TEMEL PROGRAMLAMA II
BAHAR 2020
LABORATUAR SAATİ ALIŞTIRMALARI
HAFTA 07
(25 Puan)

delicious (eski adıyla, del.icio.us, https://del.icio.us/), kullanıcıların favori bağlantılarını (bookmark'lar; yer imleri) internet üzerinden kaydetmelerini sağlayan bir web sitesidir. Her bağlantının, ait olduğu web sitesinin kategorilerini veya konularını temsil eden bir veya daha fazla "etiketi" vardır; "programlama", "yemek pişirme", "araştırma" vb gibi. Bu laboratuvar çalışmasında, delicious verilerinin küçük bir kısmını işleyerek, üzerinde kümeleme algoritmasını çalıştırabilecek hale getireceksiniz.

Bu ödev metniyle birlikte verilen bir veri seti dosyası bulacaksınız. Veri seti dosyasında geçmiş bir zaman dilimi içinde Delicious'da işaretlenmiş popüler URL'lerden bir kısmı bulunmaktadır. Her Dosyanın her satırında bir URL, o URL'nin ilk kaydedildiği tarih, kaç defa kaydedildiği, bu URL'yi etiketlemede kullanılmış ilk 10 etiket ve her bir etiket için, o URL'i etiketlemekte kaç kez kullanıldığı bilgisi yer almaktadır. Her bir satırın formatı aşağıda gösterildiği gibidir (her bir satırdaki bilgiler "tab" karakteriyle birbirinden ayrılmıştır):

- **URL**
- **Kaç defa kaydedildiği**
- **İlk kaydedildiği tarih**
- **Etiketi**
- **Kaç defa etiketlendiği**
- **[son iki alan bu yer imini etiketlemekte kullanılan 10 farklı etiket için tekrar edilir]**

Örnek olarak bir satırdaki veri şu şekildedir:

http://boingboing.net/ 11053 2002-11-15 blog 5018 news 2763 culture 2542 blogs 2475 technology 2166 fun 1525 tech 1436 daily 1016 art 641 geek 464

1. **(12,5 puan) Bu veri seti dosyasını alarak, en son derste gördüğümüz "blogdata.txt" benzeri bir matris dosyası oluşturmanızı istiyoruz. Matrisinizin satırları etiketler, kolonlarıysa URL'ler (yer im adresleri; bookmark'lar) olmalıdır. Matris hücrelerindeyse, ilgili etiketin o URL'yi etiketlemek için kaç kez kullanıldığı bilgisi yer almalıdır.**

2. **(12,5 puan) Matrisin bu halini hiyerarşik kümeleme fonksiyonumuza girdi olarak verdiğimizde algoritma kümeleme işlemini etiketler üzerinde uygulayacaktır? Yani**

etiketlemekte kullanıldıkları URL'ler ve bu URL'leri etiketleme adetleri açısından birbirlerine yakın olan etiketler kümeleme işlemi sonucunda yakın kümelerde yer alacaktır.

Kümeleme işlemini yer imleri (bookmark'lar), yani URL'ler üzerinde uygulamak istediğimizi düşünelim; yani benzer etiketlerle etiketlenmiş sayfaların kümeleme sonucunda birbirlerine yakın kümelerde olmasını istiyoruz. Kümeleme işlemini URL'ler üzerinde uygulayabilmek için matrisinizin uygun biçimi nasıl olmalıdır? Matrisinizi bu uygun biçime dönüştürün.