Word2Vec

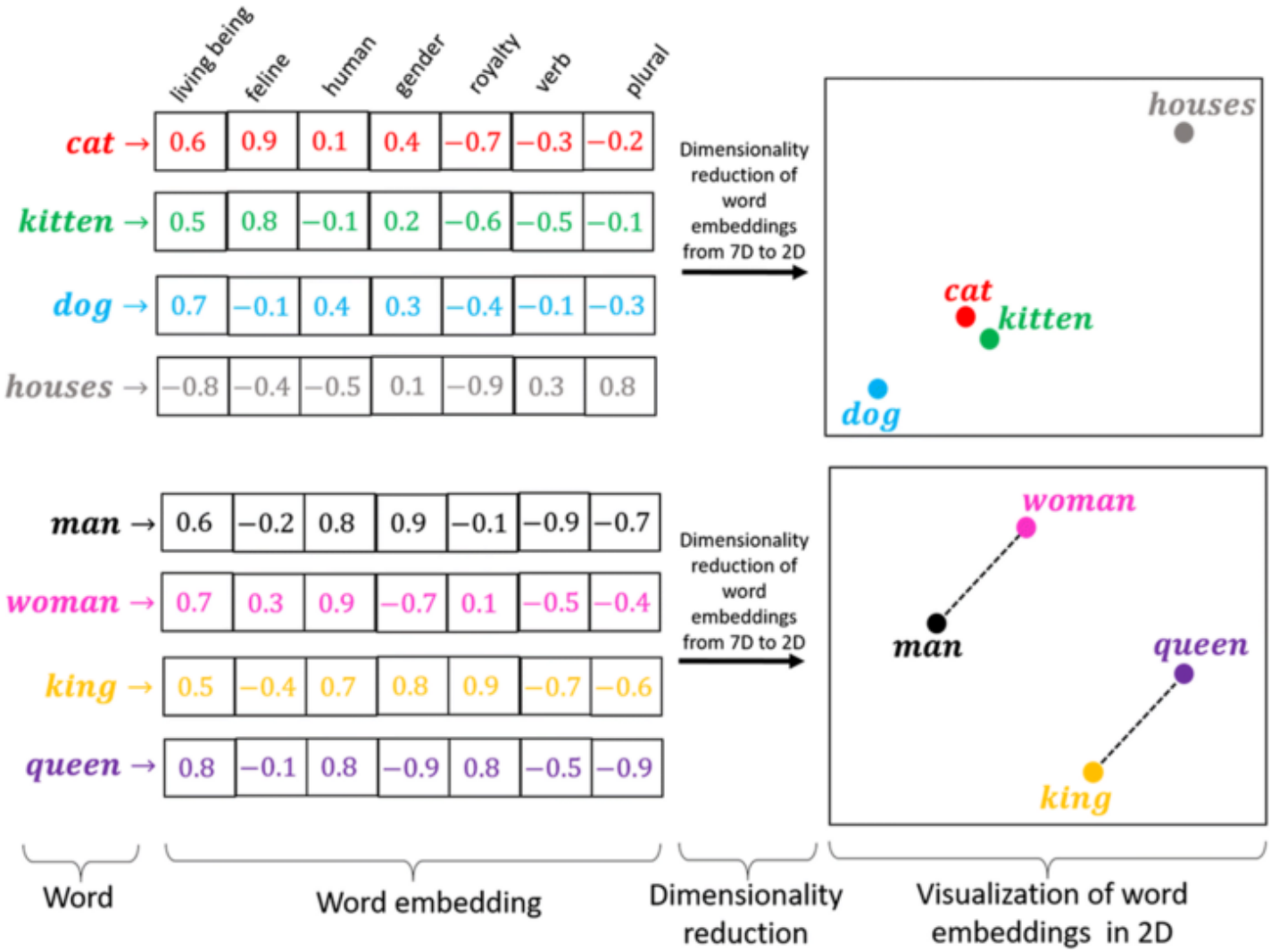
Wor2Vec, text metinlerinin numerik olarak temsil dilebilmesi için kullanılan bir yöntemdir.

Bag of Words, TF-IDF gibi metodlardan farklı olarak çalışır.

Kelimelerin vektörel olarak temsil edilmesidir.

In [2]:
```
from IPython.display import Image
Image(filename='w2v1.png')
```

Out[2]:



Yukarıdaki görselin sol tarafında bu yöntemin 7 boyutlu bir uzayda nasıl temsil edilebileceği görülüyor:

Her kelime vektörü alt alta gelecek şekilde bir matris oluşturduğumuzu düşünürsek, her sütun bir kelimenin bir özelliğini temsil ediyor olacaktır.

Öreneğin **Cat** ve **Kitten** kelimeleri için:

- **living being**
- **feline (kedicil)**

özellikleri yakın değerlere sahiptir.

- **human**

özelliği insan olmadıkları için 0'a yakın bir değere sahiptir.

- **gender**

özelliği kelimeler cinsiyet belirtmediği için 0'a yakın değerlere sahiptir.

Bu kelimelere kıyasla **Dog** kelimesi:

- **living being** özelliği yüksek olurken, **feline** özelliği, kedi türü olmadığı için 0'a yakın bir değere sahiptir.

Houses kelimesi:

- **living being** özelliği için negatif bir değere sahiptir,
- **plural (çoğul kelime)** sütunu için yüksek bir değere sahiptir.

Hemen Altındaki Matrise Bakacak Olursak:

- **man** ve **woman** kelimeleri cinsiyet sütununda zıt değerlere sahiptir.
- **king** ve **queen** kelimeleri royalty sütunlarında yüksek değerlere sahiptir.
- Bütün kelimeler **living being** sütunu için yüksek değerlere sahiptir.

Görselin sağ tarafında ise bu 7 boyutlu vektör uzayının 2 boyutlu bir düzleme indirgenerek yapılan görselleştirmesi bulunuyor:

Sağ üst görselde görüldüğü gibi,

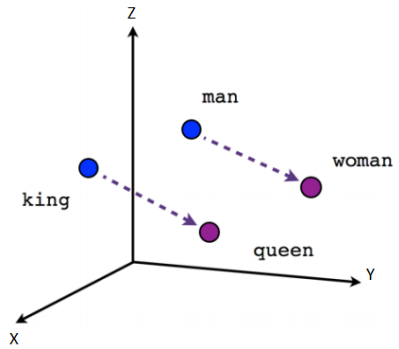
- **cat** ve **kitten** kelimeleri özellik biçiminden sadece kedi yaşına göre ayrılan kelimeler olduğu için birbirlerine çok yakın pozisyondalar.
- Bununla beraber **dog** kelimesi, bir canlı olduğu için (ve belki evcil hayvan olduğu için) **kedi** ve **kitten** kelimelerine görece yakın bir pozisyondadır, ancak tür olarak farklı bir hayvan olduğu için belli bir mesafeyi korumaktadır.
- **Houses** kelimesi ise diğerlerinde tamamen alakasız bir kelime olduğu için grafiğin görece en ters köşesindedir.

Sağ alttaki grafikte ise;

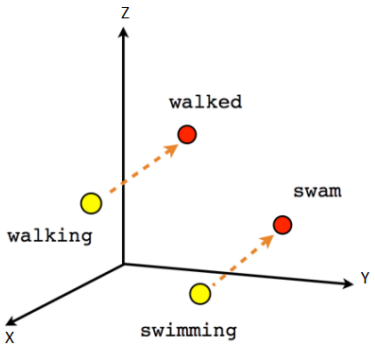
- **Man** ve **King** kelimeleri cinsiyet olarak aynı özelliğe sahip oldukları için, zıttı cinsiyet belirten kelimelere kıyasla aynı mesafeyi korumaktadırlar.

In [4]:
```
from IPython.display import Image
Image(filename='w2v2.png')
```

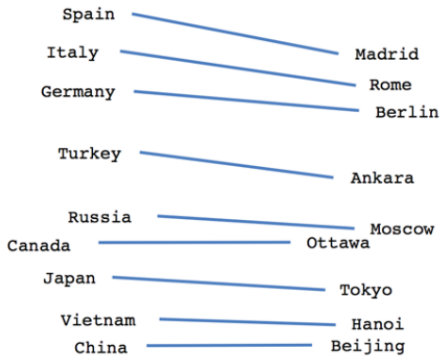
Out[4]:



Male-Female



Verb tense



Country-Capital

Benzer olarak bu grafikte 3 boyutlu bir zuaya indirgenmiş gröselleştirmeler görülebilir.

- Soldaki grafikte, **man** ve **woman** kelimeleri z ekseninde aynı yükseklikte, aynı şekilde **woamn** ve **queen** kelimleri de birbirleri ile aynı yükserkliktedirler. Bu sebeple, z ekseninin cinsiyet belirttiğini düşünebiliriz.
- Yukarıda da bahsedildiği gibi, benzer olarak, zıt cinsiyetleri belirten kelimelerY ekseninde birbirleri ile aynı mesafededir.
- X ekseninde ise **royalty** özelliğinin yerleşimleri görülüyor. Burada da **man** ve **king** ile **woman** ve **queen** birbirileri ile aynı uzaklıktadırlar.

Ortadaki grafikte ise farklı zaman belirten kelimelerin birbirlerine kıyasla ifade ettikleri anlamlara göre yerleşimleri görülmektedir.

Genel Nitelikler:

- Vektörlerin oluşturulması için cümlelerden oluşan bir **Corpus (Kelime Haznesi)**'a ihtiyaç vardır.
- Kelimelerin farklı cümlelerdeki kullanımlarına göre yakınlaklarını tespit eden bir algoritma kullanılır. Bu sayede her kelimeye bir vektör atanır.
- Kelime vektörleri kullanılan **Corpus**'a göre değişiklik gösterebilir. Örneğin Haber metinleri kullanılarak ahzırlanan bir Word2Vec modeli ile Tweetlerden veya Ansiklopedi verilerinden hazırlanan bir Word2Vec modeli, Örn. **Siyaset** kelimesine aynı vektörü atamayabilir. Örneğin, bir siyasi parti ismi Tweetler kullanılarak eğitilen bir modelde negatif veya pozitif kutuplu anlam çıkarabilir ve Ansiklopedi verilerinden hazırlanan bir Wor2Vec modeli aynı siyasi parti için Nötr anlam çıkarabilir.
- Vektör boyutları farklılık gösterilebilir. Word2Vec'te her kelime için Standart vektör boyutu 300 olarak kullanılır. Ancak istenildiği zaman değiştirilebilir. Bu vektörün boyutu, çoğunlukla eldeki verinin büyüklüğüne bağlıdır.
- Kelimlerin Vektörel olarak ifade edilmesinin en büyük avantajı, yukarıdaki örneklerde gördüğümüz gibi **cat** ve **kitten** gibi ifadelerin Kosinüs benzerliği yüksek vektörler ortaya çıkarmasıdır. Bu sayade bir yapay zeka modeli eğitilirken, model **cat** kelimesinden öğrendiği bilgiyi, **kitten** kelimesinde de uygulayabilir.

Sentence Embedding

Yukarıda açıklanan Word2Vec, bir Text Representation yöntemidir. Buna benzer bir çok farklı yöntem bulunmaktadır. Ancak bu yöntemlerin çoğu vektörleri kullanarak kelimleri/cümleleri temsil

etmektedir.

Word2Vec metodunda her kelime için farklı bir vektör oluşturuluyordu, **Sentence Embedding** ise bir cümlenin bütünün, kelime sayısı fark etmeksizin numerik olarak temsil edilmesidir. Genellikle tek bir vektörden oluşur.

Sentence Embedding'in önemi, cümlenin bağlamını(context) da içermesidir.

Örneğin, sesteş (eş sesli) bir kelime olan **Yüz** kelimesini ele aldığımızda, cümlede kuıllanımına göre bir *miktar* veya bir *organ* veya bir *fiil (yüzmek)* belirtebilir. Word2Vec modelinde, **Yüz** kelimesi için, modelin eğitildiği corpus'a (kelime haznesi) bağlı olarak bir vektör atanacaktır ve başka bir veride kullanıldığında, bağlam fark etmeksizin her zaman aynı vektöre sahip olacaktır.

- "Buraya **yüz** kere geldim"

ile

- "Kumsala yakın **yüz** dedim sana!"

cümlelerinde **Yüz** kelimesi farklı bir anlam belirtirken, Word2Vec modelinde aynı vektöre sahip olacaktır.

Sentence Embedding ise bu vektörlerin bağlamını (context) tespit edip birbirlerinden benzerliği olmayan (veya çok az olan) iki farklı vektör oluşturacaktır.

Bu durumu örnekler ile göstermek daha faydalı olacaktır:

In [1]:
```
from sentence_transformers import SentenceTransformer, util
import numpy as np
import os
import csv
import pickle
import time
model = SentenceTransformer('distiluse-base-multilingual-cased')
```

Yukarıda örneğini vertdiğmiz

- **Buraya yüz kere geldim**

ve

- **Kumsala yakın yüz dedim**

cümleleri için **Sentence Embedding** oluşturalım:

In [44]:
```
cümle_1 = model.encode("Buraya yüz kere geldim")
cümle_2 = model.encode("Kumsala yakın yüz dedim")
```

Bu modelde **Sentence Embnedding** için vektör uzunlukları 512 olarak belirlenmiştir.

In [45]:
```
print("Vektör Uzunluğu:",len(cümle_1))
print(cümle_1)
```

```
Vektör Uzunluğu: 512
[ 2.40900274e-02  1.19505841e-02  2.30297949e-02  8.87068547e-03
 -3.01123341e-03 -3.32172140e-02 -2.55911686e-02 -2.84639536e-03
 -5.80214672e-02  1.16706425e-02 -1.09724225e-02 -2.14965735e-02
 -2.01920699e-03  3.42204701e-03  7.62005290e-03  2.41464023e-02
  9.88670066e-03  2.35436261e-02  2.02406067e-02
```

```
-4.93273474e-02  3.18831950e-02 -1.24344928e-02  1.71466023e-02
 4.81941849e-02  4.08485122e-02 -1.03276698e-02 -2.85904724e-02
 2.83201388e-03 -1.18005546e-02 -5.13757169e-02  3.50094475e-02
 2.09531896e-02 -1.91960484e-02 -4.73403931e-03 -5.52146398e-02
-6.69175107e-03 -9.01556294e-03  2.56453045e-02  6.42870814e-02
 6.56180922e-03 -1.87192075e-02  2.25563720e-02 -4.28399350e-03
 3.35489251e-02  3.34759578e-02  6.23324746e-03 -1.64001286e-02
-2.52644066e-04  1.58078987e-02  2.25981721e-03 -5.46585815e-03
-1.34147555e-02 -2.75821351e-02 -1.59288086e-02  2.15599742e-02
-2.07840297e-02 -1.17642563e-02 -3.78805548e-02 -4.59833033e-02
 2.19559669e-03  2.02583950e-02 -3.83808762e-02 -5.19692823e-02
 4.56393734e-02 -2.58342922e-02  5.93757555e-02  5.34490729e-03
-4.57783602e-03 -1.36998761e-02 -6.09314023e-03  1.33667896e-02
 1.34603633e-02 -2.65577342e-04  1.67809296e-02 -2.08479054e-02
-2.17460860e-02 -1.54679203e-02 -6.86442554e-02 -4.52991314e-02
-3.00397654e-03 -1.51284710e-02  4.93012667e-02  6.35538921e-02
 2.03235690e-02  1.50408875e-02 -2.29673199e-02 -5.20997820e-03
 1.15218842e-02  4.86120256e-03 -9.10374336e-03  1.91896269e-03
 3.66594717e-02  3.35630961e-04  2.13802494e-02  6.61000656e-03
 4.69802320e-03 -9.23706125e-03  1.30896186e-02 -3.45228836e-02
-5.50784469e-02 -2.59500351e-02 -1.84984691e-02  2.64362544e-02
 6.80320896e-03  4.27296348e-02  2.08468991e-03 -1.61116458e-02
-2.09159544e-03 -2.10775342e-02  2.54462250e-02 -4.12322134e-02
-5.71270138e-02  2.74879169e-02 -1.10445045e-01  1.46277500e-02
 2.84059271e-02 -5.75602800e-03 -6.25964478e-02  5.93707338e-03
 2.04043277e-02  1.46008446e-03  1.70710701e-02  1.91819798e-02
 7.86897913e-03  3.90848331e-03 -1.57409180e-02  2.49378197e-02
-8.79051397e-04 -2.06641797e-02  2.13536415e-02  6.63169054e-03
 2.02013776e-02  7.75382761e-03  6.89602457e-03 -2.75991950e-02
-3.48018389e-03  2.31049284e-02 -1.86729878e-02  1.66046675e-02
 8.35868195e-02  2.40485650e-02 -4.18046676e-02 -3.15996371e-02
 1.27706435e-02 -1.80124342e-02  1.32619413e-02  3.43982689e-02
 2.67323535e-02  1.81915506e-03 -2.91686878e-02  2.57032271e-03
 1.06799025e-02  5.76899536e-02 -1.46420514e-02  7.84222502e-03
-1.14171878e-02  5.88225015e-02  1.25121009e-02 -5.04402816e-02
 2.34692469e-02  1.06083252e-01  1.11785857e-02 -4.23813015e-02
 6.46133441e-03 -1.71085149e-02  7.42653268e-04  3.84460352e-02
 3.70907015e-03  1.37391791e-03 -2.83818413e-02  4.61235084e-02
 6.21980242e-03  1.19454768e-02  2.35895645e-02 -6.51767850e-03
-2.94860937e-02  5.01187891e-03 -4.23242599e-02 -5.71588986e-03
 1.74756479e-02 -3.81036401e-02 -7.79538527e-02 -1.33912601e-02
 1.13306334e-02 -1.93929821e-02  5.50524108e-02 -1.86259653e-02
-2.45664101e-02 -7.70363118e-03  8.74715764e-03 -2.48034652e-02
 3.73536199e-02  3.96017507e-02 -3.14621790e-03  2.92441081e-02
-5.11951074e-02  1.49090169e-02 -5.32015860e-02  2.97975633e-03
 2.88240705e-02 -1.00333162e-01 -2.60741431e-02 -3.65557969e-02
 4.26166020e-02  1.01082232e-02 -1.73894782e-02 -1.58901345e-02
-2.28224602e-02 -2.35362314e-02  1.45675456e-02 -7.40945991e-03
-3.82107645e-02  1.97556168e-02  1.50290327e-02  3.80342081e-02
-1.36390068e-02 -1.53303356e-03 -2.19353777e-03  3.10723647e-03
 5.62288016e-02  6.66619167e-02  1.88726913e-02  2.86556575e-02
-4.65663057e-03 -5.26302606e-02  1.59642156e-02  3.16518135e-02
 1.99951250e-02  6.51072115e-02 -3.30082215e-02  2.18457524e-02
-4.55572084e-03 -3.30434255e-02 -1.79822594e-02 -9.31383297e-03
 2.22548079e-02 -1.79259609e-02 -3.46245468e-02  8.97982996e-03
-2.11946778e-02  4.17534448e-02 -8.65264889e-03 -4.48311530e-02
 1.12326220e-02  1.63106844e-02 -4.35006954e-02 -2.40382031e-02
-4.40918794e-03 -8.40807857e-04 -5.17854746e-03 -2.45911162e-02
 2.95906775e-02 -3.35962847e-02 -5.30050509e-02  4.64690337e-03
 1.30436532e-02 -1.94527209e-04  3.37423058e-04 -1.13272388e-02
-9.34171956e-03  2.17477698e-02 -9.22879030e-04 -3.31316665e-02
-5.41070895e-03 -2.67507415e-02 -2.56738644e-02 -5.90537675e-05
-2.54913568e-02 -4.67743054e-02  1.72965154e-02  1.37246093e-02
-7.50161242e-03  2.12583076e-02  2.15189792e-02 -4.12057620e-03
 2.64675412e-02  1.31225139e-02 -2.16916222e-02 -2.53018122e-02
-3.75798531e-02 -6.94770785e-03  2.70165168e-02 -3.12614068e-03
-3.30623984e-02  3.39122266e-02 -5.63484281e-02 -5.07499278e-02
-8.16203281e-03 -1.82959046e-02 -1.78233944e-02 -3.14008519e-02
 5.60090914e-02 -1.86409913e-02  1.19265234e-02  4.27996553e-02
                 2.94139106e-02  7.66699715e-03 -1.02539370e-02
```

```
-5.98079758e-03  6.79023787e-02  2.30086073e-02 -1.52109694e-02
-1.35958297e-02  3.88762192e-03 -5.20939045e-02 -1.00513678e-02
-1.83524638e-02 -1.79202426e-02 -3.43842581e-02  1.04015379e-03
 1.09481504e-02  3.50193232e-02  1.38559211e-02 -2.09327042e-03
 1.28513724e-02 -2.77588330e-02  3.64008243e-03 -2.58456822e-03
-3.94629985e-02 -8.28439277e-03 -4.78947023e-03 -1.40882423e-03
 1.93277393e-02 -1.24340747e-02 -5.55811857e-04  2.47043464e-02
-9.80751030e-03 -8.39347579e-03 -2.60346178e-02  7.22196475e-02
 7.17710471e-03 -4.33559604e-02  1.63247176e-02 -3.45839672e-02
-3.23116593e-02  7.42827123e-03 -4.69327113e-03 -7.18595609e-02
 1.14119891e-02 -2.22488493e-02 -1.80109330e-02  2.59914226e-03
-1.70614757e-03  5.26603172e-03 -5.65184513e-03 -4.21645399e-03
-6.37603865e-04 -2.17054393e-02 -3.71580049e-02 -1.62081104e-02
-4.05837037e-02 -2.47115549e-02 -4.06514071e-02 -3.58685181e-02
-2.32913587e-02 -1.27124805e-02  1.06630772e-02  1.89770237e-02
 9.88677051e-03 -1.31677799e-02 -1.13519281e-02 -2.45346911e-02
-2.30102353e-02 -1.79404337e-02 -7.83631392e-03  8.88516661e-03
 1.24351785e-03  1.76601838e-02 -1.37664648e-02 -8.69786367e-03
-4.41703424e-02  1.02809281e-04 -8.91173910e-03 -4.85877879e-03
 2.13427860e-02 -1.86798014e-02 -3.63325626e-02  1.82676762e-02
-4.26281169e-02  2.29349248e-02 -3.48594487e-02  2.30969526e-02
 2.34919246e-02 -7.05826655e-02  1.71081517e-02  2.63047852e-02
 1.06249088e-02 -2.38497183e-02 -7.91669078e-03 -4.27294075e-02
-1.80937592e-02  5.99522181e-02 -1.77837461e-02 -2.19085091e-03
 2.00691796e-03  1.89012904e-02  5.00069894e-02  1.58855077e-02
 3.49102020e-02 -4.65539721e-04  2.88394373e-02  1.69144925e-02
 3.62231508e-02  1.47856548e-02 -4.76680882e-02  2.05249544e-02
-6.36844477e-03 -2.45049107e-03 -6.50174618e-02  2.66842842e-02
-2.92182341e-02  1.14641134e-02  7.23495753e-03  2.69343052e-02
 3.17915492e-02 -4.66927923e-02 -3.00585292e-02 -3.35257500e-02
-3.58615327e-03  1.22229364e-02 -1.55391423e-02  1.35827379e-03
 3.36678810e-02 -8.94002803e-03 -4.28781919e-02 -4.05596048e-02
 4.89472691e-03 -5.86138014e-03 -3.01375836e-02  3.14018829e-03
 1.86895765e-02 -6.77238312e-03  6.68557221e-03 -2.05480549e-02
 3.65676135e-02  4.19288408e-03  2.40621399e-02 -1.27388369e-02
 1.29038217e-02  2.51556616e-02 -5.08940732e-03 -4.31576138e-03
 2.92858202e-02 -7.94319715e-03 -2.94806734e-02  2.60256859e-03
 5.86557714e-03  3.49430144e-02  1.18073088e-03 -1.55569520e-02
 4.04484197e-02  3.79238538e-02 -3.54626961e-02 -4.20616046e-02
 2.24512592e-02  3.62643450e-02  2.43666247e-02 -1.73114873e-02
-1.61283440e-03  1.50607387e-02  6.71120081e-03  5.12453727e-03
-2.12436523e-02  4.73398380e-02 -6.53607305e-03 -4.12681289e-02
 2.77876072e-02  3.51818576e-02 -1.73379313e-02  9.15566366e-03
-5.91321699e-02 -1.12758474e-02  1.90226436e-02  9.83744394e-03
-4.55000252e-02  2.77515734e-04 -6.86865719e-03  1.60268582e-02
-3.39720659e-02  2.17783395e-02  9.63545293e-02  3.57002504e-02
-2.81308703e-02  1.30292065e-02 -2.95556784e-02 -3.78856175e-02
-1.96459629e-02  8.42954218e-03  1.05170701e-02 -5.72045259e-02
 1.79564990e-02 -3.60015184e-02  1.02039920e-02  4.76389378e-03
-3.45361885e-03 -1.17568746e-02 -2.05093715e-02  1.40215000e-02
 1.78733170e-02 -3.68518941e-02  1.03771567e-01  2.13257242e-02
-2.33424976e-02  5.87667990e-03  2.38612760e-02  2.62004212e-02
 1.72066335e-02 -1.71630401e-02 -2.46229097e-02 -3.32672633e-02]
```

In [46]:
```
# İki vektörün birbirine uzaklığını hesaplamak için
from scipy import spatial
```

In [47]:
```
# Kosinüs Benzerliği = 1 - iki vektörün birbirine uzaklığı
1 - spatial.distance.cosine(cümle_1, cümle_2)
```

Out[47]:
```
0.3507169485092163
```

Bu iki cümle için oluşturulan vektörlerin kosinüs benzerliği %35'tir ve düşük bir orandır.

Başka bir örnek için;

- **Kedi ağaçtan atlayınca ayağı kırılmış**

ve

- **Kaplan ağaçtan atlayınca ayağı kırılmış**

cümlelerini ele alalım:

In [48]:
```
cümle_1 = model.encode("Kedi ağaçtan atlayınca ayağı kırılmış")
cümle_2 = model.encode("Kaplan ağaçtan atlayınca ayağı kırılmış")
```

In [49]:
```
# Kosinüs Benzerliği = 1 - iki vektörün birbirine uzaklığı
1 - spatial.distance.cosine(cümle_1, cümle_2)
```

Out[49]:
```
0.8770985007286072
```

Kedigil familyasından iki hayvan için aynı durum değerlendirildiğinde vektörlerin benzerliği %87.7 oranında epey yüksek çıkıyor.

Aynı durumu Kedi-Köpek karşılaştırması olarak yapacak olursak:

In [52]:
```
cümle_1 = model.encode("Kedi ağaçtan atlayınca ayağı kırılmış")
cümle_2 = model.encode("Köpek ağaçtan atlayınca ayağı kırılmış")

# Kosinüs Benzerliği = 1 - iki vektörün birbirine uzaklığı
1 - spatial.distance.cosine(cümle_1, cümle_2)
```

Out[52]:
```
0.7502996325492859
```

Bu sefer benzerlik görece daha düşük çıkıyor ancak yine de bağlam (context) aynı olduğu için yüksek bir değer çıkıyor.

*Not: %70 üzeri değerler genellikle yüksek olarak değerlendirilir ancak **cutoff** noktası için belirlenecek değer eldeki veriye göre farklılık gösterebilir.*

Sentence Transformer

Yukarıda örnek verilen model **sentence transformer** kütüphanesinden gelmektedir.

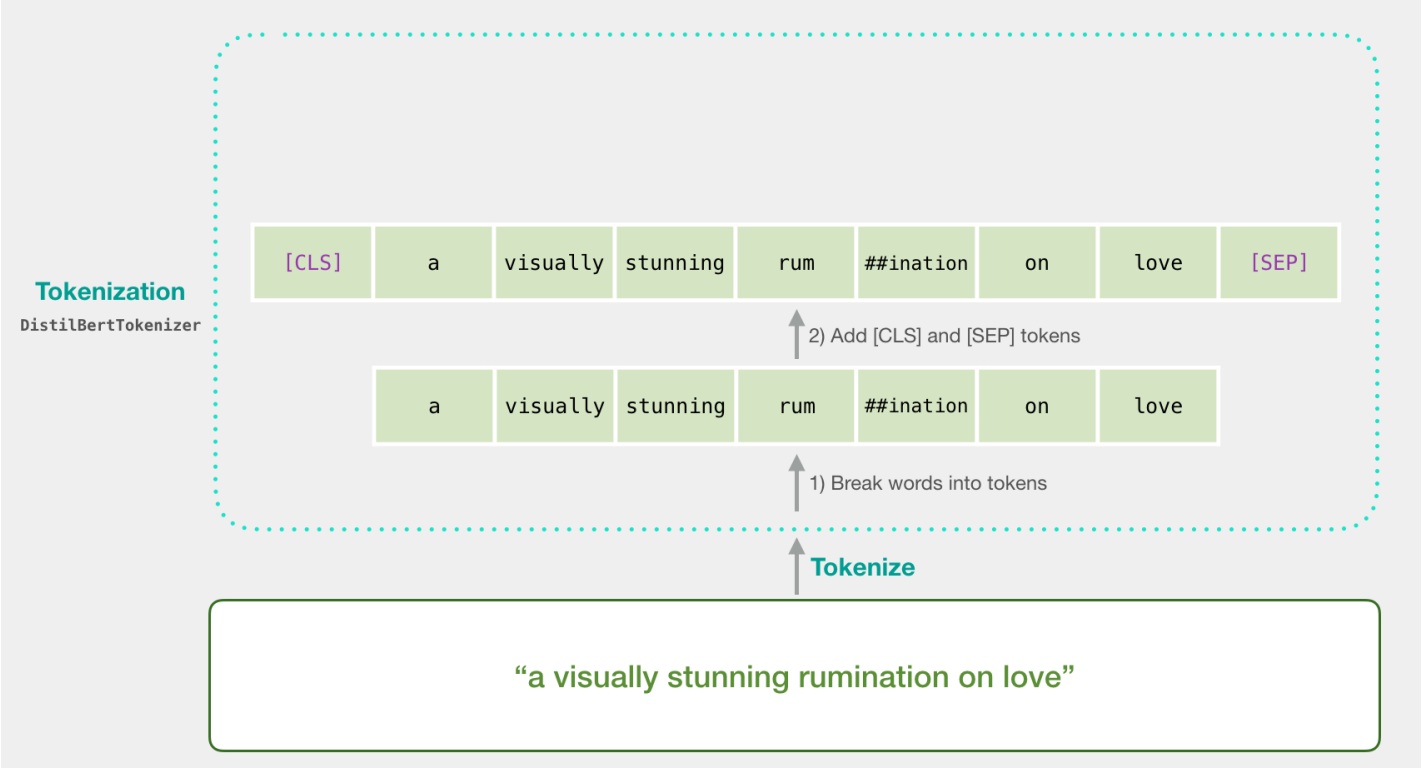
Kullanılan bu metod, Bert modelinden türemiştir.

Bert modeli, Doğal Dil İşleme alanında kullanılmak için geliştirilen, State-of-the-Art bir dil modelidir.

Burada Bert modelinin detayları anlatılmayacaktır. Ancak Bert için kullanılan **Word Representation** metodu üzerinden birkaç açıklama yapılacaktır. Bert için kullanılan vektörel temsil etme yönteminden **Tokenizer** olarak bahsedilecektir.

In [21]:
```
from IPython.display import Image
Image(filename='bert-distilbert-tokenization-1.png')
```

Out[21]:



Yukarıdaki görselde görüldüğü gibi, Bert Tokenizer'ı kelimeleri vektörlere çevirmeden önce, yukarıdaki **rumination** kelimesinde görüldüğü gibi kelimeleri **Token**lere ayırmaktadır.

Bunun sebebi Bert Tokenizer'ının sabit bir **corpus** kullanılarak eğitilmesidir. Bu sabit **corpus** Wikipedia ve Haber metinleri gibi milyonlarca cümleden oluşan çok geniş bir corpus'tur.

Bu tokenizer için kelime haznesi olarak bahsedilen **corpus**taki en sık kullanılan 50.000 kelime kullanılmıştır. Bu yüzden bir cümleyi Bert Tokenizer ile işlemek istediğimizde, eğer cümle içerisindeki kelimelerden biri, tokenizer'ın kelime haznesinde mevcut değilse, bu kelimeyi, kelime haznesinde mevcut olan diğer keliemelere bölerek tokenler oluşturmaktadır.

Buna ek olarak Bert modeli, yukarıdaki görselde görüldüğü gibi özel bir token olan **[CLS]** tokenini kullanmaktadır. Tokenizer, bu tokeni cümlenin bağlamına (context) göre her cümle için özel olarak oluşturmaktadır.

Bu **[CLS]** tokeni, tek başına bir **Sentence Embedding** vektörü olarak düşünülebilir. Ancak Bert Tokenizer'ı, algoritmasının eşsiz yapısı nedeniyle bir cümleye her zaman aynı **[CLS]** tokenini vermemektedir.

Yukarıda kullandığımız **Sentence Transformer** kütüphanesi ise, Bert ve benzeri modellerin özel **[CLS]** tokenlerinin, her defasında aynı vektörü oluşturması için **Fine-Tune*** edilmiş modeller içermektedir.

**Fine-Tune işlemi, bir modelin spesifik bir task (görev) için eğitilmesidir.*

*Not: Yuakrıdaki görseldeki **[SEP]** tokeni, cümlenin bittiğini ifade eden bir başka özel tokendir*

Bu metodun ilan verilerine uygulanması:

Verimizde bulunanlar ilanlar birden fazla cümleden oluşmaktadır, ancak kullanacağımız model, tek seferde *maksimum* 128 token kullanarak **Sentence Embedding** oluşturabilmektedir. 10 kelimelik

bir cümle minimum 10 tokenden oluşacaktır (her cümle bir token olur) veya kelime haznesinde olmayan kelimeler varsa 10'dan fazla tokene bölünecektir.

Bu yüzden ilanlar her cümle ayrı bir girdi olacak şekilde ayrılmıştır. Ardından her cümlenin **Sentence Embedding**'i oluşturulmuş ve aynı ilana tekabül eden cümlelerin **Sentence Embedding Vektörlerinin** ortalaması alınmıştır. İlanlar vektörel olarak bu şekilde temsil edilmiştir.

Ardından 499 ilan için oluşturulan 499 vektör için hiyerarşik kümeleme yöntemi kullanıldı. Bu sayede, benzer özellikleri (boyutları) içeren vektörlerden birer küme oluşturulmuş ve ilanlar toplamda 10 küme olacak şekilde gruplandırılmıştır.

Bu yöntem, özellikle vektörlerin ortalamasının alınması durumu, oldukça deneysel bir yöntem olup kesin bir geçerliliği yoktur. Bu projede ilanları gruplamak istediğim için, bu gruplamayı el ile yapmak yerine daha ilerici bir yöntemi denemek istedim. Yöntemin geçerliliğinin tespit edilmesi, ne kadar iyi gruplama yaptığı, aşağıdaki analizlerde daha net bir şekilde görülecektir.

Örnek veri aşağıdaki şekildedir:

In [23]:
```
import pandas as pd
df = pd.read_excel("ilan_metinDF.xlsx")
```

In [24]:
```
df
```

Out[24]:

	ilan_id	pozisyon	pozisyon_kisa	cumleler
0	1	Fiyat Araştırma Proje Uzman Yardımcısı	veri analisti	genel nitelikler ve is tanimi
1	1	Fiyat Araştırma Proje Uzman Yardımcısı	veri analisti	universitelerin ekonometri istatistik matemati...
2	1	Fiyat Araştırma Proje Uzman Yardımcısı	veri analisti	alanında en az 1 yıl is deneyimine sahip
3	1	Fiyat Araştırma Proje Uzman Yardımcısı	veri analisti	ms office uygulamalarına ozellikle ileri derec...
4	1	Fiyat Araştırma Proje Uzman Yardımcısı	veri analisti	tercihen iyi derecede ingilizce bilen
...	...	...	...	...
10170	499	Veri Bilimci	veri analisti	makine ogrenmesi yontemleri alanında modelleme...
10171	499	Veri Bilimci	veri analisti	acık kaynak yazılım teknolojilerini ve literat...
10172	499	Veri Bilimci	veri analisti	literatur takip edebilecek seviyede ingilizce ...
10173	499	Veri Bilimci	veri analisti	analitik dusunen sonuc odaklı calısan ve takım...
10174	499	Veri Bilimci	veri analisti	istanbul' da ikamet eden

10175 rows × 4 columns

Verideki **cumleler** sütunundaki her cümle için **Sentence Embeddinng** oluşturuluyor:

In [25]:
```
corpus_embedding = model.encode(df["cumleler"], show_progress_bar=True, convert_to_numpy=
```

Ardından verideki **ilan_id** sütunu kullanılarak her vektör ait olduğu ilana endekslenmiştir ve ilan için oluşturulan vektörlerin ortalaması alınmıştır.

In [26]:
```
ilan_dict = {}

for id in df["ilan_id"].unique():
    ilan_dict["{}".format(id)] = corpus_embedding[list(df[df["ilan_id"]==id].index)]
```

In [36]:
```
averages = []
for id in ilan_dict:
    avg = np.add.reduce(ilan_dict[id])/len(ilan_dict[id])
    averages.append(avg)
```

Vektörlerin kümelenmesi için Python'un SciKit-Learn kütüphanesinde bulunan **Agglomerative Clustering (Hiyerarşik Kümeleme)** modeli kullanılmıştır.

Bununla beraber **K-Means** yöntemi de denendi ancak **Hiyerarşik Kümeleme** yönteminin daha iyi çalıştığı gözlendi.

In [37]:
```
from sklearn.cluster import AgglomerativeClustering
```

Cluster sayısı olarak **10** seçildi.

Yöntem olarak **euclidean** ve **Ward Linkage** seçildi.

In [38]:
```
%%time
cluster = AgglomerativeClustering(n_clusters=10,affinity="euclidean",linkage="ward")

# iterasyondaki veri için fit ediyoruz
cluster.fit(averages)
```

```
Wall time: 264 ms
```

Out[38]:
```
AgglomerativeClustering(n_clusters=10)
```

Her vektörün (ilanın) ait olduğu küme:

In [39]:
```
cluster.labels_
```

Out[39]:
```
array([0, 1, 0, 1, 1, 8, 5, 8, 8, 1, 7, 3, 1, 7, 0, 0, 1, 0, 0, 1, 1, 3,
       1, 1, 1, 0, 0, 7, 7, 5, 1, 7, 6, 0, 7, 0, 7, 7, 7, 0, 1, 7, 1, 1,
       7, 0, 0, 0, 1, 0, 7, 0, 7, 0, 1, 7, 7, 7, 7, 7, 0, 7, 6, 7, 7, 2,
       2, 2, 0, 1, 7, 7, 1, 1, 0, 0, 3, 7, 7, 1, 7, 7, 7, 1, 7, 7, 7, 7,
       4, 3, 1, 5, 1, 1, 7, 1, 7, 1, 0, 7, 1, 1, 7, 2, 2, 7, 7, 2, 0, 7,
       7, 7, 2, 6, 4, 7, 1, 8, 1, 0, 1, 9, 9, 3, 3, 6, 3, 1, 3, 3, 3, 3,
       1, 3, 0, 0, 9, 1, 3, 8, 8, 0, 8, 8, 8, 8, 8, 2, 8, 0, 2, 8, 8, 8,
       5, 0, 1, 1, 8, 1, 8, 8, 2, 2, 8, 2, 0, 4, 0, 3, 4, 6, 7, 8, 2, 5,
       0, 1, 1, 4, 4, 0, 0, 1, 4, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 9, 1, 0,
       6, 0, 0, 0, 1, 0, 2, 0, 0, 3, 3, 0, 3, 0, 3, 0, 1, 7, 0, 0, 4, 1,
       8, 0, 0, 8, 3, 0, 0, 8, 7, 8, 8, 8, 8, 8, 0, 0, 8, 1, 8, 8, 9, 0,
       9, 8, 8, 0, 8, 0, 3, 8, 8, 0, 8, 8, 8, 8, 8, 0, 8, 0, 1, 1, 0, 0,
       5, 0, 1, 9, 0, 5, 2, 0, 0, 0, 7, 1, 3, 8, 5, 2, 0, 0, 0, 0, 0, 1,
       3, 0, 0, 0, 3, 3, 1, 0, 1, 0, 1, 3, 1, 3, 3, 1, 1, 0, 3, 1, 6, 0,
       1, 5, 3, 2, 3, 1, 3, 2, 0, 0, 0, 2, 2, 6, 8, 0, 0, 2, 0, 0, 3, 0,
       0, 0, 3, 0, 1, 1, 0, 1, 0, 0, 0, 3, 0, 7, 3, 3, 8, 1, 3, 0, 0, 3,
       1, 3, 3, 0, 3, 3, 9, 6, 9, 1, 4, 4, 4, 4, 4, 4, 4, 8, 4, 2, 2, 4,
       2, 3, 3, 0, 0, 0, 0, 3, 3, 3, 3, 0, 6, 3, 6, 3, 3, 3, 3, 3, 3, 3,
       3, 3, 3, 3, 3, 3, 3, 3, 1, 0, 0, 0, 1, 0, 3, 0, 0, 3, 0, 0, 3, 3,
       6, 3, 3, 3, 1, 0, 0, 8, 8, 5, 3, 3, 3, 3, 3, 3, 0, 0, 4, 1, 9, 6,
       1, 0, 0, 2, 3, 3, 9, 7, 0, 0, 0, 0, 5, 7, 0, 5, 1, 7, 1, 0, 3, 2,
       8, 1, 3, 0, 5, 0, 1, 1, 0, 0, 8, 8, 1, 1, 1], dtype=int64)
```

Bu işlemin ardından orjinal verimize geri dönüyoruz ve ilanların hangi gruba (kümeye) ait yeni bir sütun oluşturuyoruz:

In [40]:
```
import pandas as pd
df_main = pd.read_excel("mainnn.xlsx")
```

In [41]:
```
df_main["cluster"] = list(cluster.labels_)
```

İlanlardaki pozisyon isimleri kullanılarak kümelerin incelenmesi:

In [43]:
```
for küme_id in df_main["cluster"].unique():
    print("Küme {}:".format(küme_id+1))
    print(df_main.loc[df_main['cluster'] == küme_id]["pozisyon"])
    print()
```

```
Küme 1:
0            Fiyat Araştırma Proje Uzman Yardımcısı
2                         Aktüerya Kıdemli Uzmanı
14                                    CRM Uzmanı
15                          Dijital Dönüşüm Müdürü
17                              Veri Analiz Uzmanı
                           ...
481    Ticaret ve Piyasa İşlemleri Uzman Yardımcısı
487          Operasyonel Mükemmellik Kıdemli Uzmanı
489                            Proje Yönetim Uzmanı
492          Stratejik Planlama ve Raporlama Uzmanı
493            Kurumsal Verimlilik ve Analiz Müdürü
Name: pozisyon, Length: 153, dtype: object

Küme 2:
1              Aktüeryal Raporlama Uzmanı
3      Aktüeryal Rezerv Müdür Yardımcısı
4                 Aktüeryal Rezerv Uzmanı
9                        Raporlama Uzmanı
12                          Veri Analisti
                      ...
490                         Data Analisti
491      Master Data Veri Analizi Uzmanı
496                        Data Scientist
497                          Data Analyst
498                          Veri Bilimci
Name: pozisyon, Length: 83, dtype: object

Küme 9:
5                               Banka Yetkilisi/Yönetmeni
7                                       Raporlama Uzmanı
8                                   Ürün Maliyet Analisti
117                 Bireysel Tahsis ve Analiz Yöneticisi
139           Bütçe ve Raporlama Uzmanı/ Kıdemli Uzmanı
140                                        Finans Uzmanı
142                                    Finans Yöneticisi
143                             Bütçe ve Raporlama Uzmanı
144           Bütçe ve Raporlama Yöneticisi / Yönetmeni
145                                        Finans Uzmanı
146                             Bütçe ve Raporlama Uzmanı
148                              Bütçe ve Planlama Uzmanı
151                                 Kıdemli Finans Uzmanı
152                               Bütçe Raporlama Uzmanı
153                               Bütçe Raporlama Uzmanı
158                             Bütçe ve Raporlama Uzmanı
160                   Bütçe, Raporlama ve Kontrol Uzmanı
161                      Bütçe, Raporlama ve UFRS Uzmanı
164                             Bütçe ve Raporlama Uzmanı
173                             Bütçe ve Raporlama Uzmanı
242                                  Kıdemli Bütçe Uzmanı
245                   Maliyet Raporlama ve Analiz Uzmanı
249                       Maliyet Analiz - SAP CO Uzmanı
251                              Bütçe ve Raporlama Şefi
252                            Maliyet Analiz Sorumlusu
253                 Maliyet Kontrol Şefi (Cost Controller)
                                Maliyet Muhasebesi Uzmanı
```

```
255                            Bütçe ve Raporlama Uzmanı
258                                      Finans Uzmanı
260                                    Muhasebe Uzmanı
261                                Genel Muhasebe Uzmanı
265                           Bütçe ve Raporlama Sorumlusu
266                                    Muhasebe Uzmanı
268                           Mali İşler ve Finans Asistanı
271                        Bütçe Raporlama Uzmanı/Sorumlusu
272                                      Finans Uzmanı
274                                    Muhasebe Uzmanı
275                           Bütçe ve Mali Kontrol Uzmanı
276                           Maliyet Muhasebesi Yetkilisi
277                                    Muhasebe Müdürü
278                                  Mali İşler Uzmanı
280                                  Mali İşler Uzmanı
299                               Gelir Yönetimi Yönetmeni
344                  Bütçe Planlama Yöneticisi / Yönetmeni
368    Bütçe Planlama ve Yönetim Raporlama Yönetmen Yrd.
391             Masak Uyum Görevlisi & İç Kontrol Uzmanı
447                                      Satış Mühendisi
448                                Satış Destek Sorumlusu
484                                    Analist Yöneticisi
494                         Maliyet Kontrol ve Bütçe Uzmanı
495                               Resmi Raporlama Yöneticisi
Name: pozisyon, dtype: object

Küme 6:
6                         Veri Analiz Uzmanı
29           AR-GE Merkezi Süreçler Uzmanı
91                             Veri Analisti
154    Bütçe ve Raporlama Kıdemli Uzmanı
175      Hazine Dealer/Dealer Yardımcısı
286                     Endüstri Mühendisi
291     Forex Yatırım Uzmanı (Retention)
300                       Raporlama Uzmanı
331               Dijital Pazarlama Uzmanı
449                     İstatistik Elemanı
474                     Veri Tabanı Uzmanı
477                  Yazılım Destek Uzmanı
488              Planlama Uzman Yardımcısı
Name: pozisyon, dtype: object

Küme 8:
10                             Yazılım Geliştirme Mühendisi
13                             Analiz ve Entegrasyon Uzmanı
27                                           İş Analisti
28                                           İş Analisti
31                               Proje Analiz Uzmanı-Remote
34                             UX Kullanıcı Deneyimi Uzmanı
36                  Kredi Uygulamaları Yazılım Test Uzmanı
37               Hazine ve Dış Ticaret Yazılım Test Uzmanı
38                Yazılım Test Uzmanı- Katılım Bankacılığı
41                                        Yazılım Uzmanı
44                                            ERP Uzmanı
50                                     Java Yazılım Uzmanı
52                                           İş Analisti
55                                     Yazılım Test Uzmanı
56                                           ERP Sorumlusu
57                          .Net Uygulama Geliştirme Uzmanı
58               İş Analiz Uzman Yrd./Uzmanı/Kıdemli Uzman
59     Yazılım Geliştirme Uzman Yrd./Uzmanı/Kıdemli U...
61                                        SAP Yazılım Uzmanı
63                               Bilgi Teknolojileri Uzmanı
64                                           İş Analisti
70                                     Proje Analiz Uzmanı
71                                           İş Analisti
77                                     Yazılım Uzmanı
78                                           İş Analisti
80                                   Teknik Destek Uzmanı
```



```
                                       Eba Yazılım Uzmanı
```

```
82                                        İş Analisti
84                      İş Analisti (Sağlık Uygulamaları)
85                         İş Analisti (ERP Uygulamaları)
86                      İş Analisti (Eğitim Uygulamaları)
87                                      Yazılım Uzmanı
94              Bilgi Teknolojileri İş Uygulamaları Uzmanı
96                        Kanal Deneyimi ve Kontrol Uzmanı
99                                      Yazılım Uzmanı
102                                        İş Analisti
105                     Sistem Analisti (Core Bussiness)
106                                        İş Analisti
109                                     IT İş Analisti
110                               Kıdemli IT İş Analisti
111                              SAP ABAP Yazılım Uzmanı
115                                İş Analisti ve Tester
172       Kobi Bankacılığı Paz. Proje Yönetimi Yetkilisi
237                                  Yazılım Test Uzmanı
250                           Teşvik Süreçleri Yöneticisi
296                                        İş Analisti
365                                Üretim Planlama Uzmanı
469      Tedarik Zinciri İş Geliştirme Uzmanı/Uzman Yrd.
475    İş Analisti (Dijital Kanallar ve Mobil Deneyimli)
479                         Yazılım Test Uzmanı-Bankacılık
Name: pozisyon, dtype: object

Küme 4:
11                          Gelir Yönetimi Uzmanı
21                                  Gelir Uzmanı
76                       Adwords Hesap Yöneticisi
89                          Yazılım Destek Uzmanı
123                              E-Ticaret Uzmanı
                           ...
455                         Satış Operasyon Uzmanı
466                    Kategori Uzman Yardımcısı
467             Ürün Planlama Yöneticisi/Uzmanı
482                           Satın Alma Uzmanı
486    Satınalma ve Malzeme Planlama Sorumlusu
Name: pozisyon, Length: 81, dtype: object

Küme 7:
32                               Teknik Satın Alma Uzmanı
62                               Süreç İyileştirme Uzmanı
113    Kurumsal Sistemler Veri Ambarı ve Raporlama Uz...
125                  Kıdemli Performans Pazarlama Uzmanı
171                   Kurumsal ve Ticari Krediler Uzmanı
220     Kobı Pazarlama & İş Geliştirme Yetkili/Yön. Yrd.
328                         Müşteri Değer Yönetimi Uzmanı
343                Analiz ve Raporlama Uzmanı / Sorumlusu
381                   Satış Analiz ve Raporlama Sorumlusu
408                                 Satın Alma Sorumlusu
410                           Teknik Satın Alma Sorumlusu
440                                    İş Zekası Uzmanı
461                             Analiz ve Raporlama Uzmanı
Name: pozisyon, dtype: object

Küme 3:
65                               Yazılım Teknik Danışmanı
66              İş Analisti (Tahsilat, Üretim Fonksiyonu)
67                         İş Analisti (Sağlık Fonksiyonu)
103                 İş Analiz Uzman Yardımcısı ( SAP MM )
104        İş Analizi Uzman Yardımcısı (SAP SD / Retail)
107                                        İş Analisti
112    Mobil Uygulamalar Ürün Yöneticisi (Product Owner)
147     Topkapı Tüzel Bank. Ticari Portf. Yönetmeni/Yrd.
150          Müşteri Deneyimi Yöneticisi (6 Ay Dönemsel)
162     İştirakler ve Koç Holding Finansal Şirket Uzmanı
163                 Finansal Raporlama Yetkili Yardımcısı
165                                   Yaptırımlar Uzmanı
174                      Hazine Satış Yöneticisi/Yetkilisi
                 e Bankacılığı Kampanya Yönetimi Stajyeri
```

```
292                                  Portföy Yöneticisi
301                       Kargo Operasyonları Yöneticisi
333                                           CRM Uzmanı
337    Veri Analizi & Raporlama Uzmanı (6 Aylık Dönem...
341          Tarım Bankacılığı Segment Yönetimi Stajyeri
342                     Kartlı Ödeme Sistemleri Stajyeri
347                            Perakende Planlama Uzmanı
393     Ticari Kredi Risk Kontrol ve Koordinasyon Uzmanı
394     Mali Suçları Önleme Mevzuatı ve Raporlama Uzmanı
396    Karşı Taraf Kredi Riski ve Teminat Yönetimi Uz...
465                       Stok Planlama Uzman Yardımcısı
483                Ticari Kredi Risk Stratejileri Uzmanı
Name: pozisyon, dtype: object

Küme 5:
88                                BT Risk ve Uyum Uzmanı
114                      Bilgi Güvenliği Süreç Danışmanı
167                         Kredi Riski Modelleme Uzmanı
170         Büyük Ölçekli Krediler İdari ve Kanuni Takip
179                                      Risk Yöneticisi
180                                             Denetmen
184                      Bilgi Sistemleri Denetim Uzmanı
240                  Kobi Kredi Tahsis Yetkilisi - Uzmanı
384                            Risk Yönetimi Süpervizörü
385                                 Risk Yönetimi Uzmanı
386    Kredi Risk Metodolojileri ve Validasyon Yöneti...
387                                      Risk Yöneticisi
388                       Risk Yönetimi Müdür Yardımcısı
389                       Risk ve Uyum Birimi Yöneticisi
390                                Risk Yönetim Uzmanı
392                    Kredi Riski ve Modelleme Yetkilisi
395                                Risk Yönetim Müdürü
458                               Uyum ve Mevzuat Uzmanı
Name: pozisyon, dtype: object

Küme 10:
121                      Bütçe ve Mali Analiz Uzmanı
122                            Dış Ticaret Sorumlusu
136                       Ölçme Değerlendirme Uzmanı
217                                     İş Analisti
262                        Maliyet Muhasebesi Uzmanı
264                        Maliyet Muhasebesi Uzmanı
289                                     İş Analisti
380                                Planlama Elemanı
382    Medya Planlama ve Raporlama Uzman Yardımcısı
460             Veri Merkezi İzleme ve Destek Uzmanı
468              Tedarik Zinciri Yöneticisi / Uzmanı
Name: pozisyon, dtype: object
```