Dark Patterns

Petr Hanzl

January 22, 2021

1 Cíl

Celkově je za cíl přepsat veškeré kódy od vědců z Princetonu do docker kontejnerů, aby byly jednoduché na instalaci a použití. Právěže instalace a použití je zde vcelku složité, obvzláště kvůli chybějící dokumentaci. Toto se týká zejména crawlerů, které obsahují složitejší nastavení, oproti později použitím Jupyter Notebookům.

2 Popis původních crawlerů

V původní implementaci jsou v podstatě tři různé crawlery, které se liší způsobem prací, kterou vykonávají.

První je skupina programů, která získá domény tisíců eshopů podle Alexa rank. Využívají zde i placené služby na rozpoznání kategorie webu. Tento crawler je zřejmě bezpředmětný pro český trh. Na českém webu dokonce existuje katalog českých eshopů. Pro získání URL adres se tedy dá relativně jednoduše použít web Heureka.cz nebo web Asociace českých eshopů.

3 Docker Swarm

Swarm obsahuje několik druhů kontejnerů:

- **RabbitMQ server** jako message-broker middleware například pro rozdělování práce.
- **Redis databáze**, která je použitá pro ukládání získáných segmentů z webových stránek
- **Workeři**, kteří zpracovávají úkoly z fronty v RabbitMQ

Všechny kontejnery jsou propojené a pracují na společné virtuální síti vytvořené taktéž v dockeru.

Všichni workeři mají připojený sdílený virtuální diskový prostor, který se používá pro ukládání výsledků a print screenů obrazovky během procházení.

4 RabbitMQ

RabbitMQ je message-broker middleware, implementující AMPQ protokol. Nedařilo se mi nastavit hostname ve virtuální síti a tam má pevnou IP adresu **172.18.0.2**. Zároveň je možné sledovat stav fronty, připojené workery a práci ve webové administraci, která je dostupná na adrese `http://localhost:15672`. Obzvláště užitečné je zrušení nedokončené práce.

Kontejner využívá image rabbitmq:management, který používá defaultní přihlašovací údaje **guest/guest**.

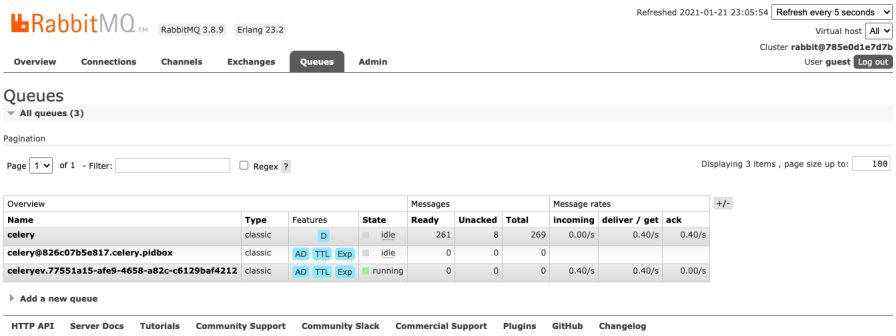

Figure 1: Ukázka přehledu rozdělaných jobů a připojených workerů ve webové administraci RabbitMQ.

5 Celery

Celery je v podstatě framework, který umožňuje v Pythonu vytvářet workery využívající AMPQ protokol pro přerozdělování práce.

6 Struktura projektu

Adresář obsahuje několik shell skriptů, pro spuštění podpůrných docker kontejnerů, pro vytvoření kontejneru workerů, pro spuštění kontejnerů a pro přidání práce do fronty. Zároveň obsahuje i skript pro vyčištění systému od provedené práce.

Projekt je rozdělen na joby a workery. Složky s workery obsahují **Dockerfile** pro vytvoření docker image a zdrojové kódy samotného workera, které se vždy nachází ve složce **src**.

Složka jobů obsahuje vstupy, které jsou přes jednoduchý shell skript, předány jednomu ze zvolených workerů, který data zpracuje a naplní frontu příslušnýma pracema.

7 Práce se swarmem

Swarm spustíme skriptem **./init.sh**, který vytvoří virtuální síť, virtuální sdílené úložiště, spustí RabbitMQ middleware a Redis databázi.

Následně musíme připojit nějaké pracovníky do swarmu. Ty najdeme ve složce **workers**.

Každá složka se rovná jeden různý typ workeru. Pokud jsme ještě tak neudělali, tak musíme workera nejdříve postavit z příslušného **Dockerfile**. Pro jednoduchost je zde však skript **./build.sh**.

Po úspěšném postavení image workera přichází na řadu samotné spuštění workera. To provedeme přes **./run_worker.sh**, kdy workerovi předáme i jméno.

Toto je velmi důležité, protože to ulechčuje následnou práci se spuštěnými kontejnery a přidáváním práce do fronty. Pracovníků můžeme přidat více, stačí vždy použít jiné jméno a budou pracovat paralelně.

Následuje už pouze přidání práce do fronty. To provedeme přechodem do složky **./jobs**, kde jsou složky pojmenované stejným jménem jako příslušní workeři. Uvnitř těchto složek je jedinný shell skript, který provede přidání práce workerovi a ten jen submitne do sdílené fronty.

Pro vyčištění systémů od provedených změn a hlavně docker kontainerů lze použít shell skript **./clear.sh**.

8 Crawler pro stažení produkčních stránek

Worker pro tento crawler se nachází ve složce **workers/extract-links** a úlohy pro tohoto workera jsou ve složce **jobs/extract-links**.

Cílem tohoto crawlera je projít seznam předaných URL adres a získání maximálně pěti produktových stránek. Produktová stránka je stránka nabízeného produktu daným eshopem.

Získávání probíhá tak, že se vytvoří spider na předané adrese, který projde až 100 webových stránek na stejné adrese.

Crawler obsahuje celkem 5 jednoduchých funkcí na jednoduchou klasifikaci, zda daná stránka je produktová. V jednodušené podobě se zkoumá existence tlačítka "Přidat do košíku".

9 Vstup

Vstupem je seznam URL adres eshopů.

10 Výstp

Existuje několik výstupů z programu:

- HTML kódy navštívených stránek
- Print screeny navštívených stránek

- JSON soubor, který obsahuje seznam nalezené odkazy na doméně
- Textový soubor, který obsahuje seznam nalezených produktových stránek
- JSON soubor, který obsahuje seznam navštívených stránek na doméně