

## Bil 211 - Lab 3

Bir metindeki N ardışık kelimelerin oluşturduğu gruplara N-gram denir ( $N > 0$ ).  $N=1$  (unigram),  $N=2$  (bigram),  $N=3$  (trigram) örneklerini aşağıda görebilirsiniz:

**n: 1** Bu makale ile ngram kavramini ogrendik  
unigrams: bu,makale,ile,ngram,kavramini,ogrendik

**n: 2** Bu makale ile ngram kavramini ogrendik  
bigrams: bu makale,makale ile,ile ngram,ngram kavramini,kavramini ogrendik

**n: 3** Bu makale ile ngram kavramini ogrendik  
trigrams: bu makale ile,makale ile ngram,ile ngram kavramini,ngram kavramini ogrendik

### Gereklilikler

- Bu labda size input olarak verilen bir dosyadaki n-gramları bulmanız ve bunları yeni bir dosyaya yazmanız beklenmektedir. Bunun için Ngram sınıfı içinde **extractNgrams()** metodu oluşturunuz.
- Output dosyasında her satırda bi ngram ve sonrasında parantez içinde o ngramın input dosyasında kaç kere bulunduğu (frekans) yazmalı.
- Dolayısıyla bir ngram input dosyasında birden fazla defa bulunsa bile output dosyasında yalnızca tek bir ngram olarak bulunmalı.
- Ngramlar, input dosyasındaki bulunma sırasına göre output dosyasına yazılmalı. Ngramların output dosyasına eklenme sırası, çıktınızın testlerle birebir örtüşmesi açısından önemli.**
- Ardışık kelimelerin aynı n-gram grubu içinde sayılması için aynı cümle içinde yer alması gerekir. Örneğin bir cümlenin son kelimesi ile sonraki cümlenin ilk kelimesi bir bigram oluşturmaz. Nokta karakteri bulunduğu cümlelerin bittiğini varsayabilirsiniz.
- Output dosyasına yeni bir n-gram eklerken kendi yazacağınız **updateNgram()** fonksiyonunu kullanın.
- Ngram'ları bulurken büyük harfler, küçük harflere çevrilmeli. Örneğin, aynı kelimelerden oluşan n-gram grupları, büyük-küçük harf olarak farklılık gösterebilir bile aynı olarak düşünülmemelidir.**
- Ngram'ları tespit ederken noktalama işaretlerini yok sayın. (Bunun için internette bulduğunuz bir yöntem/fonksiyon vs. kullanabilirsiniz)**
- Kodunuzda file IO için try/catch bloklarını mutlaka kullanın.**
- Kelimeleri tespit ederken space karakterini delimiter olarak kullanın.
- Beklenen çıktıyı daha iyi anlamak için sizinle paylaşılan örnek input ve output dosyalarını inceleyin. Kodunuzun birebir aynı çıktıyı ürettiğinden emin olun.**
- Kodunuz değerlendirilirken puan alabilmeniz için beklenen çıktı ile birebir örtüşmelidir.**

**public void extractNgrams**(String input\_file, String output\_file, int[] N\_list)

- **input\_file**: input olarak kullanılacak metin dosyası
- **output\_file**: n-gramların yazılacağı output dosyası
- **N\_list**: integer listesi. Fonksiyonunuz bu listedeki her N sayısı için metin dosyasındaki N-gramları bulacak. Örneğin [1,3,5] verildiğinde, oluşacak output dosyasında hem unigram'lar, hem trigramlar, hem five-gram'lar bulunmalı.

**Sizinle paylaşılan örnek girdi ve çıktıları inceleyebilirsiniz.**

**public void updateNgrams**(String ngram\_str, String output\_file)

- Input dosyasında bulduğunuz her n-gram'dan sonra bu fonksiyonu kullanarak output dosyasını güncellemeniz.
- Fonksiyon, string olarak verilen bir n-gram'ı output dosyasına ekler. Eğer verilen ngram, daha önce output dosyasına eklenmişse, sadece o satırdaki frekans değerini bir artırır. Eğer output dosyasında bu ngram bulunmuyorsa, en sona yeni bir satır olarak ekler.
- Bu fonksiyonu, **extractNgrams** içinde çağırarak kullanmalısınız.

## Örnek:

String ngram\_str = "hello from"

Output.txt:

hello from (1)  
from the (1)  
the other (3)  
other side (2)

updateNgram(ngram\_str, "output.txt") metodu çağırıldıktan sonra output.txt'nin son hali:

Output.txt:

hello from (2)  
from the (1)  
the other (3)  
other side (2)

---

String ngram\_str = "hello mate"

Output.txt:

hello from (1)  
from the (1)  
the other (3)  
other side (2)

updateNgram(ngram\_str, "output.txt") metodu çağırıldıktan sonra output.txt'nin son hali:

Output.txt:

hello from (1)  
from the (1)  
the other (3)  
other side (2)  
hello mate (1)

## Çıktı Formatı:

<token> <token> ... <token> (<frekans>)\n

Output dosyasında her bir satırda bir ngram ve o ngram'ın kaç defa input dosyasında bulunduğu (frekans) yazmalıdır. Ngramı oluşturan token'lar arasında ve frekans yazılmadan önce boşluk bırakılmalıdır.

Örnek:

hello from the (5)  
from the other (2)