1. Đọc và xử lý dữ liệu

In [1]:
```
import pandas as pd
f1 = open("questionsFinal.dat")
f2 = open("answersFinal.dat")
question = [line[:-1] for line in f1.readlines()]
answer = [line[:-1] for line in f2.readlines()]
data = pd.DataFrame({"question":question, "answer": answer})
data.head()
```

Out[1]:

	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. Tokenizer

In [2]:
```
def clean_data(text):
    # Chuyển sang chữ thường
    text = text.lower()
    return text
```

In [8]:
```
data['question'] = data['question'].apply(lambda x: clean_data(x))
data['answer'] = data['answer'].apply(lambda x: clean_data(x))
data.head()
```

Out[8]:

	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

In [9]:
```
from vncorenlp import VnCoreNLP
rdrsegmenter = VnCoreNLP("vncorenlp/VnCoreNLP-1.1.1.jar", annotators="ws
```

In [10]:
```
def token_vncorenlp(text):
    sentences = rdrsegmenter.tokenize(text)
    result = ""
    for sentence in sentences:
        result += " ".join(sentence)
    return result
```

In [11]:
```
data['token_question'] = data['question'].apply(lambda x: token_vncorenl
```

In [12]:
```
data.head()
```

Out[12]:

	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

In [13]:
```
# Lưu lại về` sau dùng cho nhanh
data.to_csv("data_processed.csv", index=False)
```

3. Nhúng các câu hỏi dùng tf-idf

In [15]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer()
matrix = vectorizer.fit_transform(data.token_question.to_list())
```

4. Xử lý truy vấn đầu vào

In [24]:
```
from sklearn.metrics.pairwise import cosine_similarity
```

In [25]:
```
def get_tf_idf_query(text):
    return vectorizer.transform([text])
```

In [29]:
```
def process(text):
    text = clean_data(text)
    text = token_vncorenlp(text)
    query_tfidf = get_tf_idf_query(text)
    cosineSimilarities = cosine_similarity(query_tfidf, matrix).flatten(
    indexMax = cosineSimilarities.argsort()[-5:][::-1]
    for idx in indexMax:
        print(data.iloc[idx]['question'], cosineSimilarities[idx]*100)
```

In [30]:
```
text = "tôi nghĩ mình bị HIV, tôi phải làm gì"
```

In [31]:
```
process(text)
```

```
tôi có nên làm xét nghiệm hiv? 52.53567020773382
tôi nên làm gì nếu con tôi bị ngã và gãy răng? 44.51802588345336
tôi phải làm thế´nào khi bị cảm cúm? 42.226725761420134
tại sao tôi không thể̉giảm cân mặc dù tôi chạy bộ hàng ngày? 39.248934
672893824
tôi cần làm gì khi bị phì đại cổtửcung? 31.610397630835642
```

In [32]:
```
text = "làm gì khi bị ung thư"
process(text)
```

```
bị cảm khi mang thai nên làm gì? 56.210235461032354
bị trĩ khi mang thai cần phải làm gì? 52.4994295909834
cần làm gì khi trẻ sơ sinh bị rubella? 50.73857137875929
phải làm gì khi bị nấm móng? 50.219392773154894
làm gì khi trẻ bị chảy máu cam? 49.61860879178268
```

In [ ]: