

Ders 10:

- Sepet analizi: en çok birlikte satılan ürün 2'leri nelerdir?
 - S: sepetID, ürünID'leri tutan bir matris. N satır, 2 sütun

sepetID	ÜrünID
1	2
2	1
2	789
...	...

- A: hangi ürünlerin, hangi sepetlerde olduğunu gösterebilir. ürünN: farklı ürün sayısı, SepetN: sepet sayısı, A: ürünN satır, sepetN sütunlu sadece 1 ve 0'lardan oluşan bir matris.

A	1	2	...	SepetN
1	0	1	...	0
2	1	0	...	1
...
ÜrünN	1	0	...	0

- S'i A matrisine dönüştürelim.

```
A'yi sıfırla
for i=1:N
    A(s(i,2),s(i,1))=1;
end
```

- T: hangi ürünlerin hangi ürünlerle birlikte kaç kez satıldıklarını gösterebilir. T: ürünN satır, ürünN sütunlu bir matris.

T	1	2	...	ÜrünN
1	0	12	...	0
2	23	0	...	179
...
ÜrünN	13	0	...	0

- A'dan T'yi üretelim.

```
T'yi sıfırla
for i=1:ürünN
    for j=i+1:ürünN
        for k=1:sepetN
            if A(i,k)==1 && A(j,k)==1
                T(i,j)=T(i,j)+1;
            end
        end
    end
end
```

- İf'den kurtulalım ?

```
T(i,j)=T(i,j)+A(i,k)*A(j,k);
```

- En çok birlikte satılan ürün 3'lüleri istenseydi?
- T, 3 indisli bir matris olurdu.

```
T'yi sıfırla
for i=1:ürünN
    for j=i+1:ürünN
        for k=j+1:ürünN
            for m=1:sepetN
                T(i,j,k)=T(i,j,k)+A(i,m)*A(j,m)*A(k,m);
            end
        end
    end
end
```

- Şifreleme: her harf yerine bir sayı koyalım. M harfleri içeren bir matris

A	G	İ	F	J
P	B	D	H	M
S	C	I	E	O
K	Ş	L	R	Ö
...

- Metinden sayı dizisine geçiş şifreleme, A B C → 11 22 32
- her harfin tablodaki sütunNo*10+satır No
- Sayı dizisinden metne geçiş deşifreleme 11 22 32 → A B C
- Elimizde uzunca şifreli bir metin olsa M tablosunu bilmeden şifreyi nasıl kırarız?
- O dildeki harflerin kullanım olasılıklarını kullanırız. En çok tekrar eden sayı, en çok kullanılan harfe karşılık gelir. Her harf sadece tek bir şifreye dönüşüyor.
- Buna çözüm: M matrisinde her karakterin o dildeki frekansı ile doğru orantılı bir sayıda yer alması. Türkçe'de M matrisinde bir sürü A yer alacak. Şifreleme yaparken tüm A'ların kodlarını bul, aralarından birini rasgele seç.
- Metin (M), kelimeID'lerinden oluşan bir dizi olsun. Bu metinde arka arkaya en çok geçen kelime 2'lerini bulalım.
 - KelimeN: metindeki farklı kelime sayısı
 - W: KelimeN*KelimeN'lik bir matris.
 - N: metin boyutu
 - M: N elemanlı bir sayı dizisi

```
W'yu sıfırla
for i=1:N-1
    W(M(i),M(i+1))=W(M(i),M(i+1))+1;
end
```

- Her bir kelimenin (-K,+K) etrafında en çok hangi kelimeler var.

```
W'yu sıfırla
for i=1+K:N-K
    for j=-K:K
        W(M(i),M(i+j))=W(M(i),M(i+j))+1;
    end
    // j=0 daki artımı azaltmak için
    W(M(i),M(i))=W(M(i),M(i))-1;
end
```

- Birlikte geçen kelimeler anlamsal olarak birbirine yakındır. Bu fikir kullanılarak Word2vec, FastText, Glove gibi araçlar üretilmiş. Kelimeler anlamsal uzayda

koordinatlarla temsil ediliyor. Anlamca yakın kelimeler bu uzayda yakın koordinatlara sahip. “Cimbom kazandı” ile “Galatasaray yendi” arasında benzerlik bulunabiliyor.