Duże zadanie zaliczeniowe 1. Termin 2024.05.15

Podstawową strukturą budowy żywych organizmów są białka, które powstają na podstawie sekwencji aminokwasów. Bazową metodą szukania podobnych białek jest szukanie białek o podobnej sekwencji kodującej. Jednym z projektów, które służą gromadzeniu o wszystkich białkach jest projekt UniProt (https://www.uniprot.org/), który zbiera informacje o wszystkich znanych białkach (pond 100GB skompresowanych danych) . Poza bazą wszystkich białek projekt udostępnia klastrowanie białek na podstawie podobieństwa (pod nazwą uniref).

Państwa zadaniem jest sprawdzenie, czy LSH z rozmiarem shingla 5, band 20 i row 5 pozwala na poprawne identyfikowanie kandydatów do klastrów wskazywanych przez projekt. Jako poprawność identyfikacji należy zrozumieć czy większość elementów klastra została poprawnie wskazane jako kandydaci (True positive) i czy nie jest za dużo par dla białek należących do różnych klastrów (False positive).

Ze względu na rozmiar danych przygotowaliśmy próbkę, która składa się z pliku group_definition.json zawierającego informację o klastrach i katalogu fasta zawierającego informacje o sekwencji każdej z cząstek (w postaci plików json).

Plik `group_definition.json` zawiera pojedynczy słownik, w którym jest mapowanie z nazwy grupy na listę białek należących do grupy.

Pliki w katalogu fasta są w formacie:


```
{"name": "W8RIU5", "value":
"KNSITVAWGKPIYDGGSEILGYVVELCKADEEEWQIVTPPTGLKATRFEIAKLTEHQEYKIRVCALNKVG
LGEATPVPGTVKPEDKLEAPELDLDSELRKGIVVRAGGSARIHIPFKGRPTPEITWSREEGEFTDKVQIEK
GLNYTQLSIDNCDRNDAGKYILKLENSSGSKSAFVTVKVLDTPGPPQNLAVKEVKKDSVILVWEPPIIDGG
AKVKNYVIDKRESTRKAYANVSSKCNKTSFKVENLTEGAIYYFRVMAENEFGVGVPVETVDAVKAAEPPSP
PGKVTLTDVSQTSASLMWEKPEYDGGSRVLGYVVEMQSKGTEKWS"}
```


Gdzie `name` jest nazwą białka, a `value` zawiera sekwencję aminokwasów budujących białko.

Próbka jest celowo mniejsza, aby mogli państwo swobodnie testować. Częścią rozwiązania jest pokazanie uruchomienia zadania na klastrze (albo na google cloud, albo lokalnie z użyciem dockera).