

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

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 image shows a portion of the Amazon website's homepage. On the left, there is a sidebar with various filters and search options. The main area displays a grid of products, each with a thumbnail, price, deal information, and a brief description. The products include a robotic vacuum, athletic wear, screen protectors, home sheets, battery chargers, glow-in-the-dark paint, a car charger, a digital scale, a sweatshirt, and a journal.

Filters (Left Sidebar):

- Baby Clothing & Accessories
- Beauty
- Books
- Boys' Fashion
- Camera & Photo
- Cell Phones & Accessories
- Computers & Accessories
- Costumes & Accessories
- Electronics
- [See more Department](#)
- 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**
- ★ ★ ★ ★ ★ & Up

Grid of Products:

Product Image	Price Range	Description	Rating & Reviews	Action Buttons
	\$129.99 - \$189.99	DEAL OF THE DAY Up to 35% off Yeedi Robotic Vacuums	★★★★★ 565	See details
	\$7.10 - \$58.70	Up to 15% off Active & Athleisure Wear from Our Brands	★★★★★ 439	See details
	\$6.79	iPhone 12 Mini Screen Protector-5.4"	★★★★★ 1686	Add to Cart
	\$10.82 - \$26.76	Bare Home Fitted Sheet Premium Microfiber 1800 Ultra-Soft	★★★★★ 22847	Choose options
	\$25.07 - \$169.76	Save up to 25% on sale Battery Chargers	★★★★★ 8836	See details
	\$14.44	Premium Glow in the Dark Acrylic Paint Set by neon nights - Set of...	★★★★★ 8515	Add to Cart
	\$10.43	90% Claimed Metal Car Charger by MONGORIA - Qualcomm Quick Charge 3.0 Du...	★★★★★ 4559	Add to Cart
	\$10.95	Etekcity Food Kitchen Scale, Digital Grams and Oz for Cooking, Bak...	★★★★★ 72330	Add to Cart
	\$16.13 - \$16.99	50% Claimed GADEWAKE Womens Casual Color Block Long Sleeve Round Neck Po...	★★★★★ 15266	Choose options
	\$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.

The screenshot shows the Amazon product review page for a 'Premium Glow in the Dark Acrylic Paint Set'. The top navigation bar includes 'Deliver to Hong Kong', a search bar, and account links. The main content area displays a 5-star rating of 3.9 out of 5 based on 8,515 global ratings. A color palette icon is shown next to the title. Below the rating is a distribution chart for star ratings: 5 star (52%), 4 star (14%), 3 star (14%), 2 star (8%), and 1 star (12%). There is a 'Write a review' button and a link to 'How are ratings calculated?'. Two specific reviews are highlighted: a top positive review from Jason Z. Smith and a top critical review from Kathryn. The positive review discusses the product's physics and glow characteristics. The critical review expresses disappointment with the lack of glow. A sidebar on the right contains a question-and-answer section and a 'See all 238 answered questions' link.

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,...)

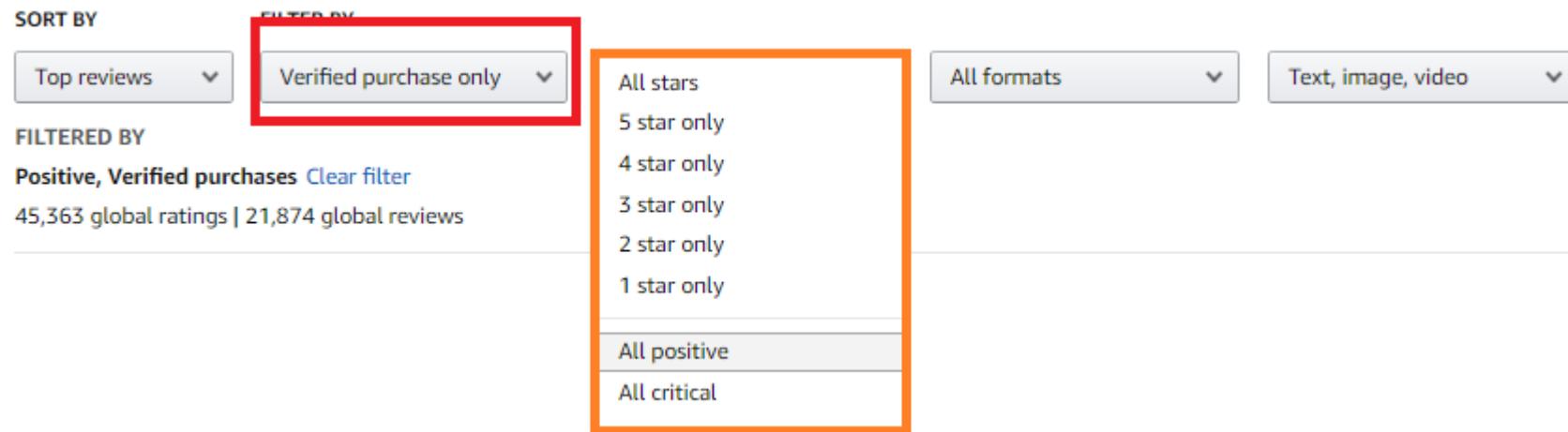
- Thu thập dữ liệu:
 - Kiểm tra số lượng đánh giá. (ít quá thì không ổn)
 - Chỉ xét đánh giá tiếng Anh.
- Tiền xử lý:
 - Loại bỏ thành phần dư thừa.
 - Tạo một bag of words.
- Mô hình hóa:
 - Áp dụng Neural Network.
 - Xác định overfitting, ...

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.



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 câu 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 stopword.
Stemming
Tạo Bag Of Word.

Tiền xử lý

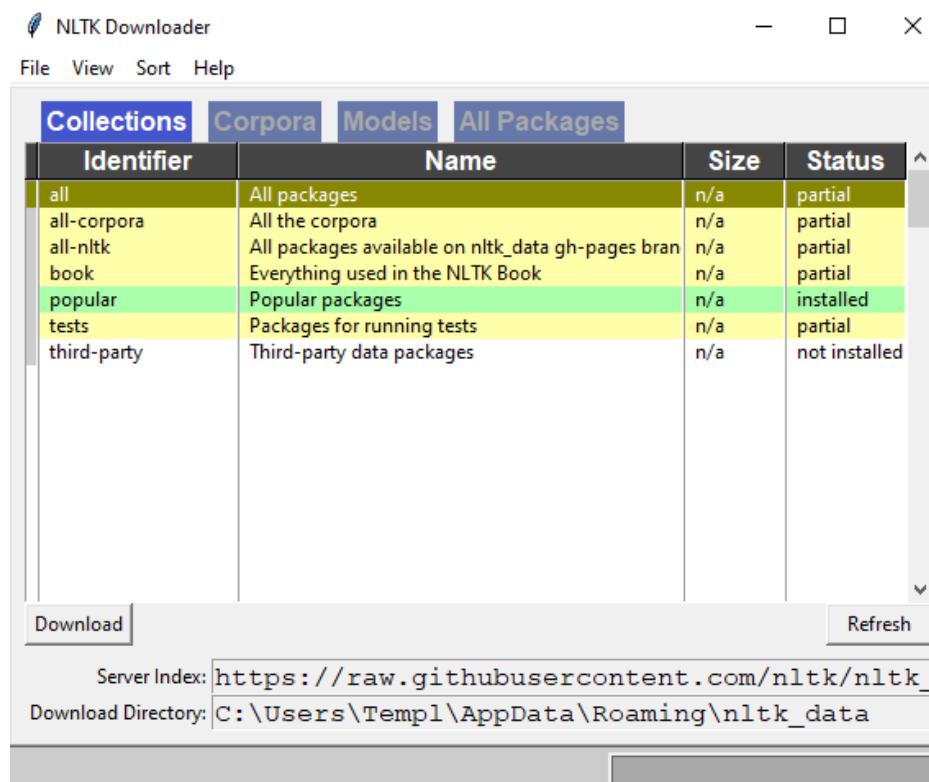
- Tạo một file: “check.txt”: gồm các từ cần tối giản.
- Tạo file “stop_word.txt” : gồm các từ không cần thiết
- Thư viện re: “*import re*” ([link](#)): Nhằm làm chuẩn, xóa sạch những thứ không thuộc a–z, A-Z (bao gồm cả số)

```
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", "fishes","fished"]  
out = []
```

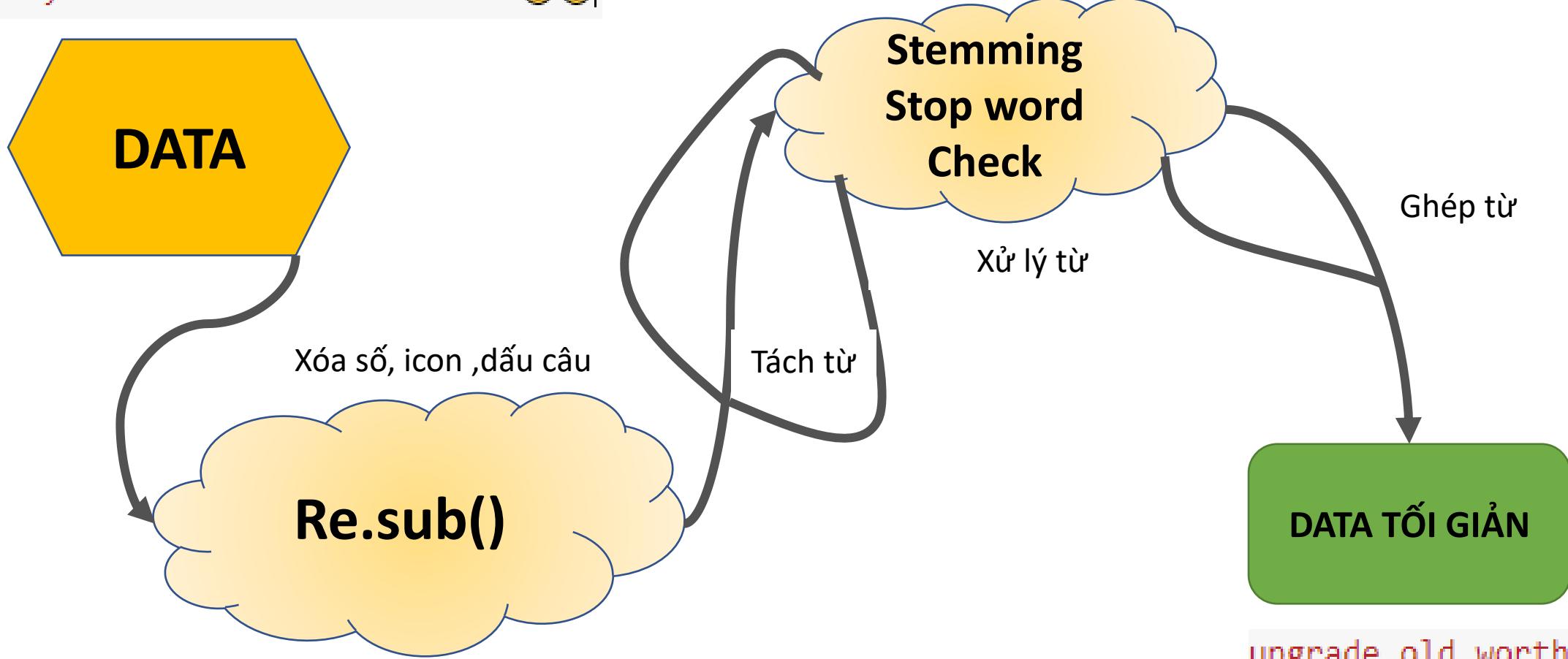
```
for word in inp:  
    out.append(englishStemmer.stem(word))
```

```
out
```

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

Tiền xử lý

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



upgrade old worth'

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.
 - Tdifvectorizer: 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
 $\log \frac{1 + n}{1 + df(d, t)} + 1$
of documents
Document frequency of the term t

(Nguồn: [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.

