

## 1 Lastno izbrana stran

Za lastni spletni strani smo izbrali dva različna paketa iger iz spletne trgovine Steam. Strani se razlikujeta v količini iger, ki so v paketu, kot tudi v opisu paketa (lahko ga ni). Slika polne velikosti na kateri smo označili podatke, ki smo jih ekstrahirali, je na voljo na našem GitHub repozitoriju v mapi `pa2/input-extraction/store.steampowered.com/Opisi.png` (povezava do slike).

**Package Details**

- BundleTitle**
- BundleTitle**: Two Point Hospital: Healthy Collection Vol. 4
- Genre**: Indie, Simulation, Strategy
- Developer**: Two Point Studios
- Publisher**: SEGA
- Languages**: English, French, Italian, German, Spanish - Spain, Polish, Russian, Simplified Chinese, Traditional Chinese, Portuguese - Brazil
- Listed languages may not be available for all games in the package. View the individual games for more details.

**Buy Two Point Hospital: Healthy Collection Vol. 4 BUNDLE (?)**

**BundleDiscount**: -59% **BundlePrice**: 32,98€ **Add to Cart**

**ABOUT THIS BUNDLE**

Discover the latest and greatest collection of Two Point Hospital content with the Healthy Collection Vol. 4 and expand your hospital horizons!

Take in new vistas, meet new characters and open up new areas of the county with the Close Encounters, Bigfoot and Pebberley Island expansions, and many more!

Build new hospitals, cure new diseases and, once you've set up a hospital for the ages, tweak it to improve your position as the best healthcare administrator in town.

**ITEMS INCLUDED IN THIS BUNDLE**

GameTitle	GameCategories	GameDiscount	GameFullPrice	GameDiscountedPrice
Two Point Hospital	Indie, Simulation	-66%	34,99€	11,89€
Two Point Hospital: Bigfoot	Indie, Simulation	-50%	8,99€	4,49€
Two Point Hospital: Pebberley Island	Indie, Simulation	-50%	8,99€	4,49€
Two Point Hospital: Close Encounters	Indie, Simulation	-50%	8,99€	4,49€
Two Point Hospital: Off the Grid	Indie, Simulation, Strategy	-50%	8,99€	4,49€
Two Point Hospital: Culture Shock	Indie, Simulation	-20%	8,99€	7,19€

Price of individual products: 37,04€  
Bundle discount: 11%  
Your cost: 32,98€

Here's what you save by buying this bundle **4,06€** **BundleSavings**

## 2 Implementacija

### 2.1 Regularni in XPath izrazi

Ekstrakcija z uporabo regularnih in XPath izrazov je bila izvedena s spodnjimi izrazi. Izrazi označeni z † se izvedejo v *single line* načinu, tj. znak pika predstavlja tudi nove vrstice.

#### 2.1.1 overstock.com

- Title:**  
`<td valign="top"> \n<a href=".+"><b>(.*?)</b></a>  
//td[@align="top"]/a/b/text()`
- Price:**  
`<span class="bigred"><b>(\$d+\.d+)</b>  
//span[@class="bigred"]/b/text()`
- ListPrice:**  
`<td[^>]*><s>(\$d+\.d+)</s>  
//td[@align="left"][@nowrap]/s/text()`
- Content:**  
`<span class="normal">([S\s]+)<br>  
//span[@class="normal"]/text()`

- *Saving:*  
You Save:</b></td><td[^>\*><span[^>\*>(\\$\\d+,?\\d+\\.\\d+)  
//td[@align="left"]/span[@class="littleorange"]/substring-before(text(), " ("
- *SavingPercent:*  
You Save:</b></td><td[^>\*><span[^>\*>.\* \\((\\d+%)\\)</span>  
//td[@align="left"]/span[@class="littleorange"]/substring-before(substring-after(text(), " ("), ")")

### 2.1.2 rtvslo.si

- *Author:*  
<div class="author-name">(.\*?)</div>  
//div[@class="author-name"]/text()
- *PublishedTime:*  
<div class="publish-meta">[\\n\\t]\*(.\*)<br>  
//div[@class="publish-meta"]/text()
- *Title:*  
<h1>(.\*?)</h1>  
//h1/text()
- *SubTitle:*  
<div class="subtitle">(.\*?)</div>  
//div[@class="subtitle"]/text()
- *Lead:*  
<p class="lead">(.\*?)</p>  
//p[@class="lead"]/text()
- *Content:*  
<article[^>\*>.\*?</figure>(.\*?)</article> † §  
//article/p/descendant-or-self::\*/text()

§: regularni izraz počistimo še s sledečim zaporedjem ukazov:

```
s = re.sub('<iframe[^>*>.*?</iframe>', '', content_text)
s = re.sub('<figure[^>*>.*?</figure>', '', s, flags=re.DOTALL)
s = re.sub('<div class="gallery">.*</div>', '', s, flags=re.DOTALL)
s = re.sub('<p[^>*>|<strong>|</strong>|<sub>|</sub>', '', s)
s = re.sub('</p>|<br>', '\\n', s)
```

### 2.1.3 store.steampowered.com

- *Description:*  
<h2>About this bundle</h2>[\\n\\t]\*<p>(.\*?)</p>  
//\*[@id="game\_area\_description"]/p/text()
- *BundleSavings:*  
bundle\_savings">(.\*?)</div>  
//div[contains(@class, "bundle\_savings")]/text()
- *Title:*  
<b>Title:</b> +(.\*?) <br>  
//div[@class="details\_block"]/p/descendant-or-self::text()[not(ancestor::b) and normalize-space(.) != ""]
- *BundleDiscount:*  
<div class=".\*game\_purchase\_discount"[^>\*>.\*<div class="discount\_pct">(.\*?)</div>  
//div[contains(@class, "game\_purchase\_discount")]/div[@class="discount\_pct"]/text()
- *FullPrice:*  
<div class=".\*game\_purchase\_discount"[^>\*>.\*<div class="discount\_original\_price">(.\*?)</div>  
//div[contains(@class, "game\_purchase\_discount")]/div/div[@class="discount\_original\_price"]/text()
- *BundlePrice:*  
<div class=".\*game\_purchase\_discount"[^>\*>.\*<div class="discount\_final\_price">(.\*?)</div>  
//div[contains(@class, "game\_purchase\_discount")]/div/div[@class="discount\_final\_price"]/text()
- *GameDiscount:*  
<div class=".\*tab\_item\_discount"[^>\*>.\*<div class="discount\_pct">(.\*?)</div>  
//div[contains(@class, "tab\_item\_discount")]/div[@class="discount\_pct"]/text()
- *GameFullPrice:*  
<div class=".\*tab\_item\_discount"[^>\*>.\*<div[^>\*>.\*<div class="discount\_original\_price">(.\*?)</div>  
//div[contains(@class, "tab\_item\_discount")]/div/div[@class="discount\_original\_price"]/text()
- *GameDiscountedPrice:*  
<div class=".\*tab\_item\_discount"[^>\*>.\*<div[^>\*>.\*<div class="discount\_final\_price">(.\*?)</div>  
//div[contains(@class, "tab\_item\_discount")]/div/div[@class="discount\_final\_price"]/text()
- *GameTitle:*  
<div class="tab\_item\_name">(.\*?)</div>  
//div[@class="tab\_item\_name"]/text()
- *GameCategories:*  
<span class="platform\_img.\*"></span>[\\n\\t]\*&nbsp; (.\*?)</div>  
//div[@class="tab\_item\_details"]/text()

## 2.2 Algoritem za samodejno ekstrakcijo podatkov

Algoritma nismo implementirali.