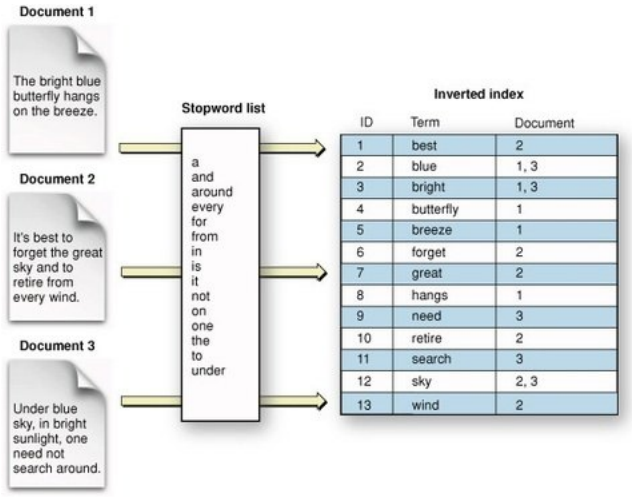
BİL-212 Veri Yapıları
Ödev - 3
Veriliş Tarihi: 03.07.2018
Teslim Tarihi: 13.07.2018
Teslim Şekli: Ödevi nasıl göndereceğiniz piazza üzerinden duyurulacaktır.
Kurallar: Geç gönderilen ödevler kabul edilmez. Kopya kesinlikle yasaktır, kopya veren ve alan öğrenciler bütün ödevlerden 0 alırlar ve ayrıca üniversite disiplin yönetmeliği kuralları bu öğrencilere uygulanır.

Bu ödevde basit bir arama motoru (search engine) kodlayacaksınız. Kodlayacağınız motor, internet üzerinde değil bilgisayarınız üzerinde hızlıca arama yapabilen bir motor olacak. Bunu yapabilmek için büyük internet arama motorlarının da kullandığı inverted index kullanacağız.

Gerçek, internet üzerindeki sayfaları indeksleyen bir inverted index aşağıdaki şekilde oluşturulur:

1. Dokümanı internetten indir.
2. Dokümandaki etiketleri (HTML tagleri mesela) ve gereksiz kelimeleri (stop words) sil. (the, a vs gibi kelimeler örneğin).
3. Kalan kelimelerin köklerini bul. (Ekler arama motorları için genelde bir anlam taşımaz.)
4. Kelimelerle birlikte doküman ID'si, dokümanın neresinde geçtikleri, dokümanda kaç kere geçtikleri bilgisini bir tabloya yaz.

Aşağıdaki şekilde bunun için basit bir örnek gösteriliyor.[1] Bu örnekte, her kelime için sadece doküman ID'leri tutuluyor. (Tablodaki ID kolonu kelime ID'lerini ifade ediyor.)



Yukarıdaki adımları internetteki bütün sayfalar için yapabilirseniz, çok temel bir arama motorunuz olur. Bu ödevde yapmanız gereken şey ise, bilgisayarınızdaki dosyalar için basit bir arama motoru (desktop search). Bunu yaparken yukarıdaki 2 ve 3. adımları yapmayacağız (NLP ile uğraşmıyoruz bu aşamada)(Yalnızca index'lenecek kelimeyi '.?!,:' karakterlerinden temizleyebilirsiniz.). İlk adımdaki dosya ise bilgisayarınızdaki diskten okuyacağınız dosya(lar) olacak. Her kelime için doküman ID'leri (kendiniz her dokümana birer eşsiz ID vereceksiniz) ve o dokümanda kaç kere geçtiği bilgilerini tutacaksınız.

Dosya sayısı ve uzunlukları fazla olduğu zaman inverted index'e ekleme yapmak uzun sürmeye başlayacaktır, çünkü tablo çok büyüyecektir. Inverted index'i hızlı ekleme ve sorgu yapabilecek şekilde tasarlamanız gerekiyor (ipucu : hash tablosu).

Kullanacağınız hash tablosunu kendiniz implement etmelisiniz. Collision handling icin separate chaining kullanmalısınız. Load factor'u 0.7 olacaktır.

Yazacağınız kod, Komut satırından ya da kullanıcıdan alınacak bir başlangıç klasörünün altındaki bütün txt uzantılı dosyaları yukarıda belirtildiği gibi bir inverted index'e yerleştirmelidir. Daha sonra oluşturulan inverted index'i başka bir dosyaya (sizin belirleyeceğiniz bir formatta) yazmalıdır. Bu dosya binary de olabilir, metin dosyası da olabilir.

Yazacağınız ikinci bir sınıf ise bu inverted index dosyasını okuyup, buna göre arama sorgularına cevap verebilmelidir. Bulunan dokümanları (kelimenin geçtigi dokümanlar), kelimenin frekansına göre sıralayıp ekrana basmalıdır.

Örnek bir program döngü içerisinde aşağıdaki gibi çalışmalıdır. Burada girilen kelimenin geçme sıklığına göre sıralanıp ilk 5 dosya frekans değeriyle birlikte verilmiştir. Çalışmalarınızda değerlerinizin örneklere benzer olmasını bekliyoruz.

[1] Quora'dan (`https://www.quora.com/Information-Retrieval-What-is-inverted-index`) alıntıdır.

'food_txt.zip' dosyasını piazza sayfasında resource kısmında bulabilirsiniz. "geçen süre" yi kullanıcıdan kelimeyi aldıktan sonra, dosyaları ne kadar sürede bulduğunuzu hesaplamak için kullanacaksınız. Yani sadece arama süresini hesaba katmalısınız.

Dosyaların bulunduğu klasör yolunu giriniz :
- /home/Desktop/food_txt
Aranacak kelime :
- bread
kashrut.txt :8
bakebred.txt :8
pot.txt :4
bread.txt :4
hitbred.txt :3
geçen süre : 2.891717 mili seconds

Aranacak kelime :
- Bread
kashrut.txt :8
bakebred.txt :8
pot.txt :4
bread.txt :4
whitbred.txt :3
geçen süre : 4.308777 mili seconds

Aranacak kelime :
- water
candy.txt :501
byfb.txt :45
chili.txt :42
beer.txt :21
meat2.txt :19
geçen süre : 4.308777 mili seconds

Aranacak kelime :
- egg
candy.txt :92
shuimai.txt :8
x-drinks.txt :6
kashrut.txt :6
wonton.txt :4
geçen süre : 2.684094 mili seconds

Aranacak kelime :
- Jalapeno
chili.txt :11
strattma.txt :3
boarchil.txt :2
oakwood.txt :2
firecamp.txt :1
geçen süre : 8.739789 mili seconds

Aranacak kelime :
- elma
scam.txt :1
geçen süre : 2.436538 mili seconds
Aranacak kelime :
- kebab
key not found
geçen süre : 7.931419 mili seconds