Domaći zadatak – JSON
(rok: 14.6.2016)

Napraviti repozitorijum na GitHub-u koji se zove „Collections-homework“ i u njega postaviti novi Eclipse projekat “CollectionsHomework” koji je povezan sa ovim repozitorijumom.

Kreirati aplikaciju koja će prebrojati pojavljivanja svih Bigrama u ulaznom stringu.

Bigram je sekvenca dva tokena, gde token može biti slovo, reč, ili proizvoljni string. Bigram je n-gram gde je n=2.

Više na: https://en.wikipedia.org/wiki/Bigram

Bigrami se koriste u raznim modelima za prepoznavanje i predikciju govora, kao i u kriptografiji.

Primer:

Petar ide u školu.

Bigrami:

1. Petar ide
2. ide u
3. u školu.

U ovom domaćem zadatku, ulaz će biti String, a bigrami će biti sastavljeni od **dva karaktera**.

Primer:

Ulazni string: asdeesdf

Bigrami:

1. as
2. sd
3. de
4. ee
5. es
6. sd
7. df

Zadatak:

U odgovarajućem paketu napraviti klasu Bigram koja ima main metodu.

Pri završetku rada programa, u konzoli bi trebalo da se ispišu frekvencije (broj pojavljivanja) za svaki od bigrama.

Ulazni String: abbcceeeeeeeabcc

Izlaz u konzoli:

bb 1
cc 2
ee 5
ab 2
bc 2
ce 1
ea 1

Sve bigrame i njihove frekvencije je potrebno čuvati u Mapi (HashMap). Pri ispisivanju frekvencija, potrebno je iterirati kroz mapu.

Uraditi commit sa jasno naznačenom porukom o tome šta je novo urađeno.

Potrebno je predvideti naredna 3 karaktera za neki ulazni String, na osnovu prethodno izračunatih frekvencija. Rezultat ispisati u konzoli.

Ulazna sekvenca za koju je potrebno predvideti 3 naredna karaktera: **ja**

Rezultat: **jabcc**

Formula po kojoj se računa naredni karakter, na osnovu prethodnog:

$$P(W_n|W_{n-1}) = \frac{P(W_{n-1}, W_n)}{P(W_{n-1})}$$

Ilustracija na navedenom primeru:

Ulazni String: **ja**

Potrebno je predvideti naredni karakter na osnovu slova **a**.

Iz frekvencija bigrama dobijenim u prethodnom delu zadatka, vidimo da samo jedan bigram počinje slovom a, i to: **ab** bigram, frekvencije **2**. Kako nema drugih bigrama, naredno slovo je **b**.

Kako sada imamo String **jab**, naredno slovo tražimo na osnovu **b** slova.

Bigrami koji počinju na slovo **b** su: **bc** frekvencije **2** i **bb** frekvencije **1**. To znači da imamo **ukupno 3** bigrama koji počinju na slovo **b**. Kako je najveća verovatnoća da se pojavi **bc** bigram (verovatnoća je **2/3**), naredno slovo je **c**.

Uraditi commit sa jasno naznačenom porukom o tome šta je novo urađeno.

Napomena: Ulazne stringove možete direktno upisati u kodu. Nije potrebno da se učitavaju sa tastature ili iz fajla. Ulazni stringovi moraju biti u jednom redu. Ulazni string za koji će se predvideti karakteri ne sme da sadrži karaktere koji se nisu pojavili u ulaznom stringu u kom ste prebrojali bigrame. Dodatne klase nije potrebno uvoditi.

Po završetku domaćeg zadatka, uraditi konačni push na novokreirani GitHub repozitorijum (sa svim commit-ovima) i dostaviti prijavu sa imenom, prezimenom i linkom na ovaj GitHub repozitorijum na sajtu http://jgrass.fon.bg.ac.rs/java-collections/. Rok za izradu ovog domaćeg zadatka je dve nedelje tj. 14.6.2016.