Spark – RDD

Wprowadzenie

Nasze wyzwanie jest z jednej strony proste, z drugiej strony dość ambitne.

Jedno z klasycznych "Hello World" świata Big Data polega na zliczaniu wystąpienia słów.
Dane wejściowe - plik tekstowy lub strumień tekstu. Dane wynikowe - liczba wystąpień każdego ze słów. Klasyka.

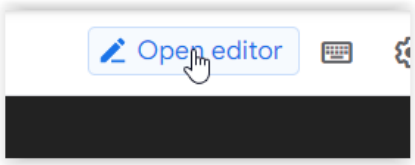
My zrobimy to samo, jednak naszymi danymi wejściowymi będą... opowiadania Artura Conan Doyla (czyli standard), ale nie w plikach tekstowych, a w formacie PDF (i to już standard nie jest).

Trudne? Nic bardziej mylnego. Python to mnogość bibliotek o niezliczonej funkcjonalności.

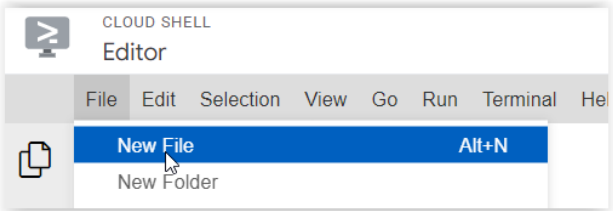
Przygotowanie środowiska i zbioru danych

1. Standardowo, w tej chwili uruchomilibyśmy po prostu klaster. My jednak musimy się do tego przygotować. Oprócz standardowych modułów Pythona, które są instalowane na klastrze *Dataproc* (klastrze Hadoop z dodatkowymi komponentami) dostępnego w ramach *Google Cloud Platform* – będziemy potrzebowali dwóch dodatkowych: `PyPDF2` oraz `pydoop`.
   Mogłoby się wydawać, że nie ma z tym problemu i można to zrobić po uruchomieniu klastra. Teoretycznie tak. Problem w tym, że klaster to nie jeden węzeł a wiele. Nasze programy nie chcemy uruchamiać na jednym węźle (np. na klastrze Sparka uruchomionym w trybie `local`). Chcemy je uruchamiać na wielu węzłach (w trybie `yarn`), i na każdym z tych węzłów te biblioteki muszą się znaleźć. Aby nie instalować ich ręcznie, łącząc się z każdym z węzłów osobno, zainstalujemy je podczas tworzenia klastra za pomocą akcji inicjalizacyjnych.

2. Uruchom terminal *Cloud Shell*, a w nim edytor (możesz także skorzystać z edytora `nano`)



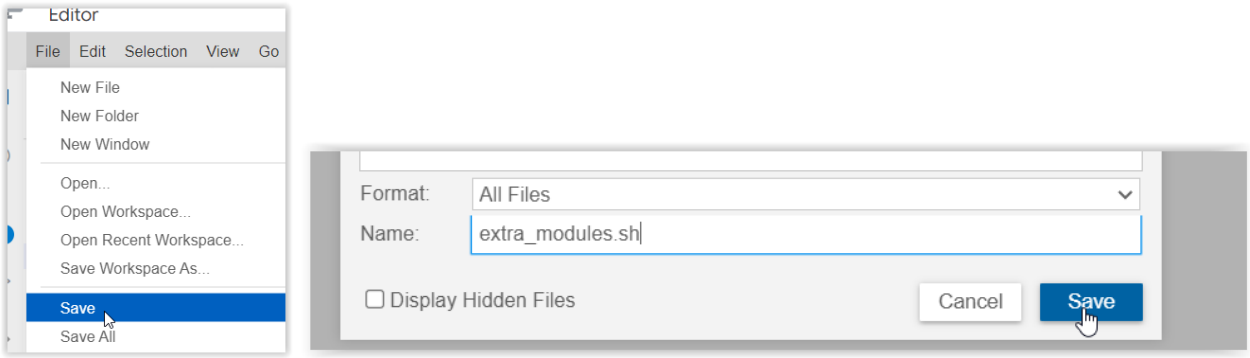
3. Utwórz nowy plik



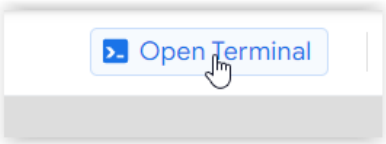
4. Wprowadź do niego następującą zawartość

```
#!/bin/sh
pip install PyPDF2
pip install pydoop
```

5. Zapisz plik jako `extra_modules.sh`



6. Powróć do widoku terminala



7. Przekopiuj plik `extra_modules.sh` do Twojego zasobnika

```
gsutil cp extra_modules.sh gs://${BUCKET_NAME}
```

8. Korzystając z poniższego polecenia i konsoli *Cloud Shell* utwórz klaster.

```
gcloud beta dataproc clusters create ${CLUSTER_NAME} \
  --enable-component-gateway --bucket ${BUCKET_NAME} \
  --region ${REGION} \
  --master-machine-type n1-standard-4 --master-boot-disk-size 50 \
  --num-workers 2 --worker-machine-type n1-standard-4 --worker-boot-disk-size 50 \
  --image-version 2.1-debian11 --optional-components JUPYTER \
  --initialization-actions gs://${BUCKET_NAME}/extra_modules.sh \
  --project ${PROJECT_ID} --max-age=3h
```

9. Pobierz eksport notatnika `SP03_w1_23-RDD-zadania.ipynb`

10. Na stronie środowiska *Jypyter* zaimportuj ten notatnik, a następnie otwórz go.
    Znajdziesz tam dalsze instrukcje.