Doküman Benzerliğinin Çıkarılması

Geliştirilen bu araç ile dokümanların birbirlerine olan benzerlikleri hesaplanmaktadır.

Program "Online" ve "Offline" olarak iki modda çalışmaktadır.

Offline Mod

Bu mod seçildiğinde ilgili ekranda "Similarity for all Documents", "Similarity Between a Document to all Documents" ve "Similarity Between Two Documents" isimli üç ayrı sekme yer almaktadır şekil-1.

Online Offline Help About Similarity For All Documents Similarity Between A Documents Open Folder Run	🕌 Offline Text Matching	
Similarity For All Documents Similarity Between A Document to All Documents Similarity Between Two Documents Open Folder Run	Online Offline Help About	
Open Folder Run	Similarity For All Documents	Similarity Between A Document to All Documents Similarity Between Two Documents
Open Folder Run		
Open Folder Run		
Run		
Run		
Run		
Run		
		Run

Şekil-1 Offline çalışma ekranı

Similarity for all Documents : Bu sekmede "Open Folder" dan seçeceğimiz bir klasör altında yer alan dosyaların birbirlerine göre benzerlikleri hesaplanarak sonuç döndürülür. Şekil-2 'de ki örnekte seçilmiş olan klasör altında 6 tane txt uzantılı dosya bulunmaktadır. Daha sonra "Run" tuşuna bastığımızda bu altı dosyanın birbirine olan benzerliklerini hesaplayarak 0-1 aralığında bir sayı ile ifade eder. Şekil-1 de iki ile dört numaralı dosyaların birbirlerine en fazla benzerlik gösterdiklerini söyleyebiliriz.



Şekil-2 Tüm dokümanların birbirlerine göre benzerlikleri

Similarity Between a Document to all Documents: Bu sekmenin görevi, seçilen tek bir dokümanın yine seçilen bir klasör içindeki dosyalar ile olan benzerliklerinin çıkarılmasını sağlamaktadır. Şekil-3 'te görüldüğü üzere seçilen klasör içerisinde 33 farklı doküman bulunmakta ve "Open File" düğmesini kullanarak seçtiğim 1 dokümanın diğer dokümanlar ile olan benzerliği sağdaki pencerede sunulmuştur. Bu örnekte 1.txt ile %100 benzerlik mevcuttur.

e Offline Help About nilarity For All Documents Similarity Between A Document t Open Folder vd Settings\ceuser\Desktop\GazeteYazarDerlemi\TuigceBar	to All Documents Similarity Between Two Documents	
nilarity For All Documents Similarity Between A Document t Open Folder vd Settings\ceuser\Desktop\GazeteYazarDerlemi\TuigceBar	to All Documents Similarity Between Two Documents	
Open Folder vd Settings\ceuser\Desktop\GazeteYazarDerlemi\TuigceBar	k construction	
	[1.00] 1.txt [0.30] 35.txt [0.30] 29.txt [0.29] 8.txt [0.29] 18.txt [0.28] 26.txt [0.28] 26.txt	
Open File		<u> </u>
Boş sayfaların kerameti Cumhuriyet'i Çok Sevmiştim HASAN CEMAL isimli kitabın sayfa adeti konusundaki anlaşmazlığı çözmüş bulunuyorum. Dünkü yazımda hatırlayacağınız veya hatırlamayacağınız üzere her köşecinin kitabın sayfa adedi konusunda ayn bir rakam vermesi üzerine isyan etmiş ve "kaç sayfa lan bu kitap?" diye kibarca hislerimi dışa vurmuştum. Yayınevi 592 diyor, yazan 600 diyor, Serdar Turgut 573 diyor, Mehmet Y. Yılmaz 588 diyor Tuhaf bir durum söz konusu ve lakın kitap elimde olmadığı için bir türlü açıp bakamıyorum.		
Bir kitabı bir gecede okumakla övünüp duran karşı oda komşum Selahartin Duman beyefendi, yazımı okumuş, gazeteye gelirken söz konusu kitabı yanına alıp öyle gelmiş. "Say bakalım kaç sayfaymış" deyip önüme attı kitabı. "Okudunuz, bitirdiniz mi hemen?" dedim "Tabii" dedi her zamanki kendinden emin haliyle, "akşam yemekten sonra başladım, sabah bitirdim"		
Beni delirtmek için mi söylüyor hakikat mi bilmiyorum.	×	

Şekil-3 Tek bir dokümanın tüm klasör içerisindeki dokümanlar ile benzerliği

Similarity Between Two Documents : Bu sekme ise "Open File-Dosya Aç" ile seçilmiş iki ayrı dosyanın birbirleri ile olan benzerliğini bulmaktadır. "Run-Çalıştır" düğmesi ile işlem başlatılır ve şekil-4'te ki ekran karşımıza gelir. Bu ekran bize karşılaştırılan bu iki dokümanın benzerliğinin 0.28 olduğunu söyler.



Şekil-4 İki dokümanın benzerliğinin çıkarılması

Online Mod

Şekil-5'te yer alan ekranın sol tarafında iki pencere yer almaktadır. Sol altta yer alan pencereden benzerleri aranacak olan doküman "Open File" düğmesi ile seçilir. Daha sonra "Analyse" düğmesi ile ilgili doküman analiz edilir. Ve en fazla sıklıkla kullanılan kelimeler çıkarılır (tek başlarına anlam ifade etmeyen=stop words ve fiiller hariç). Sol üstte yer alan pencerede Keyword ve Properties bilgileri yer almaktadır.

Properties altında bulunan:

Number Of Keywords : Dokümandan çıkarılan toplam anahtar kelime sayısı.

Keywords : Kaç tane anahtar kelime ile çalışmak istiyorsak seçme hakkını verir. Ayrıca Keywords penceresinde yer alan bir anahtar kelimeye çift tıklayarak silebilir ya da "Add" düğmesini kullanarak istediğimiz bir anahtar kelimeyi listeye ekleyebiliriz.

Search in : Seçilen dokümanın benzerlerinin html, doc ve pdf dosyalarının hangilerinde aranacağının seçilmesi için kullanılır.

Weighting : Kelimelerin ağırlıklandırılmasının hangi yönteme göre yapılacağını belirler. Pages to Search : Aramanın kaç sayfada yapılacağı.

Aynı bölümde "Find the Original File" seçeneğinin seçilmesi ile dokümanın birebir aynısının web üzerinde aranması istenebilir. Sağ taraftaki bölümde ise benzer dosyaların link adresleri ve benzerlik oranları verilmektedir.

Online Offline Help	About	
Keywords Chp Ioruttay harran siyaset	Properties Number Of Keywords : 564 Keywords : • • Search in : • Html Doc Pd Weighting : • TP • TP × IDP Pages To Search : 1 • Pun	Pittri (Doc Pdf Original Hoal Links Original Doc Links Original Pdf Links
Add GANLILIEFA UCA/GIMIZ "homom "Othelere" baildk. Otel Edesso, doku. Otel Harran, dolu. Köran Otel, dolu. Balkay Otell, dolu. Balkay Otella, dolu. Balkay Otella		

Şekil – 5 Online çalışma ekranı