11th February 2007

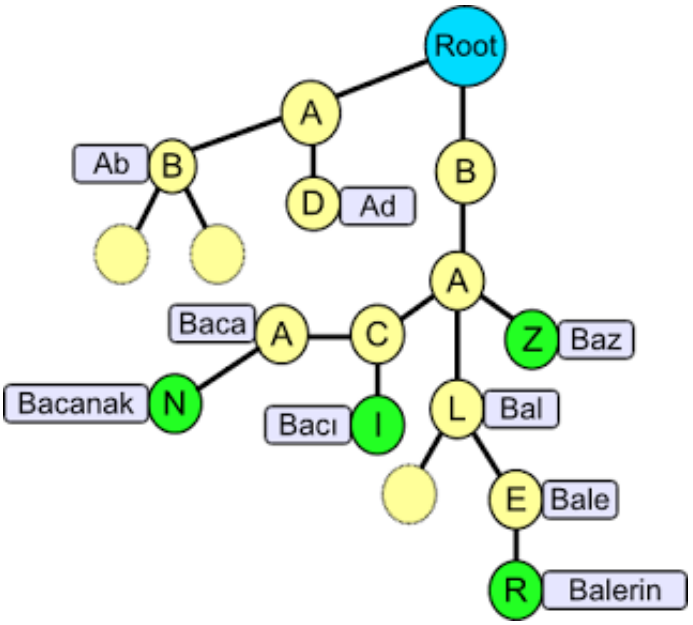
# Zemberek nasıl çalışır? 1.Sözlük ve Kök ağacı

Zemberek bir kelimenin Türkçe olup olmadığına nasıl karar veriyor? Bu basit sorunun cevabı "verilen bir kelimeyi Türkçe kök ve eklerine ayırabilirseniz Türkçedir, ayıramıyorsanız değildir". Kısacası bir kelimenin Türkçe olup olmamasını anlamak için morfolojik analiz yapabilmelisiniz. Türkçe yazım denetimi yapabilmek için önceleri en sık kullanılan kelimelerin bir dosyaya konulup gelen kelimelerin o dosyadan kontrol edilmesi gibi ilk bakışta mantıklı görünen ama biraz inceleyince pratik olmadığı anlaşılan yöntemler de düşünülmüştü. Bu tür yöntemlerin yetersiz olan %98-99 doğrulukla çalışması için bile milyonlarca kelimeyi içermesi gerekir.

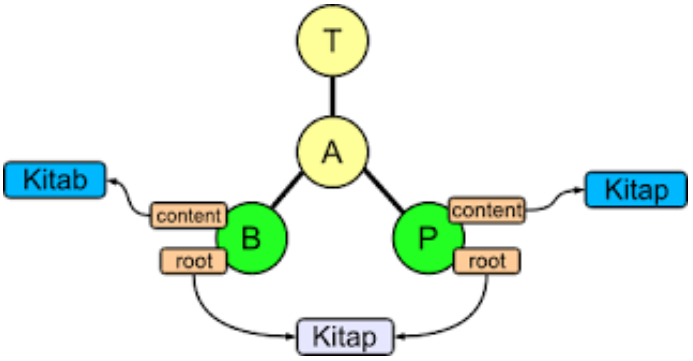
Zemberek çok basit bir sistemle verilen bir kelimeyi morfolojik olarak inceler. Önce verilen kelimenin kökü olabilecek adayları belirler, sonra da olabilecek ekleri uygun sırayla bu kök'e eklemeye çalışır. Eğer bu işlem sırasında girişteki kelimenin aynısını elde edebilmişse, o zaman uygun kök ve ekleri de bulmuş demektir ve kelime Türkçedir, eğer kök adaylarının hiçbirinden sonuç elde edilememişse o zaman kelime Türkçe değildir. bu işlemin ilk adımı olan kök adaylarının bulunması işlemini inceleyelim. Kök adaylarının bulunabilmesi için öncelikle elimizde Türkçe'deki tüm kök kelimelerinin bulunması gerekiyor. Zemberek Türkiye Türkçesi içi yaklaşık 30.000 kök içeren bir kılavuzu da beraberinde taşır, bu kılavuzda her kök tipine ve özel durumlarına göre etiketlenmiş şekilde bulunur. Diğer Türk dilleri için yapılan gerçeklemeler de benzeri bir kök sözlüğünü taşımalıdırlar. Burada bahsi geçen özel durumlar sondaki sert ünsüz harfin yumuşaması (sağlık -> sağlığa) veya bazı durumlarda sondan bir önceki sesli harfin düşmesi (burun -> burnu) gibi halleri içerir. Kökler için kullanılan etiketler de dile göre değişiklik gösterebilir. Aşağıda Azeri Türkçesi için yazılmış kök sözlüğünün küçük bir bölümü görülmektedir.

```
su AD
sağlıq AD YUM
al EY
gel EY
tuz AD
istiot AD
et AD
balıg AD
bir RA
iki RA
dörd RA
dünen ZAMAN
sabah ZAMAN
```

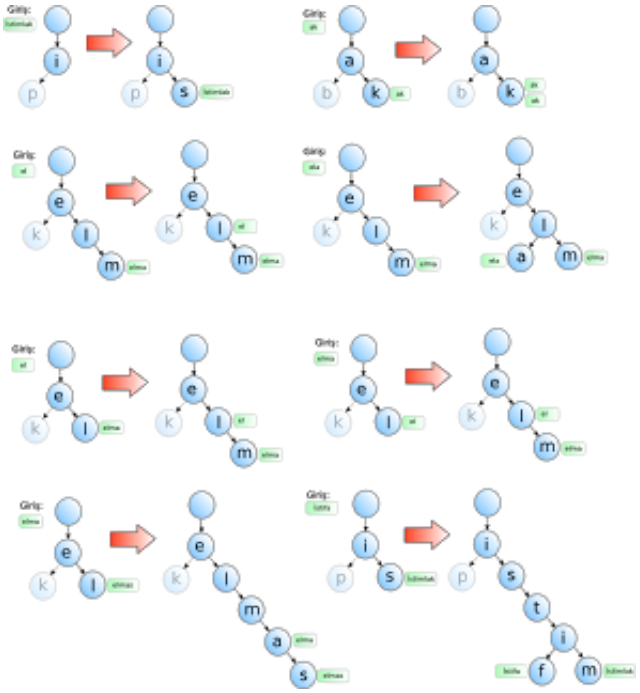
Verilen bir kelimenin kök adaylarının bulunması için zemberek bu kökleri özel bir ağaca yerleştirir. bu özel ağaç sayesinde adayların belirlenmesi son derece hızlı şekilde yapılabilmektedir. Bu ağaçta kökler içeriklerine göre yerleşirler, örneğin aşağıdaki örnekte "baz" kökü sırasıyla B- A -Z ile etiketlenmiş düğümlerin en sonuncusuna bağlanmış şekilde durur. Burada dikkati çeken bir nokta da uzun köklerin gereksiz fazladan düğüm oluşturmayacak şekilde ağaca bağlanması ve bellekten tasarruf edilmesidir. yani "Balerin" kökü B-A-L-E düğümlerinden sonra gelen R düğümüne bağlanmıştır.




[http://2.bp.blogspot.com/_jiekG5XBOyk/Rc727KV94UI/AAAAAAAAAEs/2ytqMiQAyJ4/s1600-h/agac.png]
Kök ağacı kök düğümleri ve bu düğümlerin bağlantılarından oluşur. Kitap gibi bir ek geldiği zaman yumuşama olasılığı olan köklerin değişimiş halleri de kök ağacına eklenir. Ancak bu değişmiş haller de orijinal kökü işaret ederler.




[http://4.bp.blogspot.com/_jiekG5XBOyk/Rc7yWqV94PI/AAAAAAAAAD4/S5w_kOvpVr8/s1600-h/kokdugumu.png]
Kök ağacının oluşturulması işlemi oldukça ince bazı uç durumları göz önüne almayı gerektirir. ağacın daha oluşturulurken bellek tasarrufu ve performans için yukarıda anlatılan kuralara göre davranması için aşağıdaki 7-8 durum için farklı önlemler alınması gerekmiştir. Ağaç oluşum kodları ile ilgili problemli noktalara işaret eden NZemberek [http://code.google.com/p/nzemberek/] geliştiricisi Tankut Tekeli'ye de buradan teşekkür etmek isterim.



[http://3.bp.blogspot.com/_jiekG5XBOyk/Rc70_aV94SI/AAAAAAAAAEc/_KN_44pa6IU/s1600-h/agac%282%29.png]
Bu yazıda kısaca kök sözlüğü ve kök ağacının yapısı ve oluşturulmasından bahsettim, daha sonraki yazılarda köklerin nasıl seçildiği, çözümleme işlemi gibi adımların iç yapıları hakkında biraz bilgi vermeye çalışacağım.

Daha fazla bilgi için genel olarak Zemberek ve işleyişi ile ilgili 8 sayfalık bir makaleyi [http://zemberek.googlecode.com/files/zemberek_makale.pdf] okuyabilirsiniz.

11th February 2007, M.D.A tarafından yayınlandı

14 Yorumları görüntüle

**tolga** 15 Şubat 2007 13:30:00 GMT+2

Sizce de ilk resimdeki Root yerine Kök yazsa daha güzel olmaz mıydı?

Yanıtla

---

**M.D.A** 15 Şubat 2007 14:42:00 GMT+2

Tolga, Evet haklısın, şekilleri uzun zaman önce çizmiştim, ingilizce makalelerde de kullanılabilsin diye o şekilde kalmış.

Yanıtla

---

**Rıdvan** 21 Şubat 2007 16:32:00 GMT+2

Zevkle okunabilecek bir yazı olmuş. Teşekkürler Mehmet bey...

Yanıtla

---

**ersin demirok** 14 Mart 2007 09:20:00 GMT+2

merhaba;

muhtesem bir calisma olmus,

bu kilavuzu mozilla yada thunderbird ile kullanabilmemiz icin ne yapmamiz gerekiyor?

sitelerinde su sekilde bahsediyor

http://www.mozilla.com/en-US/thunderbird/dictionaries.html

tesekkurler

ersind@gmail.com

Yanıtla

---

**M.D.A** 14 Mart 2007 09:32:00 GMT+2

Ersin, zemberek'in Firefox ve Thunderbird'de yazım denetimi servisi verebilmesi için şu anda sadece Pardus'ta yer alan zemberek sunucusunun sistemde kurulu olması ve bu iki uygulama için Pardus geliştiricilerinin hazırladığı yamaların eklenip tekrar derlenmesi gerekiyor.
Kısacası, şu anda sadecePardus altında Firefox ve Thunderbird için Türkçe yazım denetimi desteği var. diğer sistemler için bu çalışmanın yapılması gerekir.

Yanıtla

---

**Anonymous** 5 Nisan 2007 02:31:00 GMT+3

Zemberek en nihayet openoffice'deki önemli bir eksikliği giderdi. Programın geliştiricilerine müteşekkir olsam da, aspell türü her programa kolayca entegre olabilecek bir format kullanmadıklarına hayıflanıyorum. Makale gerçekten çok güzel, teşekkürler.

Yanıtla

---

**Ahmet BÜTÜN** 15 Mayıs 2009 19:24:00 GMT+3

elinize sağlık, çok güzel bir çalışma

Yanıtla

---

**Ahmet BÜTÜN** 16 Mayıs 2009 14:56:00 GMT+3

bir de bişey öğrenmek istiyorum. neden zemberek'in kendine ait bir sitesi yok, böyle blogger'larda felan işi devam ettiriyorsunuz, vakit mi bulamıyorsunuz diyicem ama sanmıyorum.

Yanıtla

---

**M.D.A** 18 Mayıs 2009 22:55:00 GMT+3

@Ahmet Butun:
Zemberek proje sayfasinda gerekli dokuman ve koda erisebilirsiniz. Google'da "zemberek" diye aratin.

Yanıtla

---

**Gönenç Ercan** 16 Kasım 2009 00:13:00 GMT+2

Biraz geç (2 yıl kadar) olsa da, ağaç yapısı kullanıldığında ortak kelime başlangıçları ortak tutulabiliyor. Fakat eğer bir automata (ki sınırlı sayıda kelime içeren bir sözlük olduğu için cycle olmayan bir automata oluyor bu) kullanılırsa kelime sonları ve ortaları da ortak olarak tutulabilir. Bu sayede hafıza kullanımı düşürülebilir. Aşağıdaki linkte bir çalışma var, Jan Daciuk'un tezi ilginizi çekebilir.

http://www.eti.pg.gda.pl/katedry/kiw/pracownicy/Jan.Daciuk/personal/adfa.html

C/C++ da gerçeklemesi var, Java bulamadim onun için ben yazmıştım. Tabi tahmin ediyorum ki ağaçtan bu yapıya geçmek size çok iş çıkarır...

Yanıtla

---

**Desteklediklerimiz** 25 Ocak 2010 02:33:00 GMT+2

cok güzel paylaşım tşklr
muhabbet
Seviyeli Sohbet
muhabbet odalari
Sohbet Alemi
Sohbet
porno izle

Yanıtla

---

**selim** 26 Şubat 2010 01:01:00 GMT+2

lisans tezimde zemberek çalışma mantığına benzeyen bir teknikle "bir textin ne kadar türkçe olduğunu tdk veritabanına göre sorguluyorum" acaba programın geliştiricisiyle nasıl irtibata geçebilirim bilgilendirirseniz çok sevinirim.

selimk_27@hotmail.com

Yanıtla

---

**selim** 26 Şubat 2010 01:02:00 GMT+2

lisans tezimde zemberek çalışma mantığına benzeyen bir teknikle "bir textin ne kadar türkçe olduğunu tdk veritabanına göre sorguluyorum" acaba programın geliştiricisiyle nasıl irtibata geçebilirim bilgilendirirseniz çok sevinirim.

selimk_27@hotmail.com

Yanıtla

---

**mervet** 28 Mart 2010 15:35:00 GMT+3

merhaba,
güzel makale. peki ben doğal dil çalışmamda kullanmak üzere sadece kökleri içeren bir sözlüğe ulaşabilir miyim?
teşekkürler

Yanıtla

Yorumunuzu girin...

Yorumlama biçimi: Kenan Savaş (

Oturumu kapat

Yayınla Önizleme

Beni bilgilendir