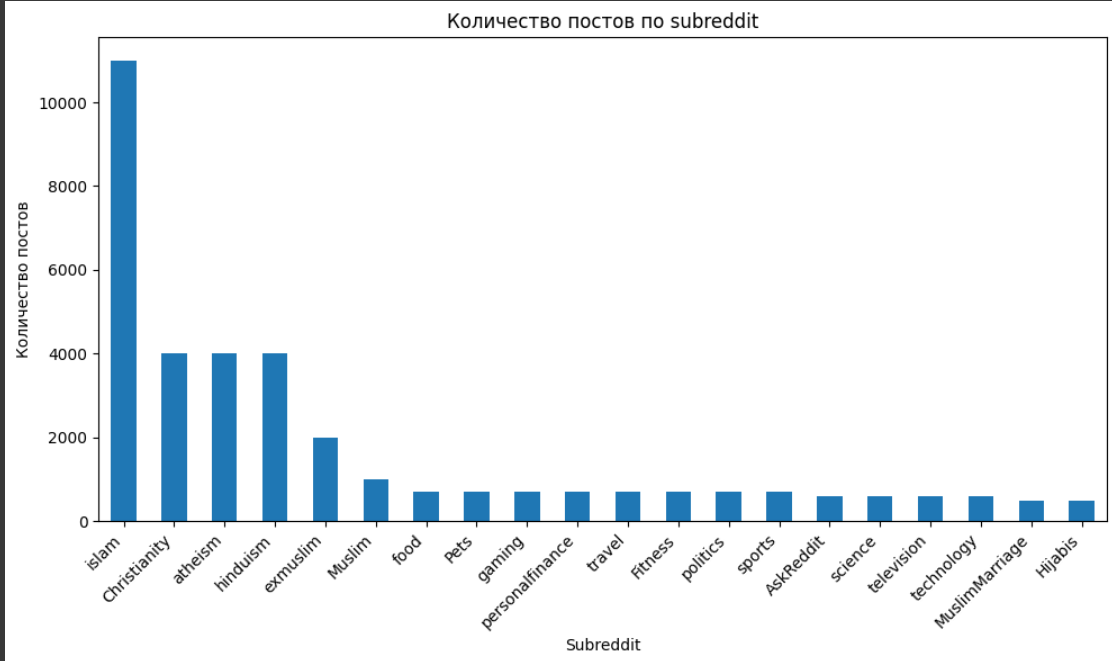
В каждом столбце датасета 35000 записей. По итогам анализа пропусков не было обнаружено.

Данные имеют следующие типы:



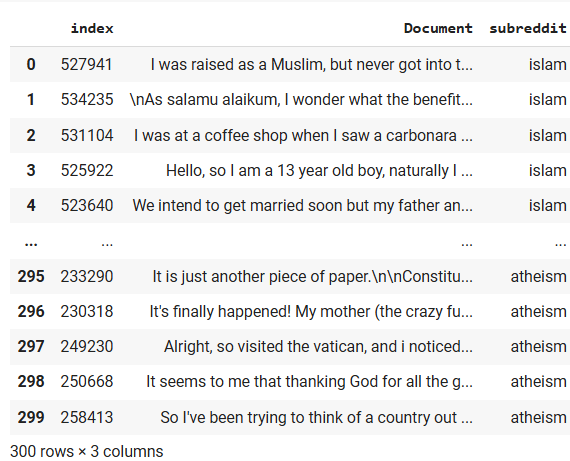
## Преобразование данных

Для создания графа знаний для начала я решил выделить 3 самых популярных «Subreddit»:

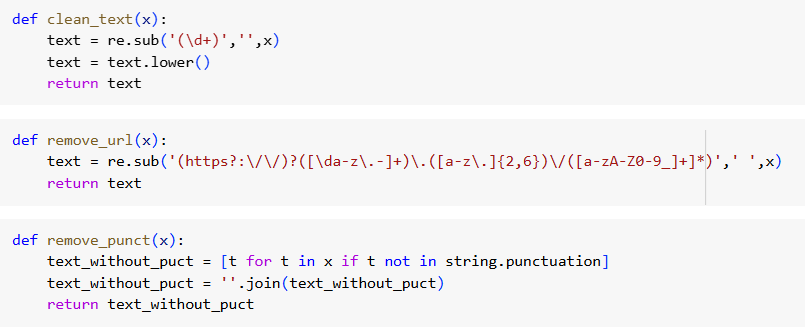


В датафрейм для графа знаний пойдут сообщения относящиеся к следующим “Subreddit”: islam, Christianity, atheism.

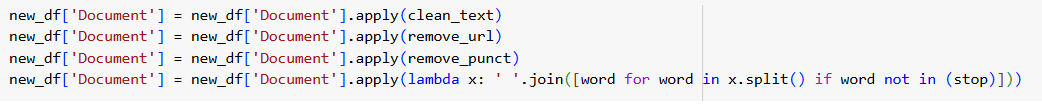
Далее создаю датасет, состоящий из сообщений, которые относятся к этим классам:



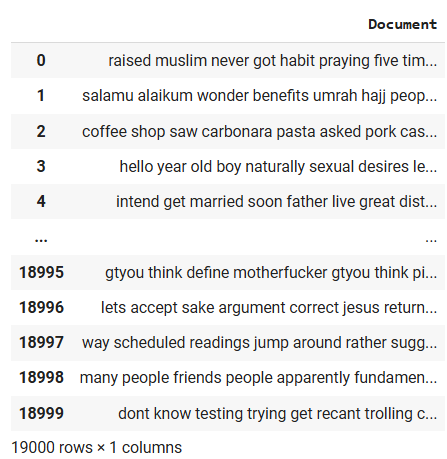
На следующем этапе нужно обработать данные, очистив текст от чисел, знаков пунктуации и ссылок, для этого я использовал следующие функции:



Далее применяем эти функции к датафрейму, а также удаляем «стоп слова».

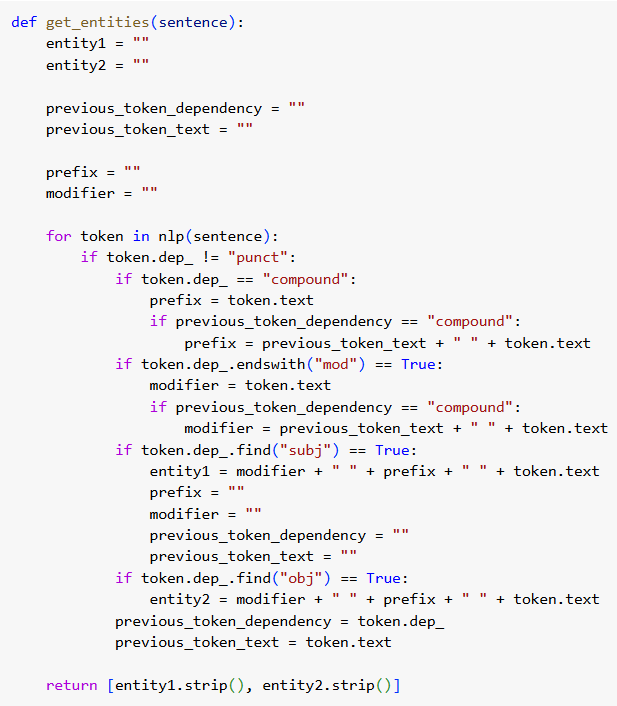


Датафрейм после преобразований:

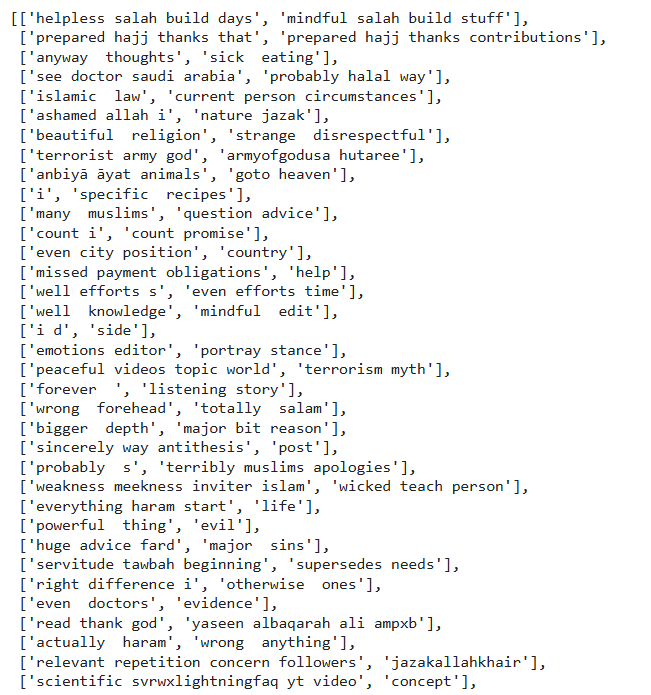


## Формирование графа знаний и векторное представление.

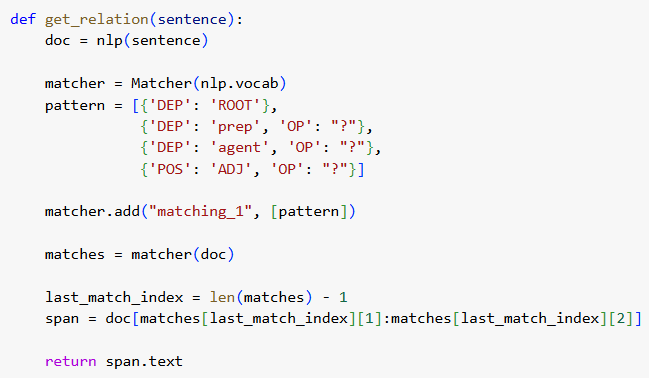
Для формирования графа знаний мы должны выделить из датафрейма именные сущности, используем для этого следующую функцию (Эта функция предназначена для извлечения двух сущностей (entity) из предложения. В ней используется библиотека SpaCy для анализа естественного языка):



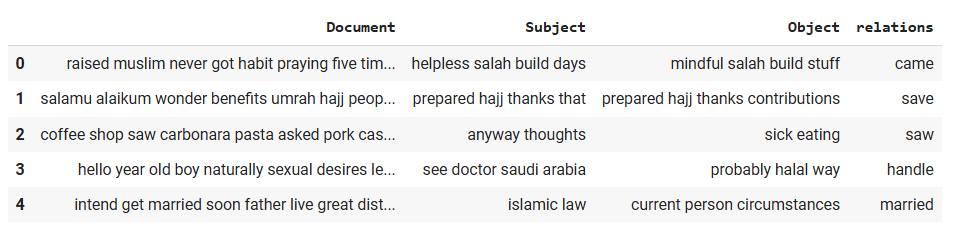
Применяя её к каждому предложению на выходе получаем список сущностей, которые из себя представляют объект и субъект в тексте сообщения:



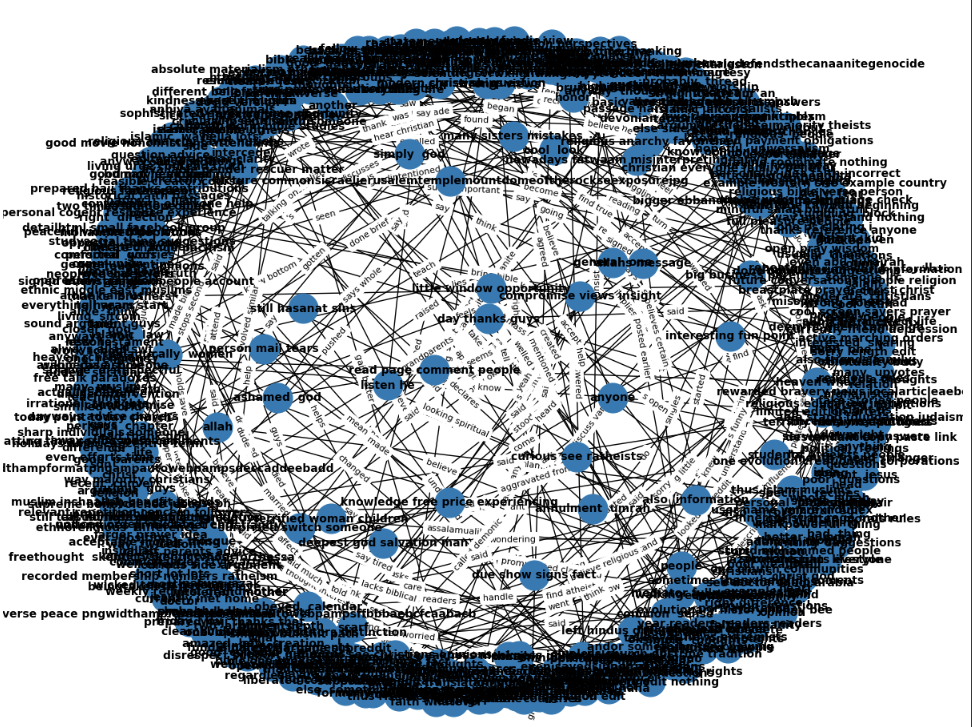
Следующим этапом будет выделение связей. Функция `get\_relation` использует библиотеку SpaCy для анализа естественного языка и поиска в предложении сущностей, соответствующих определенному шаблону зависимостей. Затем она извлекает последнее совпадение и возвращает текст этой сущности:



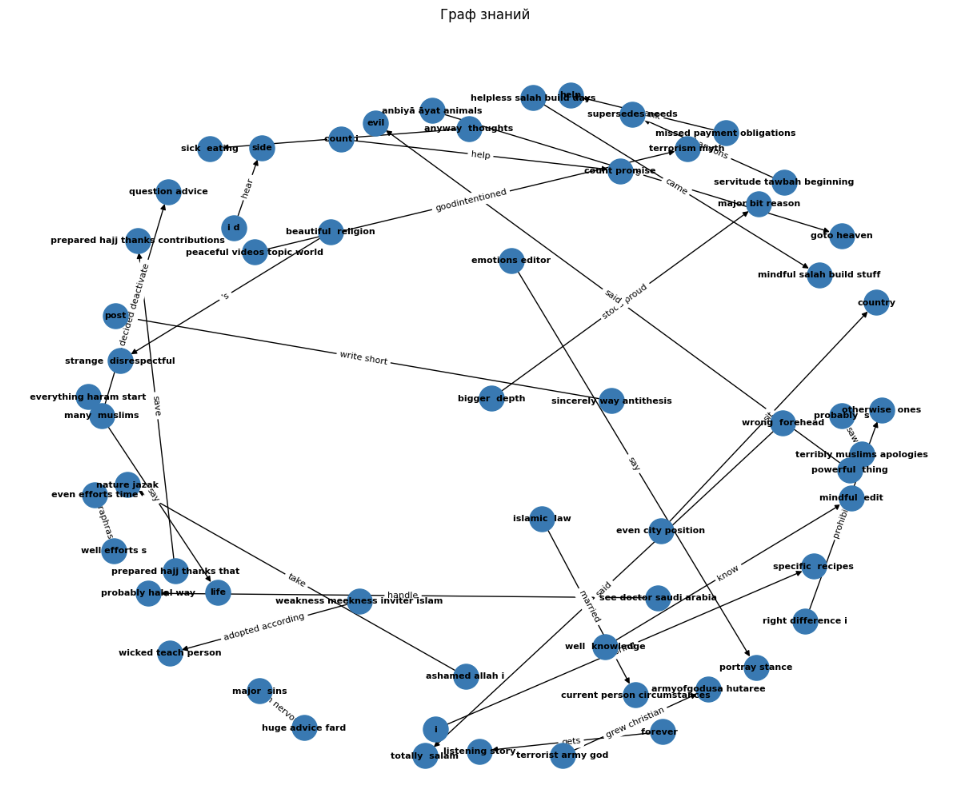
Далее добавляем сущности и связи в датафрейм и получаем готовые данные для построения графа знаний:



Теперь строим граф знаний, используя библиотеку NetworkX:



Граф с меньшим количеством вершин, чтобы наглядно были видны связи:



# Построение онтологических эмбеддингов графа знаний

## Настройка параметров модели и обучение.

Для обучения используется модель Node2Vec с гиперпараметрами:

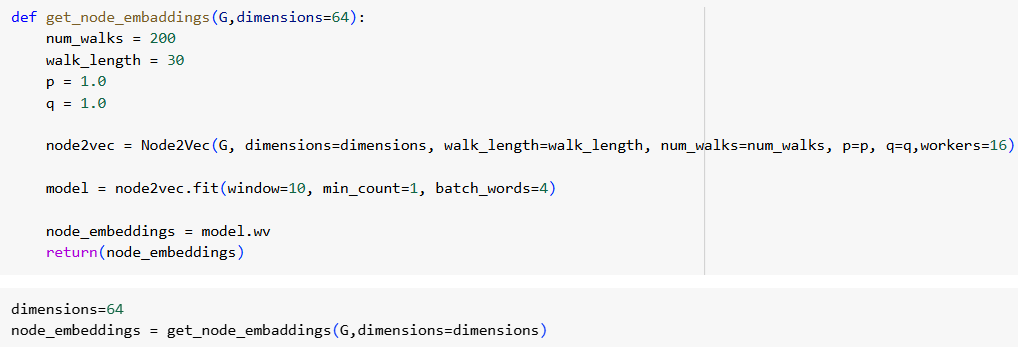
`dimensions`: Размерность векторных представлений узлов. Установлено значение 64.

`num\_walks`: Количество случайных блужданий, которые будут сгенерированы для каждого узла.(200)

`walk\_length`: Длина каждого случайного блуждания.(30)

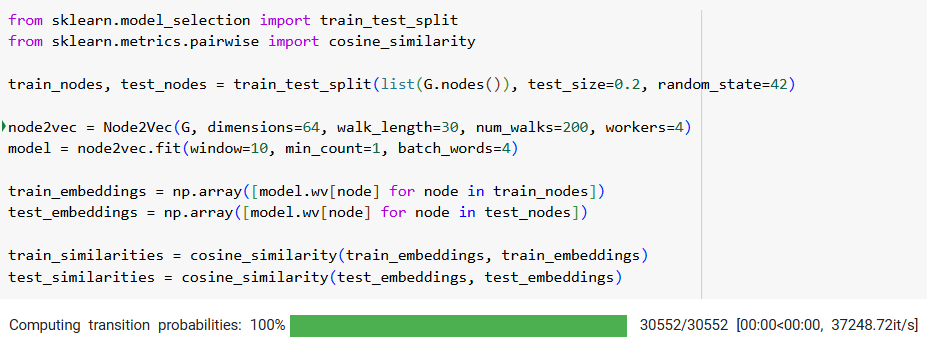
`p` и `q`: Параметры транзиции в методе Node2Vec. Они определяют вероятности перехода агента во время случайного блуждания. Параметр `p` контролирует вероятность возвращения к предыдущему узлу, а параметр `q` контролирует вероятность перехода к узлу, удаленному от текущего узла.(оба параметра равны 1)

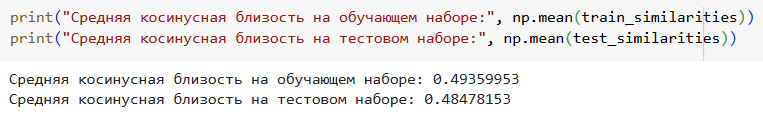
`workers`: Количество параллельных процессов, которые будут использоваться для генерации случайных блужданий. В нашем случае их 16.



## Оценка качества модели.

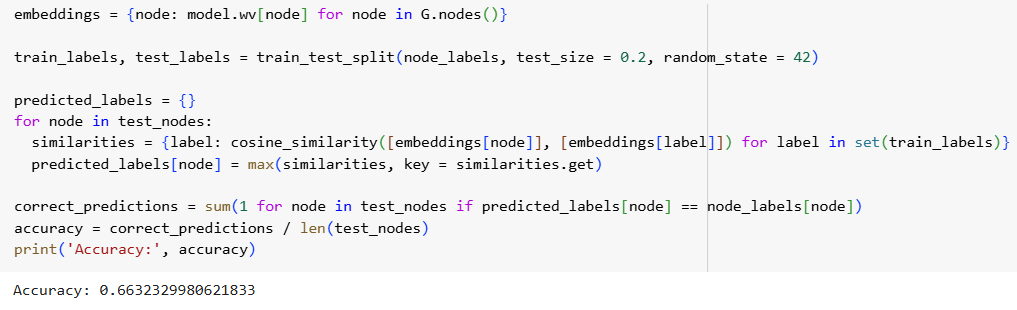
Оценка качества векторных представлений узлов графа, используя косинусную близость в качестве меры сходства между этими представлениями на обучающем и тестовом наборах:





Из этих показателей можно сделать вывод, что средняя косинусная близость на обучающем и тестовом наборах данных близка друг к другу, что может указывать на хорошую обобщающую способность модели. Однако небольшое различие между этими значениями может свидетельствовать о некоторой степени переобучения модели на обучающем наборе данных.

Оценим точность предсказаний меток узлов графа, используя меру сходства между векторными представлениями узлов и сравнивая предсказанные метки с истинными метками:



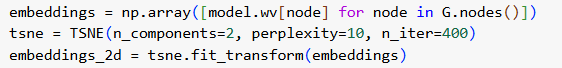
Из значения точности (Accuracy) равного 0.6632 можно сделать вывод, что модель, основанная на векторных представлениях узлов графа, способна с высокой точностью предсказывать метки для тестовых узлов.

## Представление результатов обучения в наглядном виде и анализ полученных выводов.

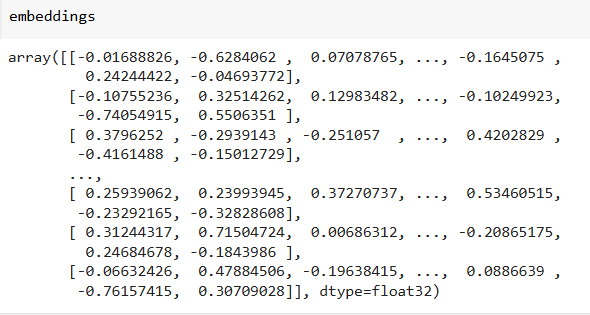
Результаты обучения модели:

Результатом обучения модели Node2Vec является получение векторных представлений узлов графа, где каждый узел представлен вектором фиксированной размерности. В данном случае, размерность векторов составляет 64. Эти векторные представления характеризуют структуру и связи между узлами в графе, отражая их семантические и топологические свойства.

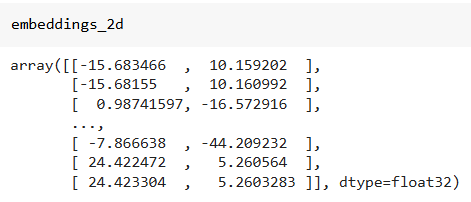
Обученная модель `model` содержит эти векторные представления, которые могут быть использованы для различных задач анализа графов, таких как поиск схожих узлов, кластеризация, классификация или визуализация. Векторные представления узлов представляют собой результат учитывающий контекстные и семантические зависимости между узлами в графе, полученные в результате обучения модели Node2Vec.



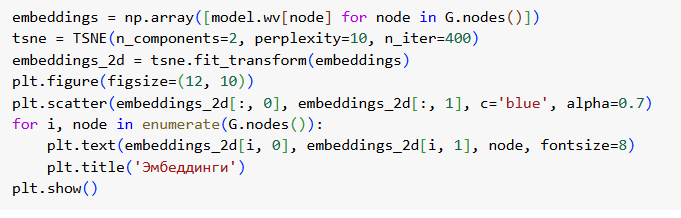
Полученные эмбеддинги:

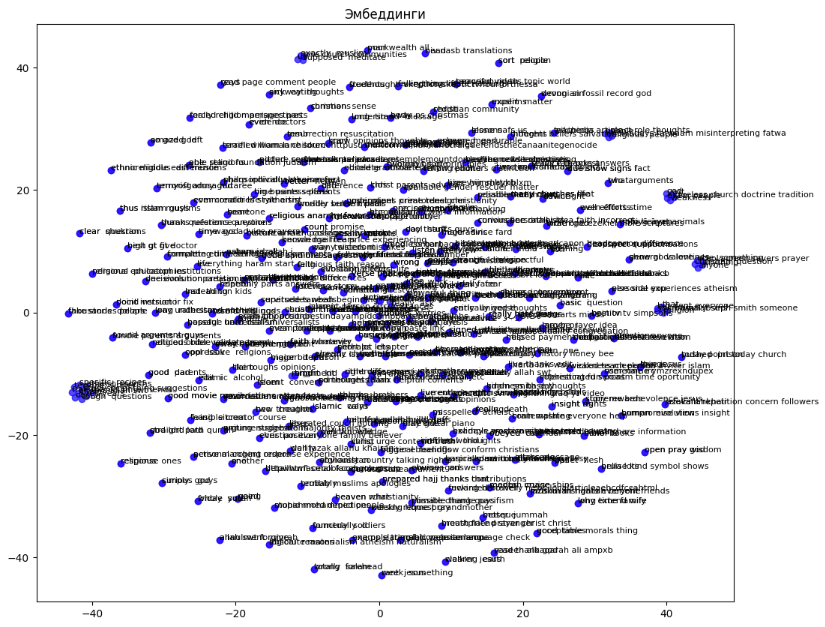


Эмбеддинги(после снижения размерности до 2, при помощи алгоритма TSNE):



Визуализация эмбеддингов:



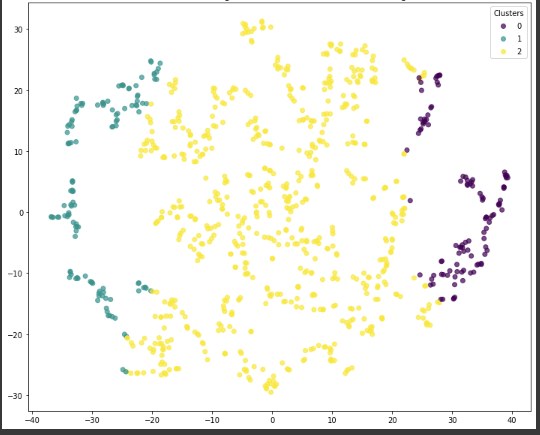


Эта визуализация показывает двумерное представление векторных эмбеддингов узлов графа. Каждый узел графа представлен точкой на двумерном графике, где координаты точек определяются значениями двух главных компонент после применения алгоритма снижения размерности.

Кроме того, для каждого узла выводится его метка (название узла), чтобы облегчить идентификацию узлов на графике.

Визуализация используется для исследования структуры и распределения узлов в графе. Она позволяет наглядно представить связи между узлами, а также выявить возможные кластеры или группы узлов с похожими свойствами или характеристиками.

Визуализация эмбеддингов узлов с использованием кластеризации K-mean:



Анализируя результаты визуализации, можно сделать следующие выводы:

1. Понимание структуры данных:

Визуализация позволяет увидеть структуру графа и связи между узлами. Кластеры представляют собой различные сообщества.

2. Качество модели:

В нашем случае кластеры узлов сформированы четко и однородно - это может свидетельствовать о хорошем качестве векторных представлений и адекватности модели.

3. Поиск аномалий:

Визуализация может помочь обнаружить аномалии или выбросы в данных. Например, узлы, которые находятся далеко от основной группы, могут указывать на необычные или непредсказуемые свойства данных.

Визуализация результатов обучения играет важную роль в понимании и анализе полученных данных. С помощью наглядных представлений мы можем лучше понять характеристики модели, выявить паттерны и тенденции, а также сделать выводы о ее эффективности

# ЗАКЛЮЧЕНИЕ

В рамках данной курсовой работы было проведено исследование по построению онтологических эмбеддингов графов знаний, с целью анализа сообщений с платформы Reddit, относящихся к различным религиям.

В первой части работы были рассмотрены теоретические основы, включая определение онтологических эмбеддингов и графов знаний, а также их применение. Это позволило установить базовое понимание предметной области и определить цели и задачи исследования.

Затем был выполнен этап создания графа знаний, который включал выбор и анализ набора данных, предварительную обработку и очистку данных, а также формирование графа и его векторное представление. Этот этап был крайне важен для подготовки данных к последующему этапу построения эмбеддингов.

Заключительным этапом стало построение онтологических эмбеддингов графа знаний, а также представление и анализ результатов.

Построение онтологических эмбеддингов графов знаний является эффективным методом анализа текстовых данных, например, таких как сообщения на платформе Reddit, относящиеся к религиозным темам. Полученные эмбеддинги могут быть использованы для различных задач, таких как классификация текстов, кластеризация сообщений и анализ семантических связей между темами.

Дальнейшие исследования в этой области могут включать более глубокий анализ, а также расширение исследования на другие типы данных и предметные области.

# 

# 

# СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ

Estelle Scifoю. Graph embedding techniques – 2020

Maël Fabien. Graph embedding

Oleg Durandin. ИСПОЛЬЗОВАНИЕ ПОДХОДА GRAPH2VEC ДЛЯ ЗАДАЧ NLP

* 2019

Hongyun Cai, Vincent W. Zheng, and Kevin Chen-Chuan Chang. A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications -2017

Mengjia Xu. Understanding graph embedding methods and their applications

- 2020

ИНТЕРНЕТ-РЕСУРСЫ

1. Лекция о графах знаний: [Лекция 1 - KG Course 2021 (migalkin.github.io)](https://migalkin.github.io/kgcourse2021/lectures/lecture1)
2. Knowledge Graph или граф знаний: что это такое и где применяется: [Knowledge Graph или граф знаний: что это такое и где применяется (sysblok.ru)](https://sysblok.ru/glossary/knowledge-graph-ili-graf-znanij-chto-jeto-takoe-i-gde-primenjaetsja/?ysclid=lv1iw4cd3z122622648)
3. Knowledge Graphs & NLP @ EMNLP 2019 Part I: [Knowledge Graphs & NLP @ EMNLP 2019 Part I | by Michael Galkin | Medium](https://mgalkin.medium.com/knowledge-graphs-nlp-emnlp-2019-part-i-e4e69fd7957c)
4. 1Датасет Religion - [Religion (kaggle.com)](https://www.kaggle.com/datasets/rayyan123z/religion)
5. Курс «Машинное обучение в семантическом и сетевом анализе 23-24 уч.г.»- [Курс: Машинное обучение в семантическом и сетевом анализе 23-24 уч.г. (fa.ru)](https://campus.fa.ru/course/view.php?id=32889)
6. [Где и как врубиться в эмбеддинги графов / Хабр (habr.com)](https://habr.com/ru/companies/ods/articles/418727/)
7. [Clustering and Classification using Knowledge Graph Embeddings — AmpliGraph 1.1.0 documentation](https://docs.ampligraph.org/en/1.1.0/tutorials/ClusteringAndClassificationWithEmbeddings.html)
8. [Онтология как инструмент создания баз знаний и баз фактов в FinTech | by Александр Морозов | Medium](https://almorozov.medium.com/%D0%BE%D0%BD%D1%82%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D1%8F-%D0%BA%D0%B0%D0%BA-%D0%B8%D0%BD%D1%81%D1%82%D1%80%D1%83%D0%BC%D0%B5%D0%BD%D1%82-%D1%81%D0%BE%D0%B7%D0%B4%D0%B0%D0%BD%D0%B8%D1%8F-%D0%B1%D0%B0%D0%B7-%D0%B7%D0%BD%D0%B0%D0%BD%D0%B8%D0%B9-%D0%B8-%D0%B1%D0%B0%D0%B7-%D1%84%D0%B0%D0%BA%D1%82%D0%BE%D0%B2-%D0%B2-fintech-609188ddad16)