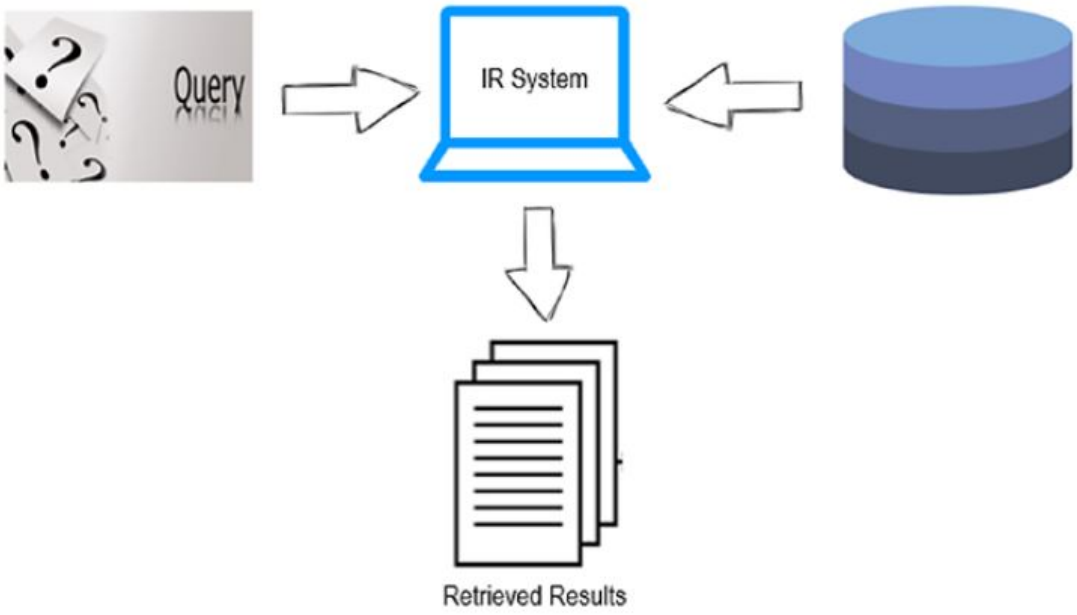
In [1]:
```
from IPython.display import Image
Image("C:\\Users\\ShahinN\\Desktop\\123423.JPG")
```

Out[1]:



ساخت / وارد کردن اسناد

In [2]:
```
Doc1 = ["With the Union cabinet approving the amendments to the Motor Vehicl
Doc2 = ["Natural language processing (NLP) is an area of computer science an

Doc3 = ["He points out that public transport is very good in Mumbai and New

Doc4 = ["But the man behind the wickets at the other end was watching just a
```

In [3]:
```
# Put all the documents in one list
fin= Doc1+Doc2+Doc3+Doc4
```

وارد کردن کتابخانه ها

In [4]:
```
1
2  import gensim
3  from gensim.models import Word2Vec
4  import numpy as np
5  import nltk
6  import itertools
7  from nltk.corpus import stopwords
8  from nltk.tokenize import sent_tokenize, word_tokenize
9  import scipy
10 from scipy import spatial
11 from nltk.tokenize.toktok import ToktokTokenizer
12 import re
13 tokenizer = ToktokTokenizer()
14 stopword_list = nltk.corpus.stopwords.words('english')
```

```
C:\Users\ShahinN\Anaconda3\lib\site-packages\gensim\utils.py:1197: UserWarning:
detected Windows; aliasing chunkize to chunkize_serial
  warnings.warn("detected Windows; aliasing chunkize to chunkize_serial")
```

بارگذاری مدل

In [6]:
```
1 model = gensim.models.KeyedVectors.load_word2vec_format('C:\\Users\\ShahinN\
2
```

ساخت سیستم IR

In [7]:
```
1  #Preprocessing
2  def remove_stopwords(text, is_lower_case=False):
3      pattern = r'[^a-zA-z0-9\s]'
4      text = re.sub(pattern, "", ",".join(text))
5      tokens = tokenizer.tokenize(text)
6      tokens = [token.strip() for token in tokens]
7      if is_lower_case:
8          filtered_tokens = [token for token in tokens if token not in stopwor
9
10     else:
11         filtered_tokens = [token for token in tokens if token.lower() not in
12
13     filtered_text = ' '.join(filtered_tokens)
14     return filtered_text
```

In [8]:
```
# Function to get the embedding vector for n dimension, we have used "300"
def get_embedding(x):
    if x in model.wv.vocab:
        return model[x]
    else:
        return np.zeros(300)

```

In [9]:
```
out_dict = {}
for sen in fin:
    average_vector = (np.mean(np.array([get_embedding(x) for x in nltk.word_

    dict = { sen : (average_vector) }
    out_dict.update(dict)
```

```
C:\Users\ShahinN\Anaconda3\lib\site-packages\ipykernel_launcher.py:3: Deprecati
onWarning: Call to deprecated `wv` (Attribute will be removed in 4.0.0, use sel
f instead).
  This is separate from the ipykernel package so we can avoid doing imports unt
il
```

In [10]:
```
def get_sim(query_embedding, average_vector_doc):
    sim = [(1 - scipy.spatial.distance.cosine(query_embedding, average_vecto
    return sim
```

In [11]:
```
# Rank all the documents based on the similarity to get
def Ranked_documents(query):
    emb = [get_embedding(x) for x in nltk.word_tokenize(query.lower())]
    query_words = (np.mean(np.array(emb,dtype=float), axis=0))

    rank = []
    for k,v in out_dict.items():
        rank.append((k, get_sim(query_words, v)))

    rank = sorted(rank,key=lambda t: t[1], reverse=True)
    print('Ranked Documents :')
    return rank
```

In [14]:
```
1 Ranked_documents("artificial intelligence")
```

```
Ranked Documents :

C:\Users\ShahinN\Anaconda3\lib\site-packages\ipykernel_launcher.py:3: Deprecati
onWarning: Call to deprecated `wv` (Attribute will be removed in 4.0.0, use sel
f instead).
  This is separate from the ipykernel package so we can avoid doing imports unt
il
```

Out[14]:
```
[('Natural language processing (NLP) is an area of computer science and artific
ial intelligence concerned with the interactions between computers and human (n
atural) languages, in particular how to program computers to process and analyz
e large amounts of natural language data.',
  [0.5173867828123971]),
 ('He points out that public transport is very good in Mumbai and New Delhi, wh
ere there is a good network of suburban and metro rail systems.',
  [0.2429788484361275]),
 ('But the man behind the wickets at the other end was watching just as keenly.
With an affirmative nod from Dhoni, India captain Rohit Sharma promptly asked f
or a review. Sure enough, the ball would have clipped the top of middle and le
g.',
  [0.2175021622311386]),
 ('With the Union cabinet approving the amendments to the Motor Vehicles Act, 2
016, those caught for drunken driving will have to have really deep pockets, as
the fine payable in court  has been enhanced to Rs 10,000 for first-time offend
ers.',
  [0.17315283313280116])]
```

In [ ]:
```
1
```