“Scrapper” le site de l’Opéra de Paris

Rodolfo Ripado | Pizza Talk Radio France | Mai 2017

Pourquoi ?

- J’aime l’opéra, mais c’est super cher …

Pourquoi ?

- J’aime l’opéra, mais c’est super cher …
- Je suis psychologiquement incapable d’acheter des places 6 mois à l’avance …

Pourquoi ?

- J’aime l’opéra, mais c’est super cher …
- Je suis psychologiquement incapable d’acheter des places 6 mois à l’avance …
- Le site de l’opéra de Paris ne fournit pas de liste claire des places les moins chères pour chaque spectacle

Quelques questions

- Combien de temps à l’avance faut-il acheter les billets pour avoir des prix “raisonnables”

Quelques questions

- Combien de temps à l’avance faut-il acheter les billets pour avoir des prix “raisonnables”

- Cela arrive-t-il que les prix descendent ?

Quelques questions

- Combien de temps à l’avance faut-il acheter les billets pour avoir des prix “raisonnables”
- Cela arrive-t-il que les prix descendent ?
- A quelle vitesse augmentent-ils ?

Quelques semaines plus tard ...

http://rodolforipado.net/operaprices

La tech

La contrainte de base : faire ça sans payer un rond

La tech

La contrainte de base : faire ça sans payer un rond

- GCE et le “always-free” tier
- Un crawler nodejs
  - ~100 requêtes, 43 spectacles, 445 représentations, ~5000 prix
- Une BD sqlite
  - 2Mb / jour
- Front : VueJS + Pages Github + API statique
- Logs envoyés à Loggly pour stockage et visu
  - 7 jours rétention de données dans le plan gratuit

La tek : guidelines pour crawler un site

- Faire un POC rapide pour tester la stabilité du HTML

La tek : guidelines pour crawler un site

- Faire un POC rapide pour tester la stabilité du HTML
- Bien logger ce qui foire, pour enrichir le modèle du site

La tek : guidelines pour crawler un site

- Faire un POC rapide pour tester la stabilité du HTML
- Bien logger ce qui foire, pour enrichir le modèle du site
- Le crawl comme pipeline
  - Pages web => items => pages web => + items, + info => … => stockage

La tek : qu’ai-je appris avec tout ça ?

Pour POC-er :

- *Scrappy* : framework python pour faire des crawlers web TRES rapidement
- *Scrappyhub.com* : héberger et programmer ses crawls gratuitement
- *Github pages* : Bon moyen d’héberger un client “lourd” JS

La tek : qu’ai-je appris avec tout ça ?

Version actuelle :

- *RxJS* : parfait pour gérer des pipelines de traitement de données, comme un crawl web
- *VueJS* : comme React mais plus simple, API plus stable. C’est top.
- *SQLite* : génial de simplicité

Et l’Opéra dans tout ça ?

- Les prix descendent régulièrement, pendant quelques heures / jours
  - jusqu’à 50 euros de moins ...

Et l’Opéra dans tout ça ?

- Les prix descendent régulièrement, pendant quelques heures / jours
  - jusqu’à 50 euros de moins ...

- La version de Carmen actuellement à l’affiche n’est pas la meilleure que j’ai vue ...

Perspectives

- Ajouter des visus sur l’évolution des prix. Mais lesquelles ?
  - Par spectacle
  - Par représentation
  - Par catégorie de prix ?
  - …

Perspectives

- Ajouter des visus sur l’évolution des prix. Mais lesquelles ?
  - Par spectacle
  - Par représentation
  - Par catégorie de prix ?
  - …
- Prendre en compte les promotions “flash”

Perspectives

- Ajouter des visus sur l’évolution des prix. Mais lesquelles ?
  - Par spectacle
  - Par représentation
  - Par catégorie de prix ?
  - …
- Prendre en compte les promotions “flash”

Perspectives

- Ajouter des visus sur l’évolution des prix. Mais lesquelles ?
  - Par spectacle
  - Par représentation
  - Par catégorie de prix ?
  - …
- Prendre en compte les promotions “flash”
- Aller plus souvent à l’Opéra !