Algoritma Analizi Dönem Projesi Raporu

Document Similarity Hesaplama Algoritması Uygulaması

Bu uygulamada verilen input dosyalarının, belirlenen k değerleri için shingleları elde edilerek Jaccard Benzerlikleri ve Minhashing yapılarak Signature Benzerlikleri oranlarına göre tablolar çıkarılıp yorumlanması işlemi gerçekleştirilmiştir.

M.Yasin SAĞLAM

15011804

ALGORİTMA ANALİZİ

GRUP 1

Algoritma Analizi Dönem Projesi Raporu

Document Similarity Hesaplama Algoritması Uygulaması

Yöntem

Uygulamada dosyaların birbirlerine olan benzerlik oranlarının K-Shingle' ları harf bazında çıkarılarak hesaplanmasının Jaccard Similarity yöntemiyle ve Minhashing yöntemiyle Signature Similarity olarak hesaplanması istenilmekte, ilgili sonuçlara ilişkin tabloların kullanıcıdan alınan threshold (eşik) değerine göre işaretlenerek farklı iki yönteme ilişkin sonuçların kıyaslanarak yorumlanması beklenmektedir. Uygulamada verilen input dosyalarının kullanıcıdan alınan k değerlerine göre shingleları hesaplanmıştır. Hesaplanan k-shingle sayıları dosya bazında ve toplamda kaç adet olduğu bilgisi ile birlikte bir tabloda ekrana yazdırılmıştır. Sonrasında ise öncelikle dosyaların birbirlerine olan benzerlik oranları Jaccard Similarity yöntemine göre hesaplanarak bir tabloda kullanıcıya gösterilmiştir. Benzerlik hesabı için kullanıcıdan alınan threshold değerine göre threshold u geçen dosya ikilileri tabloda benzer olan dosyalar olarak işaretlenmiştir ve ikililer bir tablo halinde ek olarak benzerlik oranları tablosunun altına yazılmıştır. Daha sonrasında soruda verilen hash fonksiyonu prototipi kullanılarak 100 adet hash fonksiyonu üretilmiştir. Minhashing işlemi sonucunda Signature(İmza) matrisi elde edilmiş ve bu matrise göre dosyaların birbirlerine olan Signature Similarity oranları bir tablo halinde kullanıcıdan alınan threshold değerini geçen değerler işaretli(farklı renkte) olmak kaydıyla benzer dosya ikililerini içeren tablo ile birlikte ekrana yazdırılmıştır. İmplementasyon bir adet struct kullanılarak yapılmıştır. İlgili struct aşağıdaki resimdeki gibi tanımlanmış ve açıklanmıştır.

```
typedef struct{
    char **shingles; //kiyas yapabilmek icin tum shingle lar karakter olarak saklaniyor
    int k_val; //Shingle larin k degeri tutuluyor
    int **matrix; //shingle ve file matrisi tutuluyor
    float **jaccard;//jaccard similarity oranlari tutuluyor
    float **minhash;//minhash signature similarity oranlari tutuluyor
    int file_num; //kac adet dosya bulundugu bilgisini tutar --matrix sutun sayisi
    int shingle_count; //kac adet shingle icerdigi bilgisi tutuluyor--matrix satir sayisi
    float threshold;//threshold degeri
}SHINGLE;
```

İmplementasyonda kullanılan fonksiyonlar ve prototipleri aşağıdaki resimdeki gibi tanımlanmış ve açıklanmıştır. Detaylı bilgi için kod dosyasını inceleyiniz.

```
/**
 * Ilgili structa ilk deger atamalari yapan fonksiyonlar
    @param shgl         Shingle struct pointer
*/
void initialize(SHINGLE *shgl){...}

/**
 * Dosyalari okuyarak shinglelarini cikartan ve her bir dosyanin
 * icerisinde bulunan shingle lari struct icerisindeki matrix adli
 * matriste saklayan fonksiyon
    @param fp           shingle olusturulacak dosyanin pointeri
    @param file_id      id of file
    @param shgl        Shingle struct pointer
*/
void createShingles(FILE *fp,int file_id,SHINGLE *shgl){...}

/**
 * Her bir dosya icin tekrarsiz shingle sayisini ve toplam shingle sayisini
 * tablo olarak ekrana yazan fonksiyon
    @param shgl         Shingle struct pointer
    @param files       File names
*/
void print_shingle(SHINGLE *shgl,char **files){...}

/**
 * Struct icindeki jaccard matrisinde bulunan dosyalarin benzerligini ve verilen thresholda gore,
 * benzer dosya ikililerini ekrana tablo olarak yazan fonksiyon
    @param shgl         Shingle struct pointer
    @param files       File names
*/
void printJaccard(SHINGLE *shgl,char **files){...}

/**
 * Jaccard benzerligini hesaplayarak oranlari struct icerisindeki
 * jaccard matrisine yazan fonksiyon
    @param shgl         Shingle struct pointer
    @param files       File names
*/
void calculate_jaccard(SHINGLE *shgl,char **files){...}
```

```c
/**
 * Hash degeri ureten hash fonksiyonu
    @param a   random value
    @param x   0 to hash_Size
    @param m   Shingle count
*/
int hash(int a,int x, int m){...}

/**
 * Signature benzerligine gore hesaplanan dosyalarin,
 * Benzerlik oranlari ve benzer ikililer gibi sonuclari ekrana yazdiran fonksiyon
 * @param shgl      Shingle struct pointer
 * @param files     Filenames
 */
void printSignature(SHINGLE *shgl,char **files){...}
/**
 * Imza matrisine gore benzerlik oranlarini struct icerisinde minhash matrisine yazan fonksiyon
 * @param shgl
 * @param files
 */
void calculate_minhash(SHINGLE *shgl,char **files){...}
```

Uygulama

İstenilen Sonuçlar İçin Ekran Görüntüleri

*Tablolarda sarı renk ile işaretlenen benzerlik oranları eşik değerini geçen benzerlik oranlarıdır.

Shingle Sayısı Tabloları

K=4

```
File : 1.txt --> Shingle Count: 420
File : 2.txt --> Shingle Count: 421
File : 3.txt --> Shingle Count: 418
File : 4.txt --> Shingle Count: 423
File : 5.txt --> Shingle Count: 413
File : 6.txt --> Shingle Count: 425
File : 7.txt --> Shingle Count: 423
File : 8.txt --> Shingle Count: 420
File : 9.txt --> Shingle Count: 421
File : 10.txt --> Shingle Count: 422
File : 11.txt --> Shingle Count: 420
File : 12.txt --> Shingle Count: 422
File : 13.txt --> Shingle Count: 427
File : 14.txt --> Shingle Count: 428
File : 15.txt --> Shingle Count: 422
File : 16.txt --> Shingle Count: 425
File : 17.txt --> Shingle Count: 429
File : 18.txt --> Shingle Count: 420
File : 19.txt --> Shingle Count: 418
File : 20.txt --> Shingle Count: 391
Total shingle of all files for k=4 : 540
```

K=5

```
File : 1.txt --> Shingle Count: 478
File : 2.txt --> Shingle Count: 479
File : 3.txt --> Shingle Count: 472
File : 4.txt --> Shingle Count: 476
File : 5.txt --> Shingle Count: 465
File : 6.txt --> Shingle Count: 483
File : 7.txt --> Shingle Count: 480
File : 8.txt --> Shingle Count: 477
File : 9.txt --> Shingle Count: 478
File : 10.txt --> Shingle Count: 478
File : 11.txt --> Shingle Count: 478
File : 12.txt --> Shingle Count: 480
File : 13.txt --> Shingle Count: 484
File : 14.txt --> Shingle Count: 484
File : 15.txt --> Shingle Count: 477
File : 16.txt --> Shingle Count: 481
File : 17.txt --> Shingle Count: 485
File : 18.txt --> Shingle Count: 472
File : 19.txt --> Shingle Count: 475
File : 20.txt --> Shingle Count: 447
Total shingle of all files for k=5 : 635
```

K=8

```
File : 1.txt --> Shingle Count: 558
File : 2.txt --> Shingle Count: 555
File : 3.txt --> Shingle Count: 546
File : 4.txt --> Shingle Count: 548
File : 5.txt --> Shingle Count: 537
File : 6.txt --> Shingle Count: 563
File : 7.txt --> Shingle Count: 557
File : 8.txt --> Shingle Count: 554
File : 9.txt --> Shingle Count: 552
File : 10.txt --> Shingle Count: 551
File : 11.txt --> Shingle Count: 558
File : 12.txt --> Shingle Count: 560
File : 13.txt --> Shingle Count: 561
File : 14.txt --> Shingle Count: 560
File : 15.txt --> Shingle Count: 553
File : 16.txt --> Shingle Count: 558
File : 17.txt --> Shingle Count: 559
File : 18.txt --> Shingle Count: 538
File : 19.txt --> Shingle Count: 552
File : 20.txt --> Shingle Count: 525
Total shingle of all files for k=8 : 796
```

K=4, Threshold=0.7

Jaccard Similarity

K VALUE : 4 THRESHOLD : 0.70
JACCARD SIMILARITIES OF ALL DOCUMENT COMBINATIONS

	1.txt	2.txt	3.txt	4.txt	5.txt	6.txt	7.txt	8.txt	9.txt	10.txt	11.txt	12.txt	13.txt	14.txt	15.txt	16.txt	17.txt	18.txt	19.txt	20.txt
1.txt	1.00	0.96	0.89	0.87	0.84	0.96	0.94	0.91	0.89	0.88	1.00	0.91	0.87	0.84	0.81	0.80	0.79	0.76	0.95	0.89
2.txt	0.96	1.00	0.93	0.91	0.88	0.91	0.90	0.87	0.85	0.84	0.96	0.87	0.83	0.81	0.78	0.77	0.76	0.74	0.91	0.85
3.txt	0.89	0.93	1.00	0.96	0.94	0.85	0.84	0.81	0.79	0.78	0.89	0.81	0.77	0.75	0.72	0.71	0.71	0.68	0.85	0.78
4.txt	0.87	0.91	0.96	1.00	0.97	0.84	0.82	0.80	0.78	0.76	0.87	0.79	0.77	0.74	0.72	0.71	0.70	0.68	0.82	0.76
5.txt	0.84	0.88	0.94	0.97	1.00	0.81	0.80	0.77	0.76	0.74	0.84	0.77	0.74	0.72	0.69	0.68	0.68	0.66	0.80	0.74
6.txt	0.96	0.91	0.85	0.84	0.81	1.00	0.99	0.96	0.94	0.92	0.96	0.87	0.83	0.81	0.78	0.77	0.76	0.73	0.91	0.85
7.txt	0.94	0.90	0.84	0.82	0.80	0.99	1.00	0.97	0.95	0.93	0.94	0.86	0.82	0.80	0.77	0.76	0.75	0.72	0.90	0.83
8.txt	0.91	0.87	0.81	0.80	0.77	0.96	0.97	1.00	0.98	0.96	0.91	0.83	0.79	0.77	0.75	0.74	0.73	0.71	0.87	0.81
9.txt	0.89	0.85	0.79	0.78	0.76	0.94	0.95	0.98	1.00	0.98	0.89	0.82	0.78	0.75	0.73	0.72	0.71	0.70	0.85	0.79
10.txt	0.88	0.84	0.78	0.76	0.74	0.92	0.93	0.96	0.98	1.00	0.88	0.80	0.76	0.74	0.72	0.70	0.70	0.68	0.83	0.77
11.txt	1.00	0.96	0.89	0.87	0.84	0.96	0.94	0.91	0.89	0.88	1.00	0.91	0.87	0.84	0.81	0.80	0.79	0.76	0.95	0.89
12.txt	0.91	0.87	0.81	0.79	0.77	0.87	0.86	0.83	0.82	0.80	0.91	1.00	0.95	0.92	0.89	0.88	0.87	0.84	0.90	0.84
13.txt	0.87	0.83	0.77	0.77	0.74	0.83	0.82	0.79	0.78	0.76	0.87	0.95	1.00	0.97	0.94	0.92	0.91	0.88	0.86	0.79
14.txt	0.84	0.81	0.75	0.74	0.72	0.81	0.80	0.77	0.75	0.74	0.84	0.92	0.97	1.00	0.97	0.95	0.94	0.91	0.83	0.77
15.txt	0.81	0.78	0.72	0.72	0.69	0.78	0.77	0.75	0.73	0.72	0.81	0.89	0.94	0.97	1.00	0.98	0.97	0.94	0.80	0.78
16.txt	0.80	0.77	0.71	0.71	0.68	0.77	0.76	0.74	0.72	0.70	0.80	0.88	0.92	0.95	0.98	1.00	0.99	0.96	0.79	0.77
17.txt	0.79	0.76	0.71	0.70	0.68	0.76	0.75	0.73	0.71	0.70	0.79	0.87	0.91	0.94	0.97	0.99	1.00	0.97	0.78	0.76
18.txt	0.76	0.74	0.68	0.68	0.66	0.73	0.72	0.71	0.70	0.68	0.76	0.84	0.88	0.91	0.94	0.96	0.97	1.00	0.75	0.73
19.txt	0.95	0.91	0.85	0.82	0.80	0.91	0.90	0.87	0.85	0.83	0.95	0.90	0.86	0.83	0.80	0.79	0.78	0.75	1.00	0.93
20.txt	0.89	0.85	0.78	0.76	0.74	0.85	0.83	0.81	0.79	0.77	0.89	0.84	0.79	0.77	0.78	0.77	0.76	0.73	0.93	1.00

Signature Similarity

K VALUE : 4 THRESHOLD : 0.70
SIGNATURE SIMILARITIES OF ALL DOCUMENT COMBINATIONS

	1.txt	2.txt	3.txt	4.txt	5.txt	6.txt	7.txt	8.txt	9.txt	10.txt	11.txt	12.txt	13.txt	14.txt	15.txt	16.txt	17.txt	18.txt	19.txt	20.txt
1.txt	1.00	1.00	0.99	0.99	0.97	0.99	0.98	0.98	0.97	0.97	1.00	0.98	0.97	0.97	0.96	0.96	0.96	0.95	0.99	0.98
2.txt	1.00	1.00	0.99	0.99	0.97	0.99	0.98	0.98	0.97	0.97	1.00	0.98	0.97	0.97	0.96	0.96	0.96	0.95	0.99	0.98
3.txt	0.99	0.99	1.00	1.00	0.98	0.98	0.97	0.97	0.96	0.96	0.99	0.97	0.96	0.96	0.95	0.95	0.95	0.94	0.98	0.97
4.txt	0.99	0.99	1.00	1.00	0.98	0.98	0.97	0.97	0.96	0.96	0.99	0.97	0.96	0.96	0.95	0.95	0.95	0.94	0.98	0.97
5.txt	0.97	0.97	0.98	0.98	1.00	0.96	0.95	0.95	0.94	0.94	0.97	0.95	0.94	0.94	0.93	0.93	0.93	0.92	0.96	0.95
6.txt	0.99	0.99	0.98	0.98	0.96	1.00	0.99	0.99	0.98	0.98	0.99	0.97	0.96	0.96	0.95	0.95	0.95	0.94	0.98	0.97
7.txt	0.98	0.98	0.97	0.97	0.95	0.99	1.00	1.00	0.99	0.99	0.98	0.96	0.95	0.95	0.94	0.94	0.94	0.93	0.97	0.96
8.txt	0.98	0.98	0.97	0.97	0.95	0.99	1.00	1.00	0.99	0.99	0.98	0.96	0.95	0.95	0.94	0.94	0.94	0.93	0.97	0.96
9.txt	0.97	0.97	0.96	0.96	0.94	0.98	0.99	0.99	1.00	1.00	0.97	0.95	0.94	0.94	0.93	0.93	0.93	0.92	0.96	0.95
10.txt	0.97	0.97	0.96	0.96	0.94	0.98	0.99	0.99	1.00	1.00	0.97	0.95	0.94	0.94	0.93	0.93	0.93	0.92	0.96	0.95
11.txt	1.00	1.00	0.99	0.99	0.97	0.99	0.98	0.98	0.97	0.97	1.00	0.98	0.97	0.97	0.96	0.96	0.96	0.95	0.99	0.98
12.txt	0.98	0.98	0.97	0.97	0.95	0.97	0.96	0.96	0.95	0.95	0.98	1.00	0.99	0.99	0.98	0.98	0.98	0.97	0.97	0.96
13.txt	0.97	0.97	0.96	0.96	0.94	0.96	0.95	0.95	0.94	0.94	0.97	0.99	1.00	1.00	0.99	0.99	0.99	0.98	0.96	0.95
14.txt	0.97	0.97	0.96	0.96	0.94	0.96	0.95	0.95	0.94	0.94	0.97	0.99	1.00	1.00	0.99	0.99	0.99	0.98	0.96	0.95
15.txt	0.96	0.96	0.95	0.95	0.93	0.95	0.94	0.94	0.93	0.93	0.96	0.98	0.99	0.99	1.00	1.00	1.00	0.99	0.95	0.96
16.txt	0.96	0.96	0.95	0.95	0.93	0.95	0.94	0.94	0.93	0.93	0.96	0.98	0.99	0.99	1.00	1.00	1.00	0.99	0.95	0.96
17.txt	0.96	0.96	0.95	0.95	0.93	0.95	0.94	0.94	0.93	0.93	0.96	0.98	0.99	0.99	1.00	1.00	1.00	0.99	0.95	0.96
18.txt	0.95	0.95	0.94	0.94	0.92	0.94	0.93	0.93	0.92	0.92	0.95	0.97	0.98	0.98	0.99	0.99	0.99	1.00	0.94	0.95
19.txt	0.99	0.99	0.98	0.98	0.96	0.98	0.97	0.97	0.96	0.96	0.99	0.97	0.96	0.96	0.95	0.95	0.95	0.94	1.00	0.99
20.txt	0.98	0.98	0.97	0.97	0.95	0.97	0.96	0.96	0.95	0.95	0.98	0.96	0.95	0.95	0.96	0.96	0.96	0.95	0.99	1.00

K=4, Threshold=0.8

Jaccard Similarity

K VALUE : 4 THRESHOLD : 0.80
JACCARD SIMILARITIES OF ALL DOCUMENT COMBINATIONS

	1.txt	2.txt	3.txt	4.txt	5.txt	6.txt	7.txt	8.txt	9.txt	10.txt	11.txt	12.txt	13.txt	14.txt	15.txt	16.txt	17.txt	18.txt	19.txt	20.txt
1.txt	1.00	0.96	0.89	0.87	0.84	0.96	0.94	0.91	0.89	0.88	1.00	0.91	0.87	0.84	0.81	0.80	0.79	0.76	0.95	0.89
2.txt	0.96	1.00	0.93	0.91	0.88	0.91	0.90	0.87	0.85	0.84	0.96	0.87	0.83	0.81	0.78	0.77	0.76	0.74	0.91	0.85
3.txt	0.89	0.93	1.00	0.96	0.94	0.85	0.84	0.81	0.79	0.78	0.89	0.81	0.77	0.75	0.72	0.71	0.71	0.68	0.85	0.78
4.txt	0.87	0.91	0.96	1.00	0.97	0.84	0.82	0.80	0.78	0.76	0.87	0.79	0.77	0.74	0.72	0.71	0.70	0.68	0.82	0.76
5.txt	0.84	0.88	0.94	0.97	1.00	0.81	0.80	0.77	0.76	0.74	0.84	0.77	0.74	0.72	0.69	0.68	0.68	0.66	0.80	0.74
6.txt	0.96	0.91	0.85	0.84	0.81	1.00	0.99	0.96	0.94	0.92	0.96	0.87	0.83	0.81	0.78	0.77	0.76	0.73	0.91	0.85
7.txt	0.94	0.90	0.84	0.82	0.80	0.99	1.00	0.97	0.95	0.93	0.94	0.86	0.82	0.80	0.77	0.76	0.75	0.72	0.90	0.83
8.txt	0.91	0.87	0.81	0.80	0.77	0.96	0.97	1.00	0.98	0.96	0.91	0.83	0.79	0.77	0.75	0.74	0.73	0.71	0.87	0.81
9.txt	0.89	0.85	0.79	0.78	0.76	0.94	0.95	0.98	1.00	0.98	0.89	0.82	0.78	0.75	0.73	0.72	0.71	0.70	0.85	0.79
10.txt	0.88	0.84	0.78	0.76	0.74	0.92	0.93	0.96	0.98	1.00	0.88	0.80	0.76	0.74	0.72	0.70	0.70	0.68	0.83	0.77
11.txt	1.00	0.96	0.89	0.87	0.84	0.96	0.94	0.91	0.89	0.88	1.00	0.91	0.87	0.84	0.81	0.80	0.79	0.76	0.95	0.89
12.txt	0.91	0.87	0.81	0.79	0.77	0.87	0.86	0.83	0.82	0.80	0.91	1.00	0.95	0.92	0.89	0.88	0.87	0.84	0.90	0.84
13.txt	0.87	0.83	0.77	0.77	0.74	0.83	0.82	0.79	0.78	0.76	0.87	0.95	1.00	0.97	0.94	0.92	0.91	0.88	0.86	0.79
14.txt	0.84	0.81	0.75	0.74	0.72	0.81	0.80	0.77	0.75	0.74	0.84	0.92	0.97	1.00	0.97	0.95	0.94	0.91	0.83	0.77
15.txt	0.81	0.78	0.72	0.72	0.69	0.78	0.77	0.75	0.73	0.72	0.81	0.89	0.94	0.97	1.00	0.98	0.97	0.94	0.80	0.78
16.txt	0.80	0.77	0.71	0.71	0.68	0.77	0.76	0.74	0.72	0.70	0.80	0.88	0.92	0.95	0.98	1.00	0.99	0.96	0.79	0.77
17.txt	0.79	0.76	0.71	0.70	0.68	0.76	0.75	0.73	0.71	0.70	0.79	0.87	0.91	0.94	0.97	0.99	1.00	0.97	0.78	0.76
18.txt	0.76	0.74	0.68	0.68	0.66	0.73	0.72	0.71	0.70	0.68	0.76	0.84	0.88	0.91	0.94	0.96	0.97	1.00	0.75	0.73
19.txt	0.95	0.91	0.85	0.82	0.80	0.91	0.90	0.87	0.85	0.83	0.95	0.90	0.86	0.83	0.80	0.79	0.78	0.75	1.00	0.93
20.txt	0.89	0.85	0.78	0.76	0.74	0.85	0.83	0.81	0.79	0.77	0.89	0.84	0.79	0.77	0.78	0.77	0.76	0.73	0.93	1.00

Signature Similarity

K VALUE : 4 THRESHOLD : 0.80
SIGNATURE SIMILARITIES OF ALL DOCUMENT COMBINATIONS

	1.txt	2.txt	3.txt	4.txt	5.txt	6.txt	7.txt	8.txt	9.txt	10.txt	11.txt	12.txt	13.txt	14.txt	15.txt	16.txt	17.txt	18.txt	19.txt	20.txt
1.txt	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.99	0.99	0.98	1.00	0.99	0.98	0.98	0.97	0.97	0.96	0.93	0.99	0.98
2.txt	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.99	0.99	0.98	1.00	0.99	0.98	0.98	0.97	0.97	0.96	0.93	0.99	0.98
3.txt	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.99	0.99	0.98	1.00	0.99	0.98	0.98	0.97	0.97	0.96	0.93	0.99	0.98
4.txt	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.99	0.99	0.98	1.00	0.99	0.98	0.98	0.97	0.97	0.96	0.93	0.99	0.98
5.txt	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.98	0.98	0.97	0.99	0.98	0.97	0.97	0.96	0.96	0.95	0.92	0.98	0.97
6.txt	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.99	0.99	0.98	1.00	0.99	0.98	0.98	0.97	0.97	0.96	0.93	0.99	0.98
7.txt	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.99	0.99	0.98	1.00	0.99	0.98	0.98	0.97	0.97	0.96	0.93	0.99	0.98
8.txt	0.99	0.99	0.99	0.99	0.98	0.99	0.99	1.00	1.00	0.99	0.99	0.98	0.97	0.97	0.96	0.96	0.95	0.92	0.98	0.97
9.txt	0.99	0.99	0.99	0.99	0.98	0.99	0.99	1.00	1.00	0.99	0.99	0.98	0.97	0.97	0.96	0.96	0.95	0.92	0.98	0.97
10.txt	0.98	0.98	0.98	0.98	0.97	0.98	0.98	0.99	0.99	1.00	0.98	0.97	0.96	0.96	0.95	0.95	0.94	0.91	0.97	0.96
11.txt	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.99	0.99	0.98	1.00	0.99	0.98	0.98	0.97	0.97	0.96	0.93	0.99	0.98
12.txt	0.99	0.99	0.99	0.99	0.98	0.99	0.99	0.98	0.98	0.97	0.99	1.00	0.99	0.99	0.98	0.98	0.97	0.94	0.98	0.97
13.txt	0.98	0.98	0.98	0.98	0.97	0.98	0.98	0.97	0.97	0.96	0.98	0.99	1.00	1.00	0.99	0.99	0.98	0.95	0.97	0.96
14.txt	0.98	0.98	0.98	0.98	0.97	0.98	0.98	0.97	0.97	0.96	0.98	0.99	1.00	1.00	0.99	0.99	0.98	0.95	0.97	0.96
15.txt	0.97	0.97	0.97	0.97	0.96	0.97	0.97	0.96	0.96	0.95	0.97	0.98	0.99	0.99	1.00	1.00	0.99	0.96	0.96	0.95
16.txt	0.97	0.97	0.97	0.97	0.96	0.97	0.97	0.96	0.96	0.95	0.97	0.98	0.99	0.99	1.00	1.00	0.99	0.96	0.96	0.95
17.txt	0.96	0.96	0.96	0.96	0.95	0.96	0.96	0.95	0.95	0.94	0.96	0.97	0.98	0.98	0.99	0.99	1.00	0.97	0.95	0.94
18.txt	0.93	0.93	0.93	0.93	0.92	0.93	0.93	0.92	0.92	0.91	0.93	0.94	0.95	0.95	0.96	0.96	0.97	1.00	0.92	0.91
19.txt	0.99	0.99	0.99	0.99	0.98	0.99	0.99	0.98	0.98	0.97	0.99	0.98	0.97	0.97	0.96	0.96	0.95	0.92	1.00	0.99
20.txt	0.98	0.98	0.98	0.98	0.97	0.98	0.98	0.97	0.97	0.96	0.98	0.97	0.96	0.96	0.95	0.95	0.94	0.91	0.99	1.00

K=4, Threshold=0.9

Jaccard Similarity

K VALUE : 4 THRESHOLD : 0.90
JACCARD SIMILARITIES OF ALL DOCUMENT COMBINATIONS

	1.txt	2.txt	3.txt	4.txt	5.txt	6.txt	7.txt	8.txt	9.txt	10.txt	11.txt	12.txt	13.txt	14.txt	15.txt	16.txt	17.txt	18.txt	19.txt	20.txt
1.txt	1.00	0.96	0.89	0.87	0.84	0.96	0.94	0.91	0.89	0.88	1.00	0.91	0.87	0.84	0.81	0.80	0.79	0.76	0.95	0.89
2.txt	0.96	1.00	0.93	0.91	0.88	0.91	0.90	0.87	0.85	0.84	0.96	0.87	0.83	0.81	0.78	0.77	0.76	0.74	0.91	0.85
3.txt	0.89	0.93	1.00	0.96	0.94	0.85	0.84	0.81	0.79	0.78	0.89	0.81	0.77	0.75	0.72	0.71	0.71	0.68	0.85	0.78
4.txt	0.87	0.91	0.96	1.00	0.97	0.84	0.82	0.80	0.78	0.76	0.87	0.79	0.77	0.74	0.72	0.71	0.70	0.68	0.82	0.76
5.txt	0.84	0.88	0.94	0.97	1.00	0.81	0.80	0.77	0.76	0.74	0.84	0.77	0.74	0.72	0.69	0.68	0.68	0.66	0.80	0.74
6.txt	0.96	0.91	0.85	0.84	0.81	1.00	0.99	0.96	0.94	0.92	0.96	0.87	0.83	0.81	0.78	0.77	0.76	0.73	0.91	0.85
7.txt	0.94	0.90	0.84	0.82	0.80	0.99	1.00	0.97	0.95	0.93	0.94	0.86	0.82	0.80	0.77	0.76	0.75	0.72	0.90	0.83
8.txt	0.91	0.87	0.81	0.80	0.77	0.96	0.97	1.00	0.98	0.96	0.91	0.83	0.79	0.77	0.75	0.74	0.73	0.71	0.87	0.81
9.txt	0.89	0.85	0.79	0.78	0.76	0.94	0.95	0.98	1.00	0.98	0.89	0.82	0.78	0.75	0.73	0.72	0.71	0.70	0.85	0.79
10.txt	0.88	0.84	0.78	0.76	0.74	0.92	0.93	0.96	0.98	1.00	0.88	0.80	0.76	0.74	0.72	0.70	0.70	0.68	0.83	0.77
11.txt	1.00	0.96	0.89	0.87	0.84	0.96	0.94	0.91	0.89	0.88	1.00	0.91	0.87	0.84	0.81	0.80	0.79	0.76	0.95	0.89
12.txt	0.91	0.87	0.81	0.79	0.77	0.87	0.86	0.83	0.82	0.80	0.91	1.00	0.95	0.92	0.89	0.88	0.87	0.84	0.90	0.84
13.txt	0.87	0.83	0.77	0.77	0.74	0.83	0.82	0.79	0.78	0.76	0.87	0.95	1.00	0.97	0.94	0.92	0.91	0.88	0.86	0.79
14.txt	0.84	0.81	0.75	0.74	0.72	0.81	0.80	0.77	0.75	0.74	0.84	0.92	0.97	1.00	0.97	0.95	0.94	0.91	0.83	0.77
15.txt	0.81	0.78	0.72	0.72	0.69	0.78	0.77	0.75	0.73	0.72	0.81	0.89	0.94	0.97	1.00	0.98	0.97	0.94	0.80	0.78
16.txt	0.80	0.77	0.71	0.71	0.68	0.77	0.76	0.74	0.72	0.70	0.80	0.88	0.92	0.95	0.98	1.00	0.99	0.96	0.79	0.77
17.txt	0.79	0.76	0.71	0.70	0.68	0.76	0.75	0.73	0.71	0.70	0.79	0.87	0.91	0.94	0.97	0.99	1.00	0.97	0.78	0.76
18.txt	0.76	0.74	0.68	0.68	0.66	0.73	0.72	0.71	0.70	0.68	0.76	0.84	0.88	0.91	0.94	0.96	0.97	1.00	0.75	0.73
19.txt	0.95	0.91	0.85	0.82	0.80	0.91	0.90	0.87	0.85	0.83	0.95	0.90	0.86	0.83	0.80	0.79	0.78	0.75	1.00	0.93
20.txt	0.89	0.85	0.78	0.76	0.74	0.85	0.83	0.81	0.79	0.77	0.89	0.84	0.79	0.77	0.78	0.77	0.76	0.73	0.93	1.00

Signature Similarity

K VALUE : 4 THRESHOLD : 0.90
SIGNATURE SIMILARITIES OF ALL DOCUMENT COMBINATIONS

	1.txt	2.txt	3.txt	4.txt	5.txt	6.txt	7.txt	8.txt	9.txt	10.txt	11.txt	12.txt	13.txt	14.txt	15.txt	16.txt	17.txt	18.txt	19.txt	20.txt
1.txt	1.00	0.99	0.98	0.98	0.97	0.99	0.99	0.99	0.99	0.97	1.00	0.99	0.99	0.98	0.97	0.96	0.95	0.95	0.99	0.98
2.txt	0.99	1.00	0.99	0.99	0.98	0.98	0.98	0.98	0.98	0.96	0.99	0.98	0.98	0.97	0.96	0.95	0.94	0.94	0.98	0.97
3.txt	0.98	0.99	1.00	1.00	0.99	0.97	0.97	0.97	0.97	0.95	0.98	0.97	0.97	0.96	0.95	0.94	0.93	0.93	0.97	0.96
4.txt	0.98	0.99	1.00	1.00	0.99	0.97	0.97	0.97	0.97	0.95	0.98	0.97	0.97	0.96	0.95	0.94	0.93	0.93	0.97	0.96
5.txt	0.97	0.98	0.99	0.99	1.00	0.96	0.96	0.96	0.96	0.94	0.97	0.96	0.96	0.95	0.94	0.93	0.92	0.92	0.96	0.95
6.txt	0.99	0.98	0.97	0.97	0.96	1.00	1.00	1.00	1.00	0.98	0.99	0.98	0.98	0.97	0.96	0.95	0.94	0.94	0.98	0.97
7.txt	0.99	0.98	0.97	0.97	0.96	1.00	1.00	1.00	1.00	0.98	0.99	0.98	0.98	0.97	0.96	0.95	0.94	0.94	0.98	0.97
8.txt	0.99	0.98	0.97	0.97	0.96	1.00	1.00	1.00	1.00	0.98	0.99	0.98	0.98	0.97	0.96	0.95	0.94	0.94	0.98	0.97
9.txt	0.99	0.98	0.97	0.97	0.96	1.00	1.00	1.00	1.00	0.98	0.99	0.98	0.98	0.97	0.96	0.95	0.94	0.94	0.98	0.97
10.txt	0.97	0.96	0.95	0.95	0.94	0.98	0.98	0.98	0.98	1.00	0.97	0.96	0.96	0.95	0.94	0.93	0.92	0.92	0.96	0.95
11.txt	1.00	0.99	0.98	0.98	0.97	0.99	0.99	0.99	0.99	0.97	1.00	0.99	0.99	0.98	0.97	0.96	0.95	0.95	0.99	0.98
12.txt	0.99	0.98	0.97	0.97	0.96	0.98	0.98	0.98	0.98	0.96	0.99	1.00	1.00	0.99	0.98	0.97	0.96	0.96	0.98	0.97
13.txt	0.99	0.98	0.97	0.97	0.96	0.98	0.98	0.98	0.98	0.96	0.99	1.00	1.00	0.99	0.98	0.97	0.96	0.96	0.98	0.97
14.txt	0.98	0.97	0.96	0.96	0.95	0.97	0.97	0.97	0.97	0.95	0.98	0.99	0.99	1.00	0.99	0.98	0.97	0.97	0.97	0.96
15.txt	0.97	0.96	0.95	0.95	0.94	0.96	0.96	0.96	0.96	0.94	0.97	0.98	0.98	0.99	1.00	0.99	0.98	0.98	0.96	0.97
16.txt	0.96	0.95	0.94	0.94	0.93	0.95	0.95	0.95	0.95	0.93	0.96	0.97	0.97	0.98	0.99	1.00	0.99	0.99	0.95	0.96
17.txt	0.95	0.94	0.93	0.93	0.92	0.94	0.94	0.94	0.94	0.92	0.95	0.96	0.96	0.97	0.98	0.99	1.00	1.00	0.94	0.95
18.txt	0.95	0.94	0.93	0.93	0.92	0.94	0.94	0.94	0.94	0.92	0.95	0.96	0.96	0.97	0.98	0.99	1.00	1.00	0.94	0.95
19.txt	0.99	0.98	0.97	0.97	0.96	0.98	0.98	0.98	0.98	0.96	0.99	0.98	0.98	0.97	0.96	0.95	0.94	0.94	1.00	0.99
20.txt	0.98	0.97	0.96	0.96	0.95	0.97	0.97	0.97	0.97	0.95	0.98	0.97	0.97	0.96	0.97	0.96	0.95	0.95	0.99	1.00

K=5, Threshold=0.7

Jaccard Similarity

K VALUE : 5 THRESHOLD : 0.70
JACCARD SIMILARITIES OF ALL DOCUMENT COMBINATIONS

	1.txt	2.txt	3.txt	4.txt	5.txt	6.txt	7.txt	8.txt	9.txt	10.txt	11.txt	12.txt	13.txt	14.txt	15.txt	16.txt	17.txt	18.txt	19.txt	20.txt
1.txt	1.00	0.95	0.87	0.85	0.82	0.95	0.94	0.91	0.89	0.86	1.00	0.90	0.85	0.82	0.79	0.77	0.76	0.73	0.95	0.89
2.txt	0.95	1.00	0.92	0.89	0.87	0.90	0.89	0.86	0.84	0.82	0.95	0.85	0.80	0.78	0.75	0.73	0.72	0.69	0.90	0.84
3.txt	0.87	0.92	1.00	0.95	0.92	0.83	0.82	0.79	0.77	0.75	0.87	0.79	0.74	0.71	0.69	0.67	0.66	0.63	0.83	0.77
4.txt	0.85	0.89	0.95	1.00	0.97	0.81	0.80	0.77	0.75	0.73	0.85	0.76	0.73	0.70	0.67	0.66	0.65	0.62	0.81	0.75
5.txt	0.82	0.87	0.92	0.97	1.00	0.79	0.77	0.75	0.73	0.71	0.82	0.74	0.70	0.68	0.65	0.64	0.63	0.60	0.78	0.72
6.txt	0.95	0.90	0.83	0.81	0.79	1.00	0.98	0.95	0.93	0.90	0.95	0.86	0.81	0.78	0.75	0.74	0.73	0.69	0.91	0.85
7.txt	0.94	0.89	0.82	0.80	0.77	0.98	1.00	0.97	0.94	0.92	0.94	0.85	0.80	0.77	0.74	0.73	0.71	0.68	0.89	0.83
8.txt	0.91	0.86	0.79	0.77	0.75	0.95	0.97	1.00	0.97	0.95	0.91	0.82	0.77	0.74	0.72	0.70	0.69	0.67	0.87	0.80
9.txt	0.89	0.84	0.77	0.75	0.73	0.93	0.94	0.97	1.00	0.98	0.89	0.80	0.75	0.72	0.70	0.68	0.67	0.66	0.84	0.78
10.txt	0.86	0.82	0.75	0.73	0.71	0.90	0.92	0.95	0.98	1.00	0.86	0.78	0.73	0.71	0.68	0.66	0.65	0.64	0.82	0.76
11.txt	1.00	0.95	0.87	0.85	0.82	0.95	0.94	0.91	0.89	0.86	1.00	0.90	0.85	0.82	0.79	0.77	0.76	0.73	0.95	0.89
12.txt	0.90	0.85	0.79	0.76	0.74	0.86	0.85	0.82	0.80	0.78	0.90	1.00	0.94	0.91	0.88	0.86	0.85	0.81	0.89	0.82
13.txt	0.85	0.80	0.74	0.73	0.70	0.81	0.80	0.77	0.75	0.73	0.85	0.94	1.00	0.97	0.93	0.91	0.90	0.86	0.83	0.77
14.txt	0.82	0.78	0.71	0.70	0.68	0.78	0.77	0.74	0.72	0.71	0.82	0.91	0.97	1.00	0.97	0.95	0.93	0.89	0.81	0.75
15.txt	0.79	0.75	0.69	0.67	0.65	0.75	0.74	0.72	0.70	0.68	0.79	0.88	0.93	0.97	1.00	0.98	0.96	0.92	0.78	0.76
16.txt	0.77	0.73	0.67	0.66	0.64	0.74	0.73	0.70	0.68	0.66	0.77	0.86	0.91	0.95	0.98	1.00	0.98	0.94	0.76	0.74
17.txt	0.76	0.72	0.66	0.65	0.63	0.73	0.71	0.69	0.67	0.65	0.76	0.85	0.90	0.93	0.96	0.98	1.00	0.96	0.75	0.73
18.txt	0.73	0.69	0.63	0.62	0.60	0.69	0.68	0.67	0.66	0.64	0.73	0.81	0.86	0.89	0.92	0.94	0.96	1.00	0.72	0.70
19.txt	0.95	0.90	0.83	0.81	0.78	0.91	0.89	0.87	0.84	0.82	0.95	0.89	0.83	0.81	0.78	0.76	0.75	0.72	1.00	0.93
20.txt	0.89	0.84	0.77	0.75	0.72	0.85	0.83	0.80	0.78	0.76	0.89	0.82	0.77	0.75	0.76	0.74	0.73	0.70	0.93	1.00

Signature Similarity

K VALUE : 5 THRESHOLD : 0.70
SIGNATURE SIMILARITIES OF ALL DOCUMENT COMBINATIONS

	1.txt	2.txt	3.txt	4.txt	5.txt	6.txt	7.txt	8.txt	9.txt	10.txt	11.txt	12.txt	13.txt	14.txt	15.txt	16.txt	17.txt	18.txt	19.txt	20.txt
1.txt	1.00	0.98	0.91	0.90	0.89	0.93	0.93	0.90	0.90	0.85	1.00	0.93	0.89	0.86	0.84	0.71	0.71	0.69	0.98	0.94
2.txt	0.98	1.00	0.93	0.92	0.91	0.91	0.91	0.88	0.88	0.83	0.98	0.91	0.87	0.84	0.82	0.69	0.69	0.67	0.96	0.92
3.txt	0.91	0.93	1.00	0.99	0.98	0.84	0.84	0.81	0.81	0.76	0.91	0.84	0.80	0.77	0.75	0.66	0.66	0.64	0.89	0.85
4.txt	0.90	0.92	0.99	1.00	0.99	0.83	0.83	0.80	0.80	0.75	0.90	0.83	0.79	0.76	0.74	0.65	0.65	0.63	0.88	0.84
5.txt	0.89	0.91	0.98	0.99	1.00	0.82	0.82	0.79	0.79	0.74	0.89	0.82	0.78	0.75	0.73	0.64	0.64	0.62	0.87	0.83
6.txt	0.93	0.91	0.84	0.83	0.82	1.00	1.00	0.97	0.97	0.92	0.93	0.86	0.82	0.79	0.77	0.67	0.67	0.65	0.91	0.87
7.txt	0.93	0.91	0.84	0.83	0.82	1.00	1.00	0.97	0.97	0.92	0.93	0.86	0.82	0.79	0.77	0.67	0.67	0.65	0.91	0.87
8.txt	0.90	0.88	0.81	0.80	0.79	0.97	0.97	1.00	1.00	0.95	0.90	0.83	0.79	0.76	0.74	0.65	0.65	0.64	0.88	0.84
9.txt	0.90	0.88	0.81	0.80	0.79	0.97	0.97	1.00	1.00	0.95	0.90	0.83	0.79	0.76	0.74	0.65	0.65	0.64	0.88	0.84
10.txt	0.85	0.83	0.76	0.75	0.74	0.92	0.92	0.95	0.95	1.00	0.85	0.78	0.74	0.71	0.69	0.63	0.63	0.62	0.83	0.79
11.txt	1.00	0.98	0.91	0.90	0.89	0.93	0.93	0.90	0.90	0.85	1.00	0.93	0.89	0.86	0.84	0.71	0.71	0.69	0.98	0.94
12.txt	0.93	0.91	0.84	0.83	0.82	0.86	0.86	0.83	0.83	0.78	0.93	1.00	0.96	0.93	0.91	0.75	0.75	0.73	0.93	0.89
13.txt	0.89	0.87	0.80	0.79	0.78	0.82	0.82	0.79	0.79	0.74	0.89	0.96	1.00	0.97	0.95	0.77	0.77	0.75	0.89	0.85
14.txt	0.86	0.84	0.77	0.76	0.75	0.79	0.79	0.76	0.76	0.71	0.86	0.93	0.97	1.00	0.98	0.79	0.79	0.77	0.86	0.82
15.txt	0.84	0.82	0.75	0.74	0.73	0.77	0.77	0.74	0.74	0.69	0.84	0.91	0.95	0.98	1.00	0.79	0.79	0.77	0.84	0.84
16.txt	0.71	0.69	0.66	0.65	0.64	0.67	0.67	0.65	0.65	0.63	0.71	0.75	0.77	0.79	0.79	1.00	0.99	0.97	0.70	0.69
17.txt	0.71	0.69	0.66	0.65	0.64	0.67	0.67	0.65	0.65	0.63	0.71	0.75	0.77	0.79	0.79	0.99	1.00	0.98	0.70	0.69
18.txt	0.69	0.67	0.64	0.63	0.62	0.65	0.65	0.64	0.64	0.62	0.69	0.73	0.75	0.77	0.77	0.97	0.98	1.00	0.68	0.67
19.txt	0.98	0.96	0.89	0.88	0.87	0.91	0.91	0.88	0.88	0.83	0.98	0.93	0.89	0.86	0.84	0.70	0.70	0.68	1.00	0.96
20.txt	0.94	0.92	0.85	0.84	0.83	0.87	0.87	0.84	0.84	0.79	0.94	0.89	0.85	0.82	0.84	0.69	0.69	0.67	0.96	1.00

K=5, Threshold=0.8

Jaccard Similarity

K VALUE : 5 THRESHOLD : 0.80
JACCARD SIMILARITIES OF ALL DOCUMENT COMBINATIONS

	1.txt	2.txt	3.txt	4.txt	5.txt	6.txt	7.txt	8.txt	9.txt	10.txt	11.txt	12.txt	13.txt	14.txt	15.txt	16.txt	17.txt	18.txt	19.txt	20.txt
1.txt	1.00	0.95	0.87	0.85	0.82	0.95	0.94	0.91	0.89	0.86	1.00	0.90	0.85	0.82	0.79	0.77	0.76	0.73	0.95	0.89
2.txt	0.95	1.00	0.92	0.89	0.87	0.90	0.89	0.86	0.84	0.82	0.95	0.85	0.80	0.78	0.75	0.73	0.72	0.69	0.90	0.84
3.txt	0.87	0.92	1.00	0.95	0.92	0.83	0.82	0.79	0.77	0.75	0.87	0.79	0.74	0.71	0.69	0.67	0.66	0.63	0.83	0.77
4.txt	0.85	0.89	0.95	1.00	0.97	0.81	0.80	0.77	0.75	0.73	0.85	0.76	0.73	0.70	0.67	0.66	0.65	0.62	0.81	0.75
5.txt	0.82	0.87	0.92	0.97	1.00	0.79	0.77	0.75	0.73	0.71	0.82	0.74	0.70	0.68	0.65	0.64	0.63	0.60	0.78	0.72
6.txt	0.95	0.90	0.83	0.81	0.79	1.00	0.98	0.95	0.93	0.90	0.95	0.86	0.81	0.78	0.75	0.74	0.73	0.69	0.91	0.85
7.txt	0.94	0.89	0.82	0.80	0.77	0.98	1.00	0.97	0.94	0.92	0.94	0.85	0.80	0.77	0.74	0.73	0.71	0.68	0.89	0.83
8.txt	0.91	0.86	0.79	0.77	0.75	0.95	0.97	1.00	0.97	0.95	0.91	0.82	0.77	0.74	0.72	0.70	0.69	0.67	0.87	0.80
9.txt	0.89	0.84	0.77	0.75	0.73	0.93	0.94	0.97	1.00	0.98	0.89	0.80	0.75	0.72	0.70	0.68	0.67	0.66	0.84	0.78
10.txt	0.86	0.82	0.75	0.73	0.71	0.90	0.92	0.95	0.98	1.00	0.86	0.78	0.73	0.71	0.68	0.66	0.65	0.64	0.82	0.76
11.txt	1.00	0.95	0.87	0.85	0.82	0.95	0.94	0.91	0.89	0.86	1.00	0.90	0.85	0.82	0.79	0.77	0.76	0.73	0.95	0.89
12.txt	0.90	0.85	0.79	0.76	0.74	0.86	0.85	0.82	0.80	0.78	0.90	1.00	0.94	0.91	0.88	0.86	0.85	0.81	0.89	0.82
13.txt	0.85	0.80	0.74	0.73	0.70	0.81	0.80	0.77	0.75	0.73	0.85	0.94	1.00	0.97	0.93	0.91	0.90	0.86	0.83	0.77
14.txt	0.82	0.78	0.71	0.70	0.68	0.78	0.77	0.74	0.72	0.71	0.82	0.91	0.97	1.00	0.97	0.95	0.93	0.89	0.81	0.75
15.txt	0.79	0.75	0.69	0.67	0.65	0.75	0.74	0.72	0.70	0.68	0.79	0.88	0.93	0.97	1.00	0.98	0.96	0.92	0.78	0.76
16.txt	0.77	0.73	0.67	0.66	0.64	0.74	0.73	0.70	0.68	0.66	0.77	0.86	0.91	0.95	0.98	1.00	0.98	0.94	0.76	0.74
17.txt	0.76	0.72	0.66	0.65	0.63	0.73	0.71	0.69	0.67	0.65	0.76	0.85	0.90	0.93	0.96	0.98	1.00	0.96	0.75	0.73
18.txt	0.73	0.69	0.63	0.62	0.60	0.69	0.68	0.67	0.66	0.64	0.73	0.81	0.86	0.89	0.92	0.94	0.96	1.00	0.72	0.70
19.txt	0.95	0.90	0.83	0.81	0.78	0.91	0.89	0.87	0.84	0.82	0.95	0.89	0.83	0.81	0.78	0.76	0.75	0.72	1.00	0.93
20.txt	0.89	0.84	0.77	0.75	0.72	0.85	0.83	0.80	0.78	0.76	0.89	0.82	0.77	0.75	0.76	0.74	0.73	0.70	0.93	1.00

Signature Similarity

K VALUE : 5 THRESHOLD : 0.80
SIGNATURE SIMILARITIES OF ALL DOCUMENT COMBINATIONS

	1.txt	2.txt	3.txt	4.txt	5.txt	6.txt	7.txt	8.txt	9.txt	10.txt	11.txt	12.txt	13.txt	14.txt	15.txt	16.txt	17.txt	18.txt	19.txt	20.txt
1.txt	1.00	0.96	0.91	0.88	0.85	0.95	0.95	0.92	0.92	0.89	1.00	0.94	0.88	0.87	0.86	0.71	0.71	0.69	0.93	0.90
2.txt	0.96	1.00	0.94	0.91	0.88	0.92	0.92	0.89	0.89	0.86	0.96	0.91	0.85	0.84	0.83	0.69	0.69	0.67	0.90	0.87
3.txt	0.91	0.94	1.00	0.96	0.93	0.87	0.87	0.84	0.84	0.81	0.91	0.86	0.80	0.79	0.78	0.66	0.66	0.64	0.85	0.82
4.txt	0.88	0.91	0.96	1.00	0.96	0.86	0.86	0.83	0.83	0.80	0.88	0.83	0.79	0.78	0.77	0.66	0.66	0.64	0.82	0.79
5.txt	0.85	0.88	0.93	0.96	1.00	0.83	0.83	0.80	0.80	0.77	0.85	0.80	0.76	0.75	0.74	0.63	0.63	0.61	0.79	0.76
6.txt	0.95	0.92	0.87	0.86	0.83	1.00	0.99	0.96	0.96	0.93	0.95	0.90	0.84	0.83	0.82	0.70	0.70	0.68	0.89	0.86
7.txt	0.95	0.92	0.87	0.86	0.83	0.99	1.00	0.96	0.96	0.93	0.95	0.90	0.84	0.83	0.82	0.70	0.70	0.68	0.89	0.86
8.txt	0.92	0.89	0.84	0.83	0.80	0.96	0.96	1.00	0.99	0.96	0.92	0.87	0.81	0.80	0.79	0.68	0.68	0.67	0.86	0.83
9.txt	0.92	0.89	0.84	0.83	0.80	0.96	0.96	0.99	1.00	0.96	0.92	0.87	0.81	0.80	0.79	0.68	0.68	0.67	0.86	0.83
10.txt	0.89	0.86	0.81	0.80	0.77	0.93	0.93	0.96	0.96	1.00	0.89	0.84	0.78	0.77	0.76	0.67	0.67	0.66	0.83	0.80
11.txt	1.00	0.96	0.91	0.88	0.85	0.95	0.95	0.92	0.92	0.89	1.00	0.94	0.88	0.87	0.86	0.71	0.71	0.69	0.93	0.90
12.txt	0.94	0.91	0.86	0.83	0.80	0.90	0.90	0.87	0.87	0.84	0.94	1.00	0.93	0.92	0.91	0.73	0.73	0.71	0.94	0.91
13.txt	0.88	0.85	0.80	0.79	0.76	0.84	0.84	0.81	0.81	0.78	0.88	0.93	1.00	0.98	0.97	0.76	0.76	0.74	0.88	0.85
14.txt	0.87	0.84	0.79	0.78	0.75	0.83	0.83	0.80	0.80	0.77	0.87	0.92	0.98	1.00	0.98	0.76	0.76	0.74	0.87	0.84
15.txt	0.86	0.83	0.78	0.77	0.74	0.82	0.82	0.79	0.79	0.76	0.86	0.91	0.97	0.98	1.00	0.76	0.76	0.74	0.86	0.85
16.txt	0.71	0.69	0.66	0.66	0.63	0.70	0.70	0.68	0.68	0.67	0.71	0.73	0.76	0.76	0.76	1.00	0.97	0.94	0.70	0.68
17.txt	0.71	0.69	0.66	0.66	0.63	0.70	0.70	0.68	0.68	0.67	0.71	0.73	0.76	0.76	0.76	0.97	1.00	0.95	0.70	0.68
18.txt	0.69	0.67	0.64	0.64	0.61	0.68	0.68	0.67	0.67	0.66	0.69	0.71	0.74	0.74	0.74	0.94	0.95	1.00	0.68	0.66
19.txt	0.93	0.90	0.85	0.82	0.79	0.89	0.89	0.86	0.86	0.83	0.93	0.94	0.88	0.87	0.86	0.70	0.70	0.68	1.00	0.96
20.txt	0.90	0.87	0.82	0.79	0.76	0.86	0.86	0.83	0.83	0.80	0.90	0.91	0.85	0.84	0.85	0.68	0.68	0.66	0.96	1.00

K=5, Threshold=0.9

Jaccard Similarity

K VALUE : 5 THRESHOLD : 0.90
JACCARD SIMILARITIES OF ALL DOCUMENT COMBINATIONS

	1.txt	2.txt	3.txt	4.txt	5.txt	6.txt	7.txt	8.txt	9.txt	10.txt	11.txt	12.txt	13.txt	14.txt	15.txt	16.txt	17.txt	18.txt	19.txt	20.txt
1.txt	1.00	0.95	0.87	0.85	0.82	0.95	0.94	0.91	0.89	0.86	1.00	0.90	0.85	0.82	0.79	0.77	0.76	0.73	0.95	0.89
2.txt	0.95	1.00	0.92	0.89	0.87	0.90	0.89	0.86	0.84	0.82	0.95	0.85	0.80	0.78	0.75	0.73	0.72	0.69	0.90	0.84
3.txt	0.87	0.92	1.00	0.95	0.92	0.83	0.82	0.79	0.77	0.75	0.87	0.79	0.74	0.71	0.69	0.67	0.66	0.63	0.83	0.77
4.txt	0.85	0.89	0.95	1.00	0.97	0.81	0.80	0.77	0.75	0.73	0.85	0.76	0.73	0.70	0.67	0.66	0.65	0.62	0.81	0.75
5.txt	0.82	0.87	0.92	0.97	1.00	0.79	0.77	0.75	0.73	0.71	0.82	0.74	0.70	0.68	0.65	0.64	0.63	0.60	0.78	0.72
6.txt	0.95	0.90	0.83	0.81	0.79	1.00	0.98	0.95	0.93	0.90	0.95	0.86	0.81	0.78	0.75	0.74	0.73	0.69	0.91	0.85
7.txt	0.94	0.89	0.82	0.80	0.77	0.98	1.00	0.97	0.94	0.92	0.94	0.85	0.80	0.77	0.74	0.73	0.71	0.68	0.89	0.83
8.txt	0.91	0.86	0.79	0.77	0.75	0.95	0.97	1.00	0.97	0.95	0.91	0.82	0.77	0.74	0.72	0.70	0.69	0.67	0.87	0.80
9.txt	0.89	0.84	0.77	0.75	0.73	0.93	0.94	0.97	1.00	0.98	0.89	0.80	0.75	0.72	0.70	0.68	0.67	0.66	0.84	0.78
10.txt	0.86	0.82	0.75	0.73	0.71	0.90	0.92	0.95	0.98	1.00	0.86	0.78	0.73	0.71	0.68	0.66	0.65	0.64	0.82	0.76
11.txt	1.00	0.95	0.87	0.85	0.82	0.95	0.94	0.91	0.89	0.86	1.00	0.90	0.85	0.82	0.79	0.77	0.76	0.73	0.95	0.89
12.txt	0.90	0.85	0.79	0.76	0.74	0.86	0.85	0.82	0.80	0.78	0.90	1.00	0.94	0.91	0.88	0.86	0.85	0.81	0.89	0.82
13.txt	0.85	0.80	0.74	0.73	0.70	0.81	0.80	0.77	0.75	0.73	0.85	0.94	1.00	0.97	0.93	0.91	0.90	0.86	0.83	0.77
14.txt	0.82	0.78	0.71	0.70	0.68	0.78	0.77	0.74	0.72	0.71	0.82	0.91	0.97	1.00	0.97	0.95	0.93	0.89	0.81	0.75
15.txt	0.79	0.75	0.69	0.67	0.65	0.75	0.74	0.72	0.70	0.68	0.79	0.88	0.93	0.97	1.00	0.98	0.96	0.92	0.78	0.76
16.txt	0.77	0.73	0.67	0.66	0.64	0.74	0.73	0.70	0.68	0.66	0.77	0.86	0.91	0.95	0.98	1.00	0.98	0.94	0.76	0.74
17.txt	0.76	0.72	0.66	0.65	0.63	0.73	0.71	0.69	0.67	0.65	0.76	0.85	0.90	0.93	0.96	0.98	1.00	0.96	0.75	0.73
18.txt	0.73	0.69	0.63	0.62	0.60	0.69	0.68	0.67	0.66	0.64	0.73	0.81	0.86	0.89	0.92	0.94	0.96	1.00	0.72	0.70
19.txt	0.95	0.90	0.83	0.81	0.78	0.91	0.89	0.87	0.84	0.82	0.95	0.89	0.83	0.81	0.78	0.76	0.75	0.72	1.00	0.93
20.txt	0.89	0.84	0.77	0.75	0.72	0.85	0.83	0.80	0.78	0.76	0.89	0.82	0.77	0.75	0.76	0.74	0.73	0.70	0.93	1.00

Signature Similarity

K VALUE : 5 THRESHOLD : 0.90
SIGNATURE SIMILARITIES OF ALL DOCUMENT COMBINATIONS

	1.txt	2.txt	3.txt	4.txt	5.txt	6.txt	7.txt	8.txt	9.txt	10.txt	11.txt	12.txt	13.txt	14.txt	15.txt	16.txt	17.txt	18.txt	19.txt	20.txt
1.txt	1.00	0.96	0.91	0.89	0.88	0.98	0.97	0.94	0.89	0.88	1.00	0.95	0.92	0.89	0.88	0.73	0.73	0.72	0.96	0.91
2.txt	0.96	1.00	0.95	0.93	0.92	0.94	0.93	0.90	0.85	0.84	0.96	0.91	0.88	0.85	0.84	0.71	0.72	0.71	0.92	0.87
3.txt	0.91	0.95	1.00	0.98	0.97	0.89	0.88	0.85	0.80	0.79	0.91	0.86	0.83	0.80	0.79	0.69	0.70	0.69	0.87	0.82
4.txt	0.89	0.93	0.98	1.00	0.99	0.87	0.86	0.83	0.78	0.77	0.89	0.84	0.83	0.80	0.79	0.69	0.70	0.69	0.85	0.80
5.txt	0.88	0.92	0.97	0.99	1.00	0.86	0.85	0.82	0.77	0.76	0.88	0.83	0.82	0.79	0.78	0.68	0.69	0.68	0.84	0.79
6.txt	0.98	0.94	0.89	0.87	0.86	1.00	0.99	0.96	0.91	0.90	0.98	0.93	0.90	0.87	0.86	0.72	0.72	0.71	0.94	0.89
7.txt	0.97	0.93	0.88	0.86	0.85	0.99	1.00	0.97	0.92	0.91	0.97	0.92	0.89	0.86	0.85	0.72	0.72	0.71	0.93	0.88
8.txt	0.94	0.90	0.85	0.83	0.82	0.96	0.97	1.00	0.95	0.94	0.94	0.89	0.86	0.83	0.82	0.70	0.70	0.70	0.90	0.85
9.txt	0.89	0.85	0.80	0.78	0.77	0.91	0.92	0.95	1.00	0.99	0.89	0.84	0.81	0.78	0.77	0.67	0.67	0.67	0.85	0.80
10.txt	0.88	0.84	0.79	0.77	0.76	0.90	0.91	0.94	0.99	1.00	0.88	0.83	0.80	0.77	0.76	0.67	0.67	0.67	0.84	0.79
11.txt	1.00	0.96	0.91	0.89	0.88	0.98	0.97	0.94	0.89	0.88	1.00	0.95	0.92	0.89	0.88	0.73	0.73	0.72	0.96	0.91
12.txt	0.95	0.91	0.86	0.84	0.83	0.93	0.92	0.89	0.84	0.83	0.95	1.00	0.97	0.94	0.93	0.77	0.77	0.76	0.93	0.88
13.txt	0.92	0.88	0.83	0.83	0.82	0.90	0.89	0.86	0.81	0.80	0.92	0.97	1.00	0.97	0.96	0.79	0.79	0.78	0.90	0.85
14.txt	0.89	0.85	0.80	0.80	0.79	0.87	0.86	0.83	0.78	0.77	0.89	0.94	0.97	1.00	0.99	0.81	0.81	0.80	0.87	0.82
15.txt	0.88	0.84	0.79	0.79	0.78	0.86	0.85	0.82	0.77	0.76	0.88	0.93	0.96	0.99	1.00	0.81	0.81	0.80	0.86	0.83
16.txt	0.73	0.71	0.69	0.69	0.68	0.72	0.72	0.70	0.67	0.67	0.73	0.77	0.79	0.81	0.81	1.00	0.99	0.98	0.72	0.69
17.txt	0.73	0.72	0.70	0.70	0.69	0.72	0.72	0.70	0.67	0.67	0.73	0.77	0.79	0.81	0.81	0.99	1.00	0.99	0.72	0.69
18.txt	0.72	0.71	0.69	0.69	0.68	0.71	0.71	0.70	0.67	0.67	0.72	0.76	0.78	0.80	0.80	0.98	0.99	1.00	0.71	0.68
19.txt	0.96	0.92	0.87	0.85	0.84	0.94	0.93	0.90	0.85	0.84	0.96	0.93	0.90	0.87	0.86	0.72	0.72	0.71	1.00	0.95
20.txt	0.91	0.87	0.82	0.80	0.79	0.89	0.88	0.85	0.80	0.79	0.91	0.88	0.85	0.82	0.83	0.69	0.69	0.68	0.95	1.00

K=8, Threshold=0.7

Jaccard Similarity

K VALUE : 8 THRESHOLD : 0.70
JACCARD SIMILARITIES OF ALL DOCUMENT COMBINATIONS

	1.txt	2.txt	3.txt	4.txt	5.txt	6.txt	7.txt	8.txt	9.txt	10.txt	11.txt	12.txt	13.txt	14.txt	15.txt	16.txt	17.txt	18.txt	19.txt	20.txt
1.txt	1.00	0.93	0.84	0.81	0.78	0.95	0.93	0.90	0.85	0.82	1.00	0.88	0.81	0.79	0.75	0.74	0.72	0.67	0.94	0.88
2.txt	0.93	1.00	0.91	0.87	0.84	0.88	0.87	0.83	0.79	0.76	0.93	0.81	0.75	0.73	0.70	0.68	0.66	0.62	0.88	0.81
3.txt	0.84	0.91	1.00	0.94	0.90	0.80	0.78	0.75	0.72	0.69	0.84	0.74	0.68	0.66	0.63	0.62	0.60	0.55	0.79	0.73
4.txt	0.81	0.87	0.94	1.00	0.96	0.77	0.75	0.72	0.69	0.66	0.81	0.71	0.66	0.63	0.60	0.59	0.57	0.53	0.76	0.70
5.txt	0.78	0.84	0.90	0.96	1.00	0.74	0.72	0.69	0.66	0.63	0.78	0.68	0.63	0.60	0.58	0.56	0.55	0.50	0.73	0.67
6.txt	0.95	0.88	0.80	0.77	0.74	1.00	0.97	0.93	0.89	0.85	0.95	0.83	0.77	0.75	0.71	0.70	0.68	0.63	0.90	0.83
7.txt	0.93	0.87	0.78	0.75	0.72	0.97	1.00	0.96	0.92	0.88	0.93	0.82	0.76	0.73	0.70	0.69	0.67	0.62	0.88	0.82
8.txt	0.90	0.83	0.75	0.72	0.69	0.93	0.96	1.00	0.95	0.92	0.90	0.79	0.73	0.70	0.67	0.66	0.64	0.62	0.85	0.78
9.txt	0.85	0.79	0.72	0.69	0.66	0.89	0.92	0.95	1.00	0.96	0.85	0.75	0.69	0.67	0.64	0.63	0.61	0.59	0.80	0.74
10.txt	0.82	0.76	0.69	0.66	0.63	0.85	0.88	0.92	0.96	1.00	0.82	0.72	0.66	0.64	0.61	0.60	0.58	0.56	0.77	0.71
11.txt	1.00	0.93	0.84	0.81	0.78	0.95	0.93	0.90	0.85	0.82	1.00	0.88	0.81	0.79	0.75	0.74	0.72	0.67	0.94	0.88
12.txt	0.88	0.81	0.74	0.71	0.68	0.83	0.82	0.79	0.75	0.72	0.88	1.00	0.93	0.90	0.86	0.84	0.82	0.77	0.86	0.80
13.txt	0.81	0.75	0.68	0.66	0.63	0.77	0.76	0.73	0.69	0.66	0.81	0.93	1.00	0.96	0.92	0.90	0.88	0.82	0.80	0.74
14.txt	0.79	0.73	0.66	0.63	0.60	0.75	0.73	0.70	0.67	0.64	0.79	0.90	0.96	1.00	0.96	0.94	0.91	0.86	0.77	0.71
15.txt	0.75	0.70	0.63	0.60	0.58	0.71	0.70	0.67	0.64	0.61	0.75	0.86	0.92	0.96	1.00	0.98	0.95	0.89	0.74	0.72
16.txt	0.74	0.68	0.62	0.59	0.56	0.70	0.69	0.66	0.63	0.60	0.74	0.84	0.90	0.94	0.98	1.00	0.97	0.91	0.72	0.71
17.txt	0.72	0.66	0.60	0.57	0.55	0.68	0.67	0.64	0.61	0.58	0.72	0.82	0.88	0.91	0.95	0.97	1.00	0.94	0.70	0.68
18.txt	0.67	0.62	0.55	0.53	0.50	0.63	0.62	0.62	0.59	0.56	0.67	0.77	0.82	0.86	0.89	0.91	0.94	1.00	0.65	0.64
19.txt	0.94	0.88	0.79	0.76	0.73	0.90	0.88	0.85	0.80	0.77	0.94	0.86	0.80	0.77	0.74	0.72	0.70	0.65	1.00	0.93
20.txt	0.88	0.81	0.73	0.70	0.67	0.83	0.82	0.78	0.74	0.71	0.88	0.80	0.74	0.71	0.72	0.71	0.68	0.64	0.93	1.00

Signature Similarity

K VALUE : 8 THRESHOLD : 0.70
SIGNATURE SIMILARITIES OF ALL DOCUMENT COMBINATIONS

	1.txt	2.txt	3.txt	4.txt	5.txt	6.txt	7.txt	8.txt	9.txt	10.txt	11.txt	12.txt	13.txt	14.txt	15.txt	16.txt	17.txt	18.txt	19.txt	20.txt
1.txt	1.00	0.98	0.95	0.93	0.93	0.99	0.99	0.99	0.97	0.95	1.00	0.95	0.92	0.91	0.89	0.58	0.58	0.58	0.97	0.94
2.txt	0.98	1.00	0.97	0.95	0.95	0.97	0.97	0.97	0.95	0.93	0.98	0.93	0.90	0.89	0.87	0.57	0.57	0.57	0.95	0.92
3.txt	0.95	0.97	1.00	0.96	0.96	0.94	0.94	0.94	0.92	0.90	0.95	0.90	0.87	0.86	0.84	0.54	0.54	0.54	0.92	0.89
4.txt	0.93	0.95	0.96	1.00	1.00	0.92	0.92	0.92	0.90	0.88	0.93	0.88	0.85	0.84	0.82	0.54	0.54	0.54	0.90	0.87
5.txt	0.93	0.95	0.96	1.00	1.00	0.92	0.92	0.92	0.90	0.88	0.93	0.88	0.85	0.84	0.82	0.54	0.54	0.54	0.90	0.87
6.txt	0.99	0.97	0.94	0.92	0.92	1.00	1.00	1.00	0.98	0.96	0.99	0.94	0.91	0.90	0.88	0.57	0.57	0.57	0.96	0.93
7.txt	0.99	0.97	0.94	0.92	0.92	1.00	1.00	1.00	0.98	0.96	0.99	0.94	0.91	0.90	0.88	0.57	0.57	0.57	0.96	0.93
8.txt	0.99	0.97	0.94	0.92	0.92	1.00	1.00	1.00	0.98	0.96	0.99	0.94	0.91	0.90	0.88	0.57	0.57	0.57	0.96	0.93
9.txt	0.97	0.95	0.92	0.90	0.90	0.98	0.98	0.98	1.00	0.98	0.97	0.92	0.89	0.88	0.86	0.55	0.55	0.55	0.94	0.91
10.txt	0.95	0.93	0.90	0.88	0.88	0.96	0.96	0.96	0.98	1.00	0.95	0.90	0.87	0.86	0.84	0.54	0.54	0.54	0.92	0.89
11.txt	1.00	0.98	0.95	0.93	0.93	0.99	0.99	0.99	0.97	0.95	1.00	0.95	0.92	0.91	0.89	0.58	0.58	0.58	0.97	0.94
12.txt	0.95	0.93	0.90	0.88	0.88	0.94	0.94	0.94	0.92	0.90	0.95	1.00	0.97	0.96	0.94	0.61	0.61	0.61	0.94	0.91
13.txt	0.92	0.90	0.87	0.85	0.85	0.91	0.91	0.91	0.89	0.87	0.92	0.97	1.00	0.99	0.97	0.62	0.62	0.62	0.91	0.88
14.txt	0.91	0.89	0.86	0.84	0.84	0.90	0.90	0.90	0.88	0.86	0.91	0.96	0.99	1.00	0.98	0.63	0.63	0.63	0.90	0.87
15.txt	0.89	0.87	0.84	0.82	0.82	0.88	0.88	0.88	0.86	0.84	0.89	0.94	0.97	0.98	1.00	0.64	0.64	0.64	0.88	0.87
16.txt	0.58	0.57	0.54	0.54	0.54	0.57	0.57	0.57	0.55	0.54	0.58	0.61	0.62	0.63	0.64	1.00	0.99	0.97	0.57	0.56
17.txt	0.58	0.57	0.54	0.54	0.54	0.57	0.57	0.57	0.55	0.54	0.58	0.61	0.62	0.63	0.64	0.99	1.00	0.98	0.57	0.56
18.txt	0.58	0.57	0.54	0.54	0.54	0.57	0.57	0.57	0.55	0.54	0.58	0.61	0.62	0.63	0.64	0.97	0.98	1.00	0.57	0.56
19.txt	0.97	0.95	0.92	0.90	0.90	0.96	0.96	0.96	0.94	0.92	0.97	0.94	0.91	0.90	0.88	0.57	0.57	0.57	1.00	0.97
20.txt	0.94	0.92	0.89	0.87	0.87	0.93	0.93	0.93	0.91	0.89	0.94	0.91	0.88	0.87	0.87	0.56	0.56	0.56	0.97	1.00

K=8, Threshold=0.8

Jaccard Similarity

K VALUE : 8 THRESHOLD : 0.80
JACCARD SIMILARITIES OF ALL DOCUMENT COMBINATIONS

	1.txt	2.txt	3.txt	4.txt	5.txt	6.txt	7.txt	8.txt	9.txt	10.txt	11.txt	12.txt	13.txt	14.txt	15.txt	16.txt	17.txt	18.txt	19.txt	20.txt
1.txt	1.00	0.93	0.84	0.81	0.78	0.95	0.93	0.90	0.85	0.82	1.00	0.88	0.81	0.79	0.75	0.74	0.72	0.67	0.94	0.88
2.txt	0.93	1.00	0.91	0.87	0.84	0.88	0.87	0.83	0.79	0.76	0.93	0.81	0.75	0.73	0.70	0.68	0.66	0.62	0.88	0.81
3.txt	0.84	0.91	1.00	0.94	0.90	0.80	0.78	0.75	0.72	0.69	0.84	0.74	0.68	0.66	0.63	0.62	0.60	0.55	0.79	0.73
4.txt	0.81	0.87	0.94	1.00	0.96	0.77	0.75	0.72	0.69	0.66	0.81	0.71	0.66	0.63	0.60	0.59	0.57	0.53	0.76	0.70
5.txt	0.78	0.84	0.90	0.96	1.00	0.74	0.72	0.69	0.66	0.63	0.78	0.68	0.63	0.60	0.58	0.56	0.55	0.50	0.73	0.67
6.txt	0.95	0.88	0.80	0.77	0.74	1.00	0.97	0.93	0.89	0.85	0.95	0.83	0.77	0.75	0.71	0.70	0.68	0.63	0.90	0.83
7.txt	0.93	0.87	0.78	0.75	0.72	0.97	1.00	0.96	0.92	0.88	0.93	0.82	0.76	0.73	0.70	0.69	0.67	0.62	0.88	0.82
8.txt	0.90	0.83	0.75	0.72	0.69	0.93	0.96	1.00	0.95	0.92	0.90	0.79	0.73	0.70	0.67	0.66	0.64	0.62	0.85	0.78
9.txt	0.85	0.79	0.72	0.69	0.66	0.89	0.92	0.95	1.00	0.96	0.85	0.75	0.69	0.67	0.64	0.63	0.61	0.59	0.80	0.74
10.txt	0.82	0.76	0.69	0.66	0.63	0.85	0.88	0.92	0.96	1.00	0.82	0.72	0.66	0.64	0.61	0.60	0.58	0.56	0.77	0.71
11.txt	1.00	0.93	0.84	0.81	0.78	0.95	0.93	0.90	0.85	0.82	1.00	0.88	0.81	0.79	0.75	0.74	0.72	0.67	0.94	0.88
12.txt	0.88	0.81	0.74	0.71	0.68	0.83	0.82	0.79	0.75	0.72	0.88	1.00	0.93	0.90	0.86	0.84	0.82	0.77	0.86	0.80
13.txt	0.81	0.75	0.68	0.66	0.63	0.77	0.76	0.73	0.69	0.66	0.81	0.93	1.00	0.96	0.92	0.90	0.88	0.82	0.80	0.74
14.txt	0.79	0.73	0.66	0.63	0.60	0.75	0.73	0.70	0.67	0.64	0.79	0.90	0.96	1.00	0.96	0.94	0.91	0.86	0.77	0.71
15.txt	0.75	0.70	0.63	0.60	0.58	0.71	0.70	0.67	0.64	0.61	0.75	0.86	0.92	0.96	1.00	0.98	0.95	0.89	0.74	0.72
16.txt	0.74	0.68	0.62	0.59	0.56	0.70	0.69	0.66	0.63	0.60	0.74	0.84	0.90	0.94	0.98	1.00	0.97	0.91	0.72	0.71
17.txt	0.72	0.66	0.60	0.57	0.55	0.68	0.67	0.64	0.61	0.58	0.72	0.82	0.88	0.91	0.95	0.97	1.00	0.94	0.70	0.68
18.txt	0.67	0.62	0.55	0.53	0.50	0.63	0.62	0.62	0.59	0.56	0.67	0.77	0.82	0.86	0.89	0.91	0.94	1.00	0.65	0.64
19.txt	0.94	0.88	0.79	0.76	0.73	0.90	0.88	0.85	0.80	0.77	0.94	0.86	0.80	0.77	0.74	0.72	0.70	0.65	1.00	0.93
20.txt	0.88	0.81	0.73	0.70	0.67	0.83	0.82	0.78	0.74	0.71	0.88	0.80	0.74	0.71	0.72	0.71	0.68	0.64	0.93	1.00

Signature Similarity

K VALUE : 8 THRESHOLD : 0.80
SIGNATURE SIMILARITIES OF ALL DOCUMENT COMBINATIONS

	1.txt	2.txt	3.txt	4.txt	5.txt	6.txt	7.txt	8.txt	9.txt	10.txt	11.txt	12.txt	13.txt	14.txt	15.txt	16.txt	17.txt	18.txt	19.txt	20.txt
1.txt	1.00	0.98	0.97	0.96	0.96	0.96	0.96	0.96	0.94	0.92	1.00	0.95	0.93	0.92	0.90	0.59	0.58	0.56	0.99	0.98
2.txt	0.98	1.00	0.99	0.98	0.98	0.94	0.94	0.94	0.92	0.90	0.98	0.93	0.91	0.90	0.88	0.57	0.56	0.54	0.97	0.96
3.txt	0.97	0.99	1.00	0.97	0.97	0.93	0.93	0.93	0.91	0.89	0.97	0.92	0.90	0.89	0.87	0.56	0.55	0.53	0.96	0.95
4.txt	0.96	0.98	0.97	1.00	1.00	0.92	0.92	0.92	0.90	0.88	0.96	0.91	0.89	0.88	0.86	0.57	0.56	0.54	0.95	0.94
5.txt	0.96	0.98	0.97	1.00	1.00	0.92	0.92	0.92	0.90	0.88	0.96	0.91	0.89	0.88	0.86	0.57	0.56	0.54	0.95	0.94
6.txt	0.96	0.94	0.93	0.92	0.92	1.00	0.98	0.98	0.96	0.94	0.96	0.91	0.89	0.88	0.86	0.58	0.57	0.55	0.95	0.94
7.txt	0.96	0.94	0.93	0.92	0.92	0.98	1.00	1.00	0.98	0.96	0.96	0.91	0.89	0.88	0.86	0.57	0.56	0.54	0.95	0.94
8.txt	0.96	0.94	0.93	0.92	0.92	0.98	1.00	1.00	0.98	0.96	0.96	0.91	0.89	0.88	0.86	0.57	0.56	0.54	0.95	0.94
9.txt	0.94	0.92	0.91	0.90	0.90	0.96	0.98	0.98	1.00	0.98	0.94	0.89	0.87	0.86	0.84	0.55	0.54	0.52	0.93	0.92
10.txt	0.92	0.90	0.89	0.88	0.88	0.94	0.96	0.96	0.98	1.00	0.92	0.87	0.85	0.84	0.82	0.55	0.54	0.52	0.91	0.90
11.txt	1.00	0.98	0.97	0.96	0.96	0.96	0.96	0.96	0.94	0.92	1.00	0.95	0.93	0.92	0.90	0.59	0.58	0.56	0.99	0.98
12.txt	0.95	0.93	0.92	0.91	0.91	0.91	0.91	0.91	0.89	0.87	0.95	1.00	0.98	0.97	0.95	0.61	0.60	0.58	0.94	0.93
13.txt	0.93	0.91	0.90	0.89	0.89	0.89	0.89	0.89	0.87	0.85	0.93	0.98	1.00	0.99	0.97	0.63	0.62	0.60	0.92	0.91
14.txt	0.92	0.90	0.89	0.88	0.88	0.88	0.88	0.88	0.86	0.84	0.92	0.97	0.99	1.00	0.98	0.64	0.63	0.61	0.91	0.90
15.txt	0.90	0.88	0.87	0.86	0.86	0.86	0.86	0.86	0.84	0.82	0.90	0.95	0.97	0.98	1.00	0.66	0.65	0.63	0.89	0.88
16.txt	0.59	0.57	0.56	0.57	0.57	0.58	0.57	0.57	0.55	0.55	0.59	0.61	0.63	0.64	0.66	1.00	0.99	0.94	0.59	0.59
17.txt	0.58	0.56	0.55	0.56	0.56	0.57	0.56	0.56	0.54	0.54	0.58	0.60	0.62	0.63	0.65	0.99	1.00	0.94	0.58	0.58
18.txt	0.56	0.54	0.53	0.54	0.54	0.55	0.54	0.54	0.52	0.52	0.56	0.58	0.60	0.61	0.63	0.94	0.94	1.00	0.56	0.56
19.txt	0.99	0.97	0.96	0.95	0.95	0.95	0.95	0.95	0.93	0.91	0.99	0.94	0.92	0.91	0.89	0.59	0.58	0.56	1.00	0.99
20.txt	0.98	0.96	0.95	0.94	0.94	0.94	0.94	0.94	0.92	0.90	0.98	0.93	0.91	0.90	0.88	0.59	0.58	0.56	0.99	1.00

K=8, Threshold=0.9

Jaccard Similarity

K VALUE : 8 THRESHOLD : 0.90
JACCARD SIMILARITIES OF ALL DOCUMENT COMBINATIONS

	1.txt	2.txt	3.txt	4.txt	5.txt	6.txt	7.txt	8.txt	9.txt	10.txt	11.txt	12.txt	13.txt	14.txt	15.txt	16.txt	17.txt	18.txt	19.txt	20.txt
1.txt	1.00	0.93	0.84	0.81	0.78	0.95	0.93	0.90	0.85	0.82	1.00	0.88	0.81	0.79	0.75	0.74	0.72	0.67	0.94	0.88
2.txt	0.93	1.00	0.91	0.87	0.84	0.88	0.87	0.83	0.79	0.76	0.93	0.81	0.75	0.73	0.70	0.68	0.66	0.62	0.88	0.81
3.txt	0.84	0.91	1.00	0.94	0.90	0.80	0.78	0.75	0.72	0.69	0.84	0.74	0.68	0.66	0.63	0.62	0.60	0.55	0.79	0.73
4.txt	0.81	0.87	0.94	1.00	0.96	0.77	0.75	0.72	0.69	0.66	0.81	0.71	0.66	0.63	0.60	0.59	0.57	0.53	0.76	0.70
5.txt	0.78	0.84	0.90	0.96	1.00	0.74	0.72	0.69	0.66	0.63	0.78	0.68	0.63	0.60	0.58	0.56	0.55	0.50	0.73	0.67
6.txt	0.95	0.88	0.80	0.77	0.74	1.00	0.97	0.93	0.89	0.85	0.95	0.83	0.77	0.75	0.71	0.70	0.68	0.63	0.90	0.83
7.txt	0.93	0.87	0.78	0.75	0.72	0.97	1.00	0.96	0.92	0.88	0.93	0.82	0.76	0.73	0.70	0.69	0.67	0.62	0.88	0.82
8.txt	0.90	0.83	0.75	0.72	0.69	0.93	0.96	1.00	0.95	0.92	0.90	0.79	0.73	0.70	0.67	0.66	0.64	0.62	0.85	0.78
9.txt	0.85	0.79	0.72	0.69	0.66	0.89	0.92	0.95	1.00	0.96	0.85	0.75	0.69	0.67	0.64	0.63	0.61	0.59	0.80	0.74
10.txt	0.82	0.76	0.69	0.66	0.63	0.85	0.88	0.92	0.96	1.00	0.82	0.72	0.66	0.64	0.61	0.60	0.58	0.56	0.77	0.71
11.txt	1.00	0.93	0.84	0.81	0.78	0.95	0.93	0.90	0.85	0.82	1.00	0.88	0.81	0.79	0.75	0.74	0.72	0.67	0.94	0.88
12.txt	0.88	0.81	0.74	0.71	0.68	0.83	0.82	0.79	0.75	0.72	0.88	1.00	0.93	0.90	0.86	0.84	0.82	0.77	0.86	0.80
13.txt	0.81	0.75	0.68	0.66	0.63	0.77	0.76	0.73	0.69	0.66	0.81	0.93	1.00	0.96	0.92	0.90	0.88	0.82	0.80	0.74
14.txt	0.79	0.73	0.66	0.63	0.60	0.75	0.73	0.70	0.67	0.64	0.79	0.90	0.96	1.00	0.96	0.94	0.91	0.86	0.77	0.71
15.txt	0.75	0.70	0.63	0.60	0.58	0.71	0.70	0.67	0.64	0.61	0.75	0.86	0.92	0.96	1.00	0.98	0.95	0.89	0.74	0.72
16.txt	0.74	0.68	0.62	0.59	0.56	0.70	0.69	0.66	0.63	0.60	0.74	0.84	0.90	0.94	0.98	1.00	0.97	0.91	0.72	0.71
17.txt	0.72	0.66	0.60	0.57	0.55	0.68	0.67	0.64	0.61	0.58	0.72	0.82	0.88	0.91	0.95	0.97	1.00	0.94	0.70	0.68
18.txt	0.67	0.62	0.55	0.53	0.50	0.63	0.62	0.62	0.59	0.56	0.67	0.77	0.82	0.86	0.89	0.91	0.94	1.00	0.65	0.64
19.txt	0.94	0.88	0.79	0.76	0.73	0.90	0.88	0.85	0.80	0.77	0.94	0.86	0.80	0.77	0.74	0.72	0.70	0.65	1.00	0.93
20.txt	0.88	0.81	0.73	0.70	0.67	0.83	0.82	0.78	0.74	0.71	0.88	0.80	0.74	0.71	0.72	0.71	0.68	0.64	0.93	1.00

Signature Similarity

K VALUE : 8 THRESHOLD : 0.90
SIGNATURE SIMILARITIES OF ALL DOCUMENT COMBINATIONS

	1.txt	2.txt	3.txt	4.txt	5.txt	6.txt	7.txt	8.txt	9.txt	10.txt	11.txt	12.txt	13.txt	14.txt	15.txt	16.txt	17.txt	18.txt	19.txt	20.txt
1.txt	1.00	0.95	0.91	0.89	0.87	0.99	0.99	0.99	0.98	0.98	1.00	0.96	0.94	0.94	0.92	0.52	0.52	0.50	1.00	0.95
2.txt	0.95	1.00	0.96	0.94	0.92	0.94	0.94	0.94	0.93	0.93	0.95	0.91	0.89	0.89	0.87	0.48	0.48	0.46	0.95	0.90
3.txt	0.91	0.96	1.00	0.96	0.94	0.90	0.90	0.90	0.89	0.89	0.91	0.87	0.85	0.85	0.83	0.47	0.47	0.45	0.91	0.86
4.txt	0.89	0.94	0.96	1.00	0.98	0.88	0.88	0.88	0.87	0.87	0.89	0.85	0.83	0.83	0.81	0.46	0.46	0.44	0.89	0.84
5.txt	0.87	0.92	0.94	0.98	1.00	0.86	0.86	0.86	0.85	0.85	0.87	0.83	0.81	0.81	0.79	0.44	0.44	0.42	0.87	0.82
6.txt	0.99	0.94	0.90	0.88	0.86	1.00	1.00	1.00	0.99	0.99	0.99	0.95	0.93	0.93	0.91	0.51	0.51	0.49	0.99	0.94
7.txt	0.99	0.94	0.90	0.88	0.86	1.00	1.00	1.00	0.99	0.99	0.99	0.95	0.93	0.93	0.91	0.51	0.51	0.49	0.99	0.94
8.txt	0.99	0.94	0.90	0.88	0.86	1.00	1.00	1.00	0.99	0.99	0.99	0.95	0.93	0.93	0.91	0.51	0.51	0.49	0.99	0.94
9.txt	0.98	0.93	0.89	0.87	0.85	0.99	0.99	0.99	1.00	1.00	0.98	0.94	0.92	0.92	0.90	0.50	0.50	0.48	0.98	0.93
10.txt	0.98	0.93	0.89	0.87	0.85	0.99	0.99	0.99	1.00	1.00	0.98	0.94	0.92	0.92	0.90	0.50	0.50	0.48	0.98	0.93
11.txt	1.00	0.95	0.91	0.89	0.87	0.99	0.99	0.99	0.98	0.98	1.00	0.96	0.94	0.94	0.92	0.52	0.52	0.50	1.00	0.95
12.txt	0.96	0.91	0.87	0.85	0.83	0.95	0.95	0.95	0.94	0.94	0.96	1.00	0.98	0.98	0.96	0.54	0.54	0.52	0.96	0.91
13.txt	0.94	0.89	0.85	0.83	0.81	0.93	0.93	0.93	0.92	0.92	0.94	0.98	1.00	1.00	0.98	0.55	0.55	0.53	0.94	0.89
14.txt	0.94	0.89	0.85	0.83	0.81	0.93	0.93	0.93	0.92	0.92	0.94	0.98	1.00	1.00	0.98	0.55	0.55	0.53	0.94	0.89
15.txt	0.92	0.87	0.83	0.81	0.79	0.91	0.91	0.91	0.90	0.90	0.92	0.96	0.98	0.98	1.00	0.55	0.55	0.53	0.92	0.91
16.txt	0.52	0.48	0.47	0.46	0.44	0.51	0.51	0.51	0.50	0.50	0.52	0.54	0.55	0.55	0.55	1.00	0.98	0.93	0.52	0.50
17.txt	0.52	0.48	0.47	0.46	0.44	0.51	0.51	0.51	0.50	0.50	0.52	0.54	0.55	0.55	0.55	0.98	1.00	0.95	0.52	0.50
18.txt	0.50	0.46	0.45	0.44	0.42	0.49	0.49	0.49	0.48	0.48	0.50	0.52	0.53	0.53	0.53	0.93	0.95	1.00	0.50	0.48
19.txt	1.00	0.95	0.91	0.89	0.87	0.99	0.99	0.99	0.98	0.98	1.00	0.96	0.94	0.94	0.92	0.52	0.52	0.50	1.00	0.95
20.txt	0.95	0.90	0.86	0.84	0.82	0.94	0.94	0.94	0.93	0.93	0.95	0.91	0.89	0.89	0.91	0.50	0.50	0.48	0.95	1.00

Sonuç

a- K değerlerinin 4,5 ve 8 olarak değişimine göre alınan sonuçlar Jaccard Benzerliğine göre elde edilen tablolar değerlendirilecek olunursa, k değeri arttıkça benzerlik oranlarında düşüş gözlenmektedir. Bunun sebebi k sayısının küçük olmasından kaynaklanan elde edilen toplam shingle sayısının çok sayıda tekrar içermesi durumudur. Toplam shingle sayısı küçükten büyüğe k nın 4,5,8 değerleri için gözlenmiştir. Bu kıyaslanacak nokta bakımından daha fazla bilgi kullanılması anlamına gelmektedir. Bu yüzden bununla orantılı olarak benzerlik oranları k değeri arttıkça düşmektedir.

b- Örnek veri setinde verilen dosyalara bakıldığında çok benzer oldukları görülmektedir. Derste söylendiği gibi küçük boyutlu dosyalarda k=4,5 değerleri daha iyi sonuç vermektedir. Büyük dosyalarda ise k=8,9 değerleri daha iyi sonuç vermektedir. Bu örnek veri seti için dosyaların küçük olması ve bulduğu benzerlik oranları bakımından k=5 bence daha iyi sonuç vermiştir.

c- K shingle benzerliği hesaplama bakımından similarity benzerliğinden daha maliyetlidir. Bu maliyet farkı zaman ve başarı oranı tutarlılığı açısından daha büyük veri setlerinde açıkça gözlenmektedir. Bu uygulamada dosya boyutları oldukça küçük ve benzer dosyalar oldukları (shingle sayısı az) ve 100 adet hash fonksiyonu ile imza matrisi oluşturulduğu için imza benzerliği oranları K shingle benzerliği oranlarından genellikle daha yüksek olmaktadır. Verilen threshold değerini geçen ikili sayılarına bakıldığında da bu açıkça gözlenmektedir. Bu uygulamada dosyalar birbirlerine çok benzer, küçük boyutlu ve az sayıda olmaları itibariyle en gerçekçi sonuçlar k-shingle benzerliğine göre hesapladığımız Jaccard Similarity daha gerçekçi sonuçlar vermiştir. Ancak iki algoritma arasındaki fark artan dosya boyut ve sayılarında daha belirgin olacağı için Signature Similarity

gerek K-Shingle benzerliği oranına yakınlığı gerekse çalışma zamanı olarak tercih edilmesi en mantıklı algoritmadır.

Karmaşıklık Analizi

Dosya sayısı =N
Hash Fonksiyonu Sayısı =M
Shingle Sayısı = K
Olarak alındığında K-Shingle benzerliğinin bulunması N*(N-1)/2 olacaktır. Bu 1 milyon adet dosya için yaklaşık olarak $5*10^{11}$ kıyaslama yapılacağı anlamına gelmektedir ve 5 yıl sürmektedir. Ayrıca hesaplama bakımından Shingle sayısı kadar kıyaslama yapılmaktadır. Shingle sayısından daha küçük bir M sayısının seçildiği durumlar için hesaplanan Signature similarity de algoritma daha da hızlanacaktır.

Kaynak Kod

```
1.  /**
2.  @file
3.
4.   *Bu program, input olarak verilen txt dosyalarindan shinglelar olusturularak,
5.   *dosyalarin shingle sayisi bilgilerini bir tabloda yazdirir.
6.   *jaccard yontemiyle ve minhashing sonucu olusan signature similarity yontemiyle
7.   *dokumanlarin benzerlik oranlarini tablo halinde ekrana yazdirir.
8.   *Benzerlik tablolarinda ise verilen bir threshold degeri asan benzerlikleri isaretleyerek(sar
    i renkle yazdirir)
9.   *tum bunlara ek olarak benzer dosya ikililerini bir tabloda yazdirir.
10.  * Programda 1 ile devam edilir 0 ile cikis yapilmaktadir.
11.  * Oncelikle dosya sayisi ve isimleri girilmelidir.
12.  * Daha sonrasinda ise harf bazindaki shingle lar icin k degeri kullanicidan alinmaktadir.
13.  * Son olarak ise kullanicidan threshold degeri 0.0 ile 1.0 arasinda bir deger olarak alinmakt
    adir.
14.  * Tum inputlar girildikten sonra program aciklamasinda verilen tablolar ekrana yazdirilmaktad
    ir.
15.  * Tavsiye olarak tablolarin ekrana sigmasi ve anlasilir gozukmeleri icin konsol ekranindaki f
    ont boyutu 14 olarak
16.  * ayarlanmali ve konsol ekrani tam ekran olarak kullanilmalidir.
17.
18.
19. @author
20.
21. Name             :      Muhammed Yasin SAGLAM
22. Student No       :      15011804
23. Date             :      31/12/2017
24. E-Mail           :      myasinsaglam1907@gmail.com
25. Compiler Used    :      GCC
26. IDE              :      DEV-C++(Version 5.11)
27. Operating System :      Windows 10 educational edition
28. */
```

```
29. #include <stdio.h>
30. #include <stdlib.h>
31. #include <string.h>
32. #include <ctype.h>
33. #include <time.h>
34. #define NAME_LEN 30
35. #define HASH_SIZE 100
36. 
37. typedef struct{
38.     char **shingles; //kiyas yapabilmek icin tum shingle lar karakter olarak saklaniyor
39.     int k_val; //Shingle larin k degeri tutuluyor
40.     int **matrix; //shingle ve file matrisi tutuluyor
41.     float **jaccard;//jaccard similarity oranlari tutuluyor
42.     float **minhash;//minhash signature similarity oranlari tutuluyor
43.     int file_num; //kac adet dosya bulundugu bilgisini tutar --matrix sutun sayisi
44.     int shingle_count; //kac adet shingle icerdigi bilgisi tutuluyor--matrix satir sayisi
45.     float threshold;//threshold degeri
46. }SHINGLE;
47. 
48. /**
49.     @param shgl         Shingle struct pointer
50. */
51. void initialize(SHINGLE *shgl){
52.     shgl->shingle_count=0; //toplsm shingle sayisi sifirlaniyor
53.     //Matrix ve shinglelari tutacak diziler reallocate edilebilmeleri icin ilk allocationlari
    yapiliyor.
54.     shgl->matrix=(int**)malloc(sizeof(int*));
55.     shgl->matrix[0]=(int*)calloc((size_t)shgl->file_num,sizeof(int));
56.     shgl->shingles=(char**)malloc(sizeof(char*));
57.     shgl->shingles[0]=(char*)malloc(sizeof(char)*shgl->k_val);
58. }
59. 
60. /**
61.  * Dosyalari okuyarak shinglelarini cikartan ve her bir dosyanin
62.  * icerisinde bulunan shingle lari struct icerisindeki matrix adli
63.  * matriste saklayan fonksiyon
64.     @param fp           shingle olusturulacak dosyanin pointeri
65.     @param file_id      id of file
66.     @param shgl        Shingle struct pointer
67. */
68. void createShingles(FILE *fp,int file_id,SHINGLE *shgl){
69.     int i,c,j;
70.     int s_index=shgl->shingle_count;
71.     //int total_shingle=0;
72.     char *temp_shingle=(char*)calloc(shgl->k_val,sizeof(char));
73. 
74.     int next_start=0;
75.     int control=1;
76. 
77.     next_start=ftell(fp);
78.     do{
79.         fseek(fp,next_start-1,SEEK_SET);
80.         i=0;
81.         while(control && i<shgl->k_val){
82.             c=fgetc(fp); //1 karakter okunuyor
83.             if((char)c==EOF){ //dosya sonu ise cikis
84.                 control=0;
85.             }
86.             else{ //degilse
87.                 if(isalpha(c)){ //alfabetik karakterse
88.                     temp_shingle[i]=(char)c; //tempe yaz
```

```
89.                     if(i==1) //eger tempin 1. elemanina yaziyosa 1 sonraki oradan baslayacak
90.                         next_start=ftell(fp); //dosyadaki yerini tut
91.                     i++; //indisi 1 artir
92.                 }
93.                 if((char)c==' '){ //eger bosluk gelmisse
94.                     if(i>=1 && temp_shingle[i-
    1]!=' '){ //ilk elemandan sonraki bir eleman icin bosluksa bir onceki bosluk degilse bosluk ya
    z
95.                         temp_shingle[i]=(char)c;
96.                         if(i==1)
97.                             next_start=ftell(fp);
98.                         i++;
99.                     }
100.                        else if(i==0){ //ilk eleman bosluksa direk yaz
101.                            temp_shingle[i]=(char)c;
102.                            i++;
103.                        }
104.                    }
105.                }
106.            }
107.            if(control){//dosya sonundan dolayi cikmamissa tamamen k kadarlik bir shingle a
    linmis demektir
108.                j=0;
109.                strlwr(temp_shingle); //shingle i kucuk harfe cevir
110.                while(j<shgl->shingle_count && strcmp(shgl-
    >shingles[j],temp_shingle)!=0){ //eger shingle eklenmisse kc nolu indiste
111.                    j++;
112.                }
113.                if(j==shgl->shingle_count){//yeni eklenecek demektir
114.                    shgl->matrix=(int**)realloc(shgl-
    >matrix, sizeof(int*)*(s_index+1)); //matrisi genislet
115.                    shgl->matrix[s_index]=(int*)calloc((size_t)shgl-
    >file_num,sizeof(int));
116.                    shgl->matrix[s_index][file_id]=1; //ilgili degeri 1 yap
117.                    shgl->shingles=(char**)realloc(shgl-
    >shingles, sizeof(char*)*(s_index+1)); //shigle listi genislet
118.                    shgl->shingles[s_index]=(char*)malloc(sizeof(char)*(shgl->k_val));
119.                    strcpy(shgl->shingles[s_index],temp_shingle); //shingle i yaz
120.                    //printf("\nshingle %d : %s ------ %s",s_index+1,temp_shingle,shgl-
    >shingles[s_index]);
121.                    shgl->shingle_count++;
122.                    s_index++;
123.                    temp_shingle=(char*)calloc(shgl->k_val,sizeof(char));
124.                    // total_shingle++; // dosyadaki toplam shingle sayisi
125.                }
126.                else{//daha onceden varsa matrisin ilgili gozunu 1 yap
127.                    shgl->matrix[j][file_id]=1;
128.                    temp_shingle=(char*)calloc(shgl->k_val,sizeof(char));
129.                }
130.            }
131.            //kelime alindi kontrol edilecek
132.        }while (control);
133.        //system("PAUSE");
134.        //printf("\nTotal shingle of file is : %d ",total_shingle);
135.    }
136.
137.    /**
138.     * Her bir dosya icin tekrarsiz shingle sayisini ve toplam shingle sayisini
139.     * tablo olarak ekrana yazan fonksiyon
140.        @param shgl        Shingle struct pointer
141.        @param files      File names
```

```
142.        */
143.        void print_shingle(SHINGLE *shgl,char **files){
144.            system("CLS");
145.            int i,j;
146.            int *shgl_per_file=(int*)calloc((size_t)shgl->file_num,sizeof(int));
147.            for(i=0;i<shgl->shingle_count;i++){
148.                //printf("\n%-5s -->",shgl->shingles[i]);
149.                for(j=0;j<shgl->file_num;j++){
150.                    //printf(" %d ",shgl->matrix[i][j]);
151.                    if(shgl->matrix[i][j]==1)
152.                        shgl_per_file[j]++;
153.                }
154.            }
155.            for(i=0;i<shgl->file_num;i++){
156.                printf("\nFile : %-5s --> Shingle Count: %d ",files[i],shgl_per_file[i]);
157.            }
158.            printf("\nTotal shingle of all files for k=%d : %d",shgl->k_val,shgl-
    >shingle_count);
159.            free(shgl_per_file);
160.        }
161.
162.        /**
163.         * Struct icindeki jaccard matrisinde bulunan dosyalarin benzerligini ve verilen thresh
    olda gore,
164.         * benzer dosya ikililerini ekrana tablo olarak yazan fonksiyon
165.            @param shgl        Shingle struct pointer
166.            @param files      File names
167.        */
168.        void printJaccard(SHINGLE *shgl,char **files){
169.            int i,j;
170.            //tablo ekrana yazdiriliyor...
171.            printf("\t");
172.            for(i=0;i<shgl->file_num;i++){
173.                printf("%s\t",files[i]);
174.            }
175.            printf("\n");
176.            for(i=0;i<shgl->file_num;i++){
177.                printf("%s\t",files[i]);
178.                for(j=0;j<shgl->file_num;j++){
179.                    if(shgl->jaccard[i][j]>shgl-
    >threshold){ //thresholdu gecen satirlar sari yazdiriliyor..
180.                        printf("\033[01;33m");
181.                        printf("%.2f\t",shgl->jaccard[i][j]);
182.                        printf("\033[0m");
183.                    }
184.                    else if(i==j){ //diyagonal kirmizi yazdiriliyor..
185.                        shgl->jaccard[i][j]=1;
186.                        printf("\033[1;31m");
187.                        printf("%.2f\t",shgl->jaccard[i][j]);
188.                        printf("\033[0m");
189.                    }
190.                    else{
191.                        printf("%.2f\t",shgl->jaccard[i][j]);
192.                    }
193.                }
194.                printf("\n");
195.            }
196.            //benzer ikililer yazdiriliyor
197.            printf("\nSIMILIAR DOCUMENT PAIRS ACCORDING TO JACCARD SIMILARITY");
198.            printf("\n");
199.            for(i=0;i<shgl->file_num;i++) {
```

```
200.                printf("\n%s-->>\t", files[i]);
201.                for (j = 0; j < shgl->file_num; j++) {
202.                    if (shgl->jaccard[i][j] > shgl-
    >threshold && j!=i) { //thresholdu gecen satirlar sari yazdiriliyor..
203.                        printf("\033[01;36m");
204.                        printf("(%s) ", files[j]);
205.                        printf("\033[0m");
206.                    }
207.                }
208.            }
209.        }
210.
211.        /**
212.         * Jaccard benzerligini hesaplayarak oranlari struct icerisindeki
213.         * jaccard matrisine yazan fonksiyon
214.            @param shgl        Shingle struct pointer
215.            @param files      File names
216.        */
217.        void calculate_jaccard(SHINGLE *shgl,char **files){
218.            int total_1=0;
219.            int total_2=0;
220.            int intersect=0;
221.            int union_all;
222.            int i,j,k;
223.            //all permutation loop n*(n-1)/2
224.            shgl->jaccard=(float**)malloc(shgl-
    >file_num*sizeof(float*)); //jaccard matrisi olusturuldu
225.            for(i=0;i<shgl->file_num;i++){
226.                shgl->jaccard[i]=(float*)calloc((size_t)shgl->file_num,sizeof(float));
227.            }
228.            printf("\n\nK VALUE : %d THRESHOLD : %.2f\nJACCARD SIMILARITIES OF ALL DOCUMENT COM
    BINATIONS\n",shgl->k_val,shgl->threshold);
229.            for(i=0;i<shgl->file_num;i++){
230.                for(j=i+1;j<shgl->file_num;j++){
231.                    //printf("\nFile %d (%s) - File %d (%s) : ",i+1,files[i],j+1,files[j]);
232.                    for(k=0;k<shgl->shingle_count;k++){
233.                        //printf("\n%d %d",shgl->matrix[k][i],shgl->matrix[k][j]);
234.                        if(shgl->matrix[k][i]==1)
235.                            total_1++;
236.                        if(shgl->matrix[k][j]==1)
237.                            total_2++;
238.                        if(shgl->matrix[k][i]==1 && shgl->matrix[k][j]==1)
239.                            intersect++;
240.                    }
241.                    union_all=total_1+total_2-intersect;
242.                    shgl->jaccard[i][j]=((float)intersect/(float)union_all);
243.                    shgl->jaccard[j][i]=shgl->jaccard[i][j];
244.                    //printf("%.2f ",shgl->jaccard[i][j]);
245.                    total_1=total_2=intersect=0;
246.                }
247.            }
248.            //Sonuclar ekrana yazdiriliyor
249.            printJaccard(shgl,files);
250.
251.        }
252.
253.        /**
254.         * Hash degeri ureten hash fonksiyonu
255.            @param a   random value
256.            @param x   0 to hash_Size
257.            @param m   Shingle count
```

```
258.        */
259.        int hash(int a,int x, int m){
260.            return ((a*x+1)%m);
261.        }
262.
263.        /**
264.         * Signature benzerligine gore hesaplanan dosyalarin,
265.         * Benzerlik oranlari ve benzer ikililer gibi sonuclari ekrana yazdiran fonksiyon
266.         * @param shgl      Shingle struct pointer
267.         * @param files     Filenames
268.         */
269.        void printSignature(SHINGLE *shgl,char **files){
270.            int i,j;
271.            //Hesaplanan signature similarity oranlari tablo olarak ekrana yazdiriliyor
272.            printf("\t");
273.            for(i=0;i<shgl->file_num;i++){
274.                printf("%s\t",files[i]);
275.            }
276.            printf("\n");
277.            for(i=0;i<shgl->file_num;i++){
278.                printf("%s\t",files[i]);
279.                for(j=0;j<shgl->file_num;j++){
280.                    if(shgl->minhash[i][j]>shgl->threshold){
281.                        printf("\033[01;33m");
282.                        printf("%.2f\t",shgl->minhash[i][j]);
283.                        printf("\033[0m");
284.                    }
285.                    else if(i==j){
286.                        shgl->minhash[i][j]=1;
287.                        printf("\033[1;31m");
288.                        printf("%.2f\t",shgl->minhash[i][j]);
289.                        printf("\033[0m");
290.                    }
291.                    else{
292.                        printf("%.2f\t",shgl->minhash[i][j]);
293.                    }
294.                }
295.                printf("\n");
296.            }
297.            //benzer ikililer yazdiriliyor
298.            printf("\nSIMILIAR DOCUMENT PAIRS ACCORDING TO SIGNATURE SIMILARITY");
299.            printf("\n");
300.            for(i=0;i<shgl->file_num;i++) {
301.                printf("\n%s-->>\t", files[i]);
302.                for (j = 0; j < shgl->file_num; j++) {
303.                    if (shgl->minhash[i][j] > shgl-
    >threshold && j!=i) { //thresholdu gecen satirlar sari yazdiriliyor..
304.                        printf("\033[01;36m");
305.                        printf("(%s) ", files[j]);
306.                        printf("\033[0m");
307.                    }
308.                }
309.            }
310.        }
311.        /**
312.         * Imza matrisine gore benzerlik oranlarini struct icerisinde minhash matrisine yazan f
    onksiyon
313.         * @param shgl
314.         * @param files
315.         */
316.        void calculate_minhash(SHINGLE *shgl,char **files){
```

```
317.            int intersect=0; //kesisim sayisi
318.            int i,j,k; //indis degiskenleri
319.            int temp;
320.            int a[HASH_SIZE];
321.            int rand_val; //random deger
322.            //imza matrisi olusturuluyor
323.            int **signature=(int**)malloc(sizeof(int*)*HASH_SIZE);
324.            for(i=0;i<HASH_SIZE;i++){
325.                signature[i]=(int*)malloc(sizeof(int)*shgl->file_num);
326.            }
327.            //ilk deger atamasi olarak her bir goze sonsuz niteliginde hicbir zaman alamayacagi
    bir deger olan shingle sayisi ataniyor.
328.            for(i=0;i<HASH_SIZE;i++){
329.                for(j=0;j<shgl->file_num;j++){
330.                    signature[i][j]=shgl->shingle_count;
331.                }
332.            }
333.            //hash matrisi olusturuluyor icinde hashlenmis degerleri tutar
334.            int **hash_mtr=(int**)malloc(sizeof(int*)*shgl->shingle_count);
335.            for(i=0;i<shgl->shingle_count;i++){
336.                hash_mtr[i]=(int*)calloc((size_t)HASH_SIZE,sizeof(int));
337.            }
338.            srand(time(NULL));
339.            rand_val=rand()%shgl->shingle_count-1;
340.            a[0]=rand_val;
341.            //printf("%-3d ",rand_val);
342.            for(i=1;i<HASH_SIZE;i++){ //random tekrarsiz a degerleri icin dizi olusturuluyor
343.                rand_val=rand()%shgl->shingle_count-1; //rastgele bir sayi uret
344.                j=0; //0.sayidan itibaren
345.                while(j<i){ //en son uretilen sayiya kadar bak
346.                    if(a[j]==rand_val){ //eger uretilen random sayi daha once varsa
347.                        rand_val=rand()%shgl->shingle_count-1; //yenisini uret
348.                        j=0; //en bastan kontrol etmek icin j yi sifirla
349.                    } else{
350.                        j++; //degilse bir sonraki indise
351.                    }
352.                }
353.                a[i]=rand_val; //random sayiyi ekle
354.            }
355.            //hash matrisine degerler ataniyor
356.            for(i=0;i<shgl->shingle_count;i++){
357.                for(j=0;j<HASH_SIZE;j++){
358.                    hash_mtr[i][j]=hash(a[j],i,shgl-
    >shingle_count); //hash sayisi uret ve matrise yaz
359.                }
360.            }
361.            //imza matrisi olusturuluyor
362.            for(i=0;i<shgl->shingle_count;i++){//shingle gezer
363.                for(j=0;j<shgl->file_num;j++){//dosya gezer
364.                    if(shgl-
    >matrix[i][j]==1){ //secilen shingle hangi dosyada var j de tutuluyor
365.                        for(k=0;k<HASH_SIZE;k++){ //ilgili shingle icin uretilen tum hashler
366.                            temp=hash_mtr[i][k]; //shingle icin uretilen h1,h2,h3... degerlerin
    i tempe at
367.                            if(temp<signature[k][j]){ //imza matrisindeki dosya olan j dosyasin
    in k.hash degeriyle kiyasla kucukse
368.                                signature[k][j]=temp; //imza matrisindeki hash degerlerini ilgi
    li dosya icin guncelle
369.                            }
370.                        }
371.                    }
```

```
372.                }
373.            } //imza matrisi tamamlandi boyutu [hash fonksiyonu sayisi][dosya sayisi]
374.            //minhasle bulunan benzerlik oranlarinin tutulacagi matris olusturuluyor
375.            shgl->minhash=(float**)malloc(shgl->file_num*sizeof(float*));
376.            for(i=0;i<shgl->file_num;i++){
377.                shgl->minhash[i]=(float*)calloc((size_t)shgl->file_num,sizeof(float));
378.            }
379.            //dosya ikilileri icin benzerlik oranlari signature similarity hesaplaniyor
380.            printf("\n\nK VALUE : %d THRESHOLD : %.2f\nSIGNATURE SIMILARITIES OF ALL DOCUMENT C
    OMBINATIONS\n",shgl->k_val,shgl->threshold);
381.            for(i=0;i<shgl->file_num;i++){ //i. dosyanin
382.                for(j=i+1;j<shgl->file_num;j++){ //kendisi haric diger dosyalarla
383.                    //printf("\nFile %d (%s) - File %d (%s) : ",i+1,files[i],j+1,files[j]);
384.                    for(k=0;k<HASH_SIZE;k++){ //imza matrisi boyunca
385.                        if(signature[k][i] == signature[k][j]) //benzer imza degerleri icin
386.                            intersect++; //toplami hesaplaniyor
387.                    }
388.                    shgl->minhash[i][j]=((float)intersect/(float)HASH_SIZE); //oran bulunuyor
389.                    shgl->minhash[j][i]=shgl-
    >minhash[i][j]; //bulunan oran matrisin diyagonaline de yaziliyor
390.                    intersect=0;
391.                }
392.            }
393.
394.            //Hesaplanan signature similarity oranlari tablo olarak ekrana yazdiriliyor
395.            printSignature(shgl,files);
396.
397.            //free hash matrix
398.            for(i=0;i<shgl->shingle_count;i++){
399.                free(hash_mtr[i]);
400.            }
401.            free(hash_mtr);
402.            //free signature matrix
403.            for(i=0;i<HASH_SIZE;i++){
404.                free(signature[i]);
405.            }
406.            free(signature);
407.
408.        }
409.
410.        int main() {
411.            SHINGLE *shgl=(SHINGLE*)malloc(sizeof(SHINGLE)); //shingle yapisi olusturuluyor
412.            if(!shgl){
413.                printf("Shingle struct allocation error!!! Quitting...");
414.                exit(0);
415.            }
416.            int choice=1; //programin cikisi ve modul secimini tutan degisken
417.            int i; //cevrim degiskeni
418.            char filename[NAME_LEN]; //dosya adini tutan temp dizi
419.            printf("\nPlease enter file number : "); //dosya sayisi kullanicidan aliniyor
420.            scanf("%d",&shgl->file_num);
421.            char **files=(char**)malloc(sizeof(char*)*(shgl-
    >file_num)); //dosya isimlerini tutacak dizi allocate ediliyor
422.            for(i=0;i<shgl->file_num;i++) { //dosya isimleri kullanicidan okunup kaydediliyor
423.                files[i]=(char*)malloc(NAME_LEN*sizeof(char));
424.                printf("\nPlease enter File name %d: ", i + 1);
425.                scanf("%s", filename);
426.                strcpy(files[i],filename);
427.            }
428.            printf("\nPlease enter your choice \n1-Continue with K-Value \n0-
    Exit\nChoice: ");
```

```
429.            scanf("%d",&choice); //secim kullanicidan okunuyor
430.            while(choice!=0){
431.                if(choice==1){  //devam etme secilirse
432.                    printf("\nPlease enter K value of Shingles : ");
433.                    scanf("%d",&shgl->k_val); //k degeri kullanicidan okunur
434.                    printf("\nEnter threshold value for %d shingle similarity(0.0 between 1.0)
    : ",shgl->k_val);
435.                    scanf("%f",&shgl->threshold); //threshold degeri kullanicidan okunur
436.                    initialize(shgl);//shingle struct i icin ilk deger atamasi yapiliyor
437.                    for(i=0;i<shgl->file_num;i++){
438.                        FILE *fp=fopen(files[i],"r"); //sirayla dosyalar acilir
439.                        if(!fp){
440.                            printf("File error!!! Quitting...");
441.                            exit(0);
442.                        }
443.                        createShingles(fp,i,shgl); //dosya dosya shinglelari structa olusturan
    fonksiyon
444.                        fclose(fp);
445.                    }
446.                    print_shingle(shgl,files); //dosyalarin shingle sayisi tablosunu ekrana yaz
    diran fonksiyon
447.                    calculate_jaccard(shgl,files); //jaccard matrisi olusturuluyor ve sonuclari
    ni yazdiriliyor
448.                    calculate_minhash(shgl,files); //signature matrisi olusturuluyor ve sonucla
    rini yazdiriliyor
449.                    //Free islemleri
450.                    //free jaccard result matrix
451.                    for(i=0;i<shgl->file_num;i++){
452.                        free(shgl->jaccard[i]);
453.                    }
454.                    free(shgl->jaccard);
455.                    //free minhashing results
456.                    for(i=0;i<shgl->file_num;i++){
457.                        free(shgl->minhash[i]);
458.                    }
459.                    free(shgl->minhash);
460.                    //free shgl->matrix
461.                    for(i=0;i<shgl->shingle_count;i++){
462.                        free(shgl->matrix[i]);
463.                    }
464.                    free(shgl->matrix);
465.                }
466.
467.                printf("\nPlease enter your choice \n1-Continue with K-Value \n0-
    Exit\nChoice: ");
468.                scanf("%d",&choice);
469.            }
470.
471.            system("PAUSE");
472.            return 0;
473.        }
```