

# **Text Analysis with NLP**

## **Case study: Phân tích quan điểm**

TS. Nguyen Van Vinh  
QAI, Fsoft

# Giới thiệu về giảng viên

- TS. Nguyễn Văn Vinh
- PhD tại Viện Khoa học Công nghệ tiên tiến Nhật Bản (JAIST)
- Lĩnh vực nghiên cứu: Trí Tuệ Nhân Tạo, Xử lý ngôn ngữ tự nhiên, Dữ liệu lớn
- Giảng viên Khoa CNTT, Chuyên gia AI - Trường ĐHCN – ĐHQG Hà Nội
- Senior Data Scientist, Công ty QAI, Fsoft
- Trưởng nhóm thiết kế khóa học Machine Learning của FUNiX (<https://vnexpress.net/giao-duc/funix-ra-mat-chuong-trinh-machine-learning-3986011.html>)
- Đã từng trình bày và giảng bài: Samsung Việt Nam, VNG, Ban cơ yếu chính phủ, Viện Vin BigData ...

# Adrew Ng

- “NLP is reshaping daily life. No doubt you've found valuable information using web search and the search functions found on countless websites and apps. Anti-spam systems are a critical part of the global email system. How does a smart speaker understand your commands? How does a chatbot generate relevant responses? “

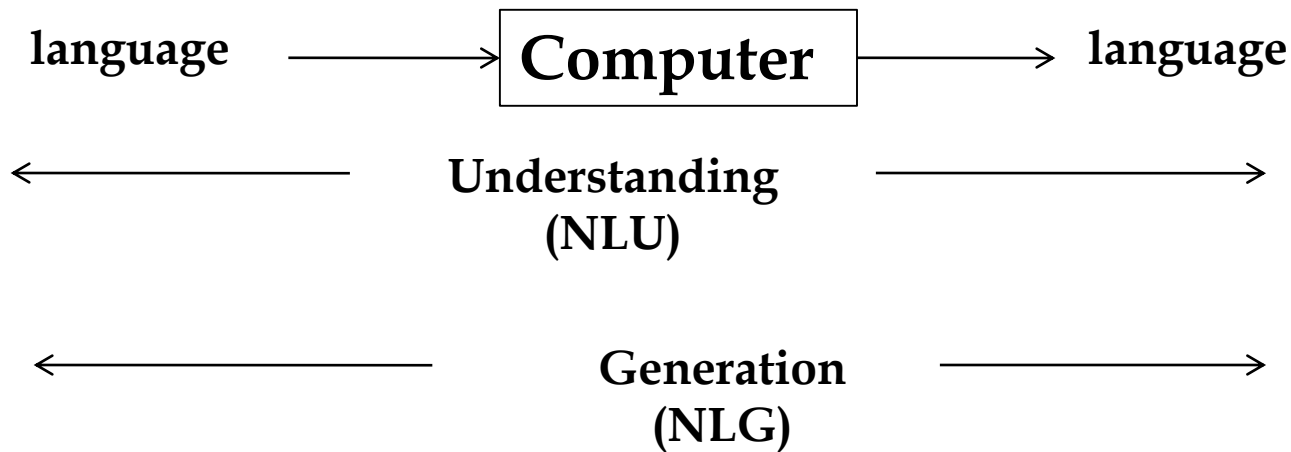
# Nội dung

- Bài toán xử lý ngôn ngữ tự nhiên
- Ứng dụng: Bài toán Phân tích quan điểm

# BÀI TOÁN XỬ LÝ NGÔN NGỮ TỰ NHIÊN (NLP)

# Xử lý ngôn ngữ tự nhiên là gì?

- Computers using natural language as input and/or output



# Go beyond the keyword matching



- Identify the **structure** and **meaning** of **words**, **sentences**, **texts** and **conversations**
- **Deep** understanding of **broad** language
- NLP is all around us

# Machine translation

[All](#)
[Images](#)
[Shopping](#)
[Apps](#)
[Videos](#)
[More](#)
[Search tools](#)

About 20,800,000 results (0.54 seconds)

Spanish

↕

↕

English

buenas noches

Edit

Goodnight

3 more translations

[Open in Google Translate](#)

Haaretz

אילי העם הוא ילילת

חיים נתניהו שוב ינסה להעביר את ד"ר טרכטנברג בממשלה

www.haaretz.co.il

אחרי הפאקסה בשבוע שעבר, בלשכת רד"מ מס' 610

להחזיק שותפה הצבעה בתום הדיון, בישראל בתום, שלם

והעמדות דורשים ריכוז או השמטה של חלק מהחלטות

Like · Comment · Translate · Share · Yesterday at 06:00

9 people like this.

View 1 share

שוק יזם עם שלם משלם את מחיר הווחות והקטניות של ראש הממשלה, שלא מציג לנכון להתגייס להבראת מערכת הבריאות. סיון דים אמנם וליאורם גרם לסב. חובר לרגש את החפץ הזה מחצה, משרד'ב כבר דרש יעוץ, ממחשבה ג. חשוי הווחות'ס בשיקפות את המדינה ממלה את אחרות. בן דר, ביקום נתניהו. נתניהו נותן: לא מתחייב להצבעה על טרכטנברג. www.ynet.co.il

חשומות הקאליציות הצליחו להטיל מורא על ראש הממשלה, שברר את מתחייב להצבעה דיום את מסקנות ד"ר טרכטנברג, שלם לם ישנה מנגנון'ס ישראל מיתו וממלת העמדות יקום בבוקר ישבת שלם ווחליו'ס קר'ב. העשרת הפוליטית, חדשת

Expand preview

Yesterday at 06:31 · Like · 2 people · Translate

עזר ודביר או שדרה'ס יעבור או שוב'ס יעבור

Yesterday at 07:10 · Like · Translate

Dalya Gumiš

שפסיון להצבעה ווחליו'ס לבצע

Yesterday at 08:11 · Like · Translate

Yuval Gilor

ממאס כבר לך הביתה

Yesterday at 08:49 · Like · Translate

Haaretz

Maybe this time he succeeds?

חיים נתניהו שוב ינסה להעביר את ד"ר טרכטנברג בממשלה

www.haaretz.co.il

אחרי הפאקסה בשבוע שעבר, בלשכת רד"מ מס' 610

להחזיק שותפה הצבעה בתום הדיון, בישראל בתום, שלם

והעמדות דורשים ריכוז או השמטה של חלק מהחלטות

Like · Comment · Original · Share · Yesterday at 06:00

9 people like this.

View 1 share

שוק יזם With full pay the price for hahihoth vakonbinot of the Prime Minister. Not find it appropriate to action hahrat the health system. Endangered animals, people and caused the suffering that must drive with ... city. Foundation of Minneapolis had already thrown him. Europe also. This man destroys the country with the citizens in its path. Contempt you Binyamin Netanyahu. Netanyahu was not committed to voting the Trachtenberg managed to impose a "partnership hloalcoalniont terror the Prime Minister, who is pledging to bring voting day the conclusions report. Trachtenberg. It's hourly opposes. Israel Beitenu-independence morning ministerial session will vialalto only it-political system, news

Expand preview

Yesterday at 06:31 · Like · 2 people · Original

עזר ודביר Or acknowledging the report moves or the chips will go

Yesterday at 07:10 · Like · Original

Dalya Gumiš

Stop vote once to

Yesterday at 08:11 · Like · Original

Yuval Gilor

Tired already go home

Yesterday at 08:49 · Like · Original

Facebook translation, image credit: Meedan.org


8



# Dialog Systems/Chatbot

## Gift shop


Items such as caps, t-shirts, sweatshirts and other miscellanea such as buttons and mouse pads have been designed. In addition, merchandise for almost all of the projects is available.



**Hi. I'm your automated online assistant. How may I help you?**


**CD or DVD**

There is a series of CDs/DVDs with selected Wikipedia content being produced by Wikipedians and [SOS Children](#).



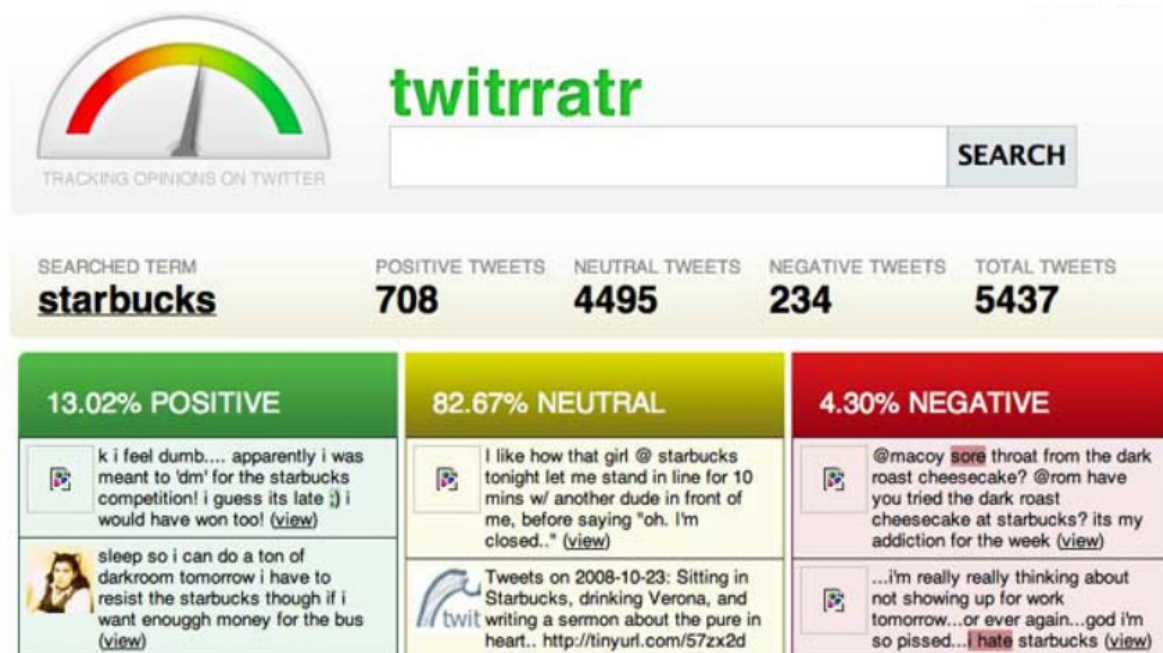
**Downloading**

Downloading content from Wikipedia is free of charge. All text content is licensed under the [GNU Free Documentation License](#).

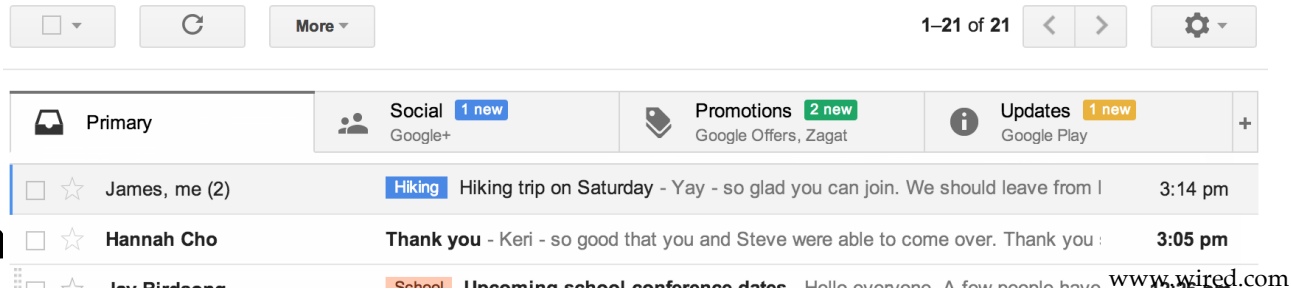


(GFDL). Images and other files are available under [different terms](#), as detailed on

# Sentiment/Opinion Analysis



# Text Classification

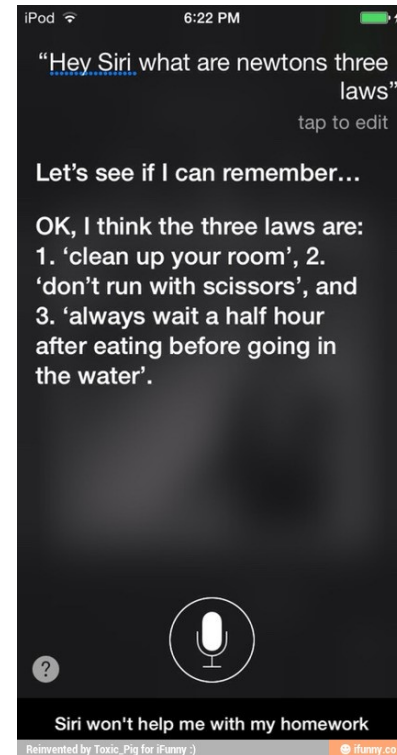


- Other a

# Question answering



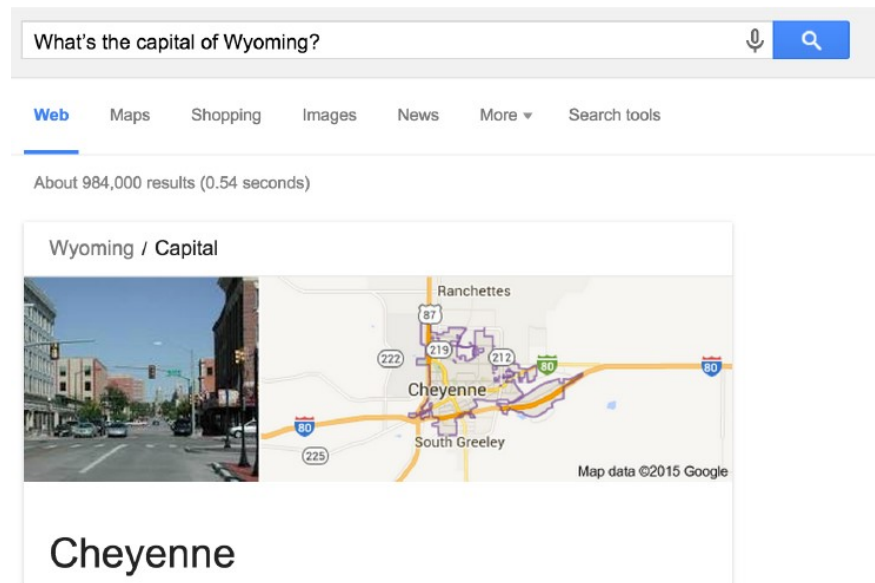
'Watson' computer wins at 'Jeopardy'



credit: ifunny.com

# Question answering

- Go beyond search



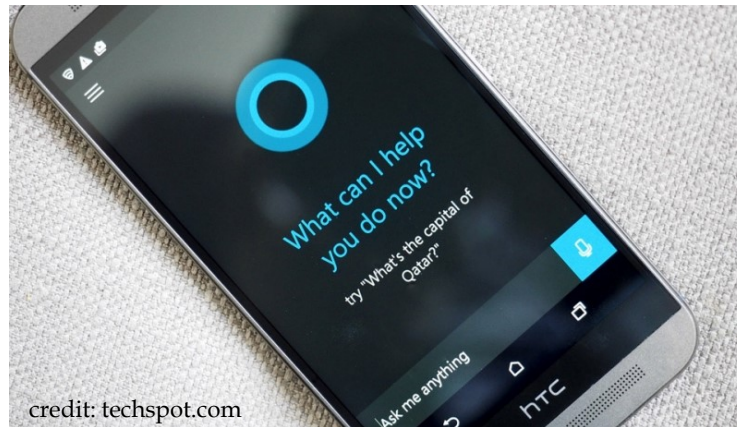
# Natural language instruction



<https://youtu.be/KkOCeAtKHlc?t=1m28s>

# Digital personal assistant

More on natural language instruction



- Semantic parsing – understand tasks
- Entity linking – “my wife” = “Kellie” in the phone book

# Information Extraction

- Unstructured text to database entries

New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

Person	Company	Post	State
Russell T. Lewis	New York Times newspaper	president and general manager	start
Russell T. Lewis	New York Times newspaper	executive vice president	end
Lance R. Primis	New York Times Co.	president and CEO	start

Yoav Artzi: Natural language processing



# **BÀI TOÁN PHÂN TÍCH QUAN ĐIỂM**

# Positive or negative movie review?

- • unbelievably disappointing
- • Full of zany characters and richly applied satire, and some great plot twists
- • this is the greatest screwball comedy ever filmed
- • It was pathetic. The worst part about it was the boxing scenes.

# Google Product Search



**HP Officejet 6500A Plus e-All-in-One Color Ink-jet - Fax / copier / printer / scanner**

**\$89 online, \$100 nearby** ★★★★★ 377 reviews

September 2010 - Printer - HP - Inkjet - Office - Copier - Color - Scanner - Fax - 250 sh

## Reviews

**Summary** - Based on 377 reviews



What people are saying

ease of use	<div><div></div><div></div><div></div><div></div><div></div></div>	"This was very easy to setup to four computers."
value	<div><div></div><div></div><div></div><div></div><div></div></div>	"Appreciate good quality at a fair price."
setup	<div><div></div><div></div><div></div><div></div><div></div></div>	"Overall pretty easy setup."
customer service	<div><div></div><div></div><div></div><div></div><div></div></div>	"I DO like honest tech support people."
size	<div><div></div><div></div><div></div><div></div><div></div></div>	"Pretty Paper weight."
mode	<div><div></div><div></div><div></div><div></div><div></div></div>	"Photos were fair on the high quality mode."
colors	<div><div></div><div></div><div></div><div></div><div></div></div>	"Full color prints came out with great quality."

# Bing Shopping

## HP Officejet 6500A E710N Multifunction Printer

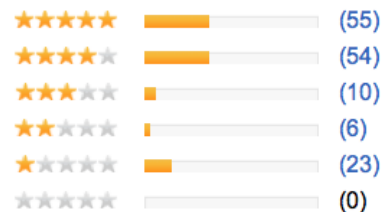
[Product summary](#) [Find best price](#) **Customer reviews** [Specifications](#) [Related items](#)



**\$121.53 - \$242.39** (14 stores)

☐ Compare

Average rating ★★★★★ (144)



Most mentioned



Show reviews by source

[Best Buy \(140\)](#)  
[CNET \(5\)](#)  
[Amazon.com \(3\)](#)

# Bài toán phân tích quan điểm

- Input: Cho một câu (đoạn) văn bản.
- Output: Xác định câu này là câu tích cực (1) hoặc tiêu cực (0).
- Thực hiện trên bộ dữ liệu review phim ảnh: IMDB

# Phân tích quan điểm

- Chuẩn bị dữ liệu: IMDB Movie Reviews
- Tiền xử lý dữ liệu
- Trích trợn đặc trưng (Vectorization)
- Chọn mô hình học máy và huấn luyện

# Tiền xử lý dữ liệu

- Đơn vị nhỏ nhất là từ. Dãy các từ:

Ví dụ:

**Text:** This is a cat. --> **Word Sequence:** [this, is, a, cat]

- Dữ liệu crawl từ web thường rất là “dirty” như mã Html, các từ viết tắt, ... nên cần phải tiền xử lý dữ liệu
- Dùng biểu thức chính qui thực hiện

# Biểu thức chính qui

- Biểu thức chính qui (or **regex**) is a sequence of characters that represent a search pattern
- Each character has a meaning; for example, “`.`” means any character that isn't the newline character: `'\n'`
- These characters are often combined with quantifiers, such as `*`, which means **zero or more**.
- Biểu thức chính qui rất có ích trong việc xử lý xâu ký tự



# Biểu thức chính qui

SYMBOL	USAGE
\$	Matches the end of the line
\s	Matches whitespace
\S	Matches any non-whitespace character
*	Repeats a character zero or more times
\S	Matches any non-whitespace character
*?	Repeats a character zero or more times (non-greedy)
+	Repeats a character one or more times
+?	Repeats a character one or more times (non-greedy)
[aeiou]	Matches a single character in the listed set
[^XYZ]	Matches a single character not in the listed set
[a-z0-9]	The set of characters can include a range

# Bài tập nhanh

- Ví dụ:

1) Trích ra số từ s:

s = 'My 2 favourite numbers are 8 and 25. My  
mobile is 0912203062'.

1) Trích ra email từ s:

s = 'Hello from shubhamg199630@gmail.com  
to priya@yahoo.com and  
bigdata@Samsung.com about the meeting  
@2PM'

# Vector đặc trưng

- Các đặc trưng sẽ hữu ích trong việc phân biệt giữa các thể loại

Table 3: Features to be computed for each text

- Counts:
  - First person pronouns
  - Second person pronouns
  - Third person pronouns
  - Coordinating conjunctions
  - Past-tense verbs
  - Future-tense verbs
  - Commas
  - Colons and semi-colons
  - Dashes
  - Parentheses
  - Ellipses
  - Common nouns
  - Proper nouns
  - Adverbs
  - *wh*-words
  - Modern slang acroynms
  - Words all in upper case (at least 2 letters long)
- Average length of sentences (in tokens)
- Average length of tokens, excluding punctuation tokens (in characters)
- Number of sentences

Higher values → this person is referring to themselves (to their opinion, too?)

Higher values → looking forward to (or dreading) some future event?

Lower values → this tweet is more formal. Perhaps not overly sentimental?

Acti

# Trích trọn đặc trưng

- Chúng ta gọi vector hóa (**vectorization**) là quá trình chung biến tập các tài liệu văn bản thành các vector đặc trưng số thực
- Kỹ thuật vector hóa đơn giản nhất là Bag Of Words (BOW). It starts with a list of words called the vocabulary (this is often all the words that occur in the training data)

# Trích trọn đặc trưng

- To use BOW vectorization in Python, we can rely on `CountVectorizer` from the `scikit-learn` library
- `scikit-learn` has a built-in list of stop words that can be ignored by passing `stop_words="english"` to the vectorizer
- Moreover, we can pass our custom pre-processing function `fromearlier` to automatically clean the text before it's vectorized.

**Training texts:** ["This is a good cat", "This is a bad day"]=>

**vocabulary:** [this, cat, day, is, good, a, bad]

**New text:** "This day is a good day" --> [1, 0, 2, 1, 1, 1, 0]

# Ứng dụng cho bài toán phân tích quan điểm

- Dữ liệu văn bản IMDB: The IMDB movie reviews dataset is a set of 50,000 reviews, half of which are positive and the other half negative
- Chúng ta có thể download dữ liệu từ:  
[http://ai.stanford.edu/~amaas/data/sentiment/aclImdb\\_v1.tar.gz](http://ai.stanford.edu/~amaas/data/sentiment/aclImdb_v1.tar.gz)
- Chúng ta có thư mục chứa dữ liệu là: `aclImdb`
- we can use the following function to load the training/test datasets from IMDB

# Ứng dụng cho bài toán phân tích quan điểm

- feature vectors that result from BOW are usually very large (80,000-dimensional vectors in this case)
- we need to use simple algorithms that are efficient on a large number of features (e.g., Naive Bayes, linear SVM, or logistic regression)

# Cải tiến mô hình hiện tại

- Trích trọng đặc trưng rất là quan trọng (Features Engineering)
- There are some biases attached with only looking at how many times a word occurs in a text. In particular, the longer the text, the higher its features (word counts) will be
- Dựa vào đặc trưng TF-IDF
- Dựa vào n-gram



# Cải tiến mô hình

- Đặc trưng TF-IDF
- Chúng ta có thể huấn luyện mô hình Linear SVM với đặc trưng TF-IDF đơn giản bằng việc thay thế hàm `CountVectorizer` bằng hàm `TfidfVectorizer`
- Kết quả tăng khoảng 2%

$$\text{TF}(\text{word}, \text{text}) = \frac{\text{number of times the word occurs in the text}}{\text{number of words in the text}}$$

$$\text{IDF}(\text{word}) = \log \left[ \frac{\text{number of texts}}{\text{number of texts where the word occurs}} \right]$$

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

# Cải tiến mô hình

- Sử dụng các từ độc lập sẽ không tốt. Ví dụ:
- if the word **good** occurs in a text, we will naturally tend to say that this text is positive, even if the actual expression that occurs is actually **not good**. **Cụm từ tốt hơn**
- Sử dụng n-gram để xử lý vấn đề này
- An N-gram is a set of N successive words (e.g., very good [ 2-gram] and not good at all [4-gram]). Using N-grams, we produce richer word sequences.
- Ví dụ với  $N = 2$ :

This is a cat. --> [this, is, a, cat, (this, is), (is, a), (a, cat)]

# Cải tiến mô hình

- In practice, including N-grams in our TF-IDF vectorizer is as simple as providing an additional parameter `ngram_range=(1, N)`.

```
vectorizer = TfidfVectorizer(stop_words="english",  
                             preprocessor=clean_text,  
                             ngram_range=(1, 2))
```

# Summary

1. Giới thiệu về bài toán NLP
2. Mô hình Logistic regression là phân lớp tuyến tính đơn giản nhưng hiệu quả
3. Thực hành với bài toán phân tích quan điểm