

# ĐỒ ÁN CUỐI KỲ

Thu thập và phân loại đánh giá về một sản phẩm trên trang Web AMAZON



Nhóm 34:

18120374 – Nguyễn Minh Hiếu

18120384 – Nguyễn Văn Hoài



[hoainguyen33/DA\\_CK\\_NMKHDL](https://github.com/hoainguyen33/DA_CK_NMKHDL)

# Giới thiệu Amazon

- Công ty là thị trường thương mại điện tử lớn nhất thế giới.
- Lúc đầu chủ yếu bán sách, rồi mở rộng thị trường ra hàng hóa tiêu dùng, phương tiện kỹ thuật.
- Được nhiều khách hàng lựa chọn, đặc biệt ở mùa dịch này.

The screenshot displays the Amazon homepage with a sidebar on the left and a grid of product deals on the right.

**Left Sidebar:**

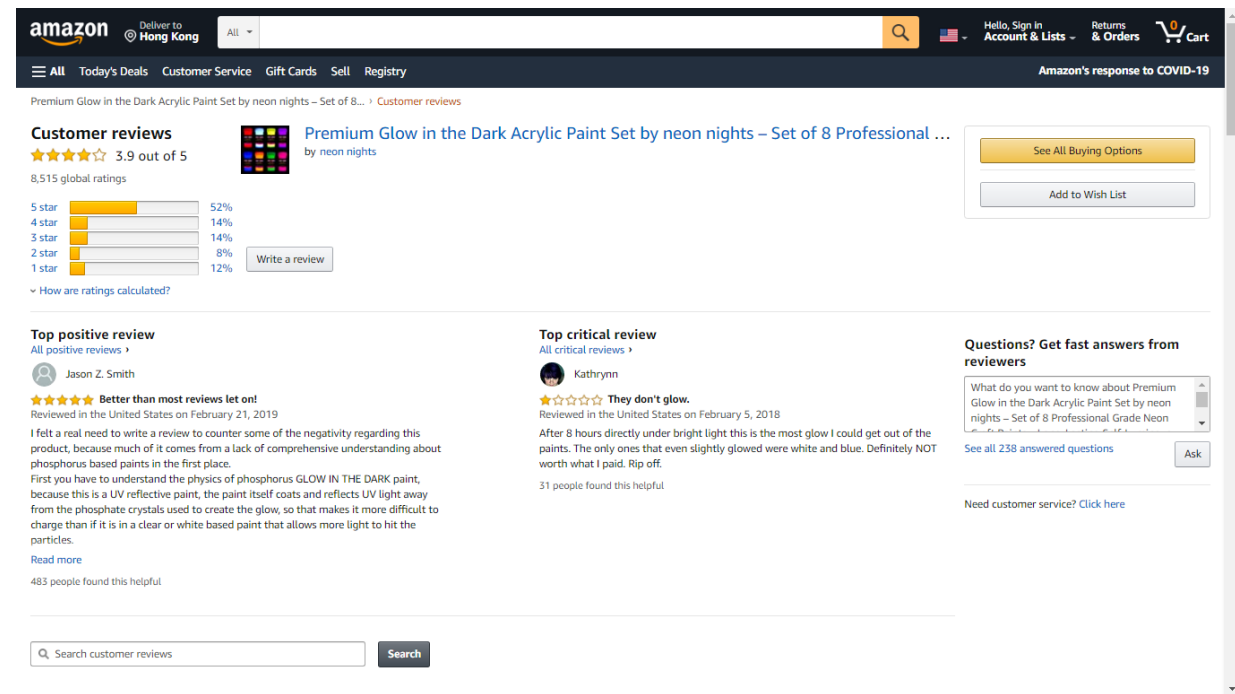
- Categories:** ☐ Baby Clothing & Accessories, ☐ Beauty, ☐ Books, ☐ Boys' Fashion, ☐ Camera & Photo, ☐ Cell Phones & Accessories, ☐ Computers & Accessories, ☐ Costumes & Accessories, ☐ Electronics.
- Deal Type:** Deal of the Day, Lightning Deals, Savings & Sales, Prime Early Access Deals.
- Availability:** ☐ Clear, ☒ Active, ☐ Upcoming, ☐ Missed.
- Price:** Under \$25, \$25 to \$50, \$50 to \$100, \$100 to \$200, \$200 & Above.
- Discount:** 10% Off or More, 25% Off or More, 50% Off or More, 70% Off or More.
- Avg. Customer Review:** 5 stars & Up, 4 stars & Up, 3 stars & Up, 2 stars & Up, 1 star & Up.

**Product Grid:**

Product	Price	Discount	Rating	Action
Yeedi Robotic Vacuum	\$129.99 - \$189.99	Up to 35% off	★★★★☆ 565	See details
Active & Athleisure Wear	\$7.10 - \$58.70	Up to 15% off	★★★★☆ 459	See details
iPhone 12 Mini Screen Protector	\$6.79	Price: \$7.99 (15% off)	★★★★☆ 1686	Add to Cart
Bare Home Fitted Sheet Premium Microfiber 1800 Ultra-Soft	\$10.82 - \$26.76	21% Claimed	★★★★☆ 22847	Choose options
Battery Chargers	\$25.07 - \$169.76	Save up to 25% on sale	★★★★☆ 8856	See details
Premium Glow in the Dark Acrylic Paint Set	\$14.44	List: \$39.99 (64% off)	★★★★☆ 8515	Add to Cart
MONGOORA Metal Car Charger	\$10.43	List: \$19.99 (48% off)	★★★★☆ 4559	Add to Cart
Esakiya Food Kitchen Scale	\$10.95	List: \$14.99 (22% off)	★★★★☆ 72350	Add to Cart
GADEWAKE Womens Casual Color Block Long Sleeve Round Neck Po...	\$16.13 - \$16.99	56% Claimed	★★★★☆ 15266	Choose options
Save on Love, Mom and Mother and Daughter K Journal and more	\$10.04 - \$11.59	Save on Love, Mom and Mother and Daughter K Journal and more	★★★★☆ 988	See details

# Trang web Amazon

- [Amazon.com](https://www.amazon.com)
- Đa dạng sản phẩm.
- Đánh giá khách quan khi có thể đăng video, ảnh...
- Có phân loại đánh giá theo ☆ hoặc tích cực và không tích cực.



# Lợi ích nhận được khi đạt được kết quả

- Nếu tìm được công thức thì công ty có thể tự động hóa quá trình rút trích ý kiến **tích cực** / **tiêu cực** từ bình luận của người dùng.
- Biết được ý kiến của người dùng về sản phẩm sẽ giúp ích cho công ty. Ở đây là AMAZON có thể kiểm tra ý kiến của khách hàng về sản phẩm để biết tình trạng khi đến tay khách hàng như thế nào (có đúng ý khách không, có lành lặn không,...)

- Thu thập dữ liệu:
  - Kiểm tra số lượng đánh giá. (ít quá thì không ổn)
  - Chỉ xét đánh giá tiếng Anh.
- Tiền xử lý:
  - Loại bỏ thành phần dư thừa.
  - Tạo một bag of words.
- Mô hình hóa:
  - Áp dụng Neural Network.
  - Xác định overfitting, ...

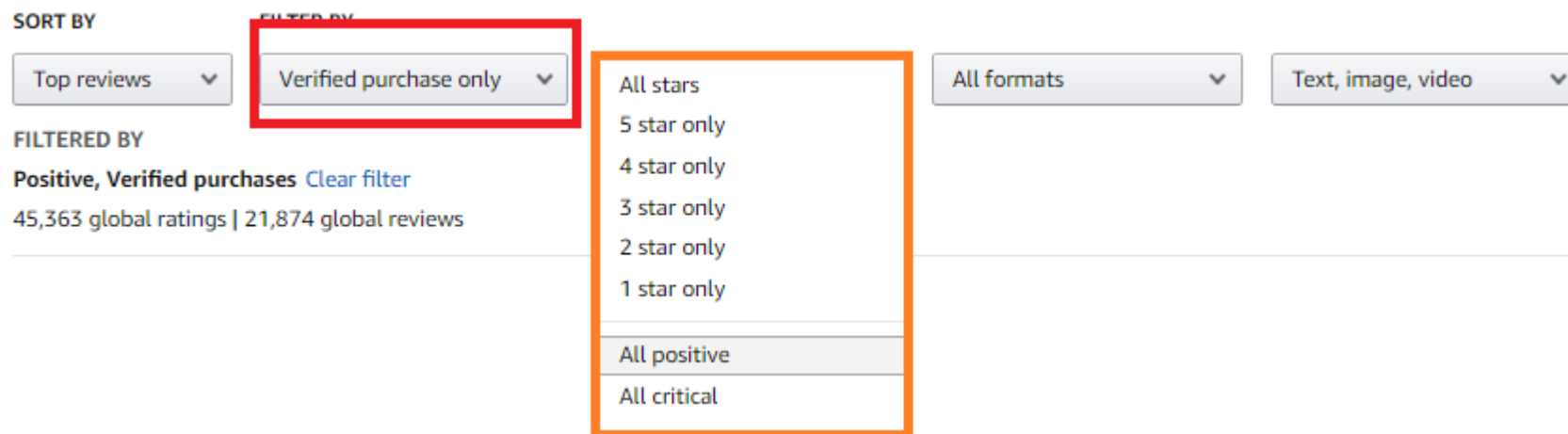
# Thu thập dữ liệu

- Truy cập sản phẩm bất kì (VD:[link](#))
- Vào link -> **Top reviews from the United States** chọn [See all reviews](#)
- Kiểm tra số lượng reviews (38,780 global reviews).
- Bên dưới có filter.

**SORT BY** **FILTER BY**

Top reviews ▼	All reviewers ▼	All stars ▼	All formats ▼	Text, image, video ▼
---------------	-----------------	-------------	---------------	----------------------

# Thu thập dữ liệu



- Tab “All reviews” thành “**Verified purchase only**”
- Tab “All stars” gồm “**All Positive**” và “**All Critical**”
- Sau đó duyệt từng trang. Mỗi trang sẽ phải chờ 3s

# Thu thập dữ liệu

- Cần thu thập Title của bình luận (in đậm)



Matt Sayar



**The Router That Was Promised**

Reviewed in the United States on February 22, 2018

Model: AC1750 WiFi | **Verified Purchase**



Jim




**Failed in seven weeks**

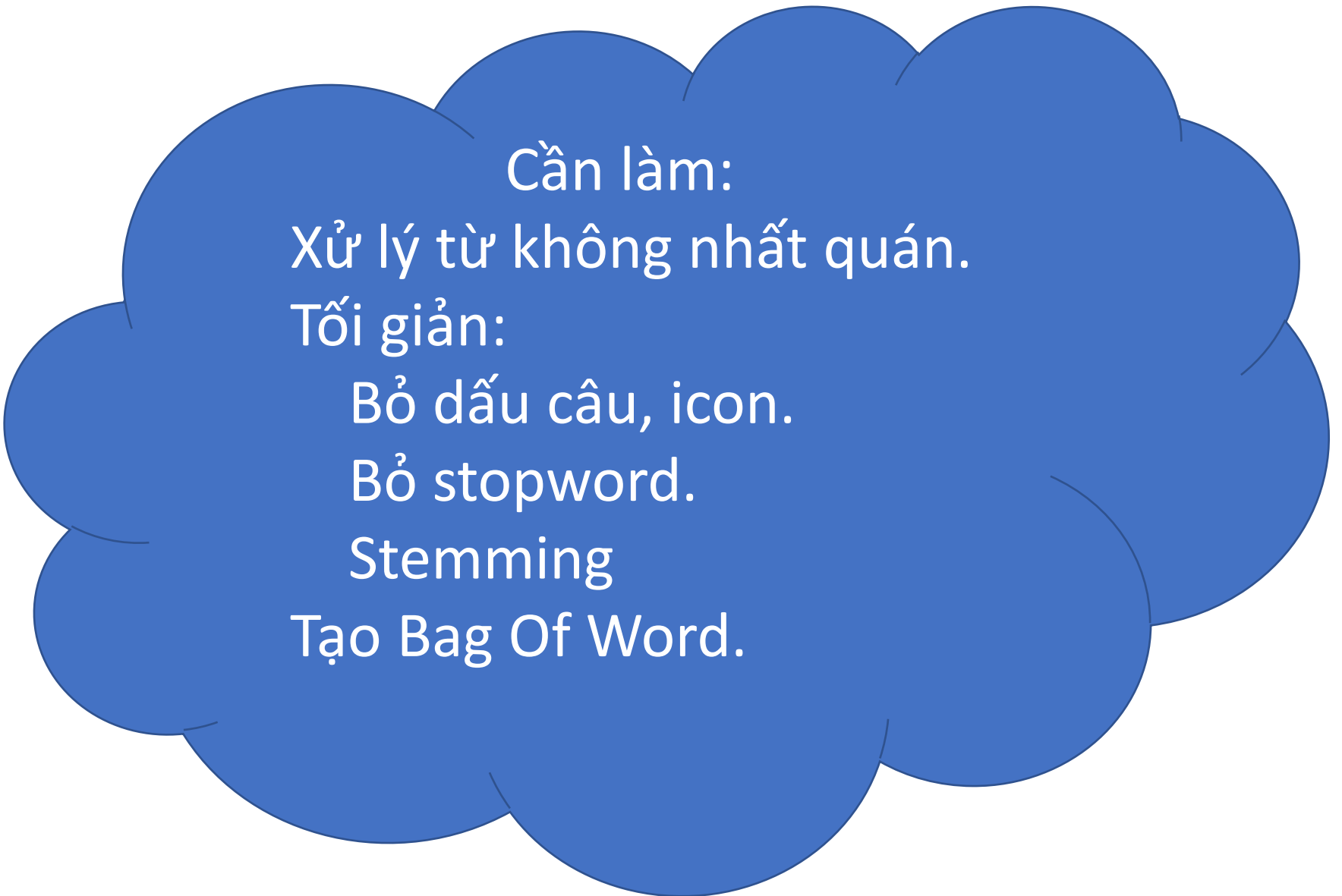
Reviewed in the United States on May 12, 2018

Model: AC2300 WiFi | **Verified Purchase**

Cần sử dụng Selenium

# Tiền xử lý

- Quan sát dữ liệu thu được (5400 mẫu)
- Nhận xét:
  - Có một số nhận xét không phải tiếng anh. (số lượng ít)
  - Có sai chính tả, viết tắt.
  - Chưa nhất quán về cách viết (VD: don't – dont).
  - Có thể chứa các icon (VD: )
  - Có dấu câu thừa (VD: ..., ?, !!!)



Cần làm:  
Xử lý từ không nhất quán.  
Tối giản:  
Bỏ dấu câu, icon.  
Bỏ stopwords.  
Stemming  
Tạo Bag Of Word.

# Tiền xử lý

- Tạo một file: “check.txt”: gồm các từ cần tối giản.
- Tạo file “stop\_word.txt” : gồm các từ không cần thiết
- Thư viện re: “*import re*” ([link](#)): *Nhằm làm chuẩn, xóa sạch những thứ không thuộc a–z, A–Z (bao gồm cả số)*

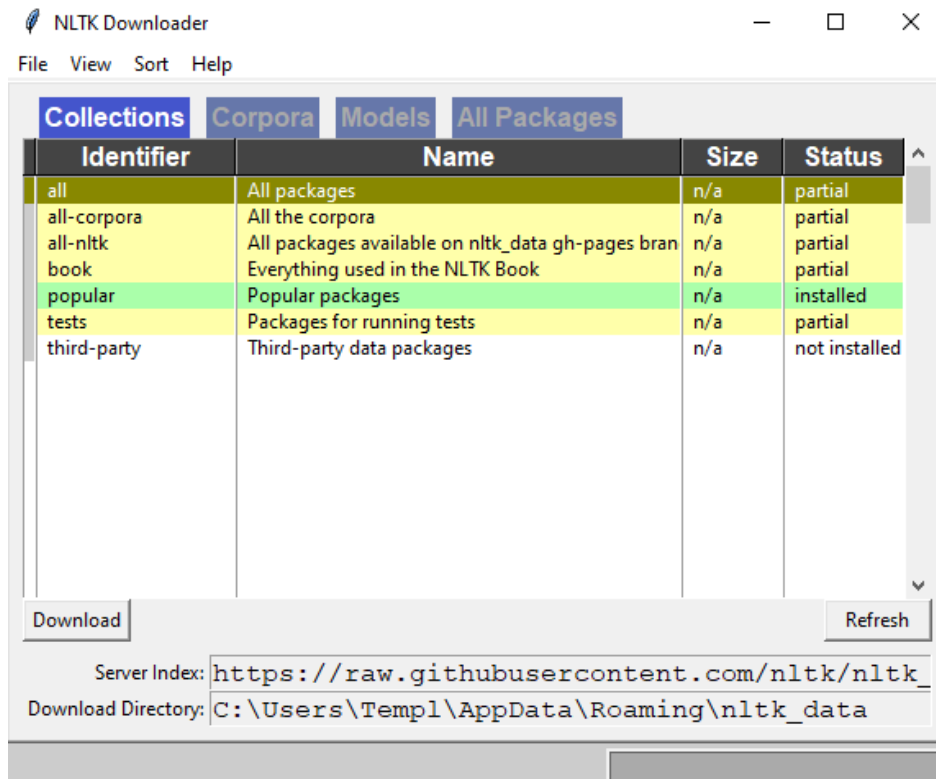
```
sentence = "Wait.... OK!!! 😊😞 Guess it gone?"
```

```
sentence = re.sub('[^A-Za-z]', ' ', sentence)
sentence
```

```
'Wait      OK      Guess it gone '
```

# Tiền xử lý

- Thư viện nltk : *nltk.download()* Chọn popular -> download  
*Nhằm dùng để Stemming (ở đây chưa nói đến thì hay ngữ cảnh)*



```
englishStemmer=SnowballStemmer("english")
```

```
inp = ["fish","fisher","fishing", "fishs","fished"]  
out = []
```

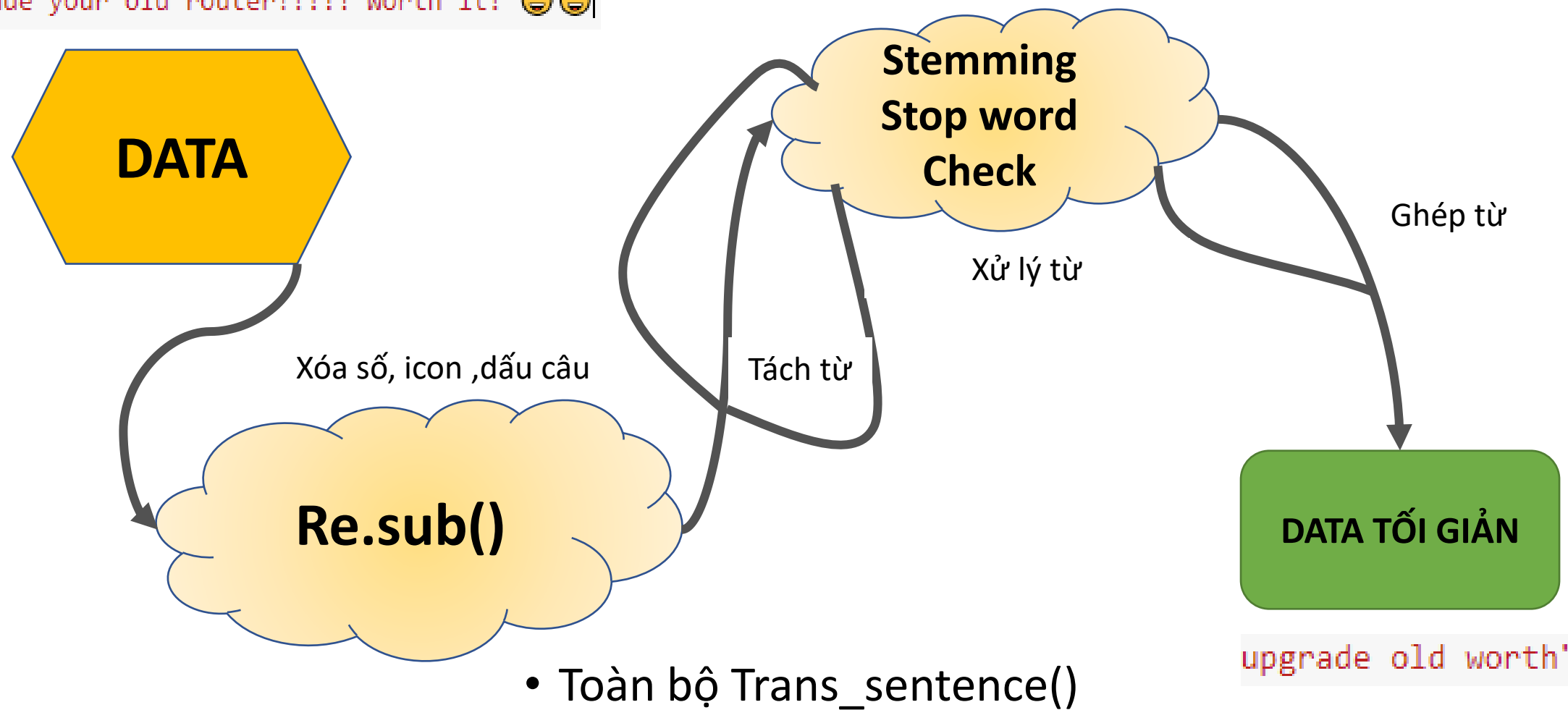
```
for word in inp:  
    out.append(englishStemmer.stem(word))
```

out

```
['fish', 'fisher', 'fish', 'fish', 'fish']
```

# Tiền xử lý

"Upgrade your old router!!!! Worth it! 😊😊"



# Tiền xử lý

- Input mong muốn của ta là có một vector số cho toàn bộ từ có trong Data. Có nhiều cách:
  - CountVector: đếm tất cả các từ có trong một câu.
  - Tfidfvectorizer: xác định các từ có giá trị nhận diện.
  - ....
- Tfidfvectorizer ([link](#))

```
sentences = ['the cat see the mouse',  
             'the house has a tiny little mouse',  
             'the mouse ran away from the house',  
             'the cat finally ate the mouse',  
             'the end of the mouse story']
```

	ate	away	cat	end	finally	from	has	house	little	mouse	of	ran	see	story	the	tiny
0	0.000	0.000	0.483	0.000	0.000	0.000	0.000	0.000	0.000	0.285	0.000	0.000	0.599	0.000	0.571	0.000
1	0.000	0.000	0.000	0.000	0.000	0.000	0.494	0.398	0.494	0.235	0.000	0.000	0.000	0.000	0.235	0.494
2	0.000	0.457	0.000	0.000	0.000	0.457	0.000	0.369	0.000	0.218	0.000	0.457	0.000	0.000	0.436	0.000
3	0.514	0.000	0.415	0.000	0.514	0.000	0.000	0.000	0.000	0.245	0.000	0.000	0.000	0.000	0.490	0.000
4	0.000	0.000	0.000	0.492	0.000	0.000	0.000	0.000	0.000	0.234	0.492	0.000	0.000	0.492	0.469	0.000

# TF-IDF

TF-IDF is a measure of originality of a word by comparing the number of times a word appears in a doc with the number of docs the word appears in.

$$\text{TF-IDF} = \text{TF}(t, d) \times \text{IDF}(t)$$

Term frequency

Number of times term  $t$  appears in a doc,  $d$

Inverse document frequency

$$\log \frac{1 + \overset{\text{\# of documents}}{n}}{1 + \underset{\text{Document frequency of the term } t}{df(d, t)}} + 1$$

(Nguồn: [medium.com](https://medium.com))

- TF: giá trị của số lần chữ 'a' xuất hiện trong câu B sẽ có ý nghĩa khác nhau khi len(B) thay đổi.
- IDF: xác định độ phổ biến của từ trên toàn bộ dữ liệu, cố gắng không bỏ sót "rare word".

	feature_name	idf_weights
0	ate	2.098612
1	away	2.098612
2	cat	1.693147
3	end	2.098612
4	finally	2.098612
5	from	2.098612
6	has	2.098612
7	house	1.693147
8	little	2.098612
9	mouse	1.000000
10	of	2.098612
11	ran	2.098612
12	see	2.098612
13	story	2.098612
14	the	1.000000
15	tiny	2.098612

IDF

X TF

	ate	away	cat	end	finally	from	has	house	little	mouse	of	ran	see	story	the	tiny
0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	2	0
1	0	0	0	0	0	0	1	1	1	1	0	0	0	0	1	1
2	0	1	0	0	0	1	0	1	0	1	0	1	0	0	2	0
3	1	0	1	0	1	0	0	0	0	1	0	0	0	0	2	0
4	0	0	0	1	0	0	0	0	0	1	1	0	0	1	2	0

||

	ate	away	cat	end	finally	from	has	house	little	mouse	of	ran	see	story	the	tiny
0	0.000	0.000	1.693	0.000	0.000	0.000	0.000	0.000	0.000	1.0	0.000	0.000	2.099	0.000	2.0	0.000
1	0.000	0.000	0.000	0.000	0.000	0.000	2.099	1.693	2.099	1.0	0.000	0.000	0.000	0.000	1.0	2.099
2	0.000	2.099	0.000	0.000	0.000	2.099	0.000	1.693	0.000	1.0	0.000	2.099	0.000	0.000	2.0	0.000
3	2.099	0.000	1.693	0.000	2.099	0.000	0.000	0.000	0.000	1.0	0.000	0.000	0.000	0.000	2.0	0.000
4	0.000	0.000	0.000	2.099	0.000	0.000	0.000	0.000	0.000	1.0	2.099	0.000	0.000	2.099	2.0	0.000

TF-IDF

## Chuẩn hóa dữ liệu: dùng euclidean norm

$$X_{the} = \frac{2.0}{\sqrt{2.0^2 + 2.0986^2 + 1.0^2 + 1.6931^2}} = 0.57094$$

$$X_{see} = \frac{2.0986}{\sqrt{2.0^2 + 2.0986^2 + 1.0^2 + 1.6931^2}} = 0.59909$$

$$X_{mouse} = \frac{1.0}{\sqrt{2.0^2 + 2.0986^2 + 1.0^2 + 1.6931^2}} = 0.28547$$

$$X_{cat} = \frac{1.6931}{\sqrt{2.0^2 + 2.0986^2 + 1.0^2 + 1.6931^2}} = 0.48334$$

	ate	away	cat	end	finally	from	has	house	little	mouse	of	ran	see	story	the	tiny
0	0.000	0.000	0.483	0.000	0.000	0.000	0.000	0.000	0.000	0.285	0.000	0.000	0.599	0.000	0.571	0.000
1	0.000	0.000	0.000	0.000	0.000	0.000	0.494	0.398	0.494	0.235	0.000	0.000	0.000	0.000	0.235	0.494
2	0.000	0.457	0.000	0.000	0.000	0.457	0.000	0.369	0.000	0.218	0.000	0.457	0.000	0.000	0.436	0.000
3	0.514	0.000	0.415	0.000	0.514	0.000	0.000	0.000	0.000	0.245	0.000	0.000	0.000	0.000	0.490	0.000
4	0.000	0.000	0.000	0.492	0.000	0.000	0.000	0.000	0.000	0.234	0.492	0.000	0.000	0.492	0.469	0.000

- Mô hình hóa dữ liệu
- Xây dựng mạng neural network : 10 tầng
- Sử dụng Tham số alpha, min\_df được dùng ở quá trình **Kiểm tra Overfitting**.
  - \* min\_df (min document frequency): chọn từ xuất hiện ít nhất min\_df lần.
  - Alpha: giảm w
- Pipeline chuyển đổi dữ liệu câu dạng tối giản sang dạng vector.

