Báo cáo tiến độ Medical Question Retrieval

1. Đọc dữ liệu

In [1]:
```
import pandas as pd
```

In [2]:
```
f1 = open("questionsFinal.dat")
f2 = open("answersFinal.dat")
question = [line[:-1] for line in f1.readlines()]
answer = [line[:-1] for line in f2.readlines()]
```

In [3]:
```
data = pd.DataFrame({"question":question, "answer": answer})
```

In [4]:
```
data.head()
```

Out[4]:

	question	answer
0	Chỉ số xét nghiệm ca 125 là gì?	Xét nghiệm CA 125 chính là chỉ số kháng nguyên...
1	3 món ăn đơn giản giúp ngực to một cách tự nhiên	Có thể bạn không biết những món ăn đơn giản th...
2	Giá khám tổng quát định kì	Như các bạn biết thì chi phí khám sức khỏe tổn...
3	Đừng để bị giảm thị lực vĩnh viễn do viêm bờ mi	Bạn có thể gặp nhiều phiền toái, mất thẩm mĩ v...
4	5 thời điểm vàng bạn nên uống mật ong	Bạn dùng mật ong vào buổi sáng để lọc sạch cặn...

2. Xử lý dữ liệu

2.1 Làm sạch dữ liệu

- Tạm xử lý: Chuyển về chữ thường

In [5]:
```
def clean_data(text):
    # Chuyển sang chữ thường
    text = text.lower()
    return text
```

In [6]:
```
data['question'] = data['question'].apply(lambda x: clean_data(x))
data['answer'] = data['answer'].apply(lambda x: clean_data(x))
data.head()
```

Out[6]:

	question	answer
0	chỉ số xét nghiệm ca 125 là gì?	xét nghiệm ca 125 chính là chỉ số kháng nguyên...
1	3 món ăn đơn giản giúp ngực to một cách tự nhiên	có thể bạn không biết những món ăn đơn giản th...
2	giá khám tổng quát định kì	như các bạn biết thì chi phí khám sức khỏe tổn...
3	đừng để bị giảm thị lực vĩnh viễn do viêm bờ mi	bạn có thể gặp nhiều phiền toái, mất thẩm mĩ v...
4	5 thời điểm vàng bạn nên uống mật ong	bạn dùng mật ong vào buổi sáng để lọc sạch cặn...

2.2 Token

- Sử dụng VncoreNLP (Theo khuyến nghị của PhoBert)

In [7]:
```
# Download VnCoreNLP-1.1.1.jar & its word segmentation component (i.e. R
!mkdir -p vncorenlp/models/wordsegmenter
!wget https://raw.githubusercontent.com/vncorenlp/VnCoreNLP/master/VnCor
!wget https://raw.githubusercontent.com/vncorenlp/VnCoreNLP/master/model
!wget https://raw.githubusercontent.com/vncorenlp/VnCoreNLP/master/model
!mv VnCoreNLP-1.1.1.jar vncorenlp/
!mv vi-vocab vncorenlp/models/wordsegmenter/
!mv wordsegmenter.rdr vncorenlp/models/wordsegmenter/
```

```
--2021-01-30 11:12:24--  https://raw.githubusercontent.com/vncorenlp/V
nCoreNLP/master/VnCoreNLP-1.1.1.jar (https://raw.githubusercontent.co
m/vncorenlp/VnCoreNLP/master/VnCoreNLP-1.1.1.jar)
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 15
1.101.192.133, 151.101.128.133, 151.101.64.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|15
1.101.192.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 27412575 (26M) [application/octet-stream]
Saving to: ‘VnCoreNLP-1.1.1.jar’

VnCoreNLP-1.1.1.jar 100%[===================>]  26,14M  3,83MB/s    in
7,0s

2021-01-30 11:12:31 (3,73 MB/s) - ‘VnCoreNLP-1.1.1.jar’ saved [2741257
5/27412575]

--2021-01-30 11:12:31--  https://raw.githubusercontent.com/vncorenlp/V
nCoreNLP/master/models/wordsegmenter/vi-vocab (https://raw.githubuserc
ontent.com/vncorenlp/VnCoreNLP/master/models/wordsegmenter/vi-vocab)
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 15
1.101.192.133, 151.101.128.133, 151.101.64.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|15
1.101.192.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 526544 (514K) [application/octet-stream]
Saving to: ‘vi-vocab’

vi-vocab            100%[===================>] 514,20K  3,19MB/s    in
0,2s

2021-01-30 11:12:32 (3,19 MB/s) - ‘vi-vocab’ saved [526544/526544]

--2021-01-30 11:12:32--  https://raw.githubusercontent.com/vncorenlp/V
nCoreNLP/master/models/wordsegmenter/wordsegmenter.rdr (https://raw.gi
thubusercontent.com/vncorenlp/VnCoreNLP/master/models/wordsegmenter/wo
rdsegmenter.rdr)
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 15
1.101.192.133, 151.101.128.133, 151.101.64.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|15
1.101.192.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 128508 (125K) [text/plain]
Saving to: ‘wordsegmenter.rdr’

wordsegmenter.rdr   100%[===================>] 125,50K  --.-KB/s    in
0,07s
```

```
2021-01-30 11:12:33 (1,72 MB/s) - ‘wordsegmenter.rdr’ saved [128508/12
8508]
```

In [8]:
```
# Load rdrsegmenter from VnCoreNLP
from vncorenlp import VnCoreNLP
rdrsegmenter = VnCoreNLP("vncorenlp/VnCoreNLP-1.1.1.jar", annotators="ws
```

In [9]:
```
def token_vncorenlp(text):
    sentences = rdrsegmenter.tokenize(text)
    result = ""
    for sentence in sentences:
        result += " ".join(sentence)
    return result
```

In [10]:
```
data['token_question'] = data['question'].apply(lambda x: token_vncorenl
```

In [11]:
```
data.head()
```

Out[11]:

	question	answer	token_question
0	chỉ số xét nghiệm ca 125 là gì?	xét nghiệm ca 125 chính là chỉ số kháng nguyên...	chỉ_số xét_nghiệm ca 125 là gì ?
1	3 món ăn đơn giản giúp ngực to một cách tự nhiên	có thể bạn không biết những món ăn đơn giản th...	3 món ăn đơn_giản giúp ngực to một_cách tự_nhiên
2	giá khám tổng quát định kì	như các bạn biết thì chi phí khám sức khỏe tổn...	giá khám tổng_quát định_kì
3	đừng để bị giảm thị lực vĩnh viễn do viêm bờ mi	bạn có thể gặp nhiều phiền toái, mất thẩm mĩ v...	đừng để bị giảm thị_lực vĩnh_viễn do viêm bờ mi
4	5 thời điểm vàng bạn nên uống mật ong	bạn dùng mật ong vào buổi sáng để lọc sạch cặn...	5 thời_điểm vàng bạn nên uống mật_ong

3. Đo độ tương đồng câu hỏi

3.1 Chuyển các câu hỏi sang vector

- Chuyển các câu hỏi sang vector để phục vụ đo độ tương đồng
- Sử dụng PhoBert, đánh đổi bộ nhớ (lưu các vector embed này lại) để lấy tốc độ

In [12]:
```
import torch
from transformers import AutoModel, AutoTokenizer

phobert = AutoModel.from_pretrained("vinai/phobert-base")
# For transformers v4.x+:
tokenizer = AutoTokenizer.from_pretrained("vinai/phobert-base", use_fast
```

```
Special tokens have been added in the vocabulary, make sure the associ
ated word embedding are fine-tuned or trained.
```

In [13]:
```
def embed(text):
    input_ids = torch.tensor([tokenizer.encode(text)])

    with torch.no_grad():
        features = phobert(input_ids)  # Models outputs are now tuples
    return features[1][0]
```

In [14]:
```
data['embed'] = data['token_question'].apply(lambda x: embed(x))
```

In [15]:
```
data.head()
```

Out[15]:

	question	answer	token_question	embed
0	chỉ số xét nghiệm ca 125 là gì?	xét nghiệm ca 125 chính là chỉ số kháng nguyên...	chỉ_số xét_nghiệm ca 125 là gì ?	[tensor(0.1633), tensor(0.1763), tensor(-0.138...
1	3 món ăn đơn giản giúp ngực to một cách tự nhiên	có thể bạn không biết những món ăn đơn giản th...	3 món ăn đơn_giản giúp ngực to một_cách tự_nhiên	[tensor(0.1557), tensor(0.0881), tensor(-0.025...
2	giá khám tổng quát định kì	như các bạn biết thì chi phí khám sức khỏe tổn...	giá khám tổng_quát định_kì	[tensor(0.2270), tensor(0.1865), tensor(-0.131...
3	đừng để bị giảm thị lực vĩnh viễn do viêm bờ mi	bạn có thể gặp nhiều phiền toái, mất thẩm mĩ v...	đừng để bị giảm thị_lực vĩnh_viễn do viêm bờ mi	[tensor(0.1749), tensor(0.1116), tensor(-0.190...
4	5 thời điểm vàng bạn nên uống mật ong	bạn dùng mật ong vào buổi sáng để lọc sạch cặn...	5 thời_điểm vàng bạn nên uống mật_ong	[tensor(0.1536), tensor(0.0923), tensor(-0.020...

In [16]:
```
import numpy as np
def consine(x, y):
    """ Tính độ tương đồng cosine """
    x = np.array(x)
    y = np.array(y)
    return x.dot(y)/(np.linalg.norm(x)*np.linalg.norm(y))
```

In [17]:
```
def get_similary(text, n):
    """ Chọn top n câu hỏi tương đồng với text nhất """
    text = clean_data(text)
    token = token_vncorenlp(text)
    input_embed = embed(token)
    result = {}
    for text, vector in zip(data['question'], data['embed']):
        result[text] = consine(input_embed, vector)
    return sorted(result.items(), key=lambda x: x[1], reverse=True)[:n]
```

In [19]:
```
%%time
get_similary("bệnh viêm gan b là gì?", 5)
```

```
CPU times: user 482 ms, sys: 3.79 ms, total: 486 ms
Wall time: 241 ms
```

Out[19]:
```
[('bệnh viêm gan a lây qua đường nào?', 0.9303732),
 ('điều trị bệnh viêm gan b ở bệnh viện nào hiệu quả?', 0.9287965),
 ('xét nghiệm viêm gan b ở đâu là tốt nhất?', 0.9268189),
 ('viêm họng xuất tiết là gì ?', 0.9254069),
 ('bệnh viêm gan a có lây không? lây qua đường nào?', 0.92391616)]
```

In [27]:
```
%%time
get_similary("làm gì khi bị ung thư", 5)
```

```
CPU times: user 424 ms, sys: 4.07 ms, total: 428 ms
Wall time: 222 ms
```

Out[27]:
```
[('làm gì để tránh khô ngứa da mùa lạnh', 0.7015503),
 ('làm thế´nào để ngăn ngừa vết chai khi tập tạ', 0.7007115),
 ('hiểu như thế´nào là đúng về ung thư', 0.6789942),
 ('phải làm gì khi bị băng huyết sau sảy thai', 0.6731989),
 ('làm thế´nào để trở lại trường học với trẻ em có những bệnh đau mạn
tính',
  0.6506179)]
```

In [25]:
```
%%time
get_similary("tôi cảm thấy mình bị HIV, tôi phải làm gì", 5)
```

```
CPU times: user 478 ms, sys: 7.98 ms, total: 486 ms
Wall time: 241 ms
```

Out[25]:
```
[('bạn có biết đâu gối của bạn bao nhiêu tuổi?', 0.75085783),
 ('tôi có nên làm xét nghiệm hiv?', 0.74550146),
 ('tại sao tôi không thể giảm cân mặc dù tôi chạy bộ hàng ngày?', 0.73
90499),
 ('bạn có thực sự cần thiết phải giảm cân?', 0.73794144),
 ('bạn có biết hormone fsh là gì?', 0.73762655)]
```

In [24]:
```
%%time
get_similary("HIV", 5)
```

```
CPU times: user 401 ms, sys: 0 ns, total: 401 ms
Wall time: 214 ms
```

Out[24]:
```
[('xét nghiệm hiv bằng phương pháp elisa', 0.7587304),
 ('đồng tính nữ không miễn nhiễm với hiv', 0.7390161),
 ('5 phó giáo sư chuyên khoa nội tiết tại hà nội', 0.7241979),
 ('hội chứng lyell - nguyên nhâ', 0.7011033),
 ('các xét nghiệm của bệnh nhân hiv', 0.7008121)]
```

In [ ]: