

Phân tích quan điểm sử dụng mạng CNN– BÀI 10

AI Academy Vietnam

Thông tin giảng viên

- TS. Phan Việt Anh
- Email: anhpv@lqdtu.edu.vn
- Mobile: 0975 639 757
- Research interest: machine learning, natural/programming language processing.

-

GIỚI THIỆU VỀ PHÂN TÍCH QUAN ĐIỂM

1. Một số ví dụ

- Cảm xúc sau khi xem xong 1 bộ phim: **tích cực** hoặc **tiêu cực**



- Thất vọng



- Nhiều nhân vật khờ khạo, châm biếm được áp dụng phong phú, và một số tình tiết tuyệt vời



- đây là bộ phim hài hay nhất từng được xem



- Thật là thảm hại. Phần tệ nhất về nó là những cảnh đấm bốc.

Tìm kiếm sản phẩm trên Google



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

Sản phẩm thương mại điện tử



5 trên 5



Tất Cả

5 Sao (9)

4 Sao (0)

3 Sao (0)

2 Sao (0)

1 Sao (0)

Có Bình Luận (9)

Có Hình Ảnh / Video (9)



thanh tuan5188



Phân loại hàng: Trắng,Không Xạc

Chất lượng rất đúng như quảng cáo, máy chính hãng, các ứng dụng hoạt động rất hiệu quả và nhanh, màn hình cảm ứng nhạy, màu sắc đẹp nhìn rất bắt mắt, camera chụp hình không chế vào đâu được, hoàn thiện rất tốt, shop giao hàng chuẩn làm mình hài lòng lắm, mọi người nên sử dụng các sản phẩm của shop.

Chất lượng sản phẩm tuyệt vời

Đóng gói sản phẩm rất đẹp và chắc chắn

Shop phục vụ rất tốt



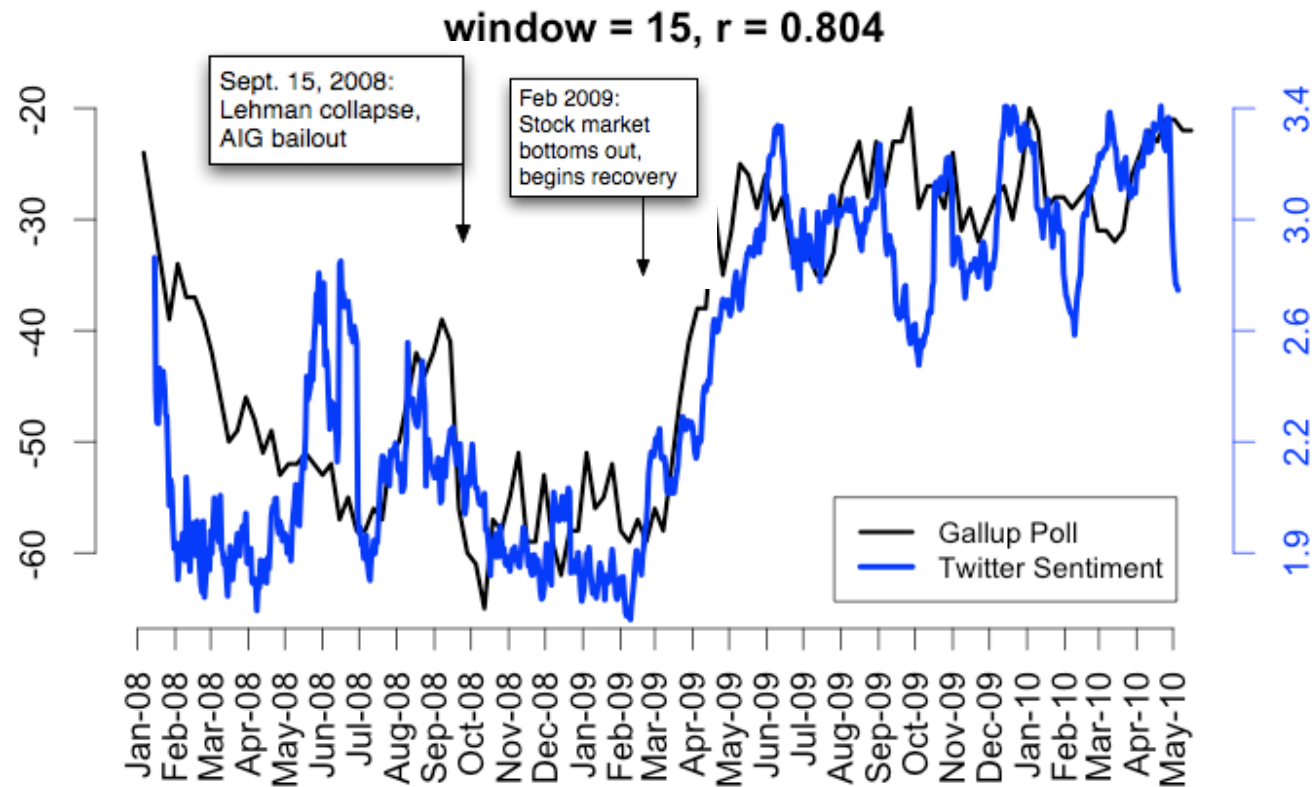
haonam61810



Phân loại hàng: Đen,Có Xạc

quá chuẩn luôn nha anh em, hàng đúng chính hãng, các chức năng cơ bản rất ổn định, màn hình siêu nét, cảm ứng nhạy lướt êm ru, chất lượng quá đỉnh, mình hài lòng lắm, shop bán hàng rất có tâm, chế độ bảo hành rõ ràng, shop đã tạo cảm giác an tâm khi sử dụng sản phẩm, mình sẽ giới thiệu bạn bè ủng hộ

Quan điểm trên Twitter so với Cuộc thăm dò lòng tin của người tiêu dùng Gallup



Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In ICWSM-2010

2. Bài toán Phân tích quan điểm

- Phân tích quan điểm là **phát hiện** các **thái độ** với *đồ vật, sự kiện hoặc con người*
- Có nhiều tên:
 - Trích xuất ý kiến (Opinion extraction)
 - Khai phá ý kiến (Opinion mining)
 - Khai thác quan điểm (Sentiment mining)
 - Phân tích chủ quan (Subjectivity analysis)

3. Một số lĩnh vực ứng dụng

- Phim: tích cực hay tiêu cực?
- Sản phẩm: Mọi người nghĩ gì về sản phẩm mới?
- Tâm lý công chúng: thế nào là niềm tin của người tiêu dùng?
Họ đang tin tưởng hay thất vọng?
- Chính trị: mọi người nghĩ gì về ứng viên?
- Dự đoán: dự đoán kết quả bầu cử hoặc xu hướng thị trường từ quan điểm xã hội.

PHÂN TÍCH QUAN ĐIỂM BẰNG PHƯƠNG PHÁP TRUYỀN THỐNG

1. Dữ liệu thực nghiệm

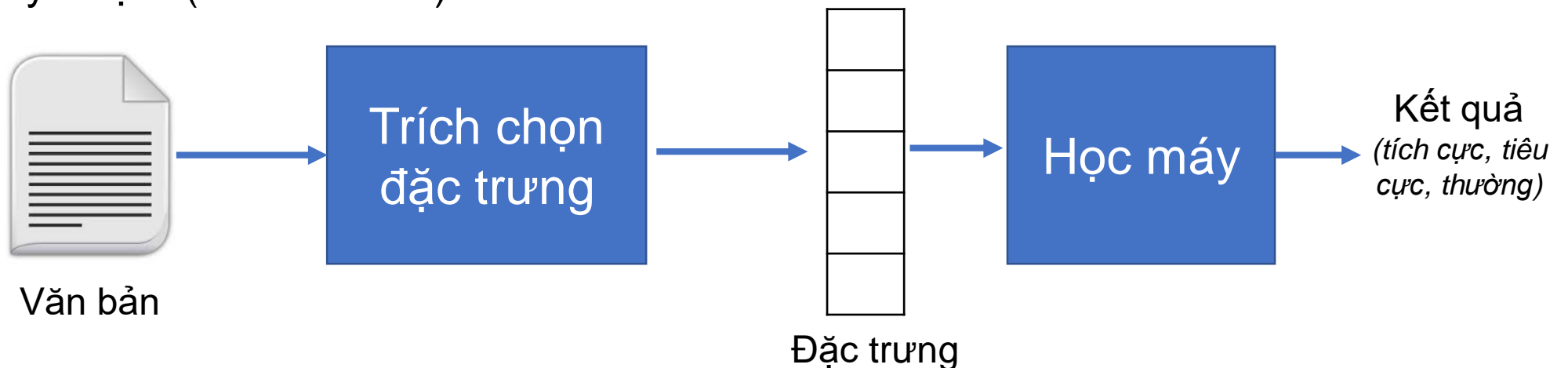
- Dự đoán quan điểm bài viết trên Twitter
- Dữ liệu:
 - Huấn luyện: 5, 971 câu (train.csv)
 - Kiểm tra: 4, 000 câu (test.csv)
 - Mỗi dòng trong file train.csv: id, quan điểm, bài viết

Id	Category	Tweet
635930169241374720	neutral	IOS 9 App Transport Security. Mm need to check if my 3rd party network pod supports it

- Nhóm quan điểm:
 - positive (tích cực)
 - negative (tiêu cực)
 - neutral (bình thường)

2. Các bước thực hiện

- Tách từ: câu/văn bản \rightarrow tập các từ
- Trích chọn đặc trưng: câu/ văn bản \rightarrow vector biểu diễn
- Các thuật toán học máy
 - Naïve Bayes
 - SVM
 - Hồi quy (regression)
 - Cây quyết định (decision tree)



Các vấn đề tách từ

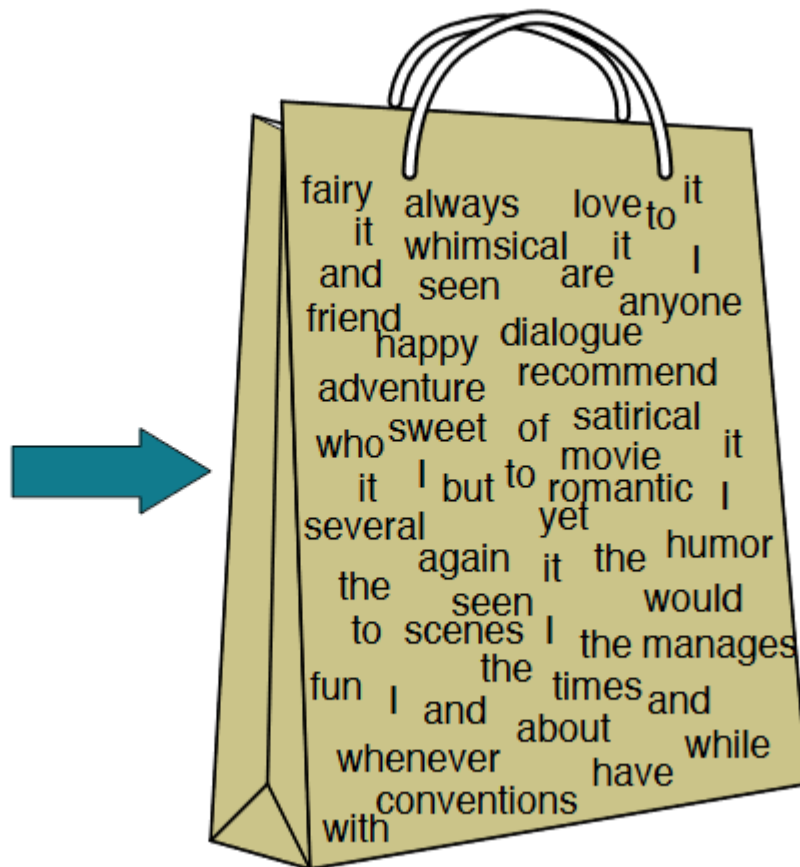
- Xử lý các thẻ HTML, XML
- Thẻ của Twitter (tên, hash tags)
- Chữ hoa – chữ thường
- Số điện thoại, ngày tháng
- Cảm xúc



- Đặc trưng
 - Từ (túi từ - bag-of-words)
 - N-grams
 - Từ loại (tính từ, trạng từ)
 - Các từ thể hiện quan điểm (lexicon-based)
 - Từ có tính chất đối nghĩa (phủ định), ví dụ “tốt” vs “*không* tốt”
 - Sự phụ thuộc cú pháp
- Trọng số đặc trưng
 - Tần suất (term frequency)
 - inverse document frequency (TF.IDF)
 - Vị trí các từ, ví dụ: tiêu đề, câu đầu, câu cuối

Bag of words

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

1. Văn bản



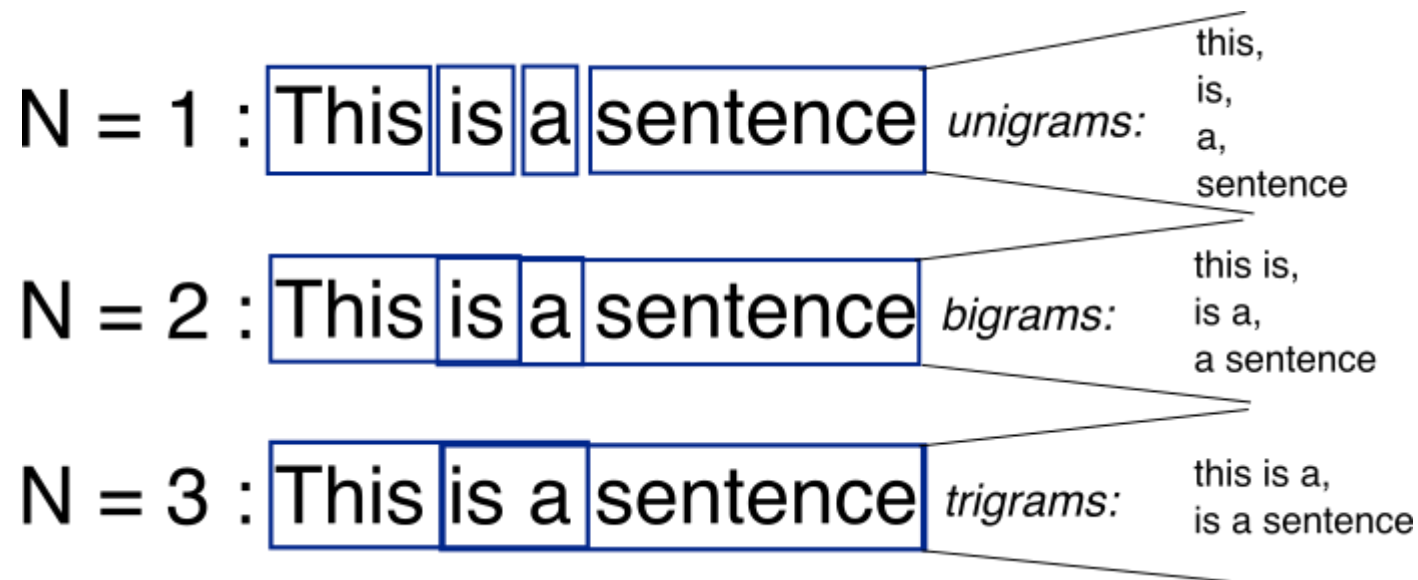
2. Từ điển



3. Vector

Bag-of-Words with N-grams

- N-grams: a contiguous sequence of n tokens from a given piece of text



TF-IDF (term frequency - Inverse Document Frequency)

- Dữ liệu gồm:
 - n tài liệu: D_1, \dots, D_n
 - t từ: T_1, \dots, T_n
- Từ xuất hiện nhiều trong tài liệu \Rightarrow quan trọng
 - $F_{ij} = w_{ij}$
- Chuẩn hóa
 - $TF_{ij} = F_{ij} / \max\{F_{ij}\}$

Trọng số các từ: IDF (Inverse Document Frequency)

- Từ xuất hiện ở nhiều tài liệu **khác nhau** => không quan trọng
 - DF_i = số tài liệu chứa từ i
 - $IDF_i = \log_2(N/DF_i)$, N = số tài liệu
- Log: giúp tỉ lệ tf và idf tương đồng.
 - $TF-IDF_{ij} = TF_{ij} * IDF_i$
- Ma trận vector biểu diễn
 - $w_{ij} = tf_{ij} * IDF_i$

$$\begin{pmatrix} & T_1 & T_2 & \dots & T_t \\ D_1 & w_{11} & w_{21} & \dots & w_{t1} \\ D_2 & w_{12} & w_{22} & \dots & w_{t2} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ D_n & w_{1n} & w_{2n} & \dots & w_{tn} \end{pmatrix}$$

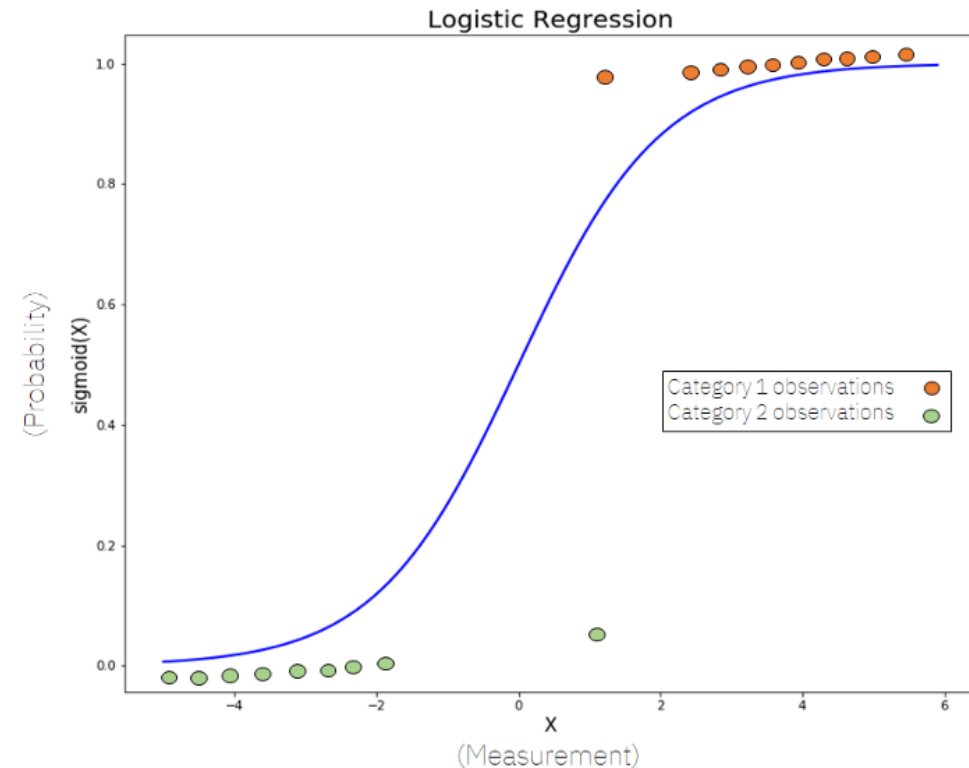
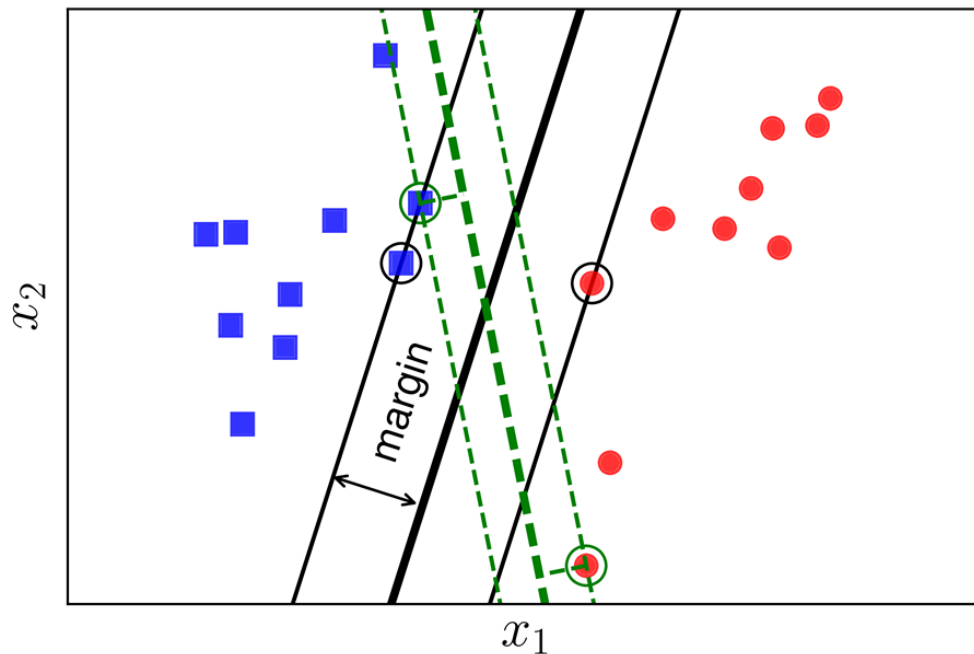
Các thuật toán học máy

- Support vector machines

$$\mathbf{w}^T \mathbf{x} + b = 0,$$

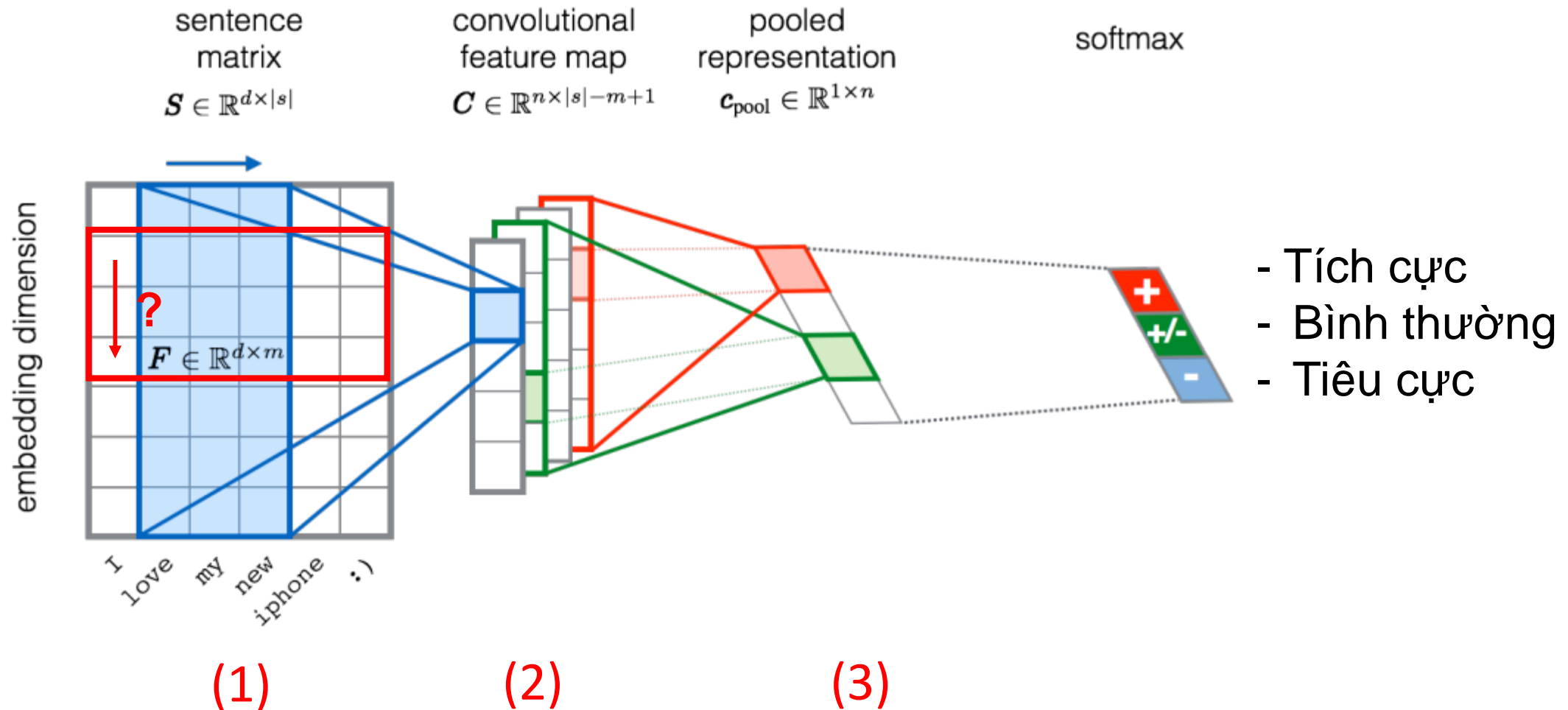
Logistic regression

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$



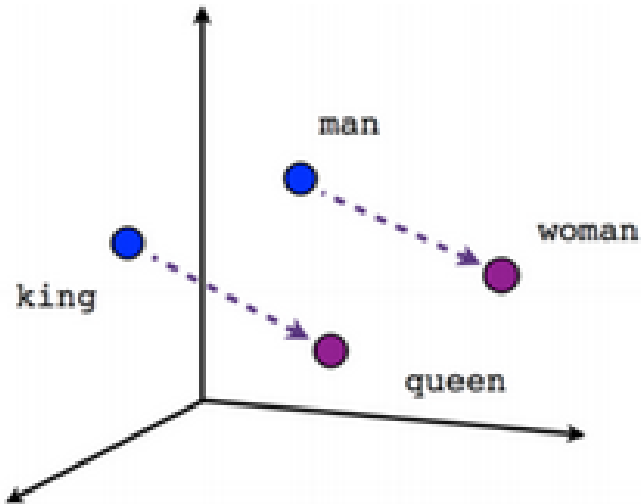
PHÂN TÍCH QUAN ĐIỂM SỬ DỤNG CNN

Mạng CNN cho Phân tích quan điểm



1. Biểu diễn vector cho từ (word2vec)

- Từ \rightarrow vector



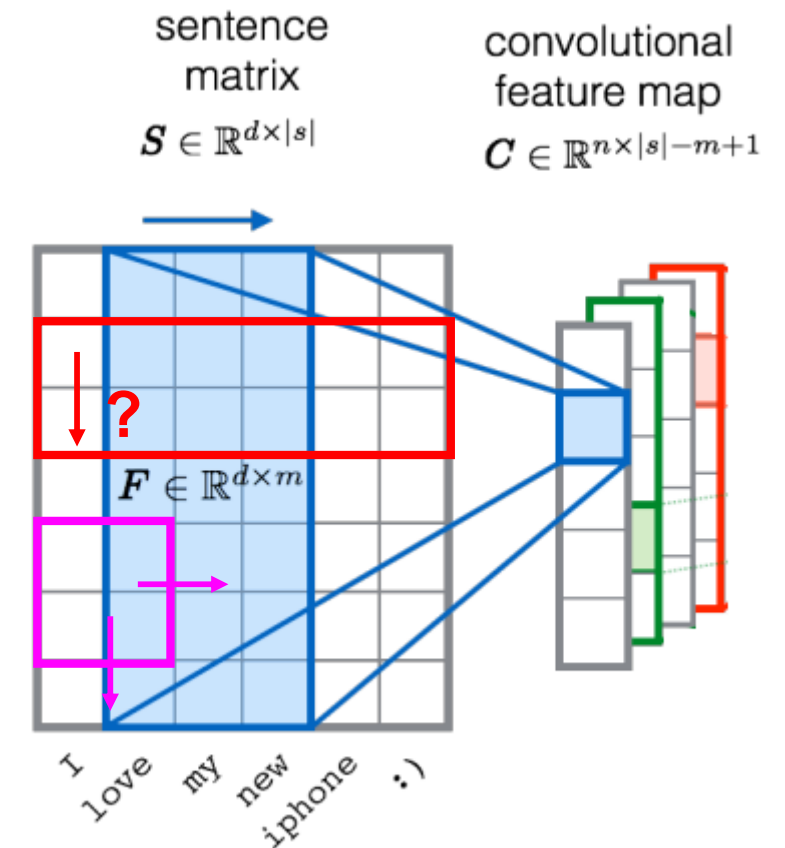
- Một số word2vec
- one-hot

- Câu \rightarrow Ma trận

0.6	0.5	0.2	-0.1	0.4
0.8	0.9	0.1	0.5	0.1
0.4	0.6	0.1	-0.1	0.7
...
...
...
...

Lớp tích chập

- Bộ lọc kích thước $k = 2, 3, 4, \dots$
- Kích thước cửa sổ trượt $k \times d$
- Cửa sổ trượt màu đỏ/tím có nên sử dụng không?

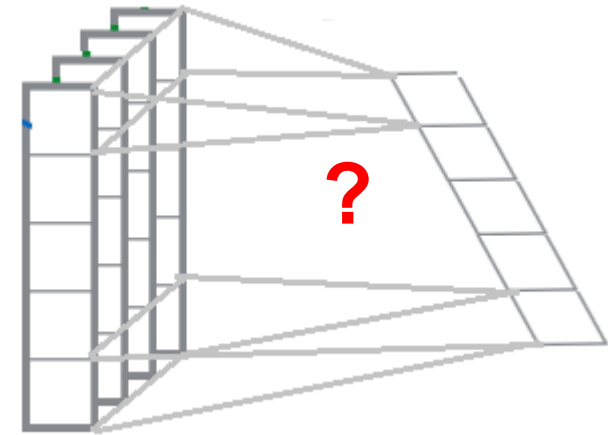
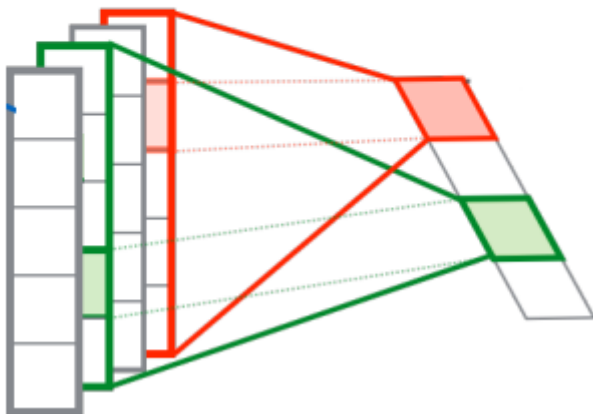


Lớp pooling

- Giảm chiều, vector hóa đặc trưng

convolutional
feature map
 $C \in \mathbb{R}^{n \times |s| - m + 1}$

pooled
representation
 $c_{\text{pool}} \in \mathbb{R}^{1 \times n}$



Lớp pooling có nên thực hiện như hình này?

PHÂN TÍCH QUAN ĐIỂM THEO KHÓA CẠNH

Quan điểm theo khía cạnh

- Trong thực tế: tìm quan điểm theo khía cạnh
- Ví dụ: *đồ ăn* rất *ngon* nhưng *dịch vụ* vô cùng *tệ*
=> *Quan điểm của câu này là gì?*

Tìm khía cạnh của quan điểm

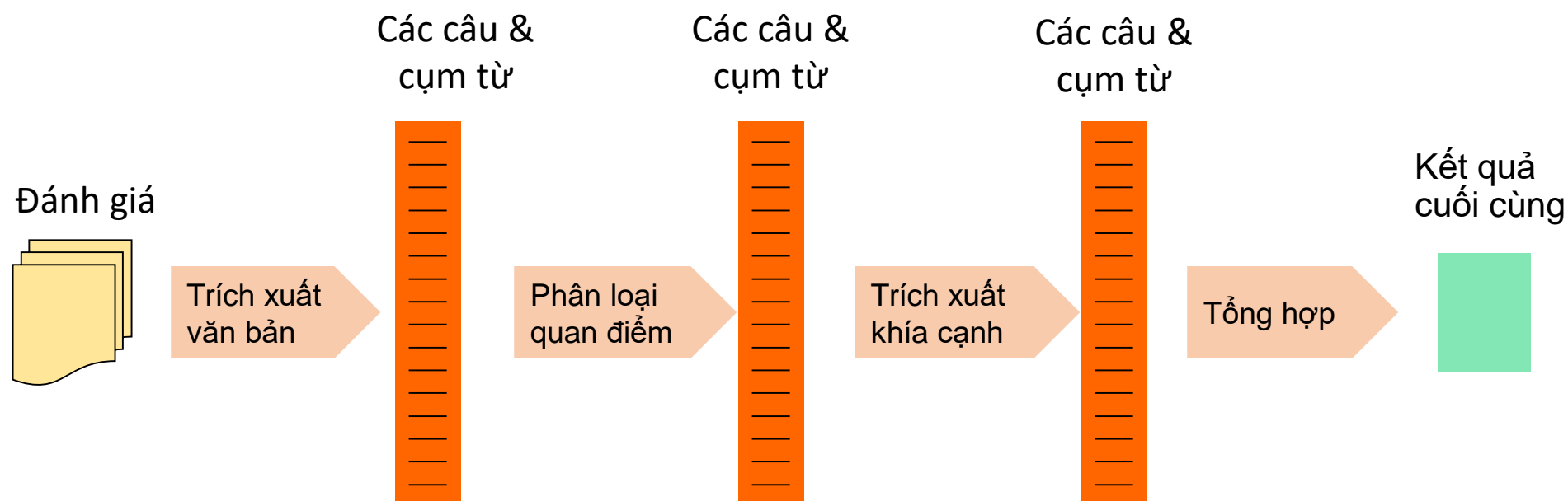
- Các cụm từ xuất hiện thường xuyên + luật (rules)
 - Tìm các cụm từ thường xuyên xuất hiện ở các câu (“*fish tacos*”)
 - Lọc bằng luật dạng “xuất hiện bên phải sau từ thể hiện quan điểm”
 - “... *great fish tacos*” => “*fish tacos*” có thể là 1 khía cạnh

Casino	casino, buffet, pool, resort, beds
Children’s Barber	haircut, job, experience, kids
Greek Restaurant	food, wine, service, appetizer, lamb
Department Store	selection, department, sales, shop, clothing

Tìm khía cạnh của quan điểm

- Tên khía cạnh có thể không ở trong câu
- Một số lĩnh vực (ví dụ nhà hàng, khách sạn), các khía cạnh tường minh
- Phân loại có giám sát:
 - Gán nhãn một tập nhỏ các câu đánh giá với các khía cạnh.
 - Ví dụ: đồ ăn, trang trí, dịch vụ, giá, NONE (không có khía cạnh)
 - Huấn luyện bộ phân lớp để phát hiện khía cạnh cho câu
 - 1 câu → khía cạnh

Tìm quan điểm đối với các khía cạnh



S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. Reis, and J. Reynar. 2008. Building a Sentiment Summarizer for Local Service Reviews. WWW Workshop

Một số kết quả của Blair-Goldensohn

Rooms (3/5 stars, 41 comments)

- (+) The room was clean and everything worked fine – even the water pressure ...
- (+) We went because of the free room and was pleasantly pleased ...
- (-) ...the worst hotel I had ever stayed at ...

Service (3/5 stars, 31 comments)

- (+) Upon checking out another couple was checking early due to a problem ...
- (+) Every single hotel staff member treated us great and answered every ...
- (-) The food is cold and the service gives new meaning to SLOW.

Dining (3/5 stars, 18 comments)

- (+) our favorite place to stay in biloxi.the food is great also the service ...
- (+) Offer of free buffet for joining the Play

Q&A

Thank you!