Zajem in analiza metapodatkov v omrežju BitTorrent s pomočjo porazdeljene razpršene tabele

Anton Luka Šijanec
izr. prof. dr. Mojca Ciglarič

Fakulteta za računalništvo in informatiko
Fakulteta za matematiko in fiziko
Univerza v Ljubljani

2026

Pozdravljeni! Danes vam bom predstavil svojo diplomsko nalogo "Zajem in analiza metapodatkov v omrežju BitTorrent s pomočjo porazdeljene razpršene tabele", ki sem jo izdelal pod mentorstvom doktorice Mojce Ciglarič.

Uvod: Motivacija

- BitTorrent (2001): protokol vsak z vsakim za izmenjavo datotek, znan po svoji uporabi za kršenje avtorskih pravic.
- Zanima nas, za kaj je uporabljen (legalnost in kategorije vsebine), prebivalci katerih držav predstavljajo največji delež uporabnikov in druge tehnične podrobnosti.
- Problem: Kako zajeti pošten vzorec prenesene vsebine?

Protokol BitTorrent, star že četrt stoletja, se uporablja za izmenjavo datotek brez potrebe po centralnem strežniku. Takim brezstrežniškim protokolom pravimo "vsak z vsakim", ker so si računalniki v njem enakovredni in se medsebojno povezujejo. Protokol je znan po tem, da se ga uporablja za kršenje avtorskih pravic – za ilegalen prenos filmov, glasbe itd.
Radi bi izvedeli, za kakšne vrste vsebin je uporabljen in koliko vsebine v omrežju je piratske narave. Radi bi izvedeli tudi, iz katerih držav so

Uvod: Motivacija

- BitTorrent (2001): protokol vsak z vsakim za izmenjavo datotek, znan po svoji uporabi za kršenje avtorskih pravic.
- Zanima nas, za kaj je uporabljen (legalnost in kategorije vsebine), prebivalci katerih držav predstavljajo največji delež uporabnikov in druge tehnične podrobnosti.
- Problem: Kako zajeti pošten vzorec prenesene vsebine?
  - Ga zajemamo iz sledilniških strežnikov, ki koordinirajo uporabnike?
  - Ga prestrezamo na usmerjevalnikih omrežnih operaterjev?

Kako priti do teh podatkov? Lahko bi recimo pridobili dnevnike sledilniških strežnikov, ki koordinirajo povezave med uporabniki in sporočajo, kdo ima kakšno datoteko. Lahko bi tudi prestrezali promet na nivoju usmerjevalnikov omrežnih operaterjev in zaznali, kaj se prenaša. Toda za ta dva načina pridobivanja podatkov bi potrebovali nek priviligiran dostop. V primeru prestrezanja prometa pri operaterju pa bi dobili geografsko popačeno sliko. Naša rešitev je drugačna.

Uvod: Motivacija

- BitTorrent (2001): protokol vsak z vsakim za izmenjavo datotek, znan po svoji uporabi za kršenje avtorskih pravic.
- Zanima nas, za kaj je uporabljen (legalnost in kategorije vsebine), prebivalci katerih držav predstavljajo največji delež uporabnikov in druge tehnične podrobnosti.
- Problem: Kako zajeti pošten vzorec prenesene vsebine?
  - Ga zajemamo iz sledilniških strežnikov, ki koordinirajo uporabnike?
    - Ne, nimamo dostopa.
  - Ga prestrezamo na usmerjevalnikih omrežnih operaterjev?
    - Ne, nimamo dostopa. Pa tudi tak vzorec bi bil geografsko pristranski.

Kako priti do teh podatkov? Lahko bi recimo pridobili dnevnike sledilniških strežnikov, ki koordinirajo povezave med uporabniki in sporočajo, kdo ima kakšno datoteko. Lahko bi tudi prestrezali promet na nivoju usmerjevalnikov omrežnih operaterjev in zaznali, kaj se prenaša. Toda za ta dva načina pridobivanja podatkov bi potrebovali nek priviligiran dostop. V primeru prestrezanja prometa pri operaterju pa bi dobili geografsko popačeno sliko. Naša rešitev je drugačna.

Uvod: Cilji

- Leta 2008 standardizirajo uporabo DHT za iskanje soležnikov. BitTorrent ne potrebuje več centralnih strežnikov. Podatki so shranjeni porazdeljeno pri uporabnikih omrežja.
- Kot uporabnik omrežja imamo shranjen majhen del podatkov.
- Cilj: izdelati program, ki beleži poizvedbe po soležnikih torrentov, ki jih prejema sodelovanjem v omrežju DHT. Te poizvedbe vsebujejo identifikator torrenta, s katerim lahko pridobimo metapodatke o torrentu (imena datotek, njihove velikosti in naslove IP prenašalcev).
- Cilj: Pridobljene metapodatke analizirati in sklepati o lastnostih omrežja BitTorrent.

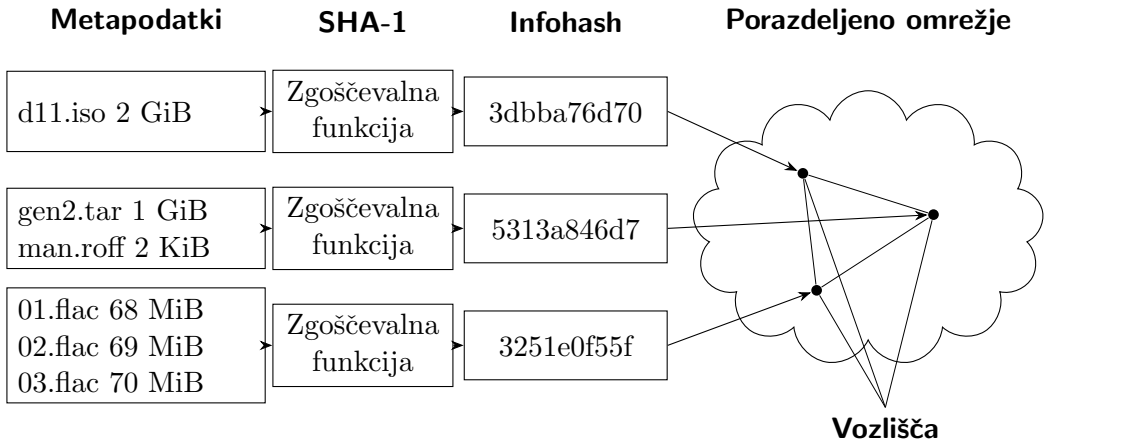
Omenili smo sledilniške strežnike, ki koordinirajo, kdo ima katero datoteko. Leta 2008 so standardizirali protokol, ki te podatke hrani v porazdeljeni razpršeni tabeli pri uporabnikih omrežja tako, da vsak uporabnik hrani nek majhen del podatkov o tem, kdo ima kakšno datoteko.
Naši iskani podatki so torej shranjeni na vseh članih omrežja (vseh odjemalcih BitTorrent), standardiziran pa je način poizvedovanja po njih. Poizvedbe vsebujejo zgoščene vrednosti (info-

Uvod: za nalogo potrebne osnovne obeh protokolov

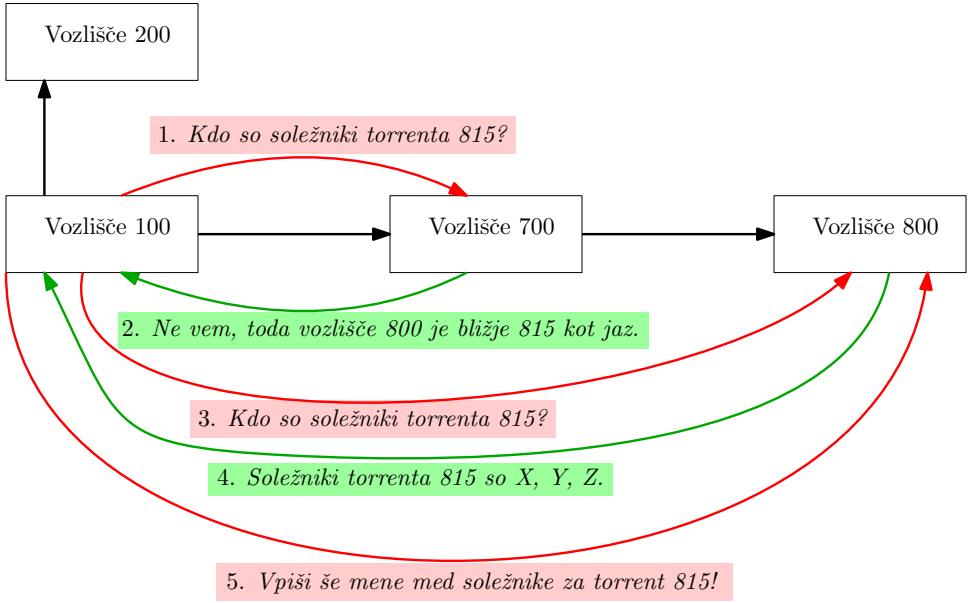
- BitTorrent (TCP): Omogoča prenos metapodatkov od odjemalca za podan *infohash*. Vsebine datotek **nismo** prenašali (bilo bi ilegalno).

Za to potrebujemo znanje o delovanju dveh ključnih protokolov: BitTorrent in DHT. BitTorrent deluje preko TCP in se uporablja za prenos imen in vsebine datotek, ko enkrat vemo, kateri računalniki na internetu te datoteke nudijo na voljo. Mi bomo prenašali samo imena datotek, vsebine pa ne.

Porazdeljena razpršena tabela (DHT) je na visokem nivoju slovar *infohash* $\mapsto$ soležniki. Vozlišča (računalniki) hranijo vrednosti pod tistim ključem, ki je blizu* fiksnega naključnega ID. Poizvedbo se po UDP vozliščem pošilja iterativno, na vsakem koraku smo bodisi preusmerjeni na *infohashu* bližje vozlišče, dokler ne najdemo vozlišč, ki hranijo iskan ključ.



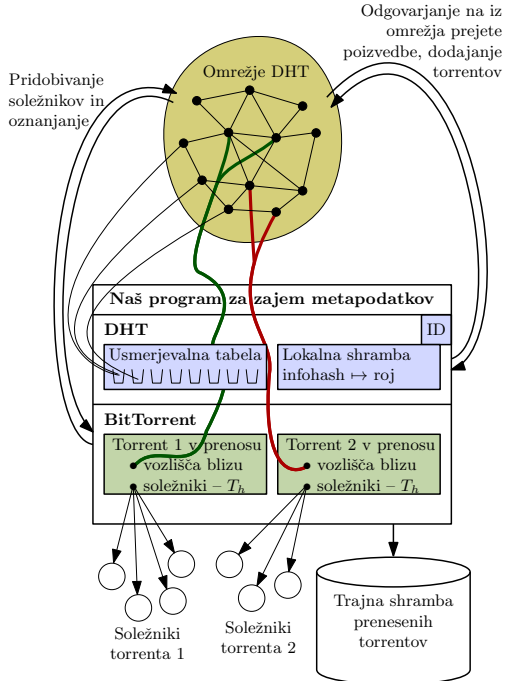
Da izvemo, kateri računalnik (naslov IP) nam lahko ponudi datoteko, pa moramo pogledati v porazdeljeno razpršeno tabelo. Ta je na visokem nivoju slovar, ki pod ključem zgoščene vrednosti torrenta hrani internetne naslove računalnikov (tako imenovani soležniki), ki datoteke nudijo na voljo ali si jih želijo prenesti.
Implementacijo DHT si lahko predstavljamo kot graf, kjer so vozlišča računalniki, vsako si izbere naključen 20-bajtni ID. Podatki o soležnikih niso shranjeni naključno, saj bi v tem primeru za



Iterativna poizvedba v povprečju obišče $\log_2 V(G)$ vozlišč.

Vozlišči sta povezani, če poznata drug drugega (torej internetni naslov in ID). Vsako vozlišče ima usmerjevalno tabelo z aktivnimi vozlišči, ki jih pozna. Ko vozlišče prejme poizvedbo po podatkih, bodisi vrne podatke bodisi napoti iskalca podatkov k drugemu, bližjemu vozlišču. Zaradi specifične strukture usmerjevalne tabele bo poizvedba preusmerjena preko logn vozlišč, preden pride do pravega vozlišča.

Zamislimo si, da smo vozlišče 700. Pravkar smo



Naš program deluje kot eno vozlišče v DHT, povezano z nekaj drugimi vozlišči. Ko dobi poizvedbo o nekem torrentu od drugega vozlišča, si, po tem ko nanjo vestno odgovori, zabeleži, da torrent obstaja, in ga doda v vrsto za prenos. Tega normalni odjemalci ne počno – oni prenašajo samo torrente, ki jih želi prenesti uporabnik. Ko naš program najde v DHT podatek o enem uporabniku, ki torrent ima, se poveže nanj in od njega prenese metapodatke – imena datotek in njihove velikosti in jih shrani v trajno shrambo.

Metodologija: Program za zajem

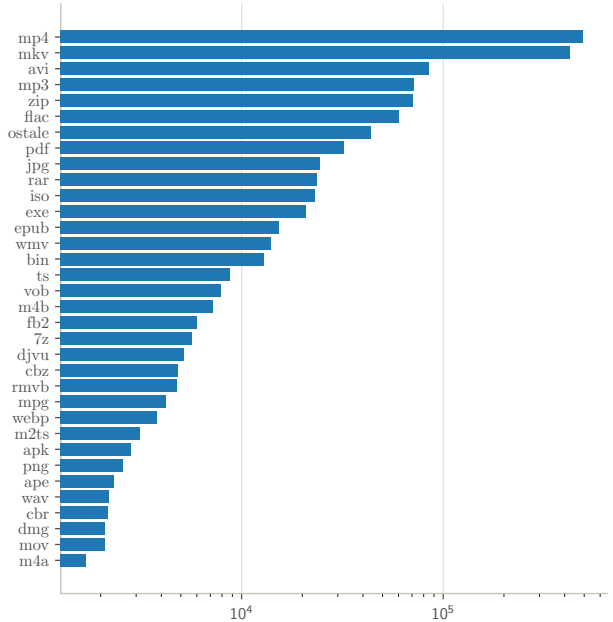
- Spisali smo odjemalec v jeziku c, ki sodeluje kot vozlišče v DHT.
- Zabeleži si *infohash* vsake poizvedbe, ki mu je poslana (in nanjo po najboljših močeh odgovori).
- Prenese metapodatke o torrentu vsakega tako najdenega *infohasha*.
- Obstoječe knjižnice za BitTorrent+DHT niso narejene za tako beleženje poizvedb. Predelave smo raje implementirali sklad BitTorrent+DHT ročno.

Program smo zaradi učinkovitosti spisali v jeziku c in smo vse dele DHT in BitTorrent spisali ročno, saj obstoječi odjemalci niso namenjeni takemu početju, predelava pa bi bila prav tako zahtevna, kot vnovično pisanje celega sklada protokolov.

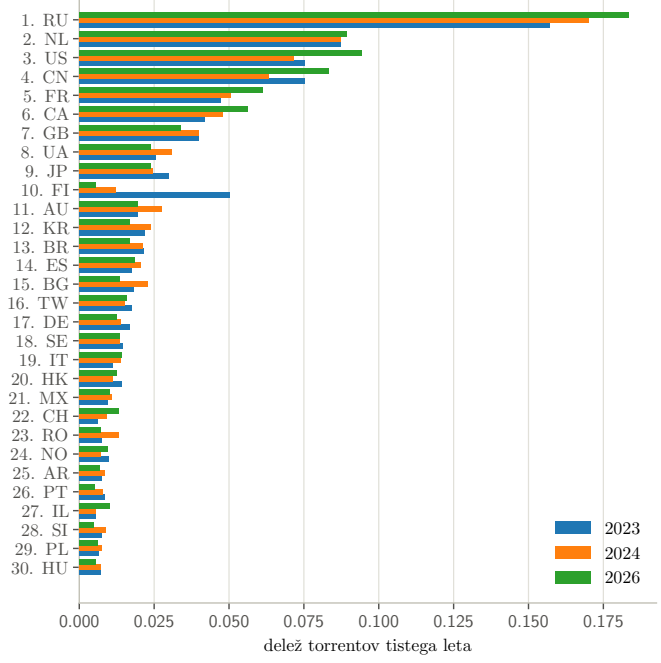
Rezultati: pregled in vsebina

- Torrente smo na predstavljen način zbirali v letih 2023, 2024 in 2026 in skupaj zbrali 1 546 563 torrentov.
- Z jezikovnim modelom smo torrentom določili legalnost in kategorijo.
  - Piratski torrenti: 93,00 %, ilegalnih še 1,86 %
  - Legitimnih: 2,36 %, nedoločljivih 2,79 %
  - Po kategorijah: 28,9 % pornografije, 15,2 % TV oddaj in serij, 14,7 % filmov, 11,5 % glasbe
- Imamo samo imena datotek, ne pa vsebine! Prenos vsebine bi zahteval 9,112 PiB prostora.
- Največ uporabnikov torrentov je iz Rusije, Nizozemske, ZDA, Kitajske, Francije, ...

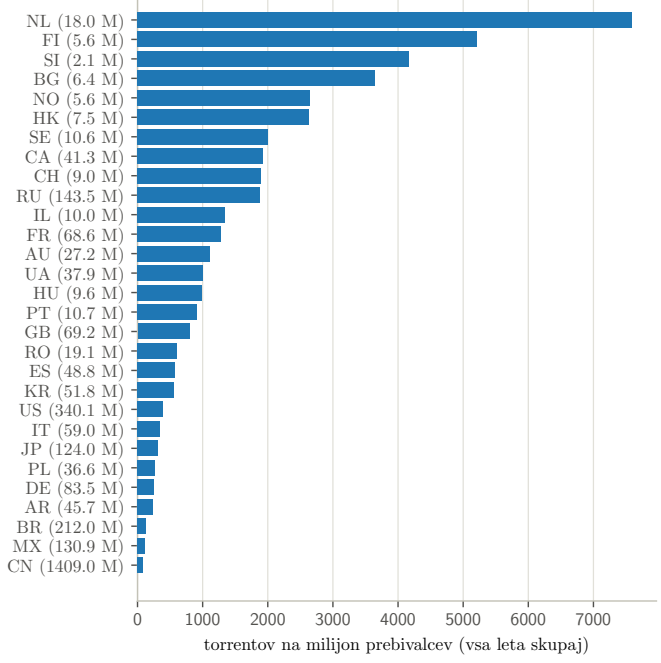
Torrente smo zbirali na strežnikih v Trzinu, Grčiji in na FRI v letih 2023, 2024 in 2026 in zajeli skupno milijon in pol torrentov.
Z jezikovnim modelom smo določili legalnost torrentov: 93 odstotkov torrentov je piratskih, dobra dva odstotka pa legalnih. Največ torrentov predstavlja pornografija, sledijo tv oddaje in serije, filmi in naposled glasba.
Prenašali smo samo imena datotek, ne pa tudi vsebine. Prenos vsebine bi zahteval preko 9 petabajtov prostora, pa tudi prenašali bi ilegalne



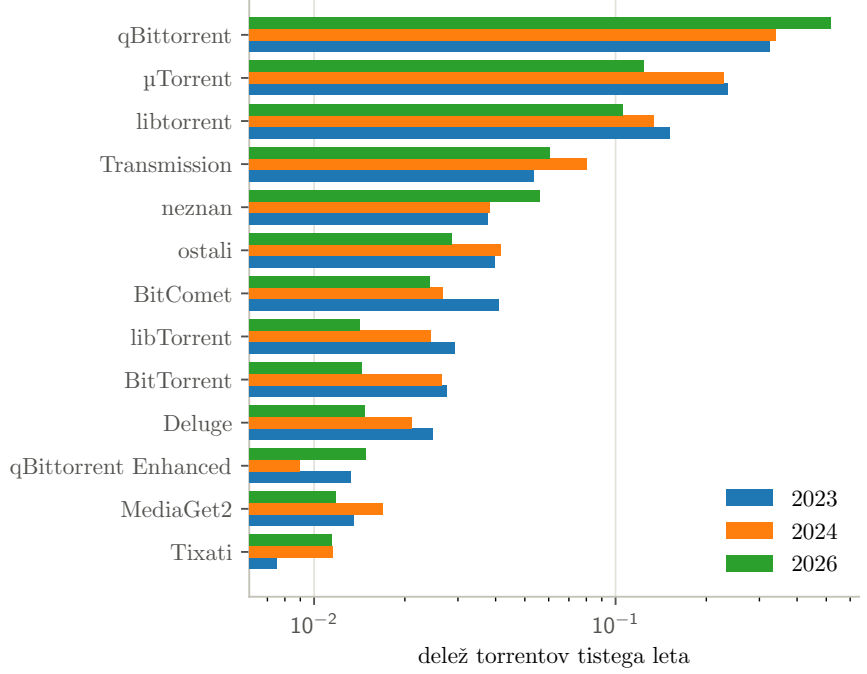
Kategorije lahko na grobo določimo tudi glede na datotečno končnico. Vidimo, da vodijo videovsebine, sledi glasba.



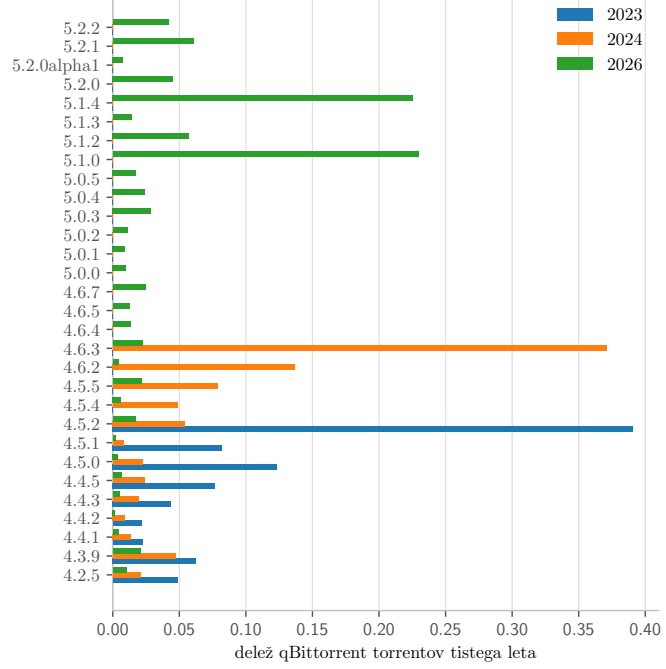
Beležili smo si tudi naslove IP, od katerih smo si metapodatke prenesli. Na podlagi tega lahko sklepamo, iz katerih držav je največ uporabnikov BitTorrenta. To so Rusija, Nizozemska, ZDA, Kitajska, Francija. Skratka, velike države.



Če pa število torrentov delimo s številom prebivalcev posamezne države, pa vidimo, da je Slovenija na tretjem mestu po prenosu torrentov na prebivalca.

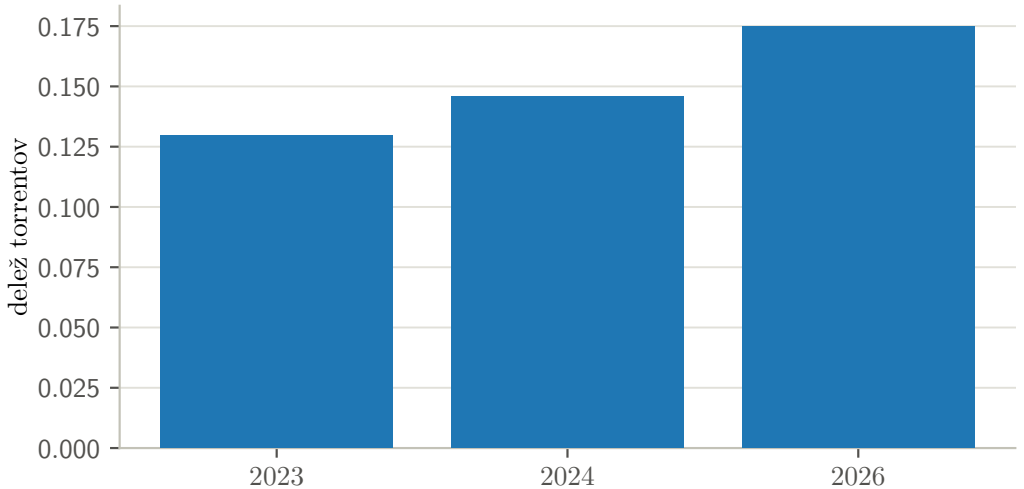


Ko se na soležnika povežemo, da bi od njega prenesli metapodatke, se nam pogosto predstavi z imenom programa, ki ga poganja. Diagram prikazuje priljubljene odjemalce. Najpopularnejši je qBittorrent, sledi mikro Torrent, ki mu popularnost pada. Sklepamo, da zato, ker ima v brezplačni različici oglase, qBittorrent pa je prosta odprtokodna programska oprema.

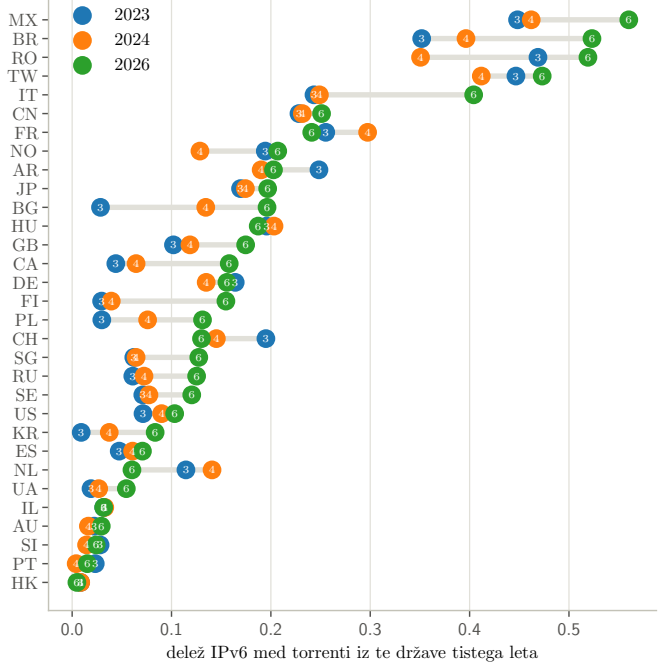


Odjemalci se nam predstavijo tudi z različico. Tule so različice programa qBittorrent. V zajemu leta 2026 je 16 odstotkov računalnikov, od katerih smo prenesli dva torrenta vsaj mesec dni narazen, posodobilo svoj odjemalec.

Rezultati: stranski produkt – razširjenost IPv6



Kot stranski produkt našega zajema lahko ocenimo tudi razširjenost modernega internetnega protokola IPv6. Opazna je rast deleža torrentov, ki smo jih prejeli preko IPv6. Tule pa je prikazana rast po državah.


2023
2024
2026
MX
BR
RO
TW
IT
CN
FR
NO
AR
JP
BG
HU
GB
CA
DE
FI
PL
CH
SG
RU
SE
US
KR
ES
NL
UA
IL
AU
SI
PT
HK
0.0
0.1
0.2
0.3
0.4
0.5
delež IPv6 med torrenti iz te države tistega leta

Zaključek

- Naredili smo program za zajem, zajemali metapodatke v treh letih in prikazali osnovne lastnosti omrežja BitTorrent.
- Potrdili smo domnevo, da je BitTorrent v glavnem uporabljen za piratsko vsebino.
- Ideje za prihodnje raziskave: opazovanje celotnega roja torrenta (skozi čas), stabilnejši in neprestan zajem, prenos vsebine datotek za določanje uporabljenih kodekov in vsebine stisnjenih arhivov.

Hvala za pozornost!

Za zaključek: Lahko rečemo, da je bil zajem torrentov uspešen in da iz pridobljenih podatkov lahko izluščimo uporabne informacije. Zajemi v različnih časovnih obdobjih so uporabni za nadaljne analize vsebin. Splošno znana domneva, da se BitTorrent uporablja pretežno za piratske vsebine, je z našo raziskavo potrjena.
Želimo si izvesti več raziskav na tem področju. Smiselno bi bilo opazovati celoten roj torrenta, to pomeni vse soležnike. Mi smo si zabeležili samo enega. Zanimivo bi bilo tudi prenašati nekaj vse-

Dodatek: Ocenjevanje pravilnosti klasifikacije (Cohenova κ)

- Ročno smo kategorije določili le 119 torrentom.
- `sklearn.metrics.cohen_kappa_score(llm_kategorije, moje_kategorije)`
- $\kappa \in [-1, 1]$, kjer je 0 naključna izbira in 1 popolno ujemanje
- $\kappa = 0,738$ za kategorije
- Večinski klasifikator, ki vedno izbere pornografijo, bi imel $\kappa = 0.294$.
- Lahko primerjamo tudi po robnih porazdelitvah, recimo: Ročno smo $\frac{113}{119} = 94{,}95\,\%$ torrentov označili za piratske, jezikovni model pa je za piratske označil $93{,}00\,\%$ torrentov (na vseh 2799).

www.musicbusinessworldwide.com/australia-worlds-biggest-pirates-torrent-sites/

Top 50 countries for total BitTorrent downloads during first half of 2012 per capita:

	Country Name	Total downloads	Approx Songs+	Population
1	Australia	19,232,252	154,242,661	21,766,711
2	Ireland	3,434,737	27,546,591	4,670,976
3	Slovenia	1,416,433	11,359,793	2,000,092
4	Canada	23,959,924	192,158,590	34,030,586
5	United Kingdom	43,263,582	346,973,928	62,698,362
6	Norway	2,827,508	22,676,614	4,691,849
7	Italy	33,158,943	265,934,723	61,016,804
8	Portugal	5,607,910	44,975,438	10,760,305
9	Croatia	2,305,095	18,486,862	4,483,804
10	Greece	4,933,478	39,566,494	10,760,136
11	Sweden	4,083,279	32,747,898	9,088,728

Dodatek: Zajem na raziskovalni infrastrukturi FRI ni bil neprekinjen