Machine Translated by Google

Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Trending YouTube Video Analysis

MD SAKIBUR HASAN (shstat10@gmail.com)

그랜드 밸리 주립대학교

Bishal Sarker

다카대학교

Diksha Shrestha

그랜드 밸리 주립대학교

Roshan Shrestha

그랜드 밸리 주립 대학교

Sajal N. Shrestha

그랜드 밸리 주립 대학교

Research Article

Keywords: YouTube, 트렌드, 좋아요, 조회수, 제목

Posted Date: 2023년 2월 7일

DOI: https://doi.org/10.21203/rs.3.rs-2548456/v1

License: 이 저작물은 Creative Commons Attribution 4.0 국제 라이선스에 따라 라이선스가 부여됩니다. 전체 라이센스 읽기

Machine Translated by Google

Abstract

YouTube는 콘텐츠 제작자가 제작한 동영상을 사용자가 보고, 공유하고, 좋아요를 누르고, 댓글을 달고, 구독할 수 있는 온라인 플랫폼입니다. 조회수, 좋아요, 동영상 연령 등을 기준으로 YouTube 인기 동영상 카테고리에 포함될 동영상을 선택합니다. 하지만 이는 콘텐츠 제작자가 직면한 주요 과제입니다. 본 연구의 주요 목표는 유튜브 인기 동영상에 영향을 미치는 요인을 분석하는 것이다. 완료하기로 결정된 5가지 작업이 있습니다. 좋아요와 조회수의 관계 결정; 다양한 카테고리가 추세를 나타내는 데 걸리는 평균 시간을 비교합니다. 가장 인기 있는 태그를 식별합니다. 이상적인 제목 길이를 결정합니다. 동영상이 인기를 끌 수 있는 가장 좋은 날을 결정합니다.

콘텐츠 제작자가 더 나은 이해를 얻는 데 도움이 될 다양한 국가의 YouTube 동영상 트렌드입니다. 매일 업데이트되는 2020년 9월부터 2022년 1월까지의 데이터세트는 좋아요와 조회수 간의 상관관계, 다양한 카테고리에 걸쳐 트렌드를 파악하는 데 소요되는 평균 시간, 국가별로 가장 많이 사용되는 태그, 최적의 제목 길이 범위, 전 세계 트렌드를 파악하는 데 도움이 됩니다. 요일과 동영상을 게시하기에 가장 좋은 요일을 지정합니다. 매일 업데이트되는 2020년 9월부터 2022년 1월까지의 데이터세트는 국가별 콘텐츠 제작자와 사용자가 좋아요와 조회수의 상관관계, 다양한 카테고리에 걸쳐 트렌드를 파악하는 데 소요되는 평균 시간, 국가별로 가장 많이 사용되는 태그, 최적의 제목 길이를 파악하는 데 도움이 됩니다. 범위, 요일별 동향, 영상을 게시하기 가장 좋은 요일 등을 파악하여 적절한 콘텐츠를 만드는 데 도움을 줍니다.

1 Introduction

YouTube는 콘텐츠 제작자가 동영상을 게시하고 사용자가 보고, 공유하고, 좋아요를 달고, 댓글을 달고, 구독할 수 있는 온라인 플랫폼입니다. YouTube는 전 세계 수백만 명의 사람들이 사용하는 성장하는 비디오 플랫폼이며 지난 몇
년 동안 YouTube에 대한 사용자 참여가 확대되었습니다. 마찬가지로 오늘날에는 비즈니스 회사에서도 제품을 광고하고 홍보하기 위해 널리 사용되어 이익 성장에 기여합니다. YouTube에는 조회수, 좋아요 및 동영상 연령에 따라 결정되는 인기 동영상 카테고리도 있습니다. 또한 인기 카테고리는 동영상 콘텐츠에 대한 사람들의 참여 선택에 따라 국가마다 다를 수 있습니다.

일반적으로 유튜브 영상 트렌딩에 대한 연구[3]는 있었지만, 유튜브 영상 트렌딩 분석을 결정하는 이유에 대해서는 국가별 분석이 이루어지지 않았다. 우리의 주요 목표는 YouTube 비디오가 여러 국가에서 인기를 끌도록 돕는 다양한 특성을 분석하여 콘텐츠 제작자가 더 잘 이해할 수 있도록 돕는 것입니다. 우리는 시청률이 인기 동영상의 좋아요 수에 어떤 영향을 미치는지 분석하고 싶습니다. 또한, 우리는 비디오가 다양한 카테고리에 걸쳐 추세를 나타내는 데 걸리는 평균 시간을 비교하는 것을 목표로 합니다. 가장 많이 사용되는 태그와 동영상의 제목 문자 수를 식별하는 것이 우리가 달성하려는 또 다른 목표입니다. 마지막으로, 인기 급상승 동영상 수가 가장 많은 요일을 정확히 찾아보고자 합니다.

우리의 동기는 최대 시청자 수에 도달하고 YouTube 채널을 성장시키는 데 필요한 정보를 제공하여 콘텐츠 제작자를 돕는 것입니다. YouTube는 총 사용자 수를 기준으로 두 번째로 인기 있는 웹사이트입니다[v7]. YouTube는 조회수에 따라 사용자에게 콘텐츠 비용을 지불합니다. YouTube는 YouTube 사용자에게 동영상에 대한 비용을 지불하기 때문에 요즘 많은 사람들이 YouTube 사용자로 경력을 시작하는 꿈을 꾸고 조회수를 쉽게 얻을 수 있는 인기 페이지에 등장하고 싶어합니다. 이로 인해 플랫폼은 전 세계의 기존 사용자와 신규 사용자에게 더욱 수익성이 높아졌습니다. 또한 현재 채널에 업로드된 동영상 콘텐츠가 어떤 추세를 보이고 있는지 이해하고, 어떤 다양한 속성이 인기 페이지에 도달하는 데 기여하는지 알아보고 싶습니다. 또한 이러한 속성이 다양한 국가 및 카테고리에 걸쳐 동일한 효과를 갖는지에 대한 정보도 얻고 싶습니다.

2 Literature Review

Cheng, Dale 및 Liu가 발표한 연구[3]는 YouTube 동영상의 특성에 대한 체계적인 측정 연구에 중점을 두었습니다. 데이터는 27개 데이터 세트가 포함된 YouTube API 및 YouTube 동영상 페이지에서 3개월 동안 수집되었습니다. 관련 영상 20개를 이용하여 성장 추세를 파악하고, 유튜브 내 영상의 수명과 길이 분포를 통해 패턴을 파악했다. 연구에서는 YouTube 크롤러를 사용하여 가장 많이 본 동영상과 최고 평점을 받은 동영상을 파악했습니다.

Machine Translated by Google

총 189개의 고유한 비디오가 포함되었으며 7개의 데이터 세트를 제공하여 매주 수행되었습니다. 연구에 따르면, 데이터세트는 음악이 22.9%로 인기 카테고리이고 엔터테인먼트가 17.8%로 뒤를 잇는 등 편향된 분포를 보여주었습니다. 가장 낮은 카테고리는 Howto와 DIY였으며, 애완동물과 동물이 그 뒤를 이었습니다. 다른 특징은 비디오 길이에 기초한 것으로, 인기 비디오 길이의 97.8%가 600초 미만인 것으로 나타났는데, 그 이유는 뮤직 비디오가 종종 해당 범위 내에 있기 때문에 음악 카테고리가 더 많은 기여를 했기 때문입니다. 특성은 또한 대부분 30MB 미만인 비디오의 파일 크기를 나열했습니다. 마찬가지로 기사에는 성장 추세를 알아보기 위해 날짜 추가 특성도 포함했는데, 그 이유는 업로드된 영상이 그다지 인기가 없었기 때문입니다. 조회수와 평점 특성은 영상의 인기도와 패턴을 파악하는 데 도움이 되므로 본 연구에서는 중요한 특성 중 하나로 간주했습니다. 연구에서 밝혀진 마지막 특징은 거듭제곱법칙을 사용한 조회수의 수명에 따른 조회수가 증가하는 추세였다. 연구에 사용된 시각화는 막대 차트, 히스토그램, 선 그래프, 산점도였습니다. 각 특성이 동영상을 인기 있게 만드는 데 어떻게 중요한 역할을 하는지 설명했습니다. 본 연구는 국가별 동영상 트렌드의 인기를 다루지는 않으며, 2007년 데이터세트를 기반으로 합니다. 우리는 2020년 9월부터 2022년 1월까지의 국가별 최근 데이터세트를 분석할 예정입니다.

Barjasteh, Liu 및 Radha의 연구[2]도 트렌드 비디오의 측정 및 분석에 중점을 두고 있습니다. 9개월 동안의 인기 동영상 시계열을 분석하고 데이터는 YouTube API에서 수집되었습니다. 이 기사에서는 시청률 라이프사이클 분석, 트렌드 동영상과 비트렌드 동영상 간의 비교 분석, 트렌드 동영상 업로더의 프로필 분석, 트렌드 동영상 카테고리와 트렌드 동영상 조회수와의 방향성 관계 분석을 보여주는 데 중점을 둡니다. 사용된 데이터 세트는 트렌딩 동영상 4000개와 비트렌딩 동영상 4000개였으며 YouTube API에서 트렌딩 동영상이 생성되면 해당 동영상에 대한 통계가 추출되었습니다. 분석은 15일, 30일, 45일, 60일에 걸쳐 데이터를 수집한 4개의 하위 집합으로 이루어졌습니다. 히스토그램 분포를 사용하여 다양한 기간에 걸쳐 동영상의 집계 조회수를 표시했으며, 누적 선 그래프를 사용하여 일수에 따른 조회수 비율을 표시했습니다. 바 차트를 이용하여 카테고리별 인기 급상승 영상과 비 인기 급상승 영상을 비교 분석하였습니다. 분석에 따르면 인기 급상승 동영상의 지속 시간 평균은 인기가 없는 동영상의 평균 지속 시간보다 길었습니다. 인기 급상승 동영상과 비 인기 동영상의 조회수를 비교 분석했습니다. 인기 급상승 동영상의 조회수는 시간이 지남에 따라 증가하는 반면, 인기가 없는 동영상의 경우 업로드된 후 특정 날짜 동안 조회수가 증가한 후 조회수가 증가함을 보여주는 선 그래프가 사용되었습니다. 포화됩니다. 트렌딩 영상의 경우 업로더의 프로필을 분석한 결과, 업로더의 86%가 남성이고 14%가 여성입니다. 또한 업로더의 84%가 구독자가 100명이 넘고 6%가 구독자가 100만 명이 넘으므로 구독자 수가 인기 동영상으로 분류되는 동영상에 어떻게 영향을 미치는지 고려하여 최종 사용자에게 동영상 도달범위를 늘렸습니다. 또한 Granger Causality를 이용하여 시계열별 트렌드 영상과 조회수 간의 방향성 관계 분석을 수행하였다. 기사의 주요 목적은 인기 급상승 동영상의 통계를 식별, 측정 및 분석하는 것이었습니다.

Gajanayake와 Sandanayake(2020)의 연구[4]에 따르면 감성 분석을 사용하여 YouTube 게임 채널의 트렌드 패턴을 식별하는 것을 목표로 합니다. 또한 이 연구에서는 추세 특징을 파악하기 위해 기계 학습 방법을 사용했습니다. 그러나 연구는 YouTube 채널에 게시된 게임 비디오에만 국한됩니다. 이번 연구에서는 YouTube API를 통해 약 1000개의 인기 게임 동영상을 수집했습니다. 마찬가지로 연구진은 유튜브 데이터를 전처리해 메타데이터로 변환한 후 유튜브 게임 영상의 트렌드 패턴, 시청자 댓글에 대한 감성 분석을 기반으로 한 영상 등을 분석했다. 저자는 분류 분석을 위해 SVM(Support Vector Machine), Naive Bayes 및 Logistic Regression을 사용했습니다. 연구에서는 선 그래프를 사용하여 동영상의 댓글 섹션에서 가장 많이 사용된 단어를 결정했습니다. 또한 저자는 YouTube 동영상의 추세 패턴을 파악하기 위해 막대 차트와 히스토그램 차트를 사용하여 게시 시간과 제목 길이를 각각 표시했습니다. 이 작업은 우리 프로젝트와 관련이 있는데, 우리는 최대 조회수를 기록하는 동영상을 게시하기에 완벽한 날을 찾고 있기 때문입니다. 하지만 이번 연구에서는 영상의 시간과 개수에 초점을 맞춘 반면, 우리 프로젝트는 조회수에 초점을 맞춰 날짜에 초점을 맞췄으며 시각화 도구로 선 그래프를 사용할 것입니다. 동시에 이 연구는 제목 길이와 비디오 수에 초점을 맞춰

Machine Translated by Google

제목 길이. 우리 프로젝트는 국가별로 인기 있는 YouTube 동영상의 제목 길이를 찾는 데 더 중점을 둡니다. 연구 결과에 따르면 동영상 게시 시간, 동영상 제목 길이, 조회수와 좋아요, 오디오 품질이 YouTube에서 동영상을 인기 급상승시키는 데 어떻게 중요한 역할을 하는지 보여주었습니다.

Hoiled, Aprem 및 Krishnamurthy의 연구[5]는 YouTube 동영상의 참여도와 인기 역학, 메타데이터에 대한 민감도에 중점을 두었습니다. 데이터 세트는 26,000개 채널의 600만 개 비디오로 구성되었습니다. 연구는 동영상의 제목, 태그, 썸네일, 설명과 같은 메타 수준 기능에 중점을 두었습니다. 마찬가지로 메타 수준 분석을 위해 첫날 조회수, 구독자 수, 동영상 썸네일 대비, 제목 길이 등을 고려하여 조회수의 인기도를 파악합니다. 이를 위해 연구자는 머신러닝 기법을 활용해 영상 조회수에 따른 영상 민감도를 알아보았다. 또한 본 연구에서는 조회수가 구독자 수에 어떤 인과관계를 갖는지 알아보기 위해 Granger 인과성 테스트를 사용했습니다. 유튜브 채널의 경우 조회수 증가에 따라 구독자 증가율이 증가했습니다. 본 연구의 또 다른 과제는 유튜브 영상의 조회수를 예측하는 것이었고, 이를 달성하기 위해 머신러닝 기법(Extreme Learning Machine)을 사용했다. 실제 조회수 데이터로 ELM을 활용하여 조회수를 예측하는 선 그래프 시각화가 구현되었습니다. 마찬가지로 조회수에 따른 사회적 상호작용을 파악하기 위해 시계열 분석 방법을 사용했습니다. 소셜 상호작용을 측정하기 위해 구독자와 조회수 간의 사상자를 이용했습니다. 또한 YouTube 동영상 업로드 일정의 역학도 연구되었습니다. 연구에 따르면 일정의 역학이 조회수와 동영상의 댓글 수에 영향을 미치는 것으로 나타났습니다. 우리의 임무와 마찬가지로, 우리는 유튜버가 자신의 동영상을 업로드하여 트렌드가 될 수 있는 통계를 보고 최적의 하루 일정을 찾아내고 있습니다. 본 연구에서는 평균 조회수에 관한 작업에 좀 더 중점을 두고 있습니다. 마찬가지로, 우리 프로젝트도 평균 조회수를 기반으로 하는 작업이 더 많습니다. 이는 동영상의 추세에 영향을 미치는 중요한 요소 중 하나이기 때문입니다.

분류, 연관 및 클러스터링을 사용한 YouTube 분석의 인기 동영상 알고리즘에 대한 Andry, Reynaldo, Lee, Christianto, Loisa 및 Manduri(2021)의 연구[1]. 본 연구에 사용된 주요 속성은 조회수, 좋아요, 싫어요, 댓글이었습니다. YouTube 알고리즘을 찾기 위해 분류, 연관, 클러스터링 등의 데이터 마이닝 기술을 사용했습니다. 연구에서는 YouTube 알고리즘의 작동 방식과 동영상이 인기 목록에 유지되는 방식을 알아냅니다.

연구자는 분류 방법을 사용하여 YouTube 알고리즘에서 인기 동영상을 찾는 데 좋아요와 조회수 속성이 중요한 역할을 한다는 사실을 발견했습니다. 이어 연구진은 연관기법을 사용해 조회수, 좋아요, 싫어요, 댓글이 관계를 나타내며, 이러한 속성도 유튜브 알고리즘에서 역할을 한다는 사실을 알아냈다. 저희 작업에서는 조회수와 좋아요가 동영상 트렌딩에 얼마나 중요한 역할을 하는지 알아보면서, 좋아요와 조회수 사이의 연관성을 알아보기 위해 조회수와 조회수의 상관 관계를 찾는 데 중점을 두었습니다. 또한, 동영상 트렌딩에 가장 큰 영향을 미치는 속성을 0부터 4까지 그룹화하는 클러스터링 기법을 사용하였고, 그 결과 조회수, 좋아요, 클릭 순으로 중요한 역할을 한다는 결론을 내렸습니다. 제목 수 및 키워드. 따라서 연구자들은 YouTube 알고리즘에 기여하는 두 가지 요소는 좋아요, 조회수, 횟수를 통한 참여이고 다른 요소는 제목 길이와 키워드인 메타데이터라고 결론지었습니다.

3 Tasks

우리의 주요 목표는 인기 YouTube 동영상에 기여하는 속성을 분석하는 것입니다. 우리가 달성하고자 하는 5가지 작업은 좋아요와 조회수 사이의 상관 관계 찾기, 카테고리가 인기를 얻는 데 걸리는 평균 시간 비교, 가장 많이 사용되는 동영상 식별입니다. 태그를 사용하여 최적의 제목 길이를 찾고 동영상이 인기를 끌 수 있는 최적의 날짜를 식별합니다.

좋아요와 조회수의 상관관계

좋아요 수와 조회수를 조사하여 2020년 9월부터 2022년 1월까지 인기 급상승 동영상의 좋아요 수와 조회수 간의 상관 관계를 파악합니다. 우리의 첫 번째 작업은 인기 급상승 동영상에 대한 좋아요와 조회수 사이의 관계를 찾는 것입니다. 이는 둘 사이에 상관관계가 있는지 식별하는 데 도움이 됩니다. 모든 작업을 수행

Machine Translated by Google

조회수가 가장 높은 동영상의 좋아요 수도 비슷할까요? 이 작업에서는 likes 속성을 사용합니다.
데이터 세트의 개수 속성을 봅니다.

여러 카테고리에 걸쳐 인기 급상승 동영상의 평균 시간을 비교하세요.

게시 날짜, 추세 날짜, 카테고리, 추세 지속 시간을 분석하여 2020년 9월부터 2022년 1월 사이에 동영상이 카테고리 전반에 걸쳐 추세가 되는 데 걸린 평균 시간을 비교합니다. 이 작업에서는 동영상 카테고리가 인기 페이지에 표시되는 데 시간이 얼마나 걸리는지 알아보고 싶습니다. 데이터 세트에는 게시 날짜 및 추세 날짜로부터 추세 시간에 대한 정보를 얻을 수 있는 31개 범주가 있습니다. 평균 소요 시간이 여러 범주에 걸쳐 일관성이 있는지 알고 싶습니다. 또한 사용자가 다른 국가 간에 전환할 수 있도록 하여 작업에 상호작용성을 추가하고 싶습니다. 마찬가지로 사용자가 시간, 일, 주 사이의 시간 범위를 변경하도록 허용할 수도 있습니다.

국가별로 가장 인기 있는 태그 식별

이 작업에서는 사용된 태그, 국가 이름, 카테고리 이름 및 카테고리 ID를 분석하여 여러 국가에서 2020년 9월부터 2022년 1월까지 최종 사용자가 동영상을 빠르게 찾을 수 있도록 인기 동영상에서 가장 많이 사용된 태그를 식별하려고 합니다. 가장 많이 사용되는 태그를 찾는 아이디어를 통해 제작자는 유사한 태그를 사용하여 트렌드 단계에 도달할 수 있습니다. 인기 급상승 동영상에서 가장 많이 사용되는 태그는 무엇인가요? 가장 많이 사용되는 태그는 국가별로 유사합니까? 모든 인기 동영상에서 태그를 추출하고 태그가 사용된 횟수를 기준으로 집계됩니다.

국가 및 카테고리를 기반으로 태그를 추가로 집계할 수 있으며, 이는 가장 많이 사용된 태그를 추가로 분석하여 비디오가 인기 페이지에 도달하는 데 도움이 됩니다.

인기 동영상의 제목 길이 빈도 분석

우리는 2020년 9월부터 2022년 1월까지의 데이터를 분석하여 크리에이터가 콘텐츠에 대한 최적의 제목 범위를 찾을 수 있도록 인기 YouTube 동영상에 대한 제목 길이의 전체 빈도 분포를 분석하려고 합니다. 이 작업을 위해 제목 속성을 사용합니다. 제목의 문자와 국가 이름의 빈도. 동영상 제목 길이가 동영상 인기를 높이는 데 어떤 역할을 합니까? 콘텐츠 제작자는 짧은 제목을 만드는 데 집중해야 할까요, 아니면 설명적인 제목을 만드는 데 집중해야 할까요? 이를 위해 각 동영상의 제목 길이를 추출하고 국가별로 집계하겠습니다.

총 인기 동영상 개수로 요일별 트렌드 파악

2020년 9월부터 2022년 1월까지의 트렌드 날짜, 조회수, 좋아요 수, 댓글, 요일 등의 속성을 분석하여 해당 영상이 트렌드를 형성하는 요일을 파악하는 것을 목표로 합니다. 어떤 요일에 가장 많은 인기가 있는지 분석하는 데 중점을 두고 있습니다. 인기 동영상 수는? 나라마다 비슷한가요? 영상의 데이터세트를 요일별 분석하여 요일별로 집계해 드립니다. 이는 추세 날짜 속성의 날짜-시간 스탬프에서 날짜를 계산하여 수행됩니다. 또한 대화형에서는 월 단위로 데이터를 추가로 집계하고 사용자가 다음을 수행하도록 허용할 수 있습니다.

다른 국가 간에 전환하여 추세를 확인하세요.

4 Dataset

YouTube의 인기 동영상 데이터세트[6]는 2020년 9월부터 2022년 1월까지 Kaggle에서 사용할 수 있으며 Rishav Sharma 사용자가 매일 업데이트합니다. 2022년 2월 25일 현재 데이터 세트 크기는 약 2.0GB입니다. 작성자는 크롤러를 사용하여 YouTube API에서 데이터를 추출했습니다. 데이터세트에는 미국, 캐나다, 영국, 일본, 독일, 프랑스, 러시아, 브라질, 멕시코, 일본, 한국의 일일 인기 동영상 정보가 포함되어 있습니다. 각 국가의 데이터는 업로드 날짜, 채널 또는 사용자, 추세 데이터 및 태그, 조회수, 좋아요 및 싫어요와 같은 기타 메타데이터에 대한 정보가 포함된 개별 csv 파일에 저장됩니다. 마찬가지로 카테고리 정보도 국가별로 다르기 때문에 json 형식으로 별도로 저장됩니다. 데이터 세트에 사용되는 속성은 다음과 같습니다.

- title - 인기 급상승 동영상의 이름입니다. 게
- 시됨 - 비디오가 게시된 날짜/시간 스탬프입니다.

Machine Translated by Google

- ChannelTitle - YouTube 채널의 이름입니다.
- CategoryTitle - 비디오 카테고리의 이름입니다.
- trendingDate - 특정 국가의 인기 페이지에 표시되는 날짜/시간 스탬프입니다. 태그: YouTube 알고리즘에서 쉽게 인식할 수 있도록
- 채널에서 사용하는 라벨입니다.
- views: 사용자의 조회수
- 좋아요: 동영상의 좋아요 수
- 싫어요: 동영상의 싫어요 수
- commentCount: 동영상의 댓글 수입니다.

게시된 At 및 trendingData에 사용되는 날짜-시간 스탬프는 UTC 시간 형식입니다. 데이터 세트의 태그 속성은 csv 파일 형식과 유사하게 각 단어가 "-"로 구분되는 텍스트 형식입니다. 분석을 위해 개별 동영상의 각 태그를 추출하겠습니다.

5 Project Design

Correlation between likes and views

이 시각화의 목표는 2020년 9월부터 2022년 1월까지 YouTube 인기 동영상에 대한 좋아요와 조회수 간의 상관 관계를 찾는 것입니다. 처음에는 시각화로 버블 차트를 사용하는 것을 고려했습니다. 수치변수나 범주형 변수 간의 관계를 분석하는 데 도움이 되기 때문이다. 이는 산점도와 매우 유사하며 산점도의 확장으로 간주될 수도 있습니다. 그러나 우리의 작업은 작업을 표현하기 위해 두 개의 숫자 변수만 고려하므로 이를 달성하기에는 버블 차트가 상당히 과도할 수 있습니다. 또한 데이터 세트가 매우 크기 때문에 거품 영역으로 인해 차트가 혼란스러워 보일 수 있습니다. 또한 영역 인코딩은 클리블랜드의 규칙에 따라 효과적인 시각적 인코딩이 아닙니다.

우리가 접한 다음 시각화는 히트맵이었습니다. 히트맵은 우리 작업에 이상적인 두 변수 간의 관계를 표시하는 데 사용할 수 있습니다. 이를 통해 데이터세트의 좋아요 수와 조회수 사이의 패턴을 발견할 수 있습니다. 그러나 우리는 대규모 데이터 세트를 보유하고 있기 때문에 사용자가 결론을 도출하기에는 부담스러울 수 있으며 세부 사항을 놓칠 수도 있습니다. 또한 클리블랜드의 규칙에 따라 우리는 히트맵의 기본 시각적 인코딩인 색상 인코딩을 잘 디코딩하지 못합니다.

우리는 위와 같은 이유로 산점도 시각화를 사용하는 경향이 있습니다. 산점도 그래프는 두 변수 사이의 관계를 탐지하는 데 도움이 될 수 있습니다. 우리는 이것이 우리의 사용 사례, 즉 인기 YouTube 비디오에 대한 좋아요와 조회수의 상관 관계를 보여주는 데 적합하다고 느꼈습니다. 마찬가지로 데이터 세트가 크기 때문에 이 시각화는 매우 잘 확장되며 사용자가 둘 사이의 상관 관계를 식별하는 데 도움이 될 수 있습니다[1]. 또한 우리는 산점도가 패턴을 발견하는 동시에 이상값을 탐지하는 데에도 도움이 될 수 있다는 사실도 발견했습니다. 산점도는 매우 간단하며 대부분의 사람들은 이를 분석하는 데 익숙합니다.

이 작업에 적용되는 시각적 인코딩은 위치와 색상입니다. 우리는 위치 인코딩을 사용하여 x축에 같은 데이터를 표시하고 y축에 조회수를 표시합니다. 클리블랜드의 규칙은 인간이 시각화에서 위치를 더 잘 인식하므로 위치 인코딩이 매우 효과적인 것으로 평가합니다. 이를 통해 사용자는 데이터 세트의 인스턴스를 빠르게 찾을 수 있습니다. 게슈탈트의 원리와 관련하여, 함께 배치된 데이터가 더 관련성이 높은 것처럼 보이면 근접성의 법칙이 적용될 수 있습니다.

색상은 이 그래프에서 사용한 또 다른 시각적 인코딩입니다. 우리는 이 작업에서 평온함을 의미하는 파란색을 선택했습니다. 결과적으로 최종 사용자의 눈에 시각화가 쉬워지고 그래프를 더 쉽게 이해할 수 있습니다. 클리블랜드의 규칙에 따르면 색상 인코딩은 효율성 순위 측면에서 그리 효과적이지 않습니다. 그러나 우리는 단일 색상을 사용하고 위치가 기본 인코딩으로 설정되어 있으므로 시각화에 영향을 미치지 않습니다. 또한 게슈탈트 유사성의 법칙은 동일한 색상을 사용한 좋아요와 조회수의 관계를 보여줍니다.

Compare average hours of trending video across mul- tiple categories

Machine Translated by Google

우리는 2020년 9월부터 2022년 1월까지 YouTube 동영상이 카테고리별 추세에 도달하는 데 소요된 평균 시간을 비교하는 것을 목표로 합니다. 이 작업을 위해 우리는 도넛 차트 및 거품형 차트와 같은 다양한 시각화를 분석했습니다. 처음에는 평균 시간을 범주 간 추세와 비교하는 작업을 시각화하기 위해 도넛형 차트를 사용할 계획이었습니다.
그럼에도 불구하고, 그래프를 분석한 결과, 도넛형 차트는 막대형 차트에 비해 순위 결정에 효과적이지 않다는 사실을 발견했습니다. 도넛 차트는 면적 및 각도와 같은 시각적 인코딩을 사용하는데, 이는 클리블랜드의 규칙에 따라 길이보다 읽고 분석하기가 더 어려울 수 있습니다. 또한 YouTube 동영상 카테고리는 약 14개이며 도넛 차트는 이 작업에 적합하지 않을 수 있습니다. 카테고리가 많으면 각 카테고리를 표현하는 데 사용할 수 있는 영역이 줄어들기 때문에 읽고 이해하기 어려울 수 있기 때문입니다. 데이터.

우리 작업에서 발견한 다음 시각화는 버블 차트입니다. 버블 차트는 범주형 변수 간의 관계를 표시하는 데 유용합니다. 독자에게도 친숙하고 이해하기 쉽습니다. 그러나 이 시각화를 사용할 때 몇 가지 문제를 발견했습니다. 우리는 이 그래프가 변수가 적을 때 이상적이라는 것을 알았습니다. 또한 클리블랜드의 법칙에 따르면 인간은 길이를 비교하는 것보다 면적을 비교하는 데 더 능숙합니다. 따라서 두 개 이상의 범주가 비슷한 값을 가지고 있으면 이를 구별하는 것이 어려울 것입니다.

따라서 우리는 마침내 작업에 대한 시각화 선택으로 막대 차트를 선택했습니다. 이 그래프는 서로 다른 값, 더 중요하게는 값 간의 차이점을 강조하는 데 사용할 수 있습니다. 각 막대는 YouTube 카테고리를 나타내고, 길이는 해당 값, 즉 평균 시간을 나타냅니다. 그런 다음 이러한 막대는 공통 척도로 표시됩니다. 이 작업에서 우리는 막대 차트가 최종 사용자가 다양한 범주 내에서 추세를 나타내는 데 소요되는 평균 시간 추세를 쉽게 인식하는 데 도움이 된다는 것을 발견했습니다. 막대 차트는 또한 다양한 범주의 데이터를 비교하는 데 도움이 되며 그래프를 쉽게 읽고 이해할 수 있도록 정렬하는 데도 도움이 됩니다. 또한 막대 차트를 사용하여 각 범주에 대한 추세로 간주되는 평균 시간의 빈도 분포를 표시할 수 있습니다.

작업에 사용되는 시각적 인코딩은 위치, 길이 및 색상입니다. 위치 인코딩을 사용하여 시각화의 x축을 따라 모든 범주를 서로 옆에 배치합니다. 클리블랜드의 규칙에 따르면 인간은 최종 사용자가 각 범주가 공통 척도에 따라 배치되는 위치를 식별하는 데 도움이 되므로 위치를 먼저 식별하는 데 가장 좋습니다. 마찬가지로, 게슈탈트의 폐쇄 원리 법칙은 카테고리 간의 구분 기호로 그래프 사이의 공백으로 반영됩니다.

우리는 시각화에서 길이 시각적 인코딩을 사용하여 YouTube 카테고리의 추세를 파악하는 데 소요되는 평균 시간을 나타냅니다.
또한 길이에 따라 카테고리의 위치를 정렬하고 있습니다. 이를 통해 빠르게 추세를 나타내는 범주와 추세를 나타내는 데 가장 오랜 시간이 걸리는 범주를 인식하는 데 도움이 됩니다. 클리블랜드의 규칙을 적용하면 시각적 인코딩의 길이는 시각화 효율성 순위의 더 높은 스펙트럼에 속합니다. 이는 길이를 사용하여 평균 시간을 나타내면 최종 결과가 의심 없이 전달될 수 있음을 확신할 수 있음을 의미합니다. 게슈탈트 원리의 관점에서 시각화 작업에 Pragnanz의 법칙을 적용할 수 있습니다. 이 법칙은 우리의 마음이 단순함을 좋아하고 규칙적이고 균일하며 질서 있는 단순한 패턴을 찾는 데 끌린다고 말합니다. 이 원리를 사용하면 훨씬 더 빠른 추세를 보이는 카테고리를 한 눈에 빠르게 식별할 수 있으며 그 반대의 경우도 마찬가지입니다.

게다가 우리가 생각한 다음 시각적 인코딩은 색상 인코딩이었습니다. 처음에는 각 시각화에 서로 다른 색상을 사용하는 아이디어가 있었습니다. 그러나 이로 인해 시각화가 상당히 혼란스러워졌고 다양한 색상 스케일이 추가적인 의미를 가져오지 못했습니다. 이러한 이유로 우리는 단일 색상을 사용하여 데이터를 시각화하기로 결정했습니다. 또한 여러 색상을 사용할 때 최종 사용자의 인지 부하를 훨씬 쉽게 비교하고 이해하고 줄일 수 있습니다. Cleveland의 규칙을 적용하면 시각화의 효율성 순위 측면에서 색상 인코딩이 훨씬 나쁩니다. 게슈탈트 원리에 따르면, 같은 색을 사용하므로 유사성의 법칙을 적용할 수 있습니다. 이를 통해 최종 사용자는 쉽게 데이터를 읽고 이해할 수 있습니다. 마찬가지로, 형상과 배경의 법칙도 적용할 수 있습니다. 이 법칙에 따르면 사물은 형상이나 배경으로 인식됩니다. 우리 작업의 맥락에서 차트의 막대는 일반 흰색 배경과 비교하여 전경 개체이므로 먼저 식별되고 표시됩니다. 그림과 지면 사이의 대비는 사용자가 두 개 이상의 물체를 식별하는 데 도움이 됩니다. 마지막으로 차분하고 뉴트럴한 컬러를 표현하기 위해 블루 컬러를 선택했습니다.

Identify most popular tags across countries

Machine Translated by Google

우리는 2020년 9월부터 2022년 1월까지 국가별 YouTube 인기 동영상에서 가장 많이 사용된 태그를 식별하려고 합니다. 시각화를 구현하기 위해 국가별로 그룹화되어 가장 많이 사용된 태그를 나타내는 트리맵 시각화를 개발하려고 합니다. 트리맵 데이터 시각화는 광범위한 계층적 데이터를 표시하고 하나 이상의 범주 간의 데이터 세트에 대한 높은 수준의 요약을 제공하는 데 적합합니다. 각 데이터 값은 크기에 비례하는 일련의 중첩된 직사각형으로 표시됩니다. 차트에서 각 직사각형 상자의 크기는 태그 수를 나타냅니다. 태그는 해당 국가를 기준으로 동봉되어 있습니다. 그래프를 통해 통역사는 시각화에 표시된 각 국가에서 가장 많이 사용되는 태그를 빠르게 찾을 수 있습니다.

작업 맥락에서 단어 클라우드 및 히스토그램과 같은 시각화 기술을 구현할 수도 있습니다. 단어 클라우드 시각화는 해석하기 쉽고 자주 사용되는 태그를 나머지 태그보다 빠르게 눈에 띄게 만드는 데 도움이 됩니다.
그러나 우리는 월드 클라우드가 방대한 데이터 세트로 가장 많이 사용되는 태그를 모두 정확하게 표시하는 효과적인 도구가 아니라는 사실을 발견했습니다.
게다가 여러 국가를 비교하는 것도 상당히 어려울 수 있습니다. 히스토그램 차트의 경우 매우 인기가 높으며 연속 빈도 분포를 나타내는 데 주로 사용됩니다. 언뜻 보면 사용된 태그의 빈도 분포를 표시할 수 있으므로 작업에 적합하다는 것을 알 수 있습니다. 그러나 이 차트를 구현할 때 태그를 나타내기에는 수직 막대가 너무 많습니다. 더욱이, 단어 클라우드 시각화와 유사하게, 다른 국가와의 교차 비교는 효과적이지 않을 수 있습니다.

이 시각화에서는 영역 및 색상 시각적 인코딩을 사용합니다. 영역 시각적 인코딩은 직사각형 상자 내부의 각 태그를 나타내는 데 사용됩니다. 또한 태그는 국가를 나타내기 위해 다른 상자 안에 중첩되어 있습니다. 각 직사각형의 크기는 태그 사용 횟수에 비례합니다. 클리블랜드의 규칙에 따라 해당 영역은 시각화 효율성 순위의 중간 스펙트럼에 속합니다. 그럼에도 불구하고 컬러 인코딩보다 훨씬 효과적인 순위입니다.

마찬가지로, 지역 내부의 모든 태그가 주변 경계로 둘러싸여 있으므로 차트에 게슈탈트의 포위 법칙을 적용할 수 있습니다. 다시 말하지만, Gestalt의 초점 법칙은 가장 많이 사용된 태그를 가장 많이 사용된 태그로 표시합니다.

태그를 시각화하기 위해 색상 시각적 인코딩이 적용됩니다. 이 작업에서는 범주형 색상 척도를 사용하여 국가를 나타냅니다. 시각화에서 각 국가에 고유한 색상을 할당하고 있으며, 국가 내부에 그룹화된 태그는 동일한 색상을 갖습니다. 이를 통해 사용자는 국가를 쉽고 빠르게 식별할 수 있습니다. 클리블랜드의 규칙은 색상 인코딩을 효율성 순위 측면에서 가장 효과적인 것으로 평가합니다. 그러나 컬러 인코딩을 올바르게 사용하면 작업에 따라 효과적인 시각화 결과를 얻을 수 있습니다. 게슈탈트 유사성의 법칙에 따르면 동일한 색상의 태그는 특정 국가에 속한다는 것을 알 수 있습니다.

Analyze the frequency of title length for the trending video

우리는 2020년 9월부터 2022년 1월 사이에 국가별로 동영상의 제목 길이 빈도를 비교하려고 합니다. 초기 단계에서는 히트맵을 사용하여 제목 길이의 빈도를 표시할 계획이었습니다. 여기서 색상 인코딩을 사용하여 제목 길이의 빈도를 표시할 수 있었습니다. . 그러나 색상 인코딩으로 인해 문자 빈도를 구별하기 어려운 몇 가지 제한 사항이 있었습니다. 클리블랜드의 법칙에 따르면 인간은 다른 시각적 인코딩과 비교하여 색상 차이를 인식하기가 어렵습니다.

또한 영상의 등장 빈도와 등장인물 수를 표현하기 위해 누적 밀도 그래프도 고려했습니다.
누적 밀도 그래프는 그래프의 곡선을 사용하여 빈도 분포를 보여줍니다. 그러나 클리블랜드 법칙에 따르면 그래프의 기울기를 위치와 길이에 비해 인간이 분석하기 어렵기 때문에 빈도분포를 해석하는 것은 꽤 어렵다. 따라서 이러한 이유로 히스토그램 차트를 구현하기로 결정했습니다. 이를 통해 비디오 제목의 문자 길이에 대한 빈도 분포를 표시할 수 있습니다. 히스토그램은 간격을 알아내는 데에도 도움이 되며, 데이터의 분포를 이해하고 찾는 것이 매우 쉽습니다. 마찬가지로, 이 시각화를 통해 사용자는 제목 길이의 높이를 빠르게 비교할 수 있습니다.

Machine Translated by Google

이 작업에는 위치, 길이, 색상이라는 세 가지 기본 시각적 인코딩이 사용됩니다. 첫 번째는 x축이 타이틀 길이를 나타내고, y축이 타이틀 길이의 빈도를 나타내는 위치이다. 또한 클리블랜드의 규칙에는 인간이 위치를 정확하게 식별할 수 있다고 명시되어 있습니다.

고려해야 할 두 번째 시각적 인코딩은 길이입니다. 길이 인코딩은 발생에 따른 비디오 제목 길이를 보여줍니다. 또한 길이는 어떤 제목 길이가 가장 빈도가 높은지 즉시 알아내는 데 도움이 됩니다. 클리블랜드 법칙에서 길이가 더 높은 순위에 속하므로 최종 사용자는 히스토그램을 보고 데이터를 빠르게 해석할 수 있습니다. 마찬가지로 Gestalt의 Pragnanz 법칙은 인기 동영상의 제목 길이 간격을 빠르게 해석할 수 있음을 보여줍니다.

사용되는 최종 시각적 인코딩은 색상 선택입니다. 제목 길이의 빈도를 표시하기 위해 파란색을 사용했습니다.
또한, 트렌드가 될 동영상에 가장 많이 사용된 제목 길이를 식별하는 데 있어 사용자의 관심을 끌기 위해 주황색을 사용했습니다. 색상은 가장 낮은 스펙트럼에 속하지만 Cleveland의 규칙에 따라 사용자는 가장 많이 사용되는 제목 길이를 주황색으로 쉽게 식별할 수 있습니다. 게슈탈트의 초점 법칙은 빈도가 높은 제목 길이가 돋보이게 하며, 길이와 색상의 차이로 인해 최종 사용자가 눈에 띄게 됩니다.

Identify the trend over the day of a week with the total number of trending videos

우리는 최대 트렌드 동영상 수로 요일을 보여주는 트렌드를 시각화하고 싶습니다. 이 시각화를 달성하기 위해 우리는 선 그래프를 구현할 계획입니다.
이 그래프는 널리 사용되는 시각화 기술이기 때문에 선택했습니다. 또한 간단하고 이해하기 쉬우며 시간이 지남에 따라 변화하는 가치를 보여주는 데 도움이 됩니다.
이 작업에서 선 그래프는 동영상이 국가별로 인기를 끌 수 있는 이상적인 요일을 식별하기 위한 추세를 보여주는 훌륭한 도구가 될 수 있습니다. 또한 다양한 선을 사용하여 여러 국가의 추세를 표시할 수 있습니다. 이 작업은 막대 차트와 원형 차트를 사용하여 구현할 수도 있습니다. 우리는 막대 차트가 데이터 값의 분포를 보여주고 여러 범주(이 경우에는 요일)의 값을 비교하는 데 도움이 된다는 것을 알았습니다. 막대 차트는 가장 기본적인 차트 중 하나이므로 간단하고 이해하기 쉽고 대부분의 사용자가 액세스할 수 있습니다. 그러나 막대 차트는 시간에 따른 추세를 표시하는 데 이상적인 선택이 아닙니다. 막대 차트와 유사하게 원형 차트도 이해하기 매우 쉽고 매일 가장 인기 있는 동영상이 있는 비율을 보여주는 데 도움이 될 수 있습니다. 그러나 원형 차트는 참조할 수 있는 척도가 없기 때문에 데이터 전달이 좋지 않을 수 있으며, 일반적으로 인간의 마음은 클리블랜드의 법칙에 따라 각도의 크기를 비교하는 데 능숙하지 않습니다.

그래프에 사용되는 시각적 인코딩은 위치, 각도 및 색상입니다. 위치 인코딩은 그래프에서 y축에 집계된 인기 동영상 수와 x축에 요일을 표시하는 데 사용됩니다.

클리블랜드의 규칙은 시각화에서 가장 효과적인 시각적 인코딩으로 순위를 매겼습니다. 이를 통해 사용자는 그래프를 쉽게 해석할 수 있습니다. 게슈탈트의 연결 법칙은 같은 나라에 속한 사물들이 하나의 집단으로 연결되어 있음을 보여줍니다. 각 지점이 다른 지점과 연결되므로 여러 국가의 한 주 동안 인기 급상승 동영상을 쉽게 찾을 수 있습니다. 게슈탈트 근접 법칙은 서로 가까운 선들이 유사한 관심과 특성을 가지고 있음을 보여줍니다.

각도 시각적 인코딩을 사용하여 동일한 국가의 개별 데이터 값을 연결하여 추세를 보여줍니다. 상승하는 기울기는 인기 급상승 동영상의 증가를 나타내고, 하락하는 기울기는 감소하는 추세를 나타냅니다. 클리블랜드의 규칙도 영상의 추세를 효과적으로 보여주기 위해 경사를 지원합니다.

우리가 사용한 또 다른 시각적 인코딩은 색상입니다. 색상 인코딩은 개별 국가를 명확하게 식별하기 위해 사용됩니다. 우리는 각 색상이 국가를 나타내는 범주형 색상 척도를 구현했습니다. 범주형 색상 척도를 통해 사용자는 국가를 빠르게 식별할 수 있습니다. 또한 국가를 대표하기 위해 블라인드 안전 색상을 사용하고 있습니다. 클리블랜드의 규칙에서는 색상 인코딩이 덜 효과적이라고 평가하지만 이를 위치 및 각도 시각적 인코딩과 결합하면 전체 시각화에 깊이가 더해졌습니다. 게슈탈트의 유사성 법칙은 데이터 값이

Machine Translated by Google

같은 색깔은 같은 나라에 속합니다. 형태와 배경의 게슈탈트 법칙은 사용자가 그래프의 정보와 범례를 한눈에 볼 수 있음을 보여줍니다.

6 Visualization And Analysis

시각화는 프런트 엔드에서 ReactJS, Nivo 차트 및 JavaScript를 사용하여 생성되었습니다. 데이터는 Python 및 FastAPI 라이브러리를 사용하여 구현된 API 서버에서 생성됩니다. 데이터 세트는 Python 명령줄 스크립트를 사용하여 MongoDB 데이터베이스로 가져왔고 여기서 데이터를 사전 처리하여 속성을 추출하고 변환했습니다. 시각화는 인기 페이지에서 비디오 기능을 만드는 데 각 특성이 어떻게 도움이 되는지 보여주는 결과를 생성합니다. 따라서 콘텐츠 제작자는 자신의 동영상을 트렌드로 만드는 방법을 더 잘 이해할 수 있습니다.

Correlation between likes and views

본 작업에서는 2020년 9월부터 2022년 1월까지 인기 급상승 동영상의 조회수와 좋아요 간의 상관관계를 분석했습니다.
우리의 주요 목표는 가장 높은 조회수와 비슷한 수의 좋아요가 있는지 파악하고 두 속성 간의 관계를 탐색하는 것이었습니다. 산점도는 그림 1에 표시된 것처럼 좋아요와 조회수 간의 관계를 효과적으로 시각화합니다.

데이터베이스에서 각 동영상의 조회수와 좋아요 수를 추출했습니다. 시각화를 생성하기 위해 Nivo 차트 산점도 구성 요소를 사용했습니다. 색상, 레이블, 배율 등의 추가 매개변수도 구성 요소에 제공되었습니다.
사용자는 국가 메뉴와 상호 작용하여 다른 국가 간에 전환하도록 선택할 수도 있습니다. React 인스턴스는 API에서 데이터를 가져오고 시각화가 새 데이터로 다시 렌더링되도록 합니다.

그림에 선형 회귀선이 표시되어 있으므로 좋아요와 조회수 사이에 강한 상관관계가 있음을 확인할 수 있습니다. 또한, 조회수와 좋아요 수가 밀집되어 있어 낮은 범위에 밀도가 나타나는 것을 확인할 수 있습니다. 게슈탈트의 근접 법칙은 시각화에 적용됩니다. 여기서 낮은 범위의 조회수와 좋아요는 멀리 있는 지점보다 조회수와 더 관련이 있습니다. 산점도를 사용하면 최종 사용자가 그림을 즉시 보고 좋아요와 조회수 사이의 관계를 알 수 있으며 여기에는 이상값이 없다는 것도 알 수 있습니다. 버블 차트와 히트맵은 좋아요와 조회수에 대한 상관관계와 이상치를 완벽하게 표시하지 못했을 것입니다. 시각화를 통해 조회수가 약

6000만 명도 대략 600만 명과 비슷한 수치를 보였습니다. 또한 Cleveland's rule에 따라 표준 눈금 간의 위치를 적용하여 최종 사용자에게 시각화를 명확하게 해줍니다. 마찬가지로, 산점도의 각 점에 대해 동일한 모양과 크기를 사용하면 시각적으로 보기에 좋습니다. 또한, 각 포인트에 동일한 색상을 사용함으로써 시각화가 단순해 보였고 최종 사용자는 조회수와 좋아요가 어떻게 연관되어 있는지에 대한 정보를 쉽게 이해할 수 있었습니다. 이는 동일한 색상과 크기를 사용하므로 게슈탈트 유사성의 법칙에도 적용됩니다. 따라서 그림 1의 시각화는 시각적 인코딩을 사용하여 조회수와 좋아요 간의 관계를 정확하게 보여 주므로 앞서 디자인 단계에서 언급한 대로 디자인을 더욱 강화합니다.

Compare average hours for trending videos across YouTube categories.

2020년 9월부터 2022년 1월까지 카테고리 전반에 걸쳐 YouTube 인기 동영상에 소요된 평균 시간을 비교하는 작업은 그림 2에서 답할 수 있습니다.
막대 차트는 미국의 모든 카테고리를 효과적으로 시각화하며 각 동영상에 소요되는 평균 시간이 인기를 끌고 있습니다. . 우리는 이 시각화에 대한 평균 소요 시간이 여러 범주에 걸쳐 일관성이 있는지 확인하는 것을 목표로 합니다. 또한 사용자가 국가 간을 전환하여 다양한 카테고리가 추세를 나타내는 데 걸리는 평균 시간이 비슷한지 다른지 확인할 수 있도록 시각화를 대화형으로 만들었습니다.

Machine Translated by Google

시각화를 위해 게시 날짜, 추세 날짜 및 카테고리 이름 속성이 데이터베이스에서 추출되었습니다.
게시 날짜와 인기 급상승 날짜 간의 차이를 계산하여 동영상이 인기 급상승하는 데 걸리는 시간을 생성했습니다. 마지막으로 계산된 시간을 각 범주별로 평균하여 집계했습니다. 그런 다음 이 데이터를 React 프런트 엔드에서 가져와 Nivo 막대 그래프 모듈을 사용하여 막대 그래프를 렌더링합니다.

막대 그래프에서 얻을 수 있는 즉각적인 통찰력은 스포츠 카테고리가 미국의 다른 카테고리에 비해 트렌드가 되기까지 최소 평균 시간(약 115시간)이 걸린다는 것입니다. 그 다음에는 비영리 단체와 활동주의가 119시간을 소요합니다. 시간, 여행 및 이벤트 카테고리는 평균 123시간입니다. 마찬가지로 음악 카테고리는 인기를 얻는 데 가장 오랜 시간이 걸리며 139시간입니다. 막대 그래프를 사용하면 클리블랜드 규칙에 따른 위치와 길이로 인해 추세가 되는 데 가장 짧은 시간과 가장 긴 시간이 걸리는 범주를 사용자가 빠르게 식별하는 데 도움이 됩니다. 또한, 파란색이라는 단일 색상을 적용하여 그래프가 더욱 투명하고 이해하기 쉬웠습니다. 이전 디자인 단계에서 설명한 것처럼 도넛 차트와 버블 차트는 영역 인코딩을 고려했으며 클리블랜드의 규칙에 따라 인간은 영역보다 길이를 가장 잘 식별합니다.

또한, 평균 근무시간이 추세와 유사한 카테고리도 찾을 수 있습니다. 막대 그래프를 보면 평균 시간이 132시간으로 유사한 교육, 엔터테인먼트, 정치 뉴스 등의 카테고리를 확인할 수 있습니다. 코미디, 애완동물, 동물 등 다른 카테고리도 평균 시간이 133시간으로 동일했습니다. 마찬가지로, 자동차와 게임 카테고리, 영화 애니메이션, 하우투 스타일 카테고리도 각각 125시간과 134시간으로 비슷한 평균 시간을 보였습니다. 이 시각화는 또한 사용자가 각 막대에 연결된 숫자와 비교하여 막대 그래프를 처음 보는 게슈탈트의 그림 및 기본 원리를 적용합니다. 마찬가지로 게슈탈트의 유사성 법칙은 유사한 범주를 식별하는 데 도움이 되었습니다. 초점 및 컨텍스트 탐색 전략을 통해 사용자는 각 범주에 소요되는 정확한 평균 시간을 더 쉽게 찾을 수 있습니다. 또한 위치, 길이 및 색상과 같은 시각적 인코딩에 대해 이전 단계에서 취한 디자인 결정은 범주 전반에 걸쳐 추세를 나타내는 데 소요되는 평균 시간을 알아내기 위해 달성하려고 했던 작업 목표를 더욱 강화하는 데 도움이 됩니다.

Identify most popular tags across countries

우리는 2020년 9월부터 2022년 1월까지 국가 전체에서 가장 많이 사용되는 태그를 분석하고 싶었습니다. 그림 3의 트리맵은 11개 국가에서 가장 많이 사용되는 상위 10개 태그를 보여줍니다. 마찬가지로 트리맵은 여러 국가에서 사용되는 각 단어 태그를 효과적으로 시각화합니다. 시각화를 통해 우리는 가장 많이 사용되는 태그가 국가별로 유사한지 확인하려고 합니다. 또한 당사의 대화형 시각화를 통해 최종 사용자는 상위 10위부터 상위 50위까지 태그를 입력할 수 있어 가장 많이 사용되는 태그에 대한 통찰력을 얻을 수 있습니다.

이 시각화를 위해 Nivo 차트 트리맵 모듈을 사용하여 그래프를 생성했습니다. 한 국가의 각 비디오에 대한 태그가 데이터베이스에서 추출되었습니다. 태그는 단일 텍스트 형식이었기 때문에 개별 단어로 더 분리되었습니다. 다음 단계에서는 태그 수를 계산하고 이를 국가별로 집계했습니다. 프런트 엔드 앱은 처리된 데이터를 수집하고 개별 국가의 색상 및 국경 크기와 같은 추가 속성이 포함된 Nivo 차트 트리맵 구성 요소를 사용하여 그래프를 초기화합니다.

시각화를 살펴보면 모든 국가에 다른 태그에 비해 빈도가 가장 높은 빈 태그가 있는 것을 확인할 수 있습니다. 그래프는 각 태그의 면적을 보여주며, 면적의 크기는 태그 사용 횟수에 비례합니다. 거의 모든 국가에는 태그가 없는 더 넓은 영역이 있었습니다. 즉, 인기 동영상에는 태그가 사용되지 않았습니다. 마찬가지로, 이 시각화는 국경이 각 국가를 둘러싸는 게슈탈트의 폐쇄 법칙을 적용한다는 것을 알 수 있습니다. 또한 대부분의 국가에서 vlog, 코미디, 재미, 축구와 같은 태그가 자주 발생한다는 통찰력도 얻을 수 있습니다.

설계 단계에서 정의한 대로 트리맵은 국가 전체에서 가장 자주 사용되는 태그를 하나의 차트에 포함하여 최종 사용자가 국가 간 태그를 쉽게 비교할 수 있기 때문에 이 작업에 완벽한 선택이라는 것을 알았습니다. 마찬가지로, 워드클라우드를 사용하는 경우에는 데이터세트의 양이 많아 가장 많이 사용되는 태그를 빠르게 식별할 수 없습니다. 또한 빈도를 표시하는 데는 히스토그램이 가장 좋았겠지만, 국가별로 태그를 비교하는 것이 더 나았을 것입니다.

Machine Translated by Google

많은 히스토그램을 만들었습니다. 단일 시각화로 국가별 태그를 비교하는 것은 쉬운 일이 아닐 것입니다.
마찬가지로, 시각적 인코딩은 명확하고 간단한 시각화를 만드는 데 도움이 되는 반면, 색상 인코딩은 여러 국가를 구분하는 데 도움이 됩니다. 범주형 색상 척도를 사용하면 최종 사용자가 다양한 국가를 빠르게 식별할 수 있고 각 국가 아래의 각 태그를 볼 수 있으므로 도움이 됩니다. 결과적으로 그림 3의 시각화는 국가 전체에서 가장 많이 사용되는 태그를 효과적으로 보여줍니다.

Analyze the frequency of title length for the trending video

그림 4의 히스토그램 시각화는 추세에 도달하기 위해 비디오의 제목 길이가 얼마나 되어야 하는지에 대한 통찰력을 제공함으로써 2020년 9월부터 2022년 1월까지 추세 비디오의 제목 길이를 보여줍니다. 색상을 사용하면 최종 사용자가 비디오의 인기를 끌기 위해 필요한 제목 길이를 빠르게 식별할 수 있습니다. 우리의 주요 동기는 최종 사용자와 콘텐츠 제작자가 최적의 제목 범위에 대한 개요를 얻을 수 있도록 제목 길이의 분포를 발견하는 것이었습니다.
그들의 비디오.

히스토그램 시각화를 위해 데이터베이스에서 각 비디오의 제목을 추출한 다음 제목의 길이를 계산했습니다.
그런 다음 제목 길이를 빈도별로 집계했습니다. 이후 ReactJS 자바스크립트 라이브러리는 API 서비스에서 처리된 데이터를 가져오고 제목 길이의 빈도 분포를 사용하여 작업을 완료하는 Data-UI 히스토그램 모듈을 사용하여 그래프를 렌더링합니다.

추세를 나타내는 비디오의 최적 제목 범위는 제목 길이가 40~50자라는 것을 히스토그램을 통해 즉시 알 수 있습니다.

Cleveland의 위치 및 길이 인코딩은 사용자가 비디오 제목의 길이를 빠르게 비교할 수 있도록 도와줍니다. 그 다음에는 제목 길이가 30-40 및 50-60 범위입니다. 가장 적게 사용되는 제목 길이는 0에서 10 사이입니다. 디자인 단계에서 설명했듯이 시각적 인코딩은 이 시각화 사용에 대한 선택을 강화했습니다. 또한 콘텐츠 제작자가 인기를 끌 수 있는 짧은 설명 제목보다 40~50 범위의 설명 제목을 가질 수 있는 두 번째 작업에 대한 통찰력도 얻었습니다.

마찬가지로, 색상 인코딩은 가장 높은 빈도의 제목 범위를 강조하여 최종 사용자가 첫눈에 통찰력을 얻을 수 있도록 시각적으로 매력적으로 만듭니다.
히트맵과 누적 밀도 그래프는 인간이 클리블랜드 법칙에 따라 길이보다 색상과 경사를 분석
하는 데 더 약하기 때문에 히스토그램이 완벽하게 시각화하므로 작업 결과를 제공할 수 없었을 것입니다. 또한, 최적의 범위를 강조하는 오렌지 컬러와 함께 여기에도 게슈탈트의 초점 법칙이 적용되어 트렌드에 맞는 최적의 타이틀 길이를 제공합니다.

Identify the trend over the day of a week with the total number of trending videos

마지막 작업을 위해 인기 급상승 동영상의 최대 조회수에 대한 최적의 요일을 식별하려는 동기는 그림 5에서 볼 수 있습니다. 우리는 동영상을 만드는 데 사용자 참여가 최대인 요일을 찾고 싶었습니다. 트렌드 목록에 있습니다. 마찬가지로 우리는 국가별로 유사한지 비교하고 싶었습니다. 선 그래프는 일주일 동안의 추세를 표시하는 데 도움이 되며 참여도가 최대인 하루에 대한 통찰력을 제공합니다.

그래프를 구현하기 위해 먼저 데이터베이스에서 트렌드 데이터와 조회수를 수집했습니다. 인기 날짜 속성은 Pandas 라이브러리를 사용하여 요일로 변환되었습니다. 다음 단계에서는 조회수를 요일별로 집계했습니다. ReactJ는 API에서 처리된 데이터를 로드한 다음 프런트 엔드에서 Nivo 라인 차트 구성 요소를 인스턴스화합니다. 꺾은선형 차트 구성 요소는 데이터, 색상, 레이블 및 축 눈금을 가져온 다음 그래프를 렌더링합니다.

Machine Translated by Google

선 그래프를 보면 브라질의 동영상 수가 11개국 중 가장 낮고 미국이 가장 높습니다.
조회수. 색상 인코딩을 사용하고 범주형 척도를 구현하면 사용자가 차별화하는 데 도움이 됩니다.
다른 나라 사이. 마찬가지로 선 그래프를 통해 미국, 러시아,
멕시코, 한국, 일본, 인도, 영국은 일요일에 사용자 참여도가 가장 높았습니다. 처음에는 디자인을 할 때
단계에서는 서로 다른 국가의 선 그래프가 서로 얽힐 것이라고 생각했습니다. 그러나 우리의 최종 시각화는
조회수 기준으로 어느 국가도 충돌하지 않는 등 우리가 생각했던 것과는 전혀 달랐습니다. 마찬가지로,
위치 인코딩을 사용하여 사용자가 x축에 요일을, y축에 조회수를 식별할 수 있도록 했습니다.
또한, 서로 닫힌 선들이 비슷한 모양을 나타내는 곳에 게슈탈트의 근접 법칙이 적용되는 것도 볼 수 있습니다.
해당 요일에 더 높은 참여에 대한 관심.

국가들이 비슷한 추세를 가지고 있는지 찾는 두 번째 작업에 대한 통찰력을 얻을 수 있습니다. 위에 나열된 국가는
금요일에 참여도가 가장 낮고 일요일에 가장 높은 유사한 경향이 있습니다. 또한 브라질의 경우 최대
참여도는 월요일에 있었고 금요일과 수요일에 가장 낮았습니다. 마찬가지로 캐나다, 독일, 프랑스도
일요일과 월요일에 가장 높았고, 금요일에 가장 낮았습니다. 11가지로
국가를 선 그래프를 사용하여 하나의 시각화로 표현함으로써 사용자는 국가 간 추세를 쉽게 비교할 수 있습니다.
반면에 막대 차트나 원형 차트를 사용했다면 각 국가마다 11개의 다른 수치가 있었을 것입니다. 따라서 우리는
선 그래프를 보면 대부분의 국가에서 인기 동영상을 게시하기에 가장 좋은 날은 일요일입니다.

7 Conclusion

이 문서를 작성하게 된 동기는 동영상이 전 세계적으로 유행하는 것으로 판단하는 특성을 조사하는 것이었습니다.
국가. 이 보고서는 좋아요와 조회수 간의 상관관계, 인기를 얻는 데 소요되는 평균 시간을 파악하는 데 도움이 되었습니다.
다양한 카테고리, 국가에서 가장 많이 사용되는 태그, 최적의 제목 길이 범위 및 당일 추세
동영상을 게시하기에 가장 좋은 날과 함께 일주일을 선택하세요. 마지막 작업에 대한 시각화는 우리를 놀라게 했습니다.
국가의 선 그래프가 서로 겹쳐져 있는 반면, 다른 작업의 결과는 전 세계에 걸쳐 알 수 있어 흥미로웠습니다.
국가. 이 보고서는 동영상을 제작하려는 국가의 콘텐츠 제작자와 사용자에게 도움이 될 것입니다.
보고서에 설명된 각 작업을 통해 인기 급상승 동영상의 특성에 대한 개요를 얻을 수 있습니다.
우리의 주요 동기였습니다.

References

1. JF Andry, SA Reynaldo, K. Christianto, FS Lee, J. Loisa 및 AB Manduro. YouTube의 인기 동영상 알고리즘 분류, 연관 및 클러스터링을 사용한 분석. ~ 안에 2021 국제회의 ~에 데이터와 소프트웨어 엔지니어링(ICoDSE) , pp. 1~6. IEEE, 2021.
2. I. Barjasteh, Y. Liu, H. Radha. 인기 동영상: 측정 및 분석. arXiv 사전 인쇄본 arXiv:1409.7733, 2014.
3. X. Cheng, C. Dale, J. Liu. 인터넷 단편영상 공유의 특징 이해: 유튜브를 사례연구로. arXiv 사전 인쇄본 arXiv:0707.3670, 2007.
4. G. 가자나야케 및 T. 산다나야케. 감성을 활용한 유튜브 게임 채널의 트렌드 패턴 식별 분석. ~ 안에 2020년 20일 국제회의 혼 진출 ICT를 위한 신흥 지역 (ICTer), pp. 149-154. IEEE, 2020.
5. W. Hoiles, A. Aprem, V. Krishnamurthy. YouTube 동영상의 참여도와 인기 역학 및 동영상에 대한 민감도 메타데이터. 지식 및 데이터 공학에 관한 IEEE 거래, 29(7):1426–1437, 2017.
6. R. 샤르마. 유튜브 트렌드 비디오 데이터세트, 2022년. https://www.kaggle.com/rsrishav/youtube-trending-video-dataset.
7. 스타티스타. 전 세계에서 가장 인기 있는 소셜 네트워크(2022년). https://www.statista.com/statistics/272014/global-social-네트워크 - 사용자 수에 따른 순위/.

Figures

Machine Translated by Google

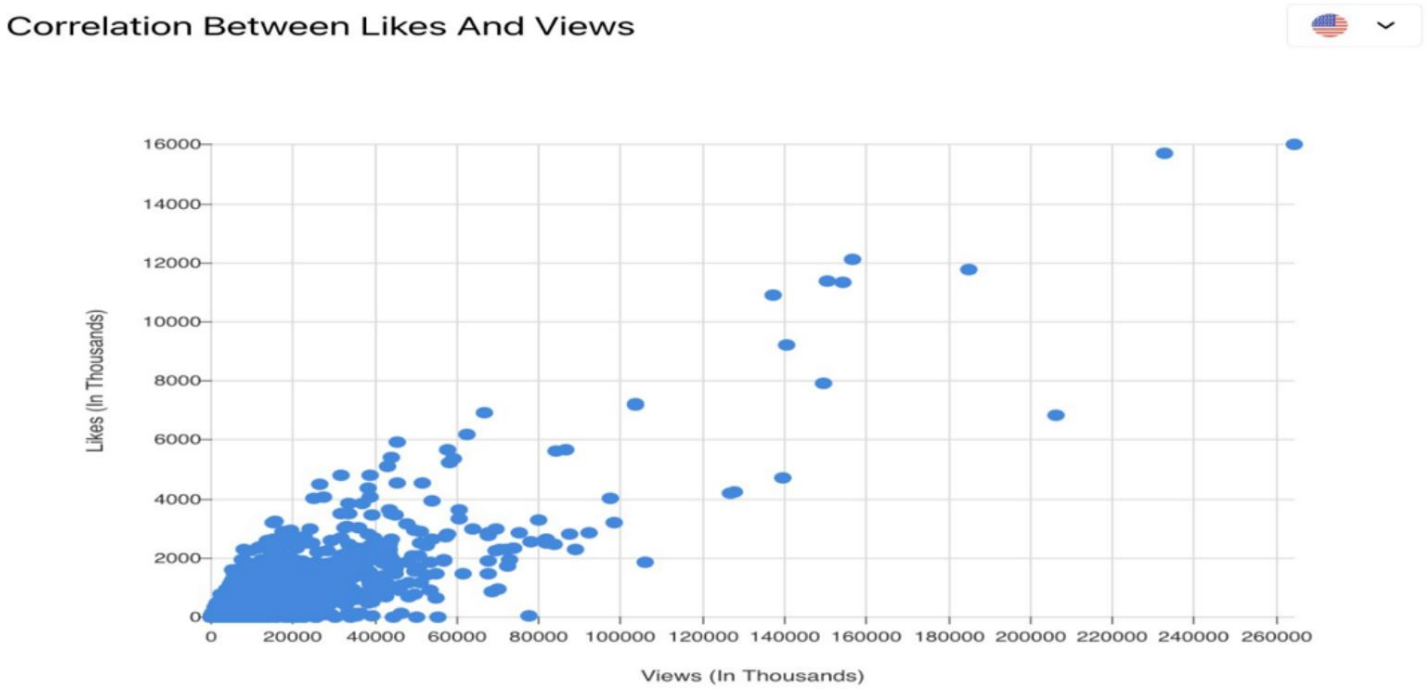

Figure 1

좋아요와 조회수의 상관관계.

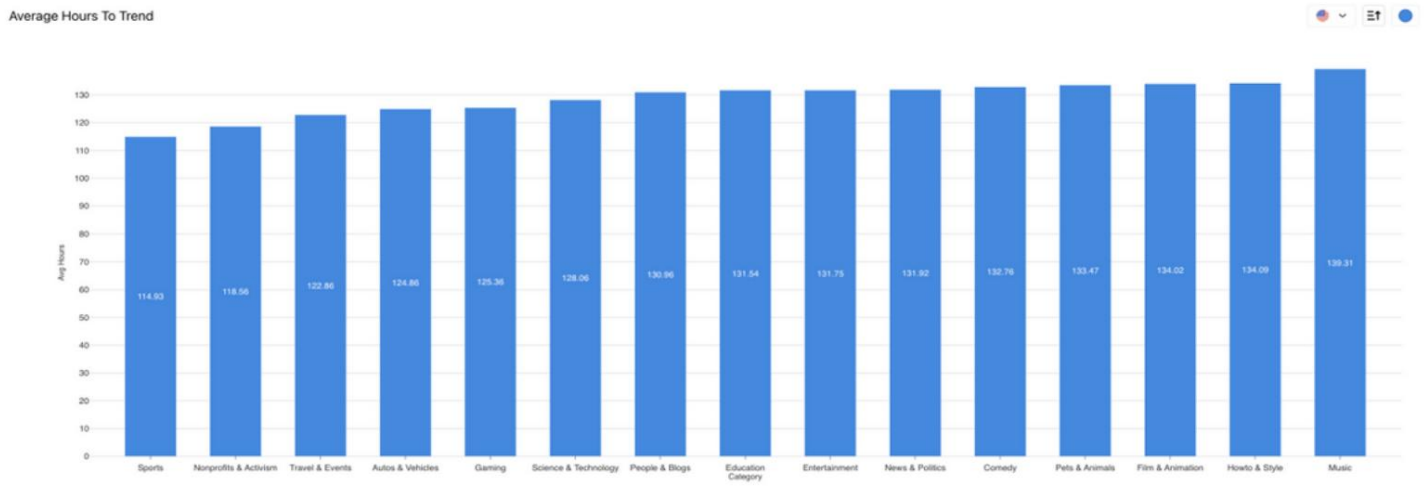

Figure 2

YouTube 카테고리 전체에서 인기 동영상의 평균 시간을 비교하는 막대 그래프입니다.

Machine Translated by Google

Figure 3

국가별로 가장 많이 사용되는 태그의 트리맵입니다.

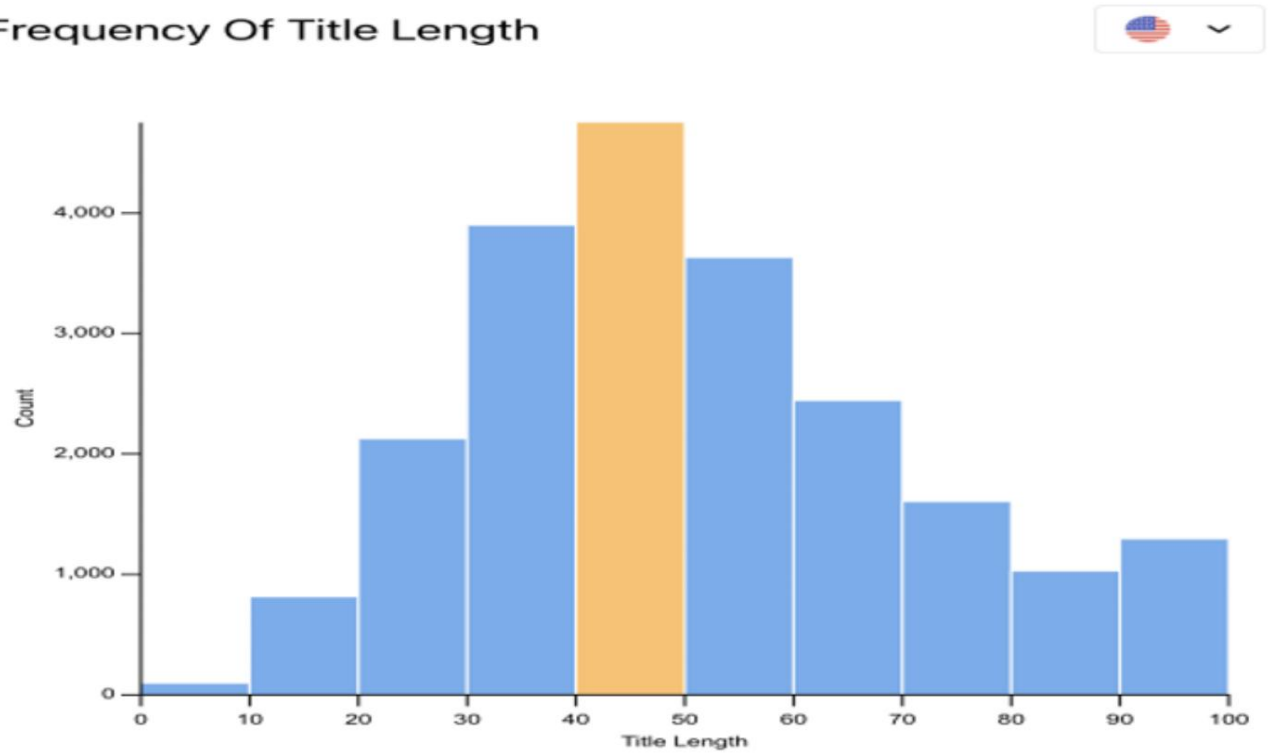

Figure 4

인기 동영상의 제목 길이 빈도 분포를 보여주는 히스토그램입니다.

Machine Translated by Google

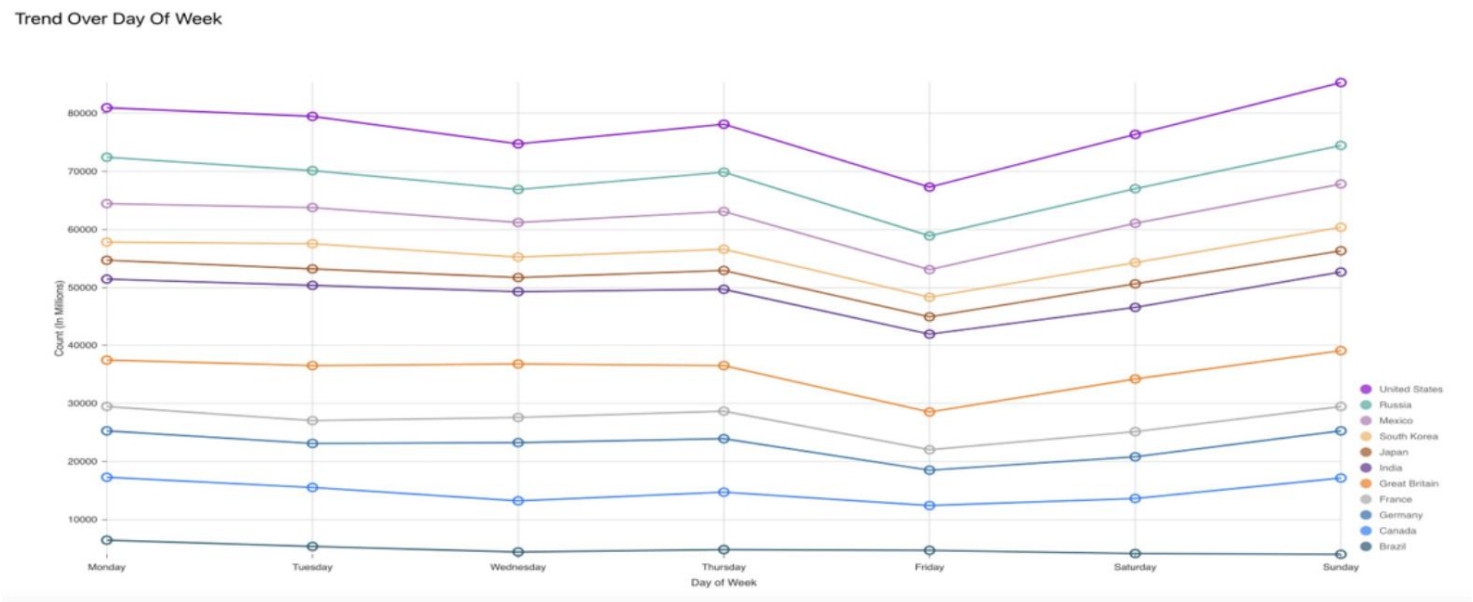

Figure 5

총 인기 동영상 수와 함께 일주일 동안의 추세를 보여주는 선 그래프입니다.