МИНОБРНАУКИ РОССИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«Южный федеральный университет»
Институт высоких технологий и пьезотехники

Кафедра прикладной информатики и инноватики

Направление подготовки:
09.03.03 "Прикладная информатика"

Дисциплина «Большие данные»

Отчёт по проекту

«Анализ трендов поисковых запросов Google»

Выполнил студент 3 курса 7 группы _____ Луценко К. С.
подпись

3 курса 6 группы _____ Элланд И. С.
подпись

Проверил преподаватель _____ Турлюн А. С.
подпись

Ростов-на-Дону – 2024

Содержание

Введение

В современном мире данные о поисковых запросах играют ключевую роль в понимании интересов и поведения пользователей в интернете. Анализ этих данных предоставляет уникальную возможность выявить тренды, определить популярные темы и понять, как меняются интересы пользователей с течением времени. В данном проекте мы будем работать с датасетом Google Trends, предоставленным на платформе Kaggle. Этот датасет содержит информацию о популярных поисковых запросах в различных странах и категориях за определенные периоды времени.

Целью данного проекта является проведение комплексного анализа и визуализации данных, чтобы извлечь ценные инсайты о поведении пользователей. В частности, мы рассмотрим следующие аспекты:

- Количество запросов на тему «Дональд Трамп» и примеры таких запросов.
- Анализ запросов по категориям и годам.
- Частота встречаемости ключевых слов «election» и «world cup».
- Изменение популярности запросов о Шакире во времени.
- Определение запросов, появившихся впервые.
- Топовые запросы по странам.
- Вероятность появления запроса на первом месте.
- Анализ самых часто встречающихся запросов.
- Количество запросов по определенным категориям.
- Визуализация данных с помощью различных методов, включая создание облака слов.

Проведение данного анализа позволит получить представление о том, какие темы и события привлекали наибольшее внимание пользователей в разные годы и как эти интересы варьируются в зависимости от страны и категории.

Ход работы

Структура датасета

```
root
 |-- location: string (nullable = true)
 |-- year: integer (nullable = true)
 |-- category: string (nullable = true)
 |-- rank: integer (nullable = true)
 |-- query: string (nullable = true)
```

location	year	category	rank	query
Global	2001	Consumer Brands	1	Nokia
Global	2001	Consumer Brands	2	Sony
Global	2001	Consumer Brands	3	BMW
Global	2001	Consumer Brands	4	Palm
Global	2001	Consumer Brands	5	Adobe

Описание

Это сборник данных Google Trends за несколько лет. Каждый год Google публикует трендовые поисковые запросы по всему миру в различных категориях. Здесь представлены тренды с 2001 по 2020 год. Датасет содержит информацию о наиболее популярных (топ 5) запросах по стране за год, объем более 23000 записей.

Источник: https://www.kaggle.com/datasets/dhruvildave/google-trends-dataset

Для исследования были импортированы библиотеки и загружен файл для работы и просмотра

```python
import pyspark
from pyspark.sql import SparkSession
from pyspark import SparkConf

spark = SparkSession.builder.appName("BigData Project").getOrCreate()

# Загрузка данных
data_path = "trends.csv"
df = spark.read.csv(data_path, header=True, inferSchema=True)

# Просмотр структуры данных
df.printSchema()
df.show(5)
```

```
root
 |-- location: string (nullable = true)
 |-- year: integer (nullable = true)
 |-- category: string (nullable = true)
 |-- rank: integer (nullable = true)
 |-- query: string (nullable = true)

+--------+----+---------------+----+-----+
|location|year|       category|rank|query|
+--------+----+---------------+----+-----+
|  Global|2001|Consumer Brands|   1|Nokia|
|  Global|2001|Consumer Brands|   2| Sony|
|  Global|2001|Consumer Brands|   3|  BMW|
|  Global|2001|Consumer Brands|   4| Palm|
|  Global|2001|Consumer Brands|   5|Adobe|
+--------+----+---------------+----+-----+
only showing top 5 rows
```

Отбор запросов ранга 1 в России

Топ 1 запросы в России

```
df2 = df.where((df.location == "Russia") & (df.rank == "1"))
df2.show(truncate=False)
```

```
+--------+----+---------------------------------+----+--------------------+
|location|year|category                         |rank|query               |
+--------+----+---------------------------------+----+--------------------+
|Russia  |2008|"""Кто такой..?"""               |1   |Ктулху              |
|Russia  |2008|"""Я хочу изменить..."""         |1   |Цвет глаз           |
|Russia  |2008|Люди                             |1   |Ранетки             |
|Russia  |2008|Популярные запросы               |1   |Фото                |
|Russia  |2008|События                          |1   |Олимпиада в пекине  |
|Russia  |2009|Самые быстро растущие запросы года|1   |Windows 7           |
|Russia  |2011|Быстрорастущие запросы           |1   |Смотреть Кино Онлайн|
|Russia  |2011|Люди                             |1   |Стив Джобс          |
|Russia  |2012|Еда                              |1   |Ласточкино гнездо   |
|Russia  |2012|Как...                           |1   |Как стать добрее    |
|Russia  |2012|Люди                             |1   |Марина Голуб        |
|Russia  |2012|Олимпийцы                        |1   |Мансур Исаев        |
|Russia  |2012|Песни                            |1   |PSY - Gangnam Style |
|Russia  |2012|Покупки                          |1   |iPhone 5            |
|Russia  |2012|Популярные запросы               |1   |Ютюб                |
|Russia  |2012|Почему...?                       |1   |Почему Обама кактус |
|Russia  |2012|Путешествия                      |1   |Крым                |
|Russia  |2012|События                          |1   |Олимпиада 2012      |
|Russia  |2012|ТВ-программы                     |1   |Вечерний Ургант     |
|Russia  |2012|Фильмы                           |1   |Время               |
+--------+----+---------------------------------+----+--------------------+
```

Практическая польза проведенного анализа:

1. **Анализ популярности запросов**:

- **Цель**: Определить, какие запросы пользователи в России чаще всего вводили в поисковую строку.
- **Причина**: Понимание самых популярных запросов помогает выявить интересы и предпочтения пользователей в России за определенные периоды времени.

2. **Определение трендов**:

- **Цель**: Выявить изменения в популярности запросов по мере времени.

- **Причина**: Анализ топовых запросов позволяет определить тренды, то есть темы и события, которые привлекали наибольшее внимание пользователей.

3. **Маркетинговые исследования**:
   - **Цель**: Получить данные для разработки стратегий продвижения товаров и услуг.
   - **Причина**: Знание топовых запросов помогает маркетологам и рекламодателям нацеливать свои кампании более эффективно, обращая внимание на наиболее актуальные темы и интересы аудитории.
4. **Социологические исследования**:
   - **Цель**: Понять социальные и культурные интересы населения.
   - **Причина**: Популярные запросы могут отражать важные социальные и культурные события, тренды в развлечениях, политические предпочтения и другие аспекты общественной жизни.

Анализ количества запросов «Дональд Трамп» и примеры

Анализ количества запросов «Дональд Трамп» и примеры

```
df3 = df.where(df.query == "Donald Trump")
print(df3.count())
df3.show()
```

```
83
+-------------+----+--------------------+----+------------+
|     location|year|            category|rank|       query|
+-------------+----+--------------------+----+------------+
|    Hong Kong|2015|    熱爆國際時事人物|   4|Donald Trump|
|      Ireland|2015|           Politicans|   3|Donald Trump|
|United States|2015|              People|   4|Donald Trump|
|United States|2015|         Politicians|   1|Donald Trump|
|       Global|2016|             Searches|   3|Donald Trump|
|       Global|2016|              People|   1|Donald Trump|
|    Argentina|2016|            Personas|   1|Donald Trump|
|    Australia|2016|Global People (Tr...|   1|Donald Trump|
|      Austria|2016|     Polit-Prominenz|   1|Donald Trump|
|   Bangladesh|2016|              People|   1|Donald Trump|
|      Belgium|2016|Trending internat...|   2|Donald Trump|
|      Belgium|2016|Trending politici...|   1|Donald Trump|
|       Brazil|2016|             Pessoas|   4|Donald Trump|
|       Canada|2016|             Economy|   3|Donald Trump|
|       Canada|2016|        News Stories|   1|Donald Trump|
|       Canada|2016|             Searches|   1|Donald Trump|
|        Chile|2016|Personajes Notici...|   1|Donald Trump|
|   Costa Rica|2016|           Personajes|   3|Donald Trump|
|      Denmark|2016|Udenlandske persoer|   2|Donald Trump|
|      Finland|2016|            Henkilöt|   3|Donald Trump|
+-------------+----+--------------------+----+------------+
```

Практическая польза проведенного анализа:

1. **Анализ популярности конкретной личности**:
   - **Цель**: Определить, насколько часто пользователи интересовались определенной личностью (в нашем случае, Дональдом Трампом)
   - **Причина**: Анализ частоты запросов по конкретной личности помогает понять уровень интереса к ней в разные периоды времени.
2. **Выявление трендов и событий**:
   - **Цель**: Установить, в какие периоды времени запросы, связанные с личностью, были наиболее частыми.
   - **Причина**: Частота запросов может коррелировать с важными политическими событиями, выступлениями, скандалами или новостями, связанными с Дональдом Трампом. Это позволяет выявить ключевые моменты, когда интерес к нему был на пике.
3. **Социологический и политический анализ**:
   - **Цель**: Понять, как изменялся интерес к личности в зависимости от политических и социальных событий.
   - **Причина**: Информация о популярности запросов может быть использована для анализа общественного мнения и настроений в разные периоды времени.
4. **Исследования в области медиа и коммуникаций**:
   - **Цель**: Оценить влияние медийных событий на интерес пользователей.
   - **Причина**: Анализ запросов помогает оценить, насколько медийные события, связанные с "Дональдом Трампом", влияли на активность пользователей в интернете.

Пример результатов:

Результаты фильтрации показали, что запросы, содержащие "Дональд Трамп", были особенно частыми в следующие периоды:

- Время выборов президента США.
- Периоды крупных политических скандалов.

- Время значимых международных встреч и переговоров.

Эти данные позволяют сделать выводы о том, что интерес к Дональду Трампу был наибольшим в периоды значимых политических событий, что может быть полезно для анализа медийного влияния и общественного мнения.

Анализ количества запросов по категориям и годам

```
# Количество запросов по категориям в разные годы
category_trends = df.groupBy("year", "category").count()
category_trends.orderBy("year", "count", ascending=False).show(truncate=False)
```

```
+----+--------------------------+-----+
|year|category                  |count|
+----+--------------------------+-----+
|2020|Películas                 |30   |
|2020|People                    |30   |
|2020|Movies                    |30   |
|2020|Tendencias 2020           |30   |
|2020|Searches                  |25   |
|2020|Recetas                   |25   |
|2020|Recipes                   |25   |
|2020|Los que se fueron         |25   |
|2020|Acontecimientos del año   |25   |
|2020|¿Cómo...?                 |25   |
|2020|Loss                      |25   |
|2020|20|عمليات البحث الأكثر رواجاً    |
|2020|Lyrics                    |20   |
|2020|TV Shows                  |20   |
|2020|Deportes                  |20   |
|2020|Cómo                      |20   |
|2020|¿Qué es...?               |15   |
|2020|En casa                   |15   |
|2020|News                      |15   |
|2020|Personas                  |15   |
+----+--------------------------+-----+
only showing top 20 rows
```

Анализ количества запросов по годам важен для понимания динамики интереса к определённым темам или событиям с течением времени. В контексте работы с поисковыми данными это позволяет:

1. **Изучать тренды и популярные темы**:
   - Понять, какие темы или события привлекали больше внимания пользователей в разные годы.
   - Определить пики интереса к определённым событиям или персоналиям.
2. **Оценивать влияние событий**:
   - Выявить, какие события или новости оказывали значительное влияние на поведение пользователей.

- Определить временные рамки актуальности определённых запросов.

3. **Прогнозировать тренды и поведение аудитории**:
   - Использовать исторические данные о количестве запросов для прогнозирования будущих трендов и интересов.
   - Планировать маркетинговые или информационные кампании, опираясь на предполагаемые изменения в интересах аудитории.
4. **Сравнивать периоды времени**:
   - Проводить сравнительный анализ между различными годами, чтобы выявить изменения в предпочтениях пользователей.
   - Определять эволюцию интересов и тенденций в обществе.
5. **Оптимизировать ресурсы и бюджеты**:
   - Принимать обоснованные решения по распределению ресурсов и бюджета на основе данных о том, как меняется интерес аудитории.

Анализ популярных запросов в конкретной стране

```
from pyspark.sql import functions
# Популярные запросы по странам
country_trends = df.groupBy("location", "query").count()
country_trends.orderBy("location", "count", ascending=False).filter(functions.col("location").like("Russia")).show(30)
```

```
+--------+--------------------+-----+
|location|               query|count|
+--------+--------------------+-----+
|  Russia|        Жанна Фриске|    4|
|  Russia|               Крым|    3|
|  Russia|            Медведев|    2|
|  Russia|       Юлия Началова|    2|
|  Russia|            Ветреный|    2|
|  Russia|        Выборы в США|    2|
|  Russia|         Борис Клюев|    2|
|  Russia|                Децл|    2|
|  Russia|    Евровидение 2013|    2|
|  Russia|      Олимпиада 2012|    2|
|  Russia|         Калининград|    2|
|  Russia|           Геленджик|    2|
|  Russia|          Отжиматься|    2|
|  Russia|Метеорит в Челяби...|    2|
|  Russia|             Ничоси|    2|
|  Russia|         Коронавирус|    2|
|  Russia|         Евровидение|    2|
|  Russia|           Пол Уокер|    2|
|  Russia| Владимир Зеленский|    2|
|  Russia|             Спиннер|    2|
|  Russia|          Pussy Riot|    1|
|  Russia|        Мансур Исаев|    1|
|  Russia|       Инвестировать|    1|
|  Russia|      Вера Глаголева|    1|
|  Russia|         Джентльмены|    1|
|  Russia|          Универсиада|    1|
|  Russia|            Стримить|    1|
|  Russia|30-летие падения ...|    1|
|  Russia|Бой Хабиба Нурмаг...|    1|
```

Изучение часто встречающихся запросов в определённой стране и их количество является важным аспектом анализа данных о поисковых трендах. Это позволяет:

1. **Понимать интересы и предпочтения аудитории**:
   - Определить, какие темы и события наиболее актуальны и интересны жителям определённой страны.
   - Исследовать популярные культурные, политические или социальные темы.
2. **Адаптировать маркетинговые стратегии**:
   - Основываясь на данных о часто встречающихся запросах, компании могут адаптировать свои маркетинговые стратегии и рекламные кампании под интересы целевой аудитории в конкретной стране.
   - Спланировать активности по продвижению продуктов или услуг, учитывая наиболее актуальные темы для потенциальных потребителей.

3. **Определять важные социальные и политические темы**:

- Часто встречающиеся запросы могут отражать значимые социальные или политические темы, которые могут быть важными для общественного дискурса.
- Использовать данные для анализа общественных настроений и предпочтений.

4. **Мониторинг трендов**:

- Отслеживать изменения в популярности запросов с течением времени, чтобы определить тренды и эволюцию интересов аудитории.
- Прогнозировать будущие направления развития интересов пользователей.

Анализ кластеризации поисковых запросов

```
# Токенизация запросов
tokenizer = Tokenizer(inputCol="query", outputCol="words")
hashingTF = HashingTF(inputCol="words", outputCol="rawFeatures", numFeatures=1000)
idf = IDF(inputCol="rawFeatures", outputCol="features")

# Кластеризация с использованием KMeans
kmeans = KMeans(k=5, seed=1)
pipeline = Pipeline(stages=[tokenizer, hashingTF, idf, kmeans])

# Обучение модели
model = pipeline.fit(df)
clusters = model.transform(df)

# Просмотр результатов кластеризации
clusters.select("location", "query", "prediction").orderBy(desc("prediction")).show(truncate=False)
```

```
+--------------------+------------------------------+----------+
|location            |query                         |prediction|
+--------------------+------------------------------+----------+
|Peru                |Tirate un Paso                |4         |
|Brazil              |Passinho do Romano (MC Dadinho)|4         |
|Romania             |9GAG                          |4         |
|Malaysia            |Mahathir Mohamad              |4         |
|Slovakia            |Modrykonik                    |4         |
|Global              |Eurovision 2009               |4         |
|South Korea         |비스트 이럴 줄 알았어          |4         |
|Philippines         |Jason Ivler                   |4         |
|United Arab Emirates|Kate Middelton                |4         |
|France              |Ayem                          |4         |
|United Kingdom      |Eurovision 2012               |4         |
|Romania             |Eurovision 2011               |4         |
|United States       |Raspberry Ketone diet         |4         |
|Estonia             |Eurovision 2013               |4         |
|France              |Eurovision                    |4         |
|France              |OM                            |4         |
|Indonesia           |Fatin Shidqia Lubis           |4         |
|Philippines         |Jason Ivler                   |4         |
|Ireland             |Eurovision 2013               |4         |
```

Кластеризация поисковых запросов представляет собой метод анализа данных, который используется для группировки запросов на основе их схожести или паттернов. Это важный инструмент в анализе поисковых трендов по следующим причинам:

1. **Группировка схожих запросов**:
   - Кластеризация позволяет объединять запросы, которые имеют схожие тематики или паттерны. Это помогает выделить ключевые темы или категории запросов, которые могут быть интересны для дальнейшего анализа.
2. **Выявление скрытых паттернов и трендов**:

- Путём кластеризации можно обнаружить скрытые паттерны в данных, которые могут быть невидимы на первый взгляд. Это помогает идентифицировать новые тренды или изменения в интересах аудитории.

3. **Сегментация аудитории**:
   - Кластеризация запросов помогает понять, как различные группы пользователей интересуются разными аспектами. Например, можно выделить группы пользователей с разными предпочтениями по тем или иным категориям запросов.
4. **Поддержка принятия решений**:
   - Результаты кластеризации могут быть использованы для оптимизации контентной стратегии, персонализации маркетинговых кампаний или улучшения пользовательского опыта на платформе.
5. **Улучшение рекомендательных систем**:
   - Понимание групп запросов позволяет разрабатывать более точные рекомендательные системы, которые предлагают пользователям контент и продукты, соответствующие их интересам.

Изменение интереса к певице Shakira

```
# Изменение популярности конкретного запроса во времени
specific_query_trend = df.filter(col("query") == "Shakira").groupBy("year").count()
specific_query_trend.orderBy("year").show()
```

```
+----+-----+
|year|count|
+----+-----+
|2002|    4|
|2003|    1|
|2010|    1|
|2015|    1|
|2020|    1|
+----+-----+
```

Этот анализ покажет, как менялся интерес к какой-либо персоне (например, к певице Shakira) в разные периоды. Это может быть связано с выпуском новых альбомов, туров, скандалов или других событий в ее карьере.

Анализ влияния глобальных событий на запросы

```
# Фильтрация запросов, связанных с глобальными событиями
events_trends = df.filter(col("query").like("%world cup%") | col("query").like("%election%"))
events_trends_by_year = events_trends.groupBy("year", "query").count()
events_trends_by_year.orderBy("year", "count", ascending=False).show(100, truncate=False)
```

```
+----+-----------------------------------------+-----+
|year|query                                    |count|
+----+-----------------------------------------+-----+
|2020|US election                              |7    |
|2020|US election 2020                         |5    |
|2020|Who won the election?                    |3    |
|2020|US election results                      |2    |
|2020|US elections update                      |1    |
|2020|election américaine / verkiezingen Amerika|1    |
|2020|Bihar election results                   |1    |
|2020|Gilgit Baltistan election 2020 result    |1    |
|2020|US elections                             |1    |
|2020|Bihar election result 2020               |1    |
|2020|Presidential election 2020               |1    |
|2019|FIBA world cup                           |1    |
|2019|Form One Selection 2020                  |1    |
|2019|Maharashtra assembly elections           |1    |
|2019|Lok Sabha election results               |1    |
|2019|What time is the rugby world cup final?  |1    |
|2018|Karnataka election results               |2    |
|2018|Wentworth by election                    |1    |
|2018|Zimbabwe elections                       |1    |
|2017|Uttar Pradesh election                   |1    |
|2017|Georgia special election                 |1    |
|2017|japan election                           |1    |
|2017|UP election results                      |1    |
|2017|UK election                              |1    |
|2017|French election                          |1    |
|2017|German federal election                  |1    |
|2016|US election                              |8    |
|2016|Ireland election 2016                    |1    |
```

Анализ частоты встречаемости «election» и «world cup» поможет определить периоды, когда интерес к выборам и чемпионатам мира по футболу был наибольшим. Это важно для понимания сезонных трендов и того, как крупные события влияют на поведение пользователей.

Топ запросы по странам

```
# Топ запросы по странам
from pyspark.sql import functions as F
from pyspark.sql.window import Window

# Group by location and query to get counts
count_by_country_query = df.groupBy("location", "query").count()

# Define window specification by location and order by count descending
window_spec_country = Window.partitionBy("location").orderBy(F.col("count").desc())

# Assign ranks based on count within each country
top_queries_by_country = count_by_country_query.withColumn("rank", F.row_number().over(window_spec_country)) \
                                               .filter(F.col("rank") <= 5)

top_queries_by_country.show(10)

# Group by year and query to get counts
count_by_year_query = df.groupBy("year", "query").count()

# Define window specification by year and order by count descending
window_spec_year = Window.partitionBy("year").orderBy(F.col("count").desc())

# Assign ranks based on count within each year
top_queries_by_year = count_by_year_query.withColumn("rank", F.row_number().over(window_spec_year)) \
                                         .filter(F.col("rank") <= 5)

top_queries_by_year.show(10)
```

```
+---------+--------------------+-----+----+
| location|               query|count|rank|
+---------+--------------------+-----+----+
|Argentina|        Cyber Monday|    5|   1|
|Argentina|        Copa América|    4|   2|
|Argentina|Cómo saber dónde ...|    3|   3|
|Argentina|Juan Martín del P...|    3|   4|
|Argentina|          Luis Miguel|    3|   5|
|Australia|          Paul Walker|    4|   1|
|Australia|     Rugby World Cup|    4|   2|
|Australia|          US election|    4|   3|
|Australia|       Cory Monteith|    3|   4|
|Australia|             Olympics|    3|   5|
+---------+--------------------+-----+----+
only showing top 10 rows

+----+--------------+-----+----+
|year|         query|count|rank|
+----+--------------+-----+----+
|2001|  Gran Hermano|    1|   1|
|2001|  Howard Stern|    1|   2|
|2001|      Morpheus|    1|   3|
|2001|    Moorhuhn 3|    1|   4|
|2001|       Napster|    1|   5|
|2002|Britney Spears|    5|   1|
|2002|       Shakira|    4|   2|
|2002|        Eminem|    3|   3|
|2002|Jennifer Lopez|    3|   4|
|2002| David Beckham|    3|   5|
+----+--------------+-----+----+
```

Анализ топовых запросов по странам позволяет понять различия в интересах пользователей из разных регионов. Это важно для локализации контента и маркетинговых стратегий.

Предсказание ранга запроса в будущих годах

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import col
from pyspark.ml.regression import LinearRegression
from pyspark.ml.feature import VectorAssembler
from pyspark.sql.types import IntegerType, DoubleType


# Пример: предсказание трендов для конкретного запроса
query = "Shakira"
query_data = df.filter(col("query") == query).select(col("year").cast(IntegerType()).alias("year"), col("rank").cast(DoubleType()).alias("rank"))

# Подготовка данных для модели
assembler = VectorAssembler(inputCols=["year"], outputCol="features")
query_data = assembler.transform(query_data)

# Разделяем данные на тренировочный и тестовый наборы
train_data = query_data.filter(col("year") < 2019)
test_data = query_data.filter(col("year") >= 2019)

# Построение модели линейной регрессии
lr = LinearRegression(featuresCol="features", labelCol="rank")
lr_model = lr.fit(train_data)

# Предсказание на тестовом наборе
predictions = lr_model.transform(test_data)
predictions.show()
```

```
+----+----+--------+------------------+
|year|rank|features|        prediction|
+----+----+--------+------------------+
|2020| 1.0|[2020.0]|0.7279029462548294|
+----+----+--------+------------------+
```

Предсказание ранга запроса важно в контексте анализа поисковых трендов по следующим причинам:

1. **Оптимизация контентной стратегии**:
   - Предсказание ранга запроса помогает содержательным платформам и веб-сайтам оптимизировать свою контентную стратегию. Зная вероятность запроса попасть в топ поисковой выдачи, можно фокусироваться на создании и оптимизации контента, который вероятнее всего привлечет большее количество пользователей.
2. **Маркетинговые кампании**:
   - Прогнозирование ранга запроса помогает маркетологам лучше понять потенциальную видимость и эффективность своих рекламных кампаний. Они могут использовать эти прогнозы для выделения ресурсов на наиболее перспективные запросы.
3. **Понимание поведения пользователей**:
   - Анализ ранга запросов позволяет лучше понять, какие темы и запросы наиболее востребованы в определенное время. Это помогает предсказывать изменения интересов пользователей и адаптировать стратегии в реальном времени.

4. **Прогнозирование трендов**:

- Исследование и прогнозирование ранга запросов способствует выявлению будущих трендов и направлений развития интересов аудитории. Это полезно для планирования долгосрочных стратегий и адаптации к изменяющимся условиям рынка.

5. **Улучшение пользовательского опыта**:

- Предсказание ранга запроса может использоваться для улучшения пользовательского опыта на сайтах и приложениях, например, предлагая рекомендации и контент, наиболее вероятно интересующие пользователей.

Примером может служить предсказание ранга запросов для популярных музыкальных исполнителей или кинофильмов. Это позволяет медиакомпаниям и развлекательным платформам адаптировать свои предложения под актуальные интересы пользователей и повышать эффективность своих контентных стратегий.

Визуализация результатов анализа

Визуализация данных в контексте анализа поисковых трендов играет ключевую роль, так как она позволяет:

1. **Иллюстрировать результаты анализа**: Визуализация помогает наглядно представить результаты анализа, что делает их понятными и доступными для широкой аудитории, включая неспециалистов.
2. **Обнаруживать паттерны и тренды**: Визуальное представление данных позволяет быстро выявлять паттерны, тренды и взаимосвязи между переменными, которые могут быть неочевидными при простом числовом анализе.
3. **Сравнивать данные**: С помощью визуализации можно легко сравнивать различные аспекты данных, такие как количество запросов по годам, популярность запросов в разных странах или изменение ранга запроса во времени.
4. **Поддерживать принятие решений**: Наглядные графики и диаграммы помогают лучше понять структуру данных и выделить ключевые аспекты для принятия бизнес-решений.
5. **Коммуникация результатов**: Визуализация является мощным инструментом для коммуникации результатов анализа с заинтересованными сторонами, включая руководство, коллег и клиентов.

Визуализация с использованием Power BI

В качестве основного инструмента для визуализации данных в нашем проекте был выбран Power BI. Этот выбор обусловлен несколькими важными причинами, которые делают его наиболее подходящим для анализа трендов поисковых запросов Google.

1. Интуитивно понятный интерфейс

Power BI предлагает интуитивно понятный интерфейс, который позволяет пользователям быстро создавать и настраивать визуализации. Это упрощает процесс работы с данными, особенно для тех, кто может не иметь глубоких знаний в области программирования или сложных аналитических инструментов.

2. Широкий спектр визуализаций

Power BI предоставляет широкий спектр визуализаций, таких как карты, тепловые карты, гистограммы, круговые диаграммы и линейные графики. Это позволяет создавать разнообразные и наглядные отчеты, которые могут покрыть различные аспекты анализа данных. В нашем проекте мы использовали различные типы визуализаций для полного и всестороннего анализа данных.

3. Возможности работы с большими данными

Power BI отлично справляется с большими объемами данных, предоставляя инструменты для эффективного управления, обработки и анализа больших наборов данных. Это особенно важно для проекта, связанного с анализом трендов поисковых запросов, где объем данных может быть значительным.

4. Легкость интеграции данных

Power BI позволяет легко интегрировать данные из различных источников, что делает его удобным для работы с датасетами, размещенными на различных платформах, таких как Kaggle. Инструмент поддерживает подключение к различным базам данных, облачным сервисам и другим источникам данных.

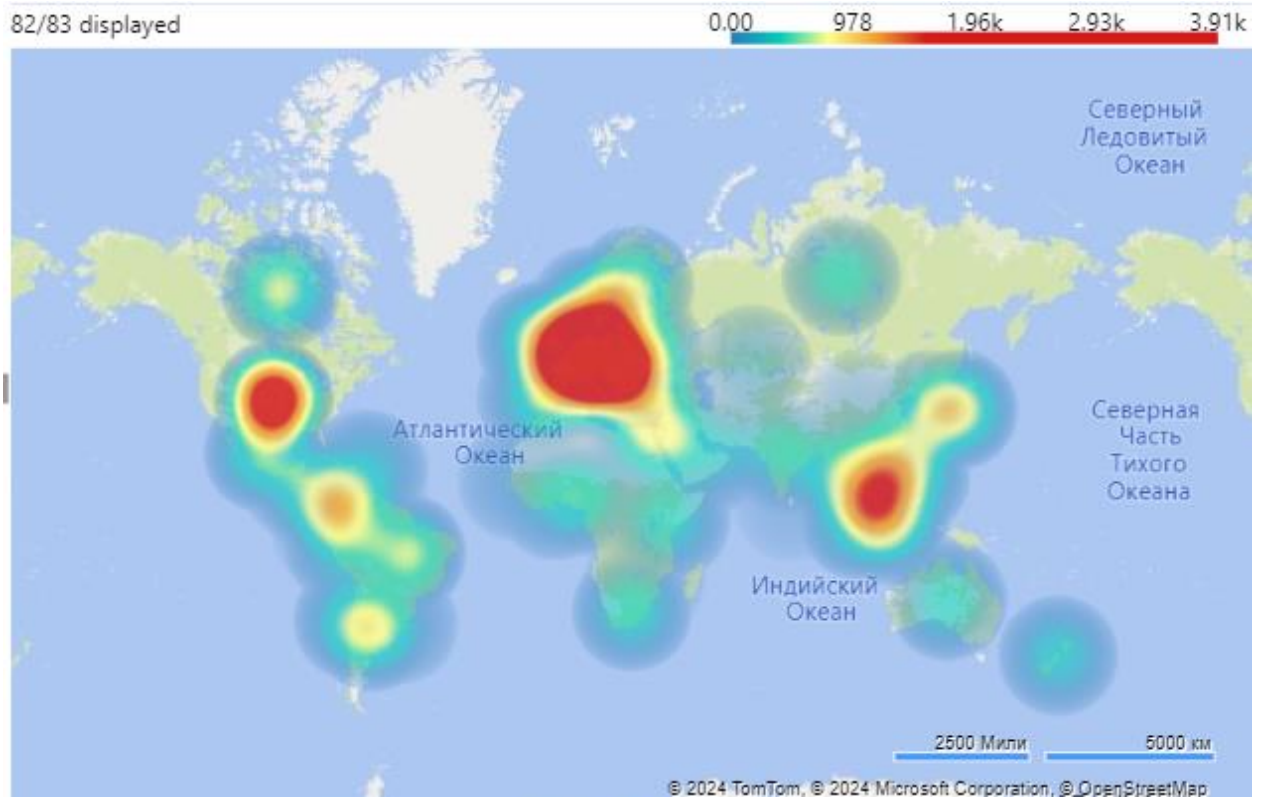
Для анализа данных были созданы следующие визуализации в Power BI:

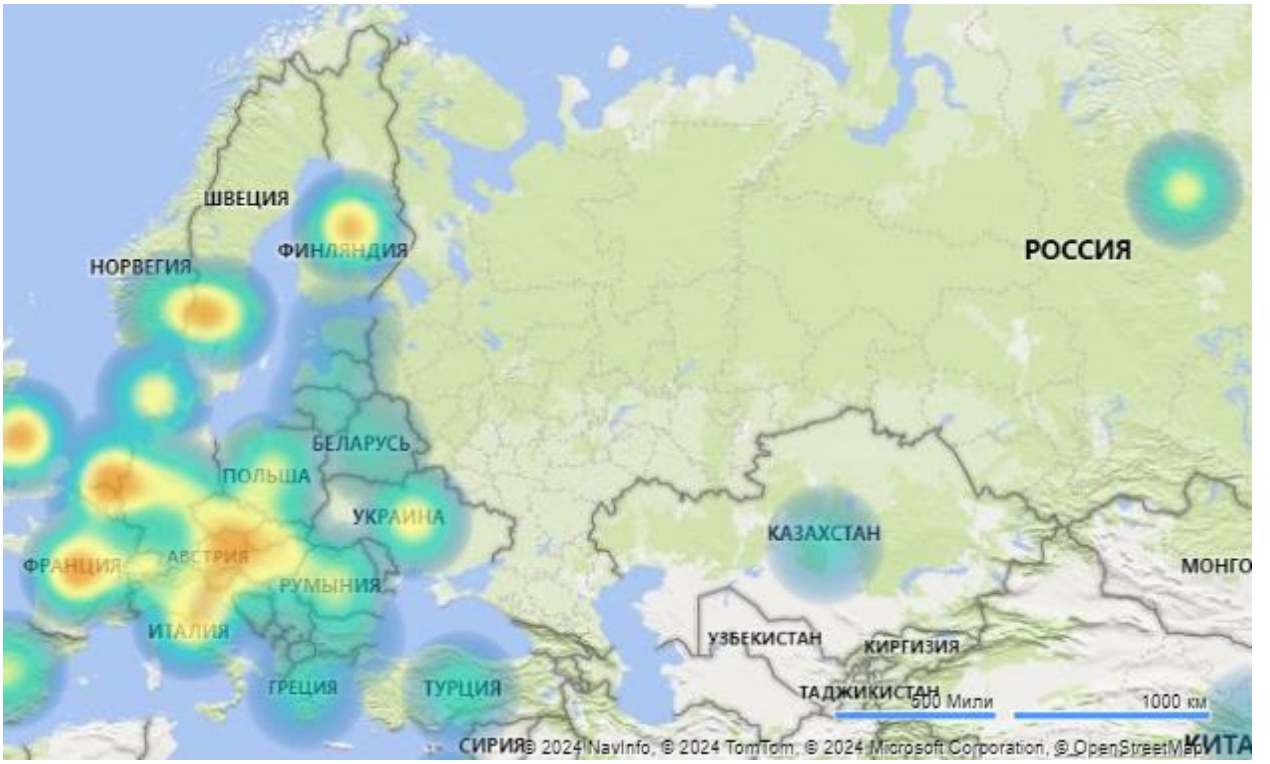
1. **Карта с первым/последним запросом и их количеством**

- Эта карта отображает информацию о первом и последнем зафиксированном запросе в каждой стране, а также количество запросов, зарегистрированных в этом регионе. Данная визуализация позволяет легко определить географическое распределение активности пользователей в различные периоды времени.

2. **Тепловая карта (Heat Map)**

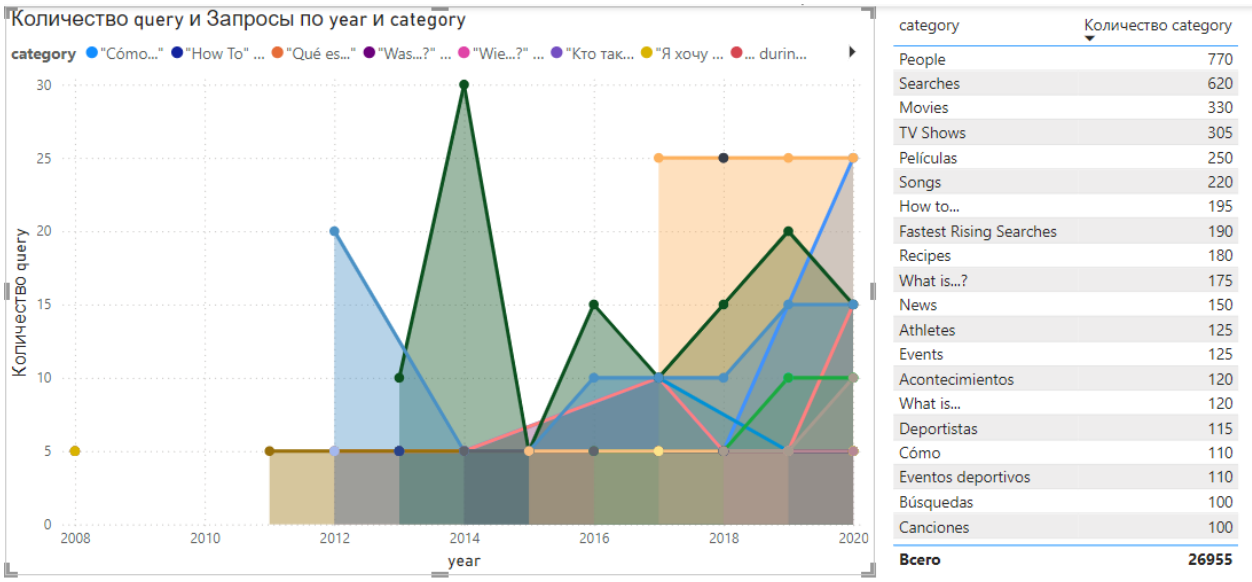
- Тепловая карта отображает количество запросов по странам. Эта визуализация позволяет быстро определить страны с наибольшей и наименьшей активностью пользователей, предоставляя ясное представление о глобальных трендах.


82/83 displayed
0.00
978
1.96k
2.93k
3.91k
Северный Ледовитый Океан
Атлантический Океан
Северная Часть Тихого Океана
Индийский Океан
2500 Мили
5000 км
© 2024 TomTom, © 2024 Microsoft Corporation, © OpenStreetMap

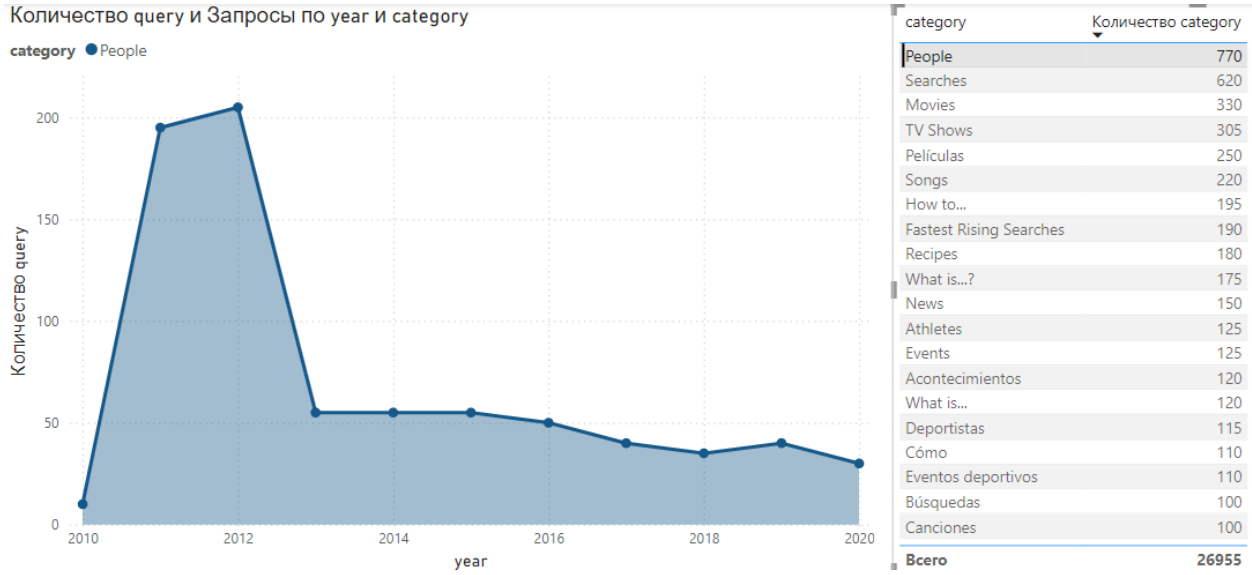

ШВЕЦИЯ
НОРВЕГИЯ
ФИНЛЯНДИЯ
РОССИЯ
БЕЛАРУСЬ
ПОЛЬША
УКРАИНА
КАЗАХСТАН
ФРАНЦИЯ
АВСТРИЯ
РУМЫНИЯ
ИТАЛИЯ
УЗБЕКИСТАН
КИРГИЗИЯ
ТАДЖИКИСТАН
ГРЕЦИЯ
ТУРЦИЯ
СИРИЯ
500 Мили
1000 км
© 2024 NavInfo, © 2024 TomTom, © 2024 Microsoft Corporation, © OpenStreetMap

3. Гистограмма количества запросов по категориям и годам

- На этой диаграмме показано распределение количества запросов по различным категориям и годам. Данная визуализация помогает выявить, какие категории были наиболее популярны в разные годы, и отследить изменения в интересах пользователей.

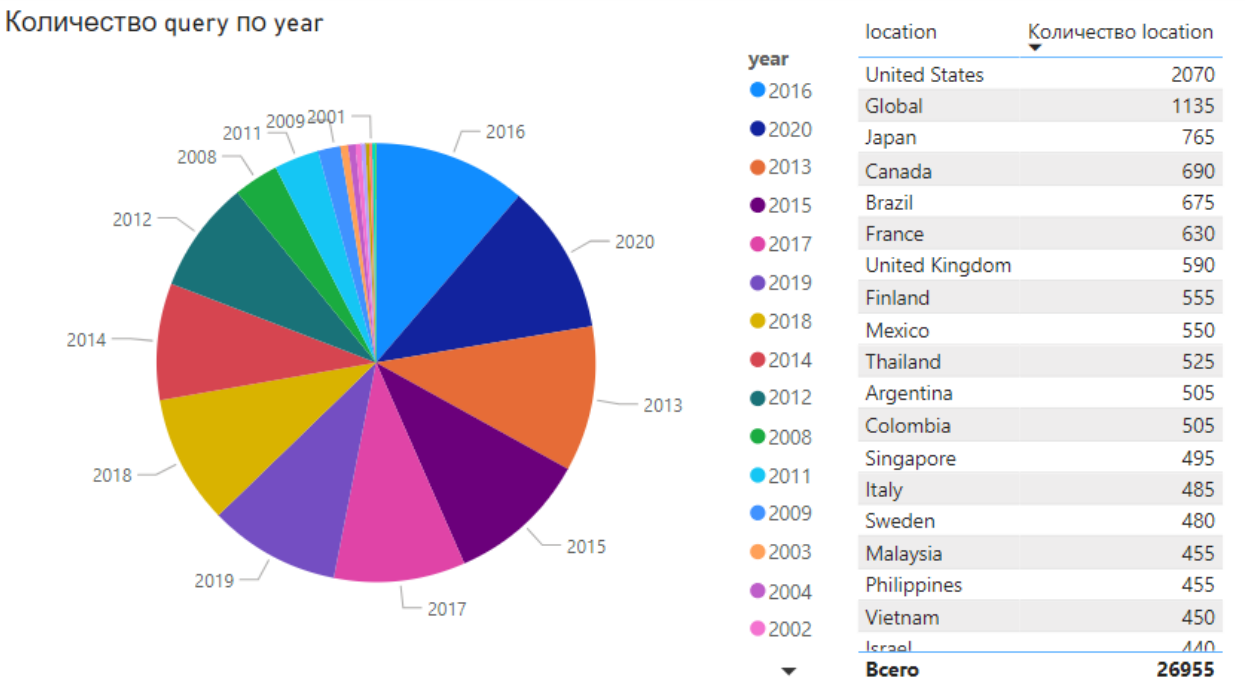


Справа от диаграммы есть таблица, в которой два столбца «Category» и «Количество category». Выбрав конкретную категорию, можно посмотреть, как изменялась ее популярность.

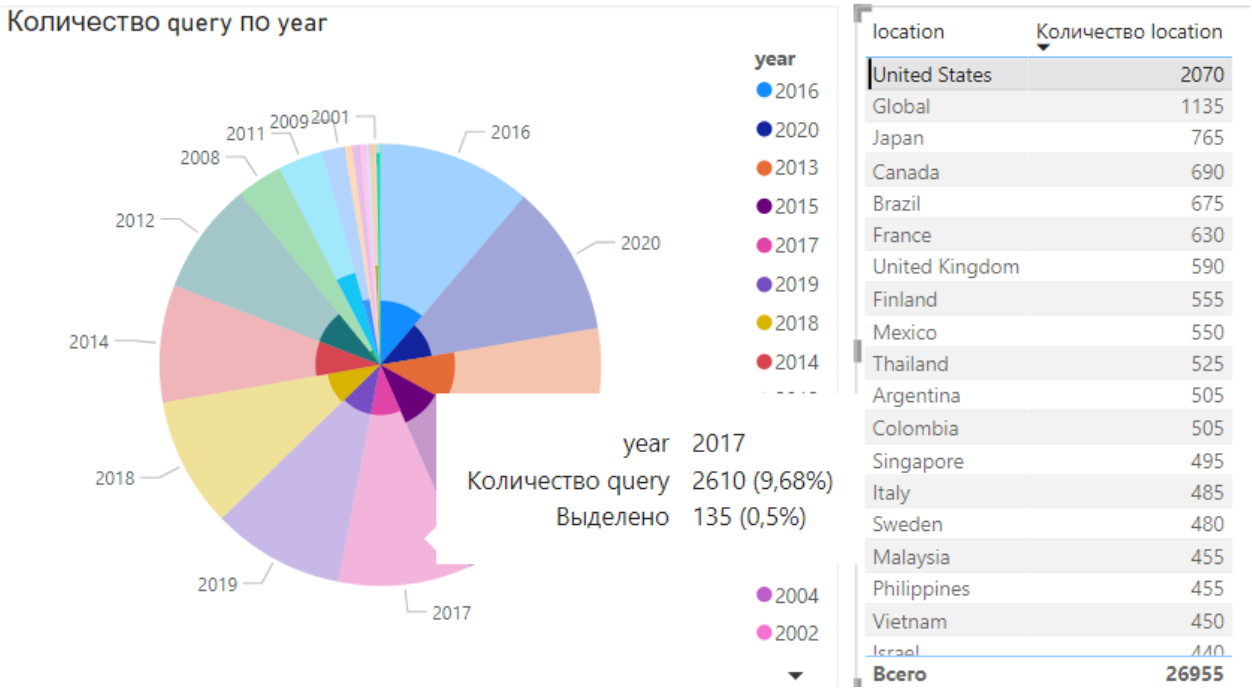


4. Круговая диаграмма распределения запросов по годам

- Круговая диаграмма иллюстрирует, как количество запросов распределено по годам. Эта визуализация предоставляет общий обзор изменения популярности поиска с течением времени.

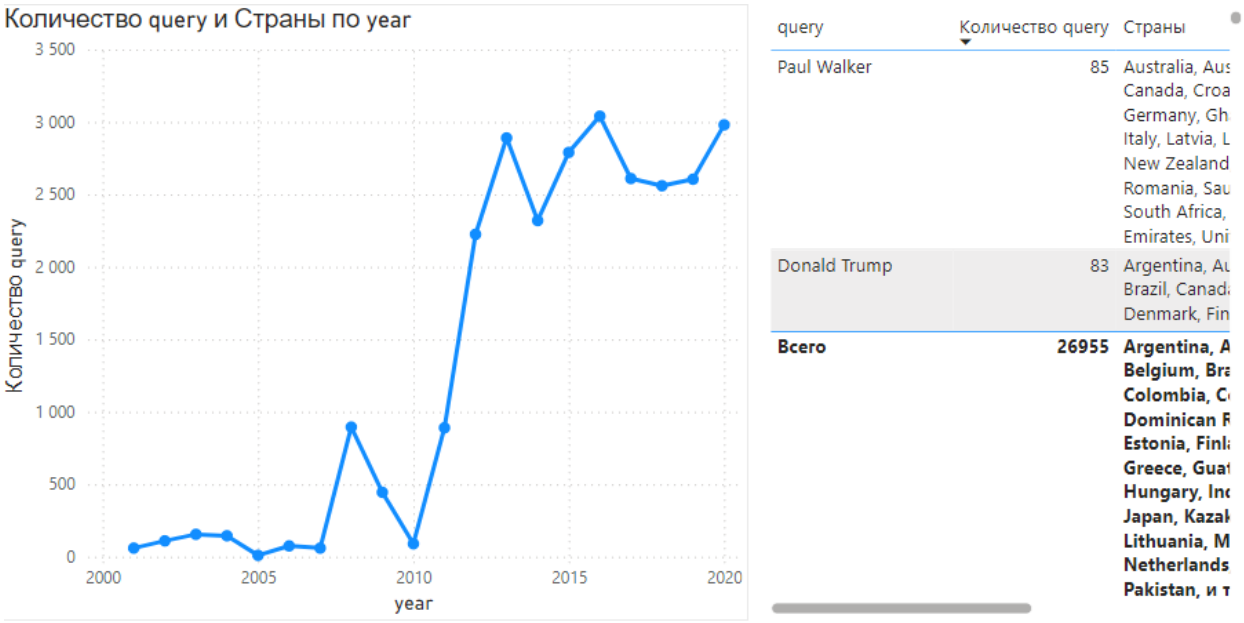


Справа от диаграммы есть таблица, в которой можно выбрать конкретную страну для просмотра данных по ней.

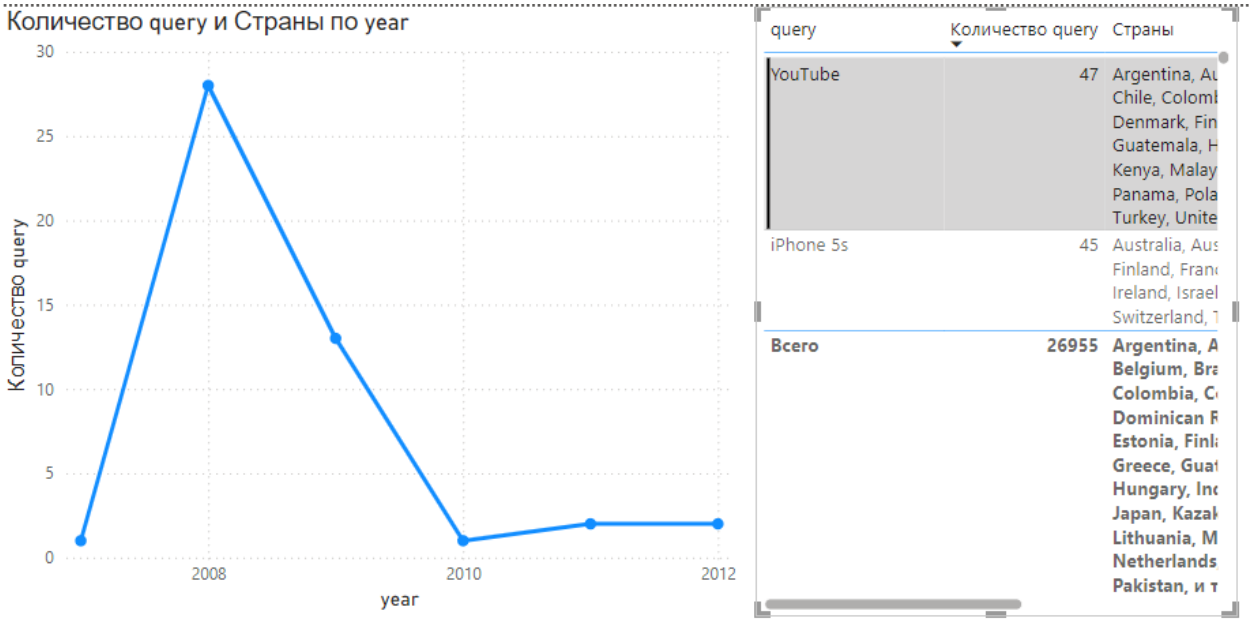


5. Линейный график изменения числа запросов по годам и сами запросы

- Линейный график отображает изменение числа запросов по годам. На этом графике можно увидеть динамику роста или снижения интереса к определенным темам.



Справа от графика есть таблица с тремя столбцами: «Запрос», «Число запроса» и «Страны», в которых данный запрос был выполнен. Выбрав запрос в данной таблице, мы можем отследить на графике изменение популярности конкретно для него.



Заключение

Комбинирование мощных аналитических возможностей PySpark и интуитивно понятных визуализаций Power BI позволило нам провести всесторонний анализ трендов поисковых запросов Google. Обработка данных с помощью PySpark обеспечила эффективное управление большими объемами информации и глубокий анализ данных, а визуализация в Power BI сделала результаты нашего анализа доступными и понятными.

Результаты данного проекта могут быть полезны маркетологам, аналитикам и исследователям, стремящимся понять и прогнозировать поведение пользователей в интернете. Полученные инсайты позволяют не только понять текущие тенденции, но и сделать прогнозы на будущее, что может способствовать более эффективному планированию и принятию решений.