Tehnologije Hadoop I Spark

- **Opis problema**

  Problem koji će biti analiziran u ovom radu je analiza veza između država sveta pomoću bratskih gradova. Bratstvo između gradova je dogovor o saradnji između različitih gradova sa ciljem širenja kulture i komercijalne saradnje.

- **Skup podataka**

  Skup podataka koji će se koristiti u projektu su skinuti sa sajta www.wikidata.org. Koristeći SPARQL upit dobijeni su podaci o gradovima, njima bratskim gradovima I podaci o njihovim zemljama. Na kraju su dobijena četiri fajla sa Kojima će biti odgovarajući grafovi. Podaci se nalate u csv fajlovima I njihova imena su :
  - *gradoviSaDržavama.csv* sa ID projevima I imenima gradovima kao I ID brojevima I imenima država Kojima pripadaju
  - grane.csv sa po dva ID broja gradova koji predstavljaju bratski odnos
  - drzave.csv sa ID brojevima I imenima država
  - gradDrzavaGrane.csv sa po dva ID broja koji predstavljaju vezu između država I gradova

  ID-jevi gradova I država su međusobno jedinstveni pošto dolaze sa WikiData. Svaki fajl sadrži početnu header liniju. Svaka sledeća linija sadrži podatke gde se IDjevi nalaze na početnu linije.

- **Algoritmi**

  Analiza će se sprovesti korišćenjem PySpark I GraphFrames biblioteka za rad sa podacima u grafovima. Kako bismo pronašli grupe država sa dobrim trenutnim ili istorijskim odnosima, potrebni su izmenjeni grafovi država. Te grafove ćemo dobiti kada izbacimo duplicate veza između država I kada uklonimo vezu država koje dele premalo bratskih gradova. Pronalaženje grupa će biti rađeno pomoću algoritma iz GraphFrame paketa. Ti algoritmi su:
  - *Strongly connected components* – algoritam za analizu grafova koji identifikuje skupove čvorova u gradu u kojem postoji put između svaka dva čvora u skupu. Primenom SCC na graf bratskih veza gradova, identifikovaće se skupovi država u Kojima postoji snažna povezanost između njihovih grafova. Pomaže u grupisanju država koje imaju istorisjki jake veze.
  - *Label Propagation Algorithm* – algoritam za razvrstavanje čvorova u grafu na osnovu labela koje se šire kroz graf na osnovu lokalnih prava. Koristi se za otkrivanje zajednica država koje dele slične veze između svojih gradova. Label se šire kroz graf na osnovu bratskih veza, a grupisanjem čvorova prema ovim labelama dobija se zajednice država sa sličnim karakteristikama saradnje.

- *Hadoop MapReduce* – preprocesiranje I obrada podataka pre nego što se podaci učitaju u GraphFrames koji omogućavaju efikasno skladištenje I analizu velikih skupova podataka

- **Tehnologije**

  Programski jezik kojie je odlučen da se koristi u projektu je python. Izabran je zbog raznovrsnih biblioteka. Koristiće se PySpark, koji je Python API koji omogućava rad sa velikim skupovima podataka I analizu grafova korišćenjemm Spark-ovog GraphFrames paketa. GraphFrames je biblioteka za Apache Spark koja pruža podršku za rad sa grafovima. Pandas je popularna biblioteka za vizualizaciju podataka. MRJob je Python biblioteka koja omogućava pisanje Hadoop MapReduce dela. Okruženje koje će biti upotrebljeno u projektu je Visual Studio Code.

- **Cilj**

  Cilj projekta je pronalaženje I analiziranje veza između država sveta pomoću bratskih gradova. Krajnji rezultat ovog rada treba da budu grupe država koje imaju ili su istorisjki imale dobre odnose što je dovelo do pravljenja bratskih odnosa između njihovih gradova.

Nataša Gavrilović E5 18/2023