ქართული ენის ენტროპია და შეკუმშვა

ამ დავალების მიზენია ქართული ტექსტის შემკუმშველი პროგრამის დაწერა. თავიდან შევისწავლით ქართულ ენაში შეხვედრილი ასოების განაწილებას, და შემდეგ ავაგებთ შესაბამის შემკუმშველ პროგრამას.

ჩვეულებრივ ტექსტში ქართული ასოების და ცარიელი ადგილების გარდა ბევრი სხვა აღნიშვნა გამოიყენება (სასვენი ნიშნები, რიცხვები, …), მაგრამ ჩვენ ამოცანის გამარტივებისთვის მხოლოდ ქართულ ანბანს და ცარიელ ადგილს “space”-ს გამოვიყენებთ. შესაბამისად დავალების უმეტეს ნაწილში 34 სიმბოლო გექნებათ.

ტექსტური ფაილები UFT-8 კოდირებით არის შენახული. ქართულ ასოებს შეესაბამებათ 3 ბაიტიანი კოდები რომლებც თექვსმეტობითში გამოისახება: “E1-83-90”-დან “E1-83-B0”-ს ჩათვლით კოდებით. “space”-ს კი შეესაბამება ერთ ბაიტიანი კოდი თექვსმეტობითში: “20” (ანუ ათობითში: “32”).

ყველა პროგრამას უნდა დაარქვათ ის სახელი, რაც შესაბამის საკითხში იქნება მითითებული. დაწერილი პროგრამები (source ფაილები) მოათავსეთ ერთ folder-ში. ამ folder-ს დაარქვით თქვენი freeuni ელ-ფოსტის მისამართი სიმბოლო "@"-მდე. ეს folder-ი შეკუმშეთ zip ფაილად და არქივი ატვირთეთ google classroom-ზე.

თქვენი დაწერილი პროგრამები არ უნდა მუშაობდეს ძალიან ნელა, და არ უნდა იყენებდეს ძალიან ბევრ მეხსიერებას. კონკრეტულად, პროგრამა შემოწმდება 1 ბირთვიან (საკმაოდ ნელ) პროცესორზე, ნახევარ გიგაბაიტიანი ოპერატიული მეხსიერებით. საჯარო ტესტებში მოცემული თითოეული შემთხვევისთვის თქვენს პროგრამას არ უნდა სჭირდებოდეს 15 წამზე მეტი დრო.

1. ასოების განაწილება [1 ქულა]

დაწერეთ პროგრამა, რომელიც ფაილიდან (რომლის სახელიც პროგრამას პირველ არგუმენტად გადმოეცემა) წაიკითხავს ქართულ ტექსტს (სასვენი ნიშნების გარეშე) და დათვლის შემდეგ სიხშირეებს:

1. თითოეული სიმბოლოს ალბათობა
2. სიმბოლოთა წყვილების ალბათობები

გარანტირებულია რომ ფაილში შეგხვდებათ მხოლოდ ქართული ასოები და ცარიელი ადგილები (space-ები). არ შეგხვდებათ სასვენი ნიშნები, არაქართული ასოები და ნებისმიერი სხვა სიმბოლოები. Space ლექსიკოგრაფიულად ყველაზე მცირე სიმბოლოდ ჩათვალეთ და მისი ალბათობაც გაითვალისწინეთ. თუ რომელიმე სიმბოლო ან სიმბოლოების წყვილი არ გხვდებათ, მისი ალბათობა ნულად ჩათვალეთ. სიმბოლოების წყვილების (და მხოლოდ წყვილების) დათვლის დროს, ჩათვალეთ რომ ფაილის პირველი სიმბოლოს წინა სიმბოლო არის ცარიელი ადგილი (Space).

შედეგი ჩაწერეთ ფაილში რომლის სახელიც მეორე არგუმენტად გადმოეცემა პროგრამას. პირველ ხაზზე ჩაწერეთ space-ებით გამოყოფილი სიმბოლოების ალბათობები ანბანის მიხედვით (პირველი სიმბოლო space არის), მეორე ხაზზე კი წყვილების ალბათობები დალაგებული ლექსიკოგრაფიულად {__, _ა, _ბ, … _ჰ, ა_, აა, აბ, … აჰ, ბ_, ბა, ბბ, … ბჰ, … ჰ_, ჰა, ჰბ, … ჰჰ}. შედეგები გამოიტანეთ მინიმუმ 7 ციფრის სიზუსტით წერტილის მერე. მაგალითისთვის იხილეთ საცდელი მონაცემები და პასუხები.

პროგრამას დაარქვით “Distrib.xxx”, სადაც xxx თქვენს მიერ არჩეულ პროგრამირების ენაზე დამოკიდებული გაფართოებაა.

2. ენტროპია, წყვილების ენტროპია და პირობითი ენტრობია **[1 ქულა]**

დაწერეთ პროგრამა (სავარაუდოდ წინა პუნქტში დაწერილ პროგრამაზე დაშენებით), რომელიც მოცემული (პირველ არგუმენტად გადმოცემული) ტექსტური ფაილისთვის დათვლის და გამოიტანს შემდეგ მაჩვენებლებს (თითო რიცხვი თითო ხაზზე):

1. სიმბოლოების განაწილების ენტროპია
2. სიმბოლოების წყვილების განაწილების ენტროპია
3. სიმბოლოების განაწილების წინა სიმბოლოთი პირობითი ენტროპია. ანუ $H(X_n|X_{n-1})$

შედეგები ჩაწერეთ მეორე არგუმენტად გადმოცემულ ფაილში.

პროგრამას დაარქვით "Entropy.xxx" (xxx ისევ ფაილის სახელის გაფართოებაა).

3. უპრეფიქსო კოდის აგება სიგრძეებით **[1.5 ქულა]**

დაწერეთ პროგრამა, რომელიც პირველ არგუმენტად გადმოცემული ფაილის პირველი ხაზიდან წაიკითხავს რიცხვ n-ს, რომელიც აღნიშნავს შესაძლო შეტყობინებების რაოდენობას და ამავე ფაილის მეორე ხაზიდან წაიკითხავს n ცალ მთელ რიცხვს (space-ებით გამოყოფილს) რომლებიც კოდური სიტყვების სიგრძეებს აღნიშნავენ. შემდეგ თქვენი პროგრამა შეამოწმებს აკმაყოფილებენ თუ არა ეს რიცხვები კრაფტის უტოლობას (ანბანის ზომა a = 2) და თუ აკმაყოფილებენ, ააგებს ამ სიგრძეების მქონე უპრეფიქსო კოდს (ანბანით {0,1}). ეს კოდი უნდა გამოიტანოთ მეორე არგუმენტად გადმოცემულ ფაილში (თითო ხაზზე თითო კოდური სიტყვა, ყველა სხვა მონაცემის გარეშე). თუ რიცხვები კრაფტის უტოლობას არ აკმაყოფილებენ და შესაბამისად უპრეფიქსო კოდი არ არსებობს მეორე არგუმენტად გადმოცემული ფაილი ცარიელი დატოვეთ.

კოდური სიტყვები უნდა იყოს დალაგებული იმ მიმდევრობით რა მიმდევრობითაც სიგრძეები გადმოგეცათ და არა სიგრძის ზრდადობის მიხედვით. ანუ რიცხვებს [2, 1, 2] შეიძლება შეესაბამებოდეს კოდი [10, 0, 11] მაგრამ არა [0, 10, 11]. გაითვალისწინეთ რომ სიტყვების სიგრძეები შეიძლება საკამოდ დიდი იყოს (მაგ. 30-40) და ექსპონენტიალური ალგორითმი არ გამოიყენოთ. ასეთი სიგრძეების მაგალითები სატესტო მონაცემებშიც შეგხვდებათ.

პროგრამას დაარქვით "PrefCode.xxx".

4. ჰაფმანის კოდი **[1.5 ქულა]**

დაწერეთ პროგრამა, რომელიც მოცემული ალბათობებისთვის აგებს ორობით ჰაფმანის კოდს. პირველ არგუმენტად გადმოგეცემათ ფაილის სახელი, რომელშიც პირველ ხაზზე ეწერება შეტყობინებების ვარიანტების რაოდენობა n. შემდეგ ხაზზე კი ეწერება n ცალი არაუარყოფითი რიცხვი - შეტყობინებების ალბათობები. აგებული კოდი ზუსტად იგივე ფორმატით უნდა გამოიტანოს პროგრამამ, როგორც წინა პუნქტში. კოდური სიტყვები დაალაგეთ იმავე მიმდევრობით რა მიმდევრობითაც შეტყობინებების ალბათობები გადმოგეცათ. შედეგი ჩაწერეთ მეორე არგუმენტად გადმოცემულ ფაილში.

პროგრამას დაარქვით "Huffman.xxx".

5. შეკუმშვა **[1.5 ქულა]**

დაწერეთ პროგრამა, რომელიც პირველ არგუმენტად გადმოცემული ფაილიდან წაიკითხავს მესამე ან მეოთხე პუნქტში გამოტანილი ფორმატით ჩაწერილ კოდს, რომელსაც 34 სიტყვა ექნება, მეორე არგუმენტად გადმოცემული ფაილიდან წაიკითხავს ქართულ ტექსტს და ამ ტექსტს წაკითხული კოდის საშუალებით ჩაწერს მესამე არგუმენტად გადმოცემულ ფაილში. იგულისხმება რომ პირველი კოდური სიტყვა შეესაბამება space-ს, ხოლო შემდეგი სიტყვები ქართულ ასოებს ანბანის მიხედვით. გარანტია გაქვთ, რომ მოცემული კოდი უპრეფიქსო იქნება და მეორე ფაილში მხოლოდ ქართული ასოების და space-ების შემცველი ტექსტი იქნება.

მიღებული კოდური სიტყვების კონკატენაცია ჩაწერეთ ფაილში არა სტრინგად, არამედ “დავალება 0”-ის მესამე ნაწილში გაკეთებული მეთოდით გადაიყვანეთ ბაიტების მიმდევრობაში (თავისი დაბოლოებით) და ისე ჩაწერეთ.

პროგრამას დაარქვით “Compress.xxx”

6. გახსნა **[1.5 ქულა]**

დაწერეთ პროგრამა, რომლიც პირველ არგუმენტად გადმოცემული ფაილიდან წაიკითხავს კოდს და ამ კოდის საშუალებით გახსნის მეორე არგუმენტად გადმოცემულ ფაილში ჩაწერილ კოდირებულ (შეკუმშულ) მონაცემებს. შედეგად აღდგენილი ტექსტი მესამე არგუმენტად გადმოცემულ ფაილში უნდა ჩაწეროთ.

პროგრამას დაარქვით “Decompress.xxx”