

## google-news-scraper

Во нашиот проект го користиме следниот пребарувач (scraper) за новости: [github.com/lewisdonovan/google-news-scraper](https://github.com/lewisdonovan/google-news-scraper).

Претставува Node.js пакет за пребарување на статии на Google News платформата врз основа на search query или специфични теми на Google News. Овој пакет користи Puppeteer за преземање на статиите и може да се прилагоди со различни опции за конфигурација.

Датотеката package.json е основен дел од Node.js проекти. Содржи метаподатоци и претставува конфигурациска датотека за проектот, вклучувајќи детали за проектот, неговите зависимости, скрипти и друго. package.json датотеката на овој пребарувач ја има на следниот линк: [github.com/lewisdonovan/google-news-scraper/blob/master/package.json](https://github.com/lewisdonovan/google-news-scraper/blob/master/package.json).

За користење на google-news-scraper потребно е да се направи следново:

1. Клонирај го GitHub репозиториумот  
`git clone https://github.com/lewisdonovan/google-news-scraper`
2. Влези во датотеката google-news-scraper  
`cd google-news-scraper`
3. Изврши ја следната команда за инсталација на пакетот  
`npm install google-news-scraper`  
\*доколку немаш npm инсталирај од [nodejs.org](https://nodejs.org).
4. Креирај едноставна scrape.js скрипта за пребарување на статии. Пример:

```
const googleNewsScraper = require('google-news-scraper');
const fs = require('fs');

(async () => {
  try {
    const config = {
      searchTerm: "Tesla",
      prettyURLs: true,
      timeframe: '1y',           // Articles from the past 1 year
      getArticleContent: false
    };

    const articles = await googleNewsScraper(config);
    console.log('Fetched articles:', articles.length);

    fs.writeFileSync('tesla_news.json', JSON.stringify(articles, null, 2));
    console.log('News for Tesla saved to tesla_news.json');
  } catch (error) {

    console.error('Error:', error);
  }
})();
```

5. Изврши ја скриптата  
`node scrape.js`
6. Во директориумот во кој се наоѓаш пронајди ја излезната датотека tesla\_news.json.  
Изглед на една статија од json датотетеката:

```
{
  "title": "I've driven tons of EVs. My first time driving a Tesla was still difficult.",
  "link": "https://www.businessinsider.com/tesla-model-3-review-quirks-difficulties-experienced-autos-reporter-2024-1",
  "image": "https://news.google.com/api/attachments/CC8iJ0NnNW5NbWRJTUdoTVJHbHpPR1IwVFJEQ0FSaURBaWdCTWdNkFCQQ=-w400-h224-p-df-rw",
  "source": "Business Insider",
  "datetime": "2024-01-21T08:00:00.000Z",
  "time": "Jan 21",
  "articleType": "regular"
},
```

Некои од својствата и филтрите кои може да се користат при пребарување:

- searchTerm (опционално): search query за наоѓање артикли. Пример: searchTerm: "The Oscars".
- baseUrl (опционално): Алтернативно URL за пребарување, корисно за пребарување специфични новини. Пример: baseUrl: "https://news.google.com/topics/CAAqJggKIiBDQkFTRWdvSUwyMHZNRGRqTldJb0VnVmxi aTFIU2pFQVAB".
- prettyURLs (опционално, default: true)
- timeframe (опционално): Филтрирај според временска рамка. Формат: број + оператор (пр., 1y за 1 година, 12h за 12 часа). Валидни оператори: h (hours), d (days), m (months), y (years).
- getArticleContent (опционално, default: false): Добивање на целата содржина на статиите.