2017-2018 Güz Yarıyılı
Algoritma Analizi Final Projesi

Konu : *Locality Sensitive Hashing* Yöntemi ile Benzer Dokümanların Tesbiti

Problem : Doküman sayısının milyonlar mertebesinde olduğu uygulamalarda benzer(tamamen aynı olmayan) dokümanları bulmak için kelime veya karakter bazında karşılaştırma yapıldığında N doküman için Nx(N-1)/2 karşılaştırma yapılması gerektiği düşünüldüğünde zaman alıcı bir iştir. Bu projede *Locality Sensitive Hashing* yönteminin ilk 2 aşamasını gerçekleştirerek benzer dokümanları bulan bir sistem tasarlamanız istenmektedir.

İşlem Adımları: Ödev 2 ana bölümden oluşmaktadır.

1. **Shingling :** Dokümanların k=4,5 ve 8 değerleri için *k-shingle'larını* elde ediniz. Her k-shingle için aşağıdaki bilgileri ekrana yazdırınız:
   a. Her doküman kaç adet k-shingle'dan oluşuyor ?
   b. Her doküman çifti için Jaccard benzerliğini hesaplayarak dokümanların benzerlik oranını veriniz.

   Shingleri elde ederken şunlara dikkat ediniz.
   a. Sadece harfleri ve boşlukları değerlendiriniz.
   b. Harfleri küçük harf olarak değerlendiriniz.
   c. Kelimeler arasında boşluk sayısı 1'den fazla ise sadece 1 boşluk olarak alınız.
   d. Bir shingle'ı sadece 1 defa saklayınız.

2. **Min-hashing :** Hash fonksiyonu olarak $\mathbf{h_i(x) = (a*x + 1) \bmod m}$ fonksiyonunu kullanınız. Fonksiyondaki **a** değerini m'den küçük bir rasgele(random) değer olarak belirleyerek **i=1..100** olarak değişirken 100 adet hash fonksiyonu elde ediniz. Min hashing yöntemi ile imza matrisini(signature matrix) elde ediniz. Her k-shingle için aşağıdaki bilgileri ekrana yazdırınız :
   a. Her doküman çiftinin imza değerlerini kullanarak Jacard benzerliğini hesaplayarak imza benzerlik oranlarını veriniz.
   b. Dışardan okuyacağınız bir eşik seviyesi değeri için imza benzerlik oranı bu eşik seviyesinden büyük olan doküman çiftlerinin isimlerini ekrana yazdırınız.

Teslim Edilecekler: Aşağıda verilen bütün bilgileri içeren tek bir doküman hazırlayınız.
Yaptığınız çalışmayı yöntem ve uygulama bölümlerinden oluşan bir raporda anlatınız.

1. **Yöntem** bölümünde problemi kısaca anlatıp, algoritmanıza ait **ana adımları** yazınız.
2. **Uygulama** bölümünde k=4,5 ve 8 değerleri için aşağıdaki bilgileri dokümanınıza ekleyiniz.
   a. Her k-shingle için her dokümanın kaç farklı shingle'ı olduğunu ve toplam shingle sayısını 1 tablo olarak veriniz.
   b. Her doküman çifti için k-shingle benzerliği oranı ve imza matrisi benzerliği oranı değerlerini bir tablo olarak veriniz.
   c. Eşik seviyesi oranını 0.7, 0.8 ve 0.9 aldığınızda k-shingle benzerliği oranına göre hangi dokümanlar, imza matrisi benzerliği oranına göre hangi dokümanlar benzer kabul ediliyorsa tablo üzerinde gösteriniz.
3. **Sonuç** : Uygulama bölümünde elde ettiğiniz sonuçları aşağıdaki sorulara cevap verecek şekilde yorumlayınız.
   a. k'nın farklı değerleri için benzerlik oranları nasıl değişti?
   b. Sizce hangi k değeri için en doğru sonuçlar alındı?
   c. Bu uygulama için k-shingle benzerliği ile imza matrisi benzerliği sonuçlarını karşılaştırarak başarılarını 1-2 cümle ile yorumlayınız.
4. Algoritmanızın **C dilinde** programını hazırlayarak dokümana ekleyiniz.

Önemli : Algoritmanızı kendiniz tasarlayınız. İstenilen algoritmaların hazır kodunu internetten bulabilirsiniz. Fakat tasarımı kendiniz yaparsanız hem daha iyi öğrenirsiniz hem de edindiğiniz tecrübe öğrendiklerinizin kalıcı olmasını sağlar. Ayrıca internette bulunan hazır bir koda belli bir eşik seviyesinden fazla benzeyen kodlar **kopya** olarak değerlendirilecektir ☹ .

Teslim İşlemleri:
Ödevlerinizi 7 Ocak 2018 Pazar akşamı 23.59'a kadar e-mail ile göndermeniz gerekmektedir. **Ödev teslimi ile ilgili açıklamalar için** Arş. Grv. Zeynep Banu Özger'in sayfasını takip ediniz.

Değerlendirme: Ödeviniz aşağıdaki gibi değerlendirilecektir:

Algoritma Tasarımı ve Programın Çalışması: (%60)

1. Ödev, istenilen işlerin tamamını yerine getirmelidir.
2. Gereksiz kontrollerden ve işlemlerden arınmış bir tasarım yapılmalıdır.
3. Programda gerekli alt modüller belirlenerek her modül ayrı fonksiyon olarak yazılmalıdır.
4. Program hatasız çalışmalıdır.
5. Programın çalışması sırasında, konuyu bilmeyen kişilerin rahatlıkla anlayabilmesi için, giriş ve çıkışlarda mesajlarla bilgi verilmelidir.

Rapor Dokümantasyonu: (%40)

1. Raporun ilk sayfasında, dersin adı, öğrencinin ad, soyad ve numarası, ödev konusu bilgileri yer almalıdır.
2. Kaynak kodda değişken deklerasyonu yapılırken her değişken tek satırda tanımlanmalı, tanımın yanına değişkenin ne için kullanılacağı açıklama olarak yazılmalıdır.
3. Değişken ve fonksiyon(veya metod) isimleri anlamlı olmalıdır.
4. Her fonksiyonun (veya metodun) yaptığı iş, parametreleri ve dönüş değeri açıklanmalıdır.
5. Gerekli yerlerde açıklama satırları ile kodda yapılan işlemler açıklanmadır.
6. Gereksiz kod tekrarı olmamalıdır.
7. Kaynak kodun formatı düzgün olmalıdır.