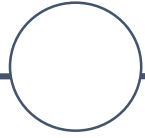


# Using CV model to enrich training corpus for NLP model

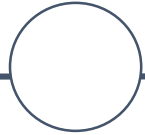


Presenter: Bao-Dai



# Content

- 1. Introduction**
- 2. Methods**
  - a. Genre classification**
  - b. Content extraction**
- 3. Evaluation result**



# Content

- 1. Introduction**
2. Methods
  - a. Genre classification
  - b. Content extraction
3. Evaluation result

1

## Problem definition




# Two genres of webpage

Tuesday 6 November 2018

**News Health**

**Harris has no date for extra beds as trolley numbers rise**



Unwiling: Simon Harris with Doreen Kelly of CHM2 Day Ward at the opening of the cataract centre at Nenagh Hospital. Photo: Brian Arthur

David Raleigh and Eilish O'Regan  
November 6 2018 2:30 AM

The schedule for opening extra winter beds to ease hospital overcrowding still remains unclear despite a worrying rise in patients on trolleys yesterday.

Health Minister Simon Harris was unable to say when the promised 79 extra beds would be available other than to repeat that "they will come on stream between now and early 2019".

He admitted he was worried about the inability of hospitals to cope with an influx of winter patients.

It comes as the HSE's own figures show 312 patients were on trolleys yesterday, a 35.6pc rise over the same day last year.

Of these, 163 were waiting over nine hours for a bed - a worsening scenario compared to 115 last year.

Figures from the Irish Nurses and Midwives Organisation (INMO), which include patients on trolleys who were moved to wards, showed 449 in all were in need of a bed with 51 patients on trolleys at the worst hit, University Hospital Limerick.

Others struggling to find beds were Cork University Hospital, where 42 were on trolleys, and Letterkenny Hospital, which had 38 patients suffering delays. Mr Harris who was speaking in Nenagh, Co Tipperary, where he opened a new cataract surgical centre, told reporters: "I say this respectfully, everybody asks am I worried about winter approaching.

"I'm actually worried about the capacity of the health service every day of the week.

**Most Read** **Most Shared**

Only four hospitals get extra beds to help ease trolley crisis **Read**

Thousands caught in health 'twilight zone' over private insurance **Read**

GPs to hold EGM on new abortion law **Read**

Women can ring 24/7 line to seek abortion **Read**

Harris has no date for extra beds as trolley numbers rise **Read**

**Promoted links**

Mac Users Guide (2018) - The Only Antivirus Provider You Should Use. **Read**

Ông lưng đanh cho iPhone Xs Max RINGKE Fusion **Read**

How to update your home using a colour palette **Read**

Lion Air crash puts focus on safety in Asia's booming air travel market **Read**

**Editor's Choice**

David E. Wade: 'Will midterm elections embolden or undermine President Trump?' **Read**

Brendan O'Connor: 'Why does everyone hate me for being calm?' **Read**

EXCLUSIVE: U2 feared they were finished: the inside story of the night Bono lost his voice **Read**

## Article

**Headphones**

**Elcster 136 Kids Headphones Children Girls Boys Teens Adults Foldable Adjustable On Ear Headsets 3.5mm Jack Compatible iPad Cellphones Computer Kindle MP3/4 Airplane School Tablet Black/Orange**  
by ELCESTER  
\$14<sup>99</sup>

**Product Features**  
... Durable & Not Tangle. The headphone cord length is 4.9 ft (1.5m), using ...

**Refine by**

**Brand**

- ☐ Panasonic
- ☐ AmazonBasics
- ☐ Sennheiser
- ☐ Sony
- ☐ Bose
- ☐ Audio-Technica
- ☐ ALIEN
- ☐ Beats
- ☐ Mpow
- ☐ Anker
- ☐ Sound Intense
- ☐ Behringer
- ☐ Headphones
- ☐ Headphones

**Headphone Feature**

Microphone  
Sports & Exercise  
Lightweight  
Noise-Cancelling  
Foldable  
Tangle-Free Cord  
DJ Style  
Water-Resistant  
Phone Control  
Noise-Isolating  
Volume Control

**Headphone Wireless Type**

- ☐ Bluetooth
- ☐ RF
- ☐ Infrared
- ☐ NFC

**Condition**

New  
Used

**Headphone Earcup Style**

- ☐ Closed-Back
- ☐ Open-Back
- ☐ Semi-Open Back

**Color**

**Avg. Customer Review**

★★★★★ & Up  
★★★★ & Up  
★★★ & Up  
★★ & Up  
★ & Up

**Ear Buds Wired Earphones Earbuds with Remote and Mic 3.5mm in Ear Earbud Headphones with Microphone and Volume Control Stereo Noise Isolating Compatible Android Phones, iPhone, iPod, iPad, Samsung**  
by Ueader  
\$12<sup>99</sup>

**Product Features**  
... with coupon  
... all device. In ear earbud headphones with microphone compatible ...

**Bluetooth Over Ear Headphones, Foldable Hi-Fi Deep Bass Wireless Headphones with SD Card Slot Wired Headset with Mic for Airplane Travel Office Work (Blue)**  
by Dron  
\$21<sup>99</sup>

**Product Features**  
... please take off the bluetooth headphones every 2-3 hrs to get your ears ...

**Shalle Magnetic Wireless In-Ear Earphones, Sport Fit Design Bluetooth Earbuds Headphones, Sweatproof Headsets(Bluetooth 4.2 Super Sound Quality)**  
by Shalle  
\$22<sup>99</sup>

**Product Features**  
... magnets let you attach the two headphones together when not using and ...

**Sponsored: Betron B510 Earphones Headphones, Powerful Bass Driven Sound, 12mm Large Drivers, Ergonomic Design for iPhone, iPad, iPod, Samsung and MP3 Players**  
by Betron  
\$11<sup>99</sup>

**Product Features**  
... magnets let you attach the two headphones together when not using and ...

**Sponsored: Arctic Foldable Headphones with Microphone and Volume Control | NRGSound CL750 On-Ear Stereo Earphones | Great for Kids/Teens/Adults (Black/Red)**  
by NRGSound  
\$18<sup>99</sup>

**Product Features**  
... magnets let you attach the two headphones together when not using and ...

## List-viewed

## 6

1

# Why do we need to extract web content?

- 20 News group
- Reuters News
- Sentiment140
- Yelp reviews



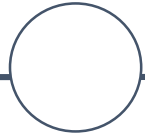
Academia

There is **no suitable dataset** for this problem. I must build this **manually**



Industry

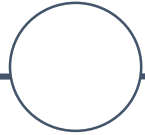
Image source: Moscow Institute of Physics and Technology



# Content

1. Introduction
- 2. Methods**
  - a. Genre classification
  - b. Content extraction
3. Evaluation result



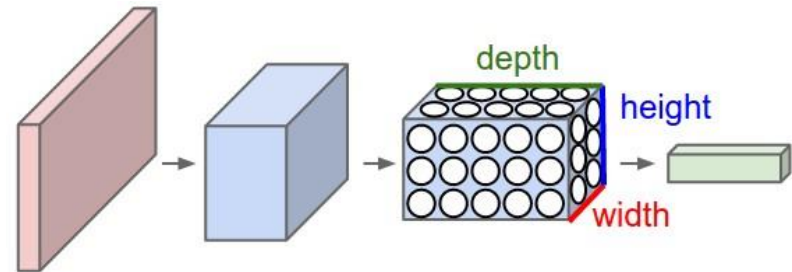
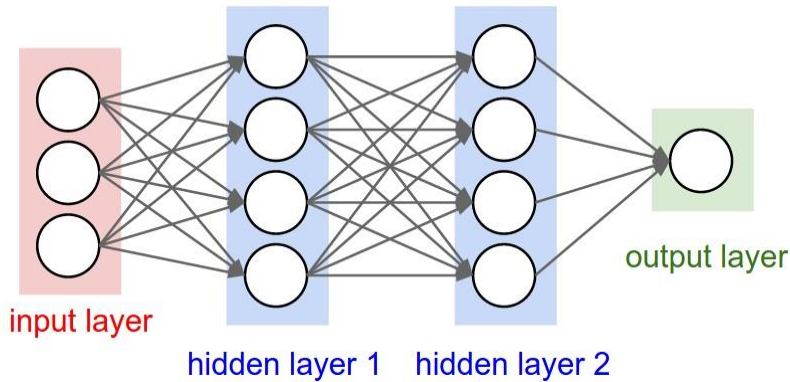


# Content

1. Introduction
- 2. Methods**
  - a. Genre classification**
  - b. Content extraction
3. Evaluation result

2

# Convolutional Neural Network



## Feedforward Neural Network

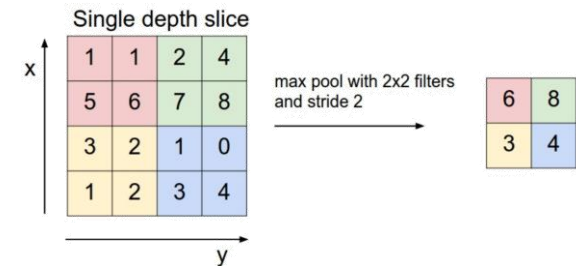
## Convolutional Neural Network

1 <sub>x1</sub>	1 <sub>x0</sub>	1 <sub>x1</sub>	0	0
0 <sub>x0</sub>	1 <sub>x1</sub>	1 <sub>x0</sub>	1	0
0 <sub>x1</sub>	0 <sub>x0</sub>	1 <sub>x1</sub>	1	1
0	0	1	1	0
0	1	1	0	0

Image

4		

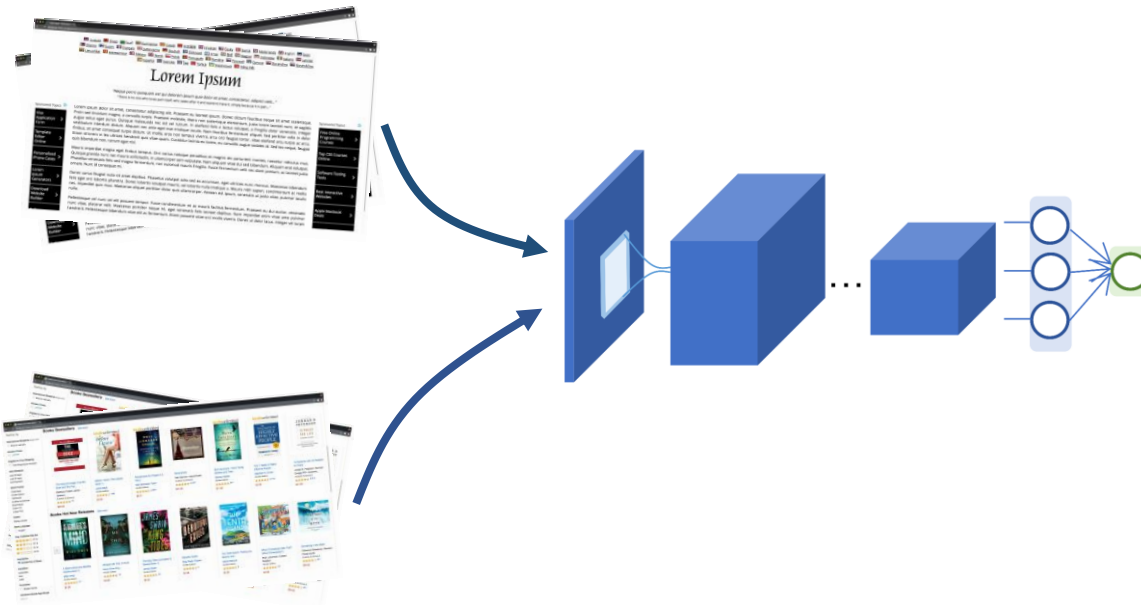
Convolved Feature



2

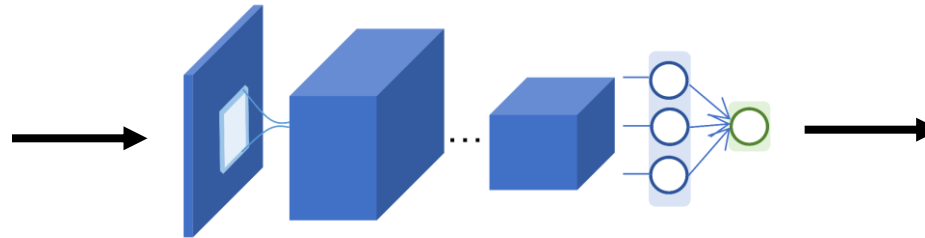
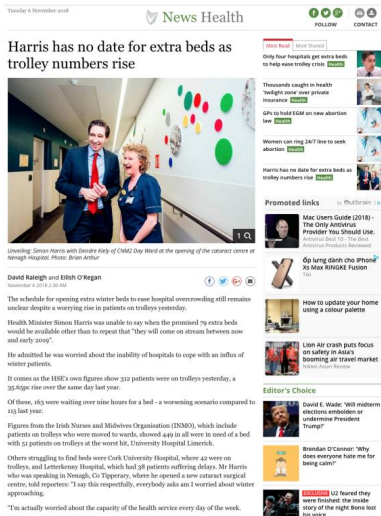
## Genre classification

- Use CNN architecture



Training and testing  
phase

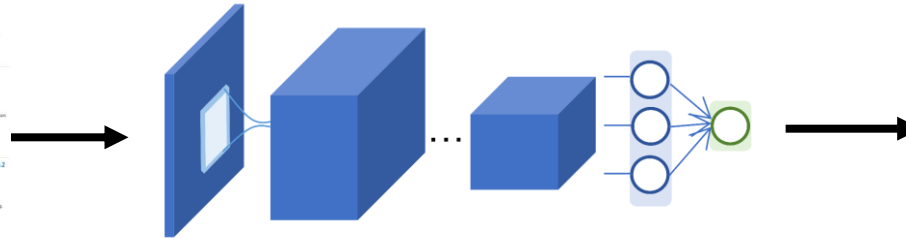
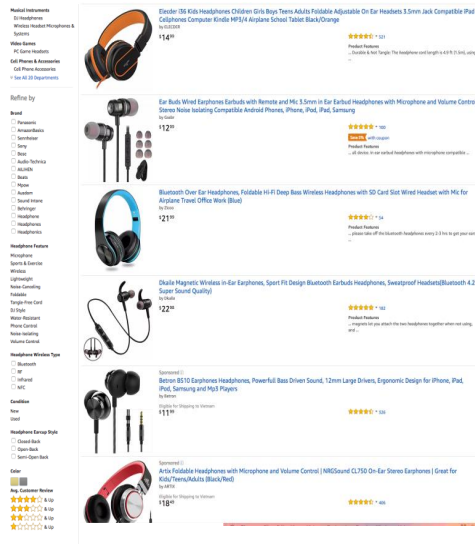
# Genre classification



Predicting phase

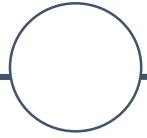
2

# Genre classification



List-viewed

Predicting phase

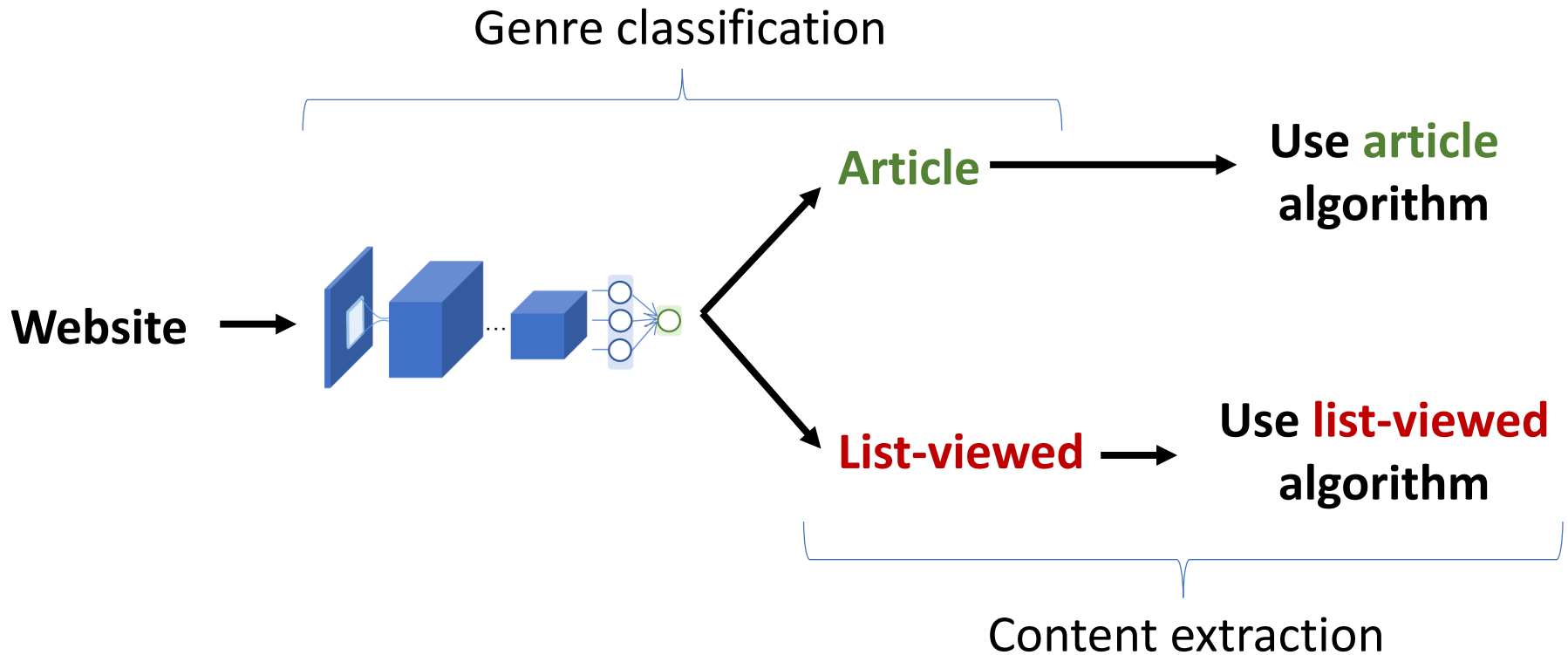


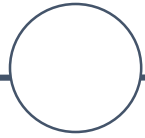
# Content

1. Introduction
- 2. Methods**
  - a. Genre classification
  - b. Content extraction**
3. Evaluation result

2

## The purpose of web genre





# Content

1. Introduction
- 2. Methods**
  - a. Genre classification
  - b. Content extraction**
    - i. Article**
    - ii. List-viewed
3. Evaluation result





# Article page extraction

Video Events Viewpoints World Business Entertainment Sports Law Education Health Living Travel Science Digitization Vehicles Community Center of

Documentary Analysis of the Vietnamese 5 Life That Here Military

World Vietnamese 5


Friday, 9/11/2018, 16:30 (GMT + 7)

## The Vietnamese guy who becomes a Washington state legislator

Joe Nguyen was one of the first two Vietnamese-Americans to be elected to the Washington state legislature following the mid-term US vote.

• Many Vietnamese-Americans win midterm elections / 13 Vietnamese-Americans win the US midterm elections

Northwest Asian Weekly on Nov. 8 reported that Joe Nguyen won against rival Shannon Braddock in the 34th race. With 57.4% of the votes, he became the first Vietnamese in the upper Washington State. This is one of the highlights of the midterm elections here.



Joe Nguyen, the first Vietnamese lawmaker in the Senate of Washington. Photo: People for Joe Nguyen

Like many Vietnamese who came to the United States a few decades ago, Joe's family life was not easy. They relied heavily on social security and when Joe's father was paralyzed and died after an accident, his mother became the main caretaker for four children. The situation makes Joe brothers soon self-aware of their responsibilities and do not mind any work to have money to cover life.

In an interview in May, Joe said he had slept on a filthy cushion in the basement of his family's shabby home. His mother worked as a tailor and Joe often woke up when she sewed backpacks.

In spite of his hard work, he was consistently at the helm and a senior in high school. He then entered Seattle University and served as president of the student body for two years, a rare event in school history.

Joe is very grateful to the people who helped his family, so he later became actively involved in social activities as well as advocating affordable housing and health services for the people. Joe is a senior executive at Microsoft Corporation, in addition to the president of Wellspring Family Services, which helps children and homeless families.

Most view

The needle holder into the Australian strawberry is Vietnamese women

Japanese arrested a group of young Vietnamese who steal cosmetics in 3 minutes

Vietnamese suspects may plan to insert needles into strawberries for months

The Vietnamese woman became an American MP at the age of 31

Vietnamese people spend billions to trade in markets in Vietnam

Proficiency in English by practicing speaking with Western teachers every

Sponsorship TOPICA NATIVE

Tom's utterances criticize America

Experts say he has not done anything to calm the tension when speaking at an international conference in Shanghai.

China may be in awe of the mid-term US election

The Korean intentions of publishing the portrait of Kim Jong-un

Justice Minister Trump, Trump could start a 'bloodshed' administration

This life

Dogs sit down on the roadside where the owner dies more than 80 days in China

Death Hole 'swallows' the woman walking on the sidewalk in China

American boys were investigated because of the fascist picture

Chinese boy crouched in the car for his mother in the rain

Category: Hepatitis B

Victory of millions of

89% of people with

PowerPoint File Edit View Insert Format Arrange

Elements Console Sources Network Performance Memory Application Security

```

<div id="myvne_taskbar"></div>
<!-- end taskbar -->
<header class="p_header"></header>
<header id="header" class="section_m_header"></header>
<!--end header-->
<!--main_menu menu PC-->
<!-- Start Menu -->
<nav id="main_menu" class="p_menu"></nav>
<!-- End Menu -->
<!-- Start SUB MENU -->
<div class="wrap_sub_menu"></div>
<section class="cat_header clearfix"></section>
<!--End main_menu-->
<!-- Breadcrumb -->
<!--End Breadcrumb-->
<!-- CONTENT -->
<section class="container" data-component-modulejs="detail" data-component-page-type="text" data-component-page-config="{"article_id":"3836552"}">
<!--wrap_sidebar_12-->
<section class="wrap_sidebar_12">
<section class="sidebar_1">
<header class="clearfix"></header>
<h1 class="title_news_detail mb10"></h1>
<h2 class="description"></h2>
<p class="related_news"></p>
<!--article class="content_detail fck_detail width_common block_ads_connect"> == $0
<p class="Normal">
<em></em>
<font style="vertical-align: inherit;">
<font style="vertical-align: inherit;"></font>
<font style="vertical-align: inherit;">This is one of the highlights of the midterm elections here.</font>
</p>
<table class="tplCaption" cellpadding="3" cellspacing="0" border="0" align="center" style="width: 350px;">
<tbody></tbody>
</table>
<p class="Normal">
<font style="vertical-align: inherit;">
<font style="vertical-align: inherit;">
"
Like many Vietnamese who came to the United States a few decades ago, Joe's family life was not easy. "
</font>
<font style="vertical-align: inherit;"></font>
<font style="vertical-align: inherit;"></font>
</p>
<p class="Normal"></p>
<p class="Normal"></p>
<p class="Normal"></p>
<table class="tplCaption" cellpadding="3" cellspacing="0" border="0" align="center" style="width: 350px;">

```

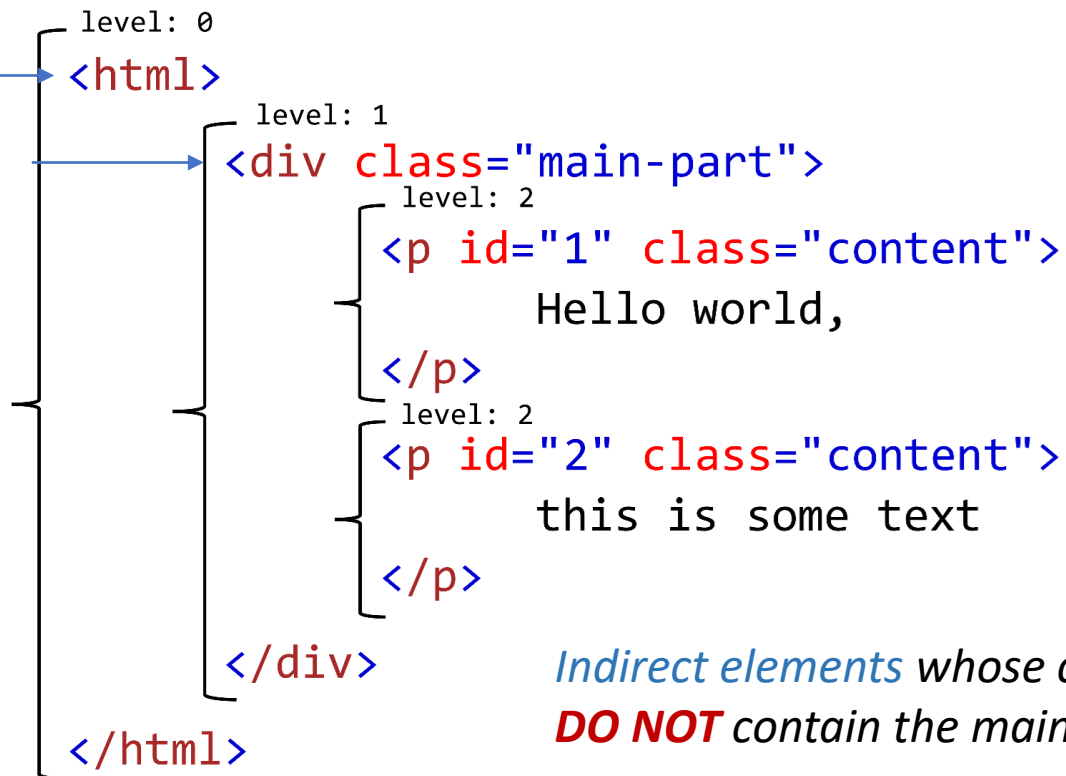
html body section section section article.content\_detail.fck\_detail.width\_common.block\_ads\_connect

This **tag** contains main content

## Indirect element

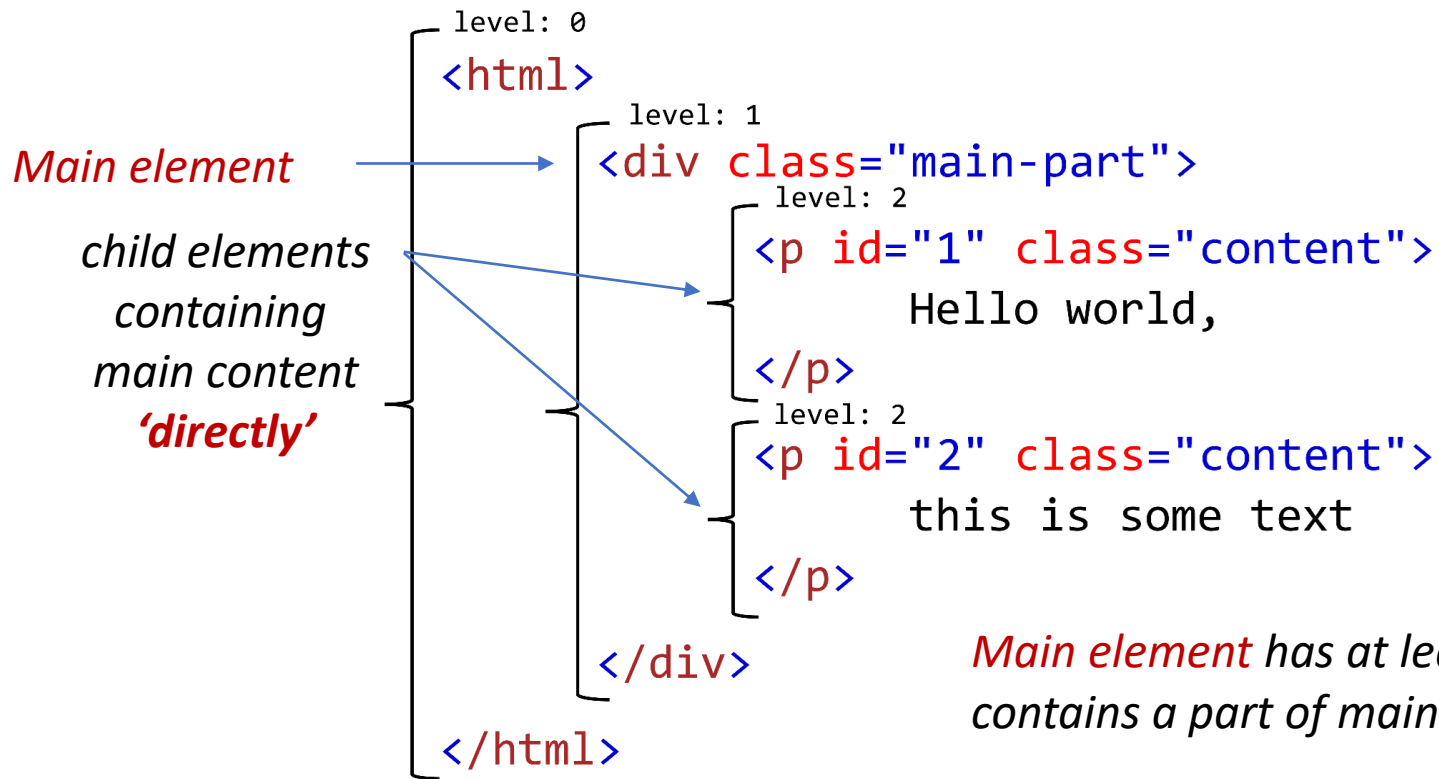
*"Indirect"  
element*

*child elements  
of <html>*



*Indirect elements* whose child elements  
**DO NOT** contain the main content directly

## Main element



*Main element has at least one child that contains a part of main content directly*



[Video](#)
[Events](#)
[Viewpoints](#)
[World](#)
[Business](#)
[Entertainment](#)
[Sports](#)
[Law](#)
[Education](#)
[Health](#)
[Living](#)
[Travel](#)
[Science](#)
[Digitization](#)
[Vehicles](#)
[Community](#)
[Center of](#)

Joe Nguyen was one of the first two Vietnamese-Americans to be elected to the Washington state legislature following the mid-term US vote.

*Northwest Asian Weekly* on Nov. 8 reported that Joe Nguyen won against rival Shannon Braddock in the 34th race. With 57.4% of the votes, he became the first Vietnamese in the upper Washington State. This is one of the highlights of the midterm elections here.



Joe is very grateful to the people who helped his family, so he later became actively involved in social activities as well as advocating affordable housing and health services for the people. Joe is a senior executive at Microsoft Corporation, in addition to the president of Wellspring Family Services, which helps children and homeless families.



The needle holder into the Australian strawberry is Vietnamese women

Japanese arrested a group of young Vietnamese who steal cosmetics in 3 minutes

Vietnamese suspects  
may plan to insert  
needles into strawberries  
for months

The Vietnamese woman became an American MP at the age of 31

### Vietnamese people spend billions to trade in markets in Vientiane



**TRƯỢT TUYẾT**  
TẠI HÀN QUỐC  
*Hơn cả một bài nghiệm!*  
Từ 12/11 - 31/12/2018

Tom's utterances criticize America

Experts say he has not done anything to calm the tension when speaking at an international conference in Shanghai.

- China may be in awe of the mid-term US election
- The Korean intentions of publishing the portrait of Kim Jong-un
- Justice Minister Trump, Trump could start a 'bloodshed' administration

Dogs sit down on the roadside where the owner dies more than 80 days in China

American boys were investigated because of the fascist picture

advertisement

Category: Hepatitis B



Victory of millions of people



89% of people with

```

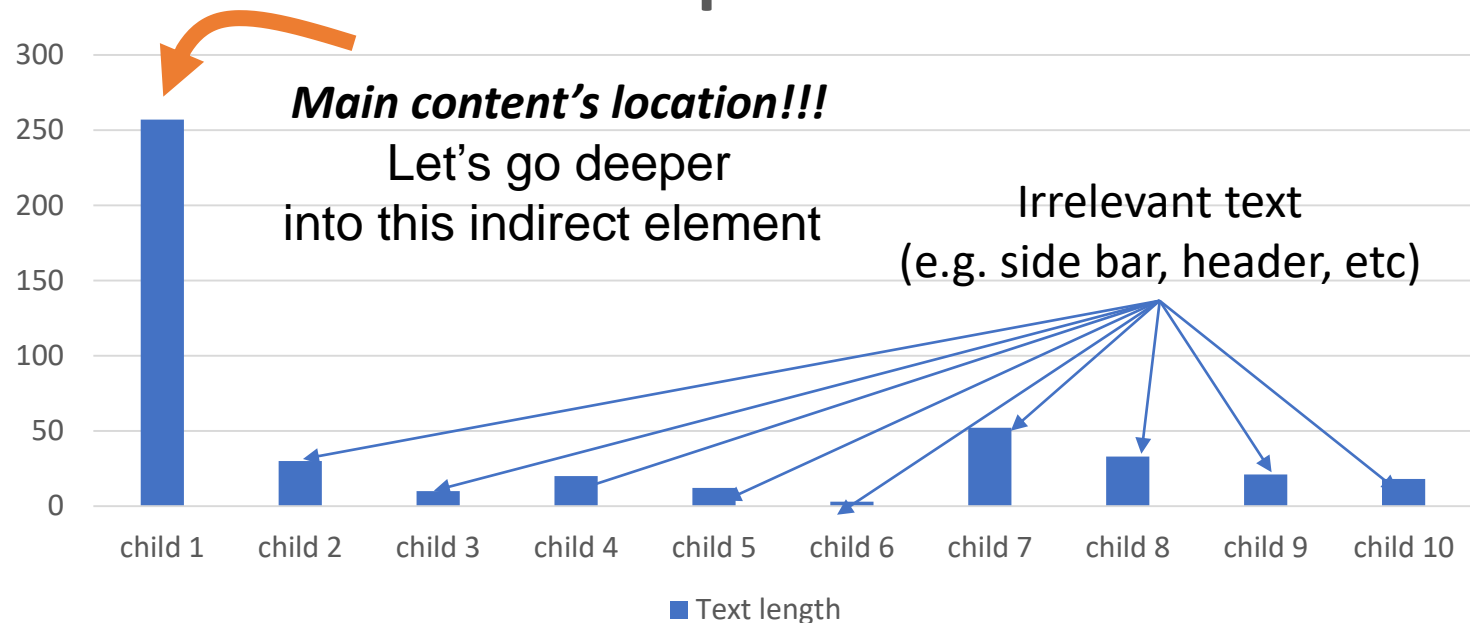
Elements Console Sources Network Performance Memory History Security
<!--</main_sidebar_12-->
v id="myvne_taskbar"></div>
- end taskbar -->
ader class="p_header"></header>
ader id="header" class="section_m_header"></header>
-end header-->
-main_menu menu PC-->
- Start Menu -->
v id="main_menu" class="p_menu"></nav>
- End Menu -->
- Start SUB MENU -->
v class="wrap_sub_menu"></div>
ction class="cat_header clearfix"></section>
-End main_menu-->
- Breadcrumb -->
-End Breadcrumb-->
- CONTENT -->
ction class="container" data-component-modulejs="detail" data-component-page-type="text"
component-page-config="{\"article_id\":\"3836552\"}"
!->wrap_sidebar_12-->
section class="wrap_sidebar_12">
<section class="sidebar_1">
▶<header class="clearfix"></header>
▶<h1 class="title_news_detail mb10"></h1>
▶<h2 class="description"></h2>
▶<p class="related_news"></p>
▼<article class="content_detail fck_detail width_common block_ads_connect"> == $0
  ▼<p class="Normal">
    ▶<em></em>
    ▼<font style="vertical-align: inherit;">
      ▶<font style="vertical-align: inherit;"></font>
      <font style="vertical-align: inherit;">This is one of the highlights of the midterm
        elections here.</font>
    </font>
  </p>
  ▼<table class="tplCaption" cellpadding="3" border="0" align="center"
    style="width: 350px;">
    ▶<tbody></tbody>
  </table>
  ▼<p class="Normal">
    ▼<font style="vertical-align: inherit;">
      ▼<font style="vertical-align: inherit;">
        "
        Like many Vietnamese who came to the United States a few decades ago, Joe's family
        life was not easy. "
      </font>
      ▶<font style="vertical-align: inherit;"></font>
      ▶<font style="vertical-align: inherit;"></font>
    </font>
  </p>
  ▶<p class="Normal"></p>
  ▶<p class="Normal"></p>
  ▶<p class="Normal"></p>
  ▼<table class="tblCaption" cellpadding="3" border="0" align="center"

```

## Indirect elements

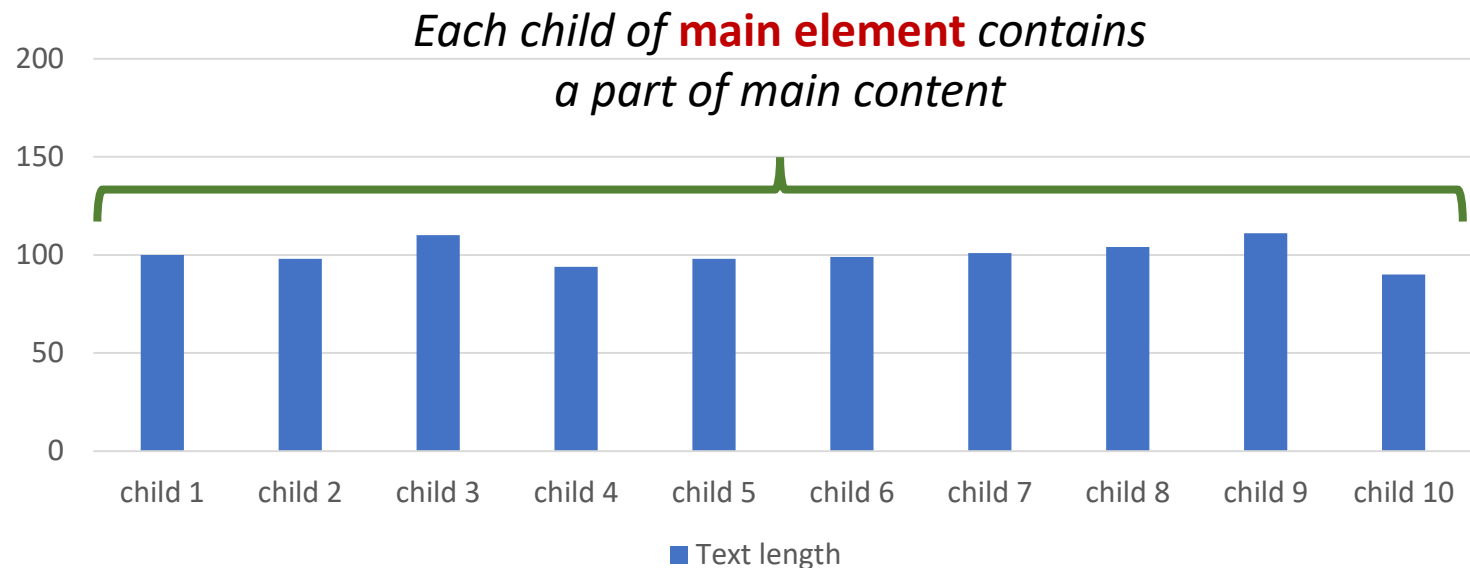
*Main  
element*

## Child elements text length comparison



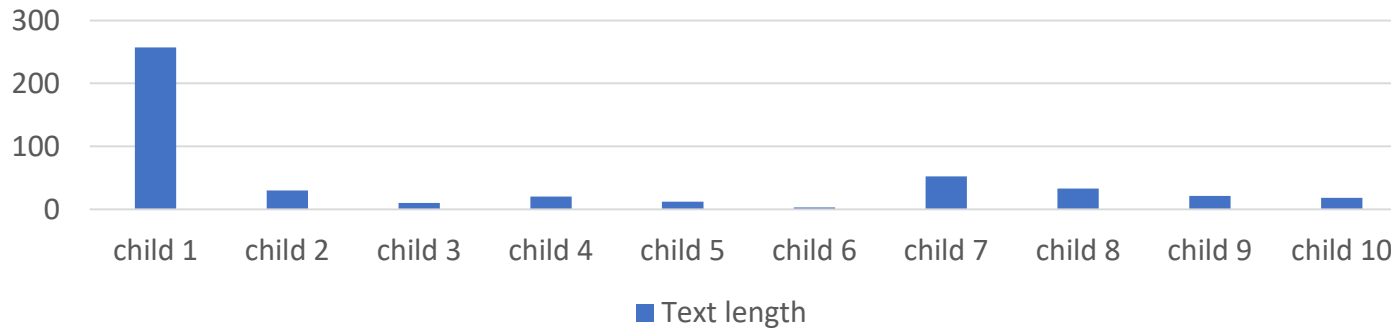
## Observation: **Main element**

### Child elements text length comparison



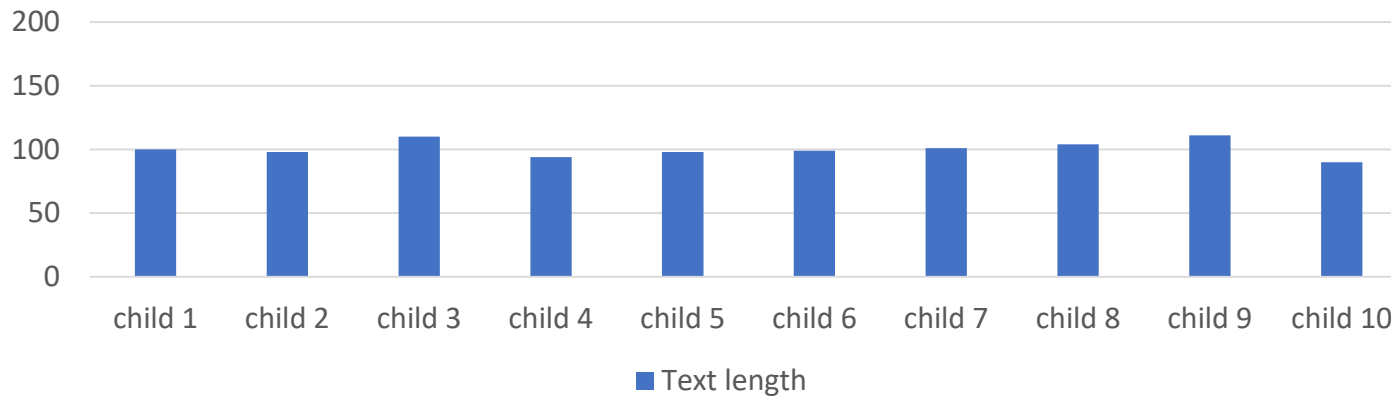
2

## Speaking of standard deviation



Indirect element

High standard deviation



Main element

Low standard deviation



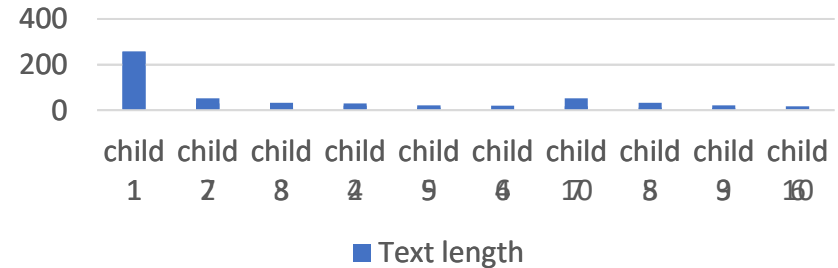
## Article algorithm

Traverse the HTML document recursively, for each element:

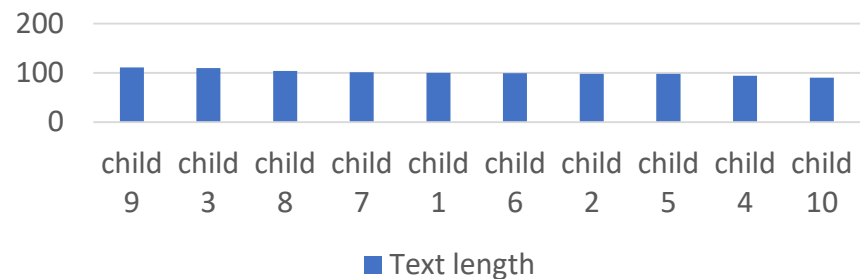
- **Step 1:** Get all child's text lengths and sort
- **Step 2:** Compute standard deviation ( $S$ )
- **Step 3:** Compute the subtraction between 2 highest children  $\Delta$
- **Step 4:** Compare:

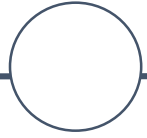
If  $\Delta \geq S$ : indirect element  $\rightarrow$  continue and repeat with its highest child

If  $\Delta < S$ : main element  $\rightarrow$  extract the content of this element



$$s = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N - 1}}$$





# Content

1. Introduction
- 2. Methods**
  - a. Genre classification
  - b. Content extraction**
    - i. Article
    - ii. List-viewed**
3. Evaluation result

## OFFICE SOLUTIONS



THIẾT BỊ VĂN PHÒNG

OFFICE EQUIPMENT

From ₫ 1,000



BÚT - VIẾT CÁC LOẠI

PEN - WRITING TYPES

From ₫ 1,000



DỤNG CỤ HỌC SINH

SCHOOL SUPPLIES

From ₫ 1,000



SỔ TAY - GIẤY GHI CHÚ

HANDWARE - NOTE P...

From ₫ 1,000



QUÀ TẶNG TỔNG HỢP

GIFTS

From ₫ 1,000

## All Of The Files

Book Online  
 Stationery  
 Flashcards  
 Musical  
 Remembrance  
 Book of Foreign Languages  
 More

## SEARCH FILTER

### Place Of Sale

- ☐ Hanoi  
☐ TP. Ho Chi Minh  
☐ Thai Nguyen  
☐ Vinh Phuc  
 More

### Transporter


- ☐ Delivery of Savings  
☐ Fast delivery  
☐ Viettel Post  
☐ Quick VNPost  
 More

### Shop Type

- ☐ Shopee Mall

Sorted by **Popular** Latest Selling Price

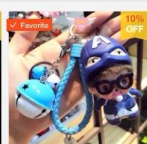
1 / 100



HOW TO PLAY GAME PUBG

₫ 80,000 ₫ 29,000


1696 ★★★★★ (first)



Funny Babies Lock Button ✓ Baby Captain ✓ Uniqu...

₫ 60,000 ₫ 45,000


1634 ★★★★★ (3)



FREESHIP 99K NATIONALITY

₫ 6,000 ₫ 2,500


633 ★★★★★ (3804)



Sports Cup

₫ 100,000

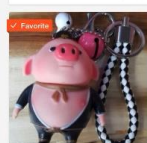
2171 ★★★★★ (4)



Keychain - 5 samples

₫ 3,000 - ₫ 12,500


1671 ★★★★★ (3)



Pork chop slices

₫ 45,000


6316 ★★★★★ (first)



PUBG metal very beautiful PIGG

₫ 45,000


1311 No reviews yet



Beautiful sand clock - Beautiful birthday gifts

₫ 350,000


2779 ★★★★★ (6)



Hook BT21

₫ 60,000 ₫ 30,000

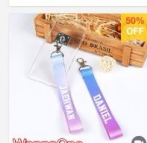
582 ★★★★★ (432)



Photobook Love Yourself Answer BTS (BTS picture)

₫ 79,000


1201 ★★★★★ (650)



Wanna One Padlock F124

₫ 10,000

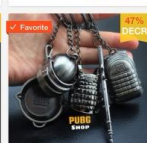
1000 ★★★★★ (1000)



Bubg simulation game

₫ 10,000

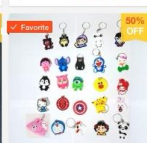
1000 ★★★★★ (1000)



THE GAME OF PUBG

₫ 10,000

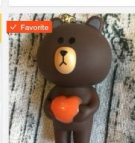
1000 ★★★★★ (1000)



Cute cute silicon hanging

₫ 10,000

1000 ★★★★★ (1000)



Brown bear plastic

₫ 10,000

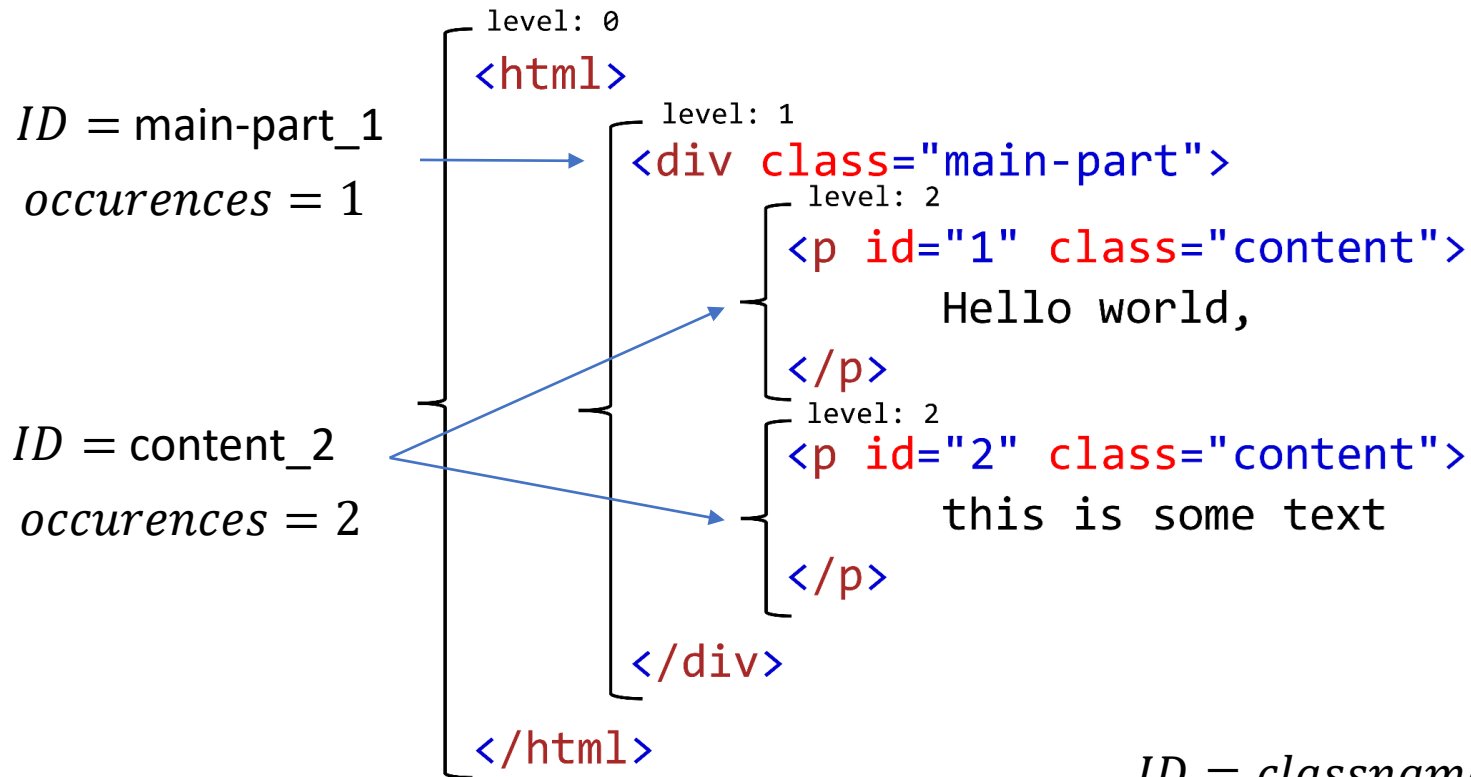
1000 ★★★★★ (1000)

## 27-Jun-19

[illegible]

28

## Element's ID and occurrence



## Observation: list-viewed page

- Same classname and same level (ID)
- Has many occurrences
- Their text length might be higher than other irrelevant part's, but not too long

→ Use the equation below:

$$R = \frac{2 \times \text{occurrences} \times \text{text\_length}}{\text{occurrences} + \text{text\_length}}$$

...to balance both features

[illegible]

## List-viewed algorithm

Traverse the HTML document recursively, for each ID:

- **Step 1:** Get its occurrences and cumulative text length
- **Step 2:** Compute its ranking score  $R$

If done:

- **Step 3:** Rank the IDs based on their  $R$  scores and get  $n$  highest IDs
- **Step 4:** In  $n$  highest IDs, pick the ID which has the **highest average text length**  
→ Extract all tags of the chosen ID

$$R = \frac{2 \times \text{occurrences} \times \text{text\_length}}{\text{occurrences} + \text{text\_length}}$$

$$ATL = \frac{\text{text\_length}}{\text{occurrences}}$$



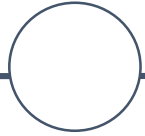
---

2

## Content extraction pipeline

Web page





# Content

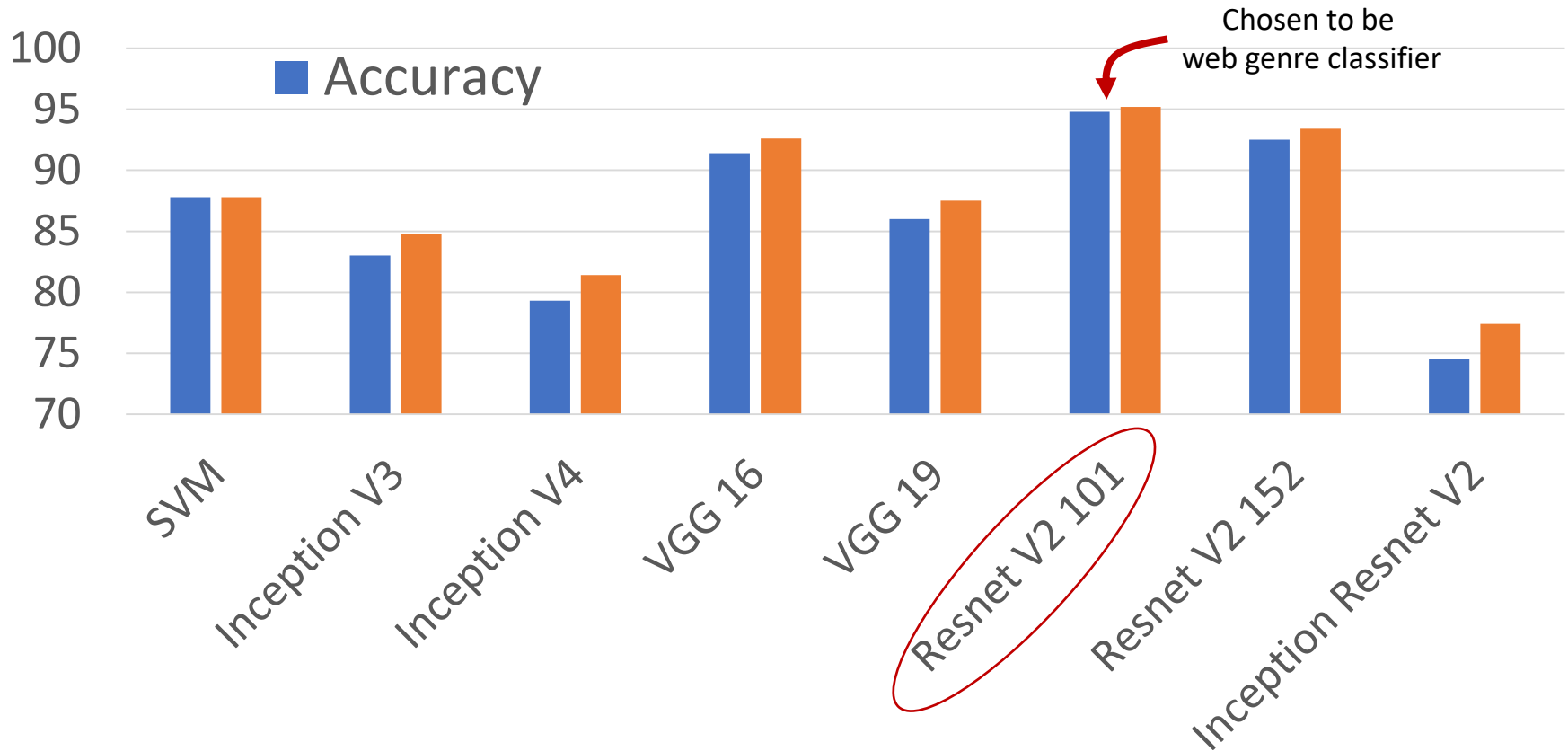
1. Introduction
2. Methods
  - a. Genre classification
  - b. Content extraction
- 3. Evaluation result**

## Evaluation: Web genre classification

- Manually annotated images rendered from random URLs
- The dataset contains 2,372 article pages and 3,350 list-view pages
- 90% for train, 10% for test
- Compared different CNN-based models
- Baseline: SVM
- SVM was trained on features derived from HTML document of the web, including:
  - text length
  - number of identical elements, clusters, images, digits
  - number of various marks (. , ; : ? !)

3

## Evaluation: Web genre classification



## Evaluation: Content extraction

- Manually extract content from randomly 100 articles page and 100 list-viewed pages to build a **gold corpus** (each document is a **gold sequence**)
- Compute precision, recall and F1 score based on Longest Common Subsequence between **gold sequence** and **extracted sequence**

$$p = \frac{\text{length}(\text{gold} \cap \text{extracted sequence})}{\text{length}(\text{extracted sequence})}$$

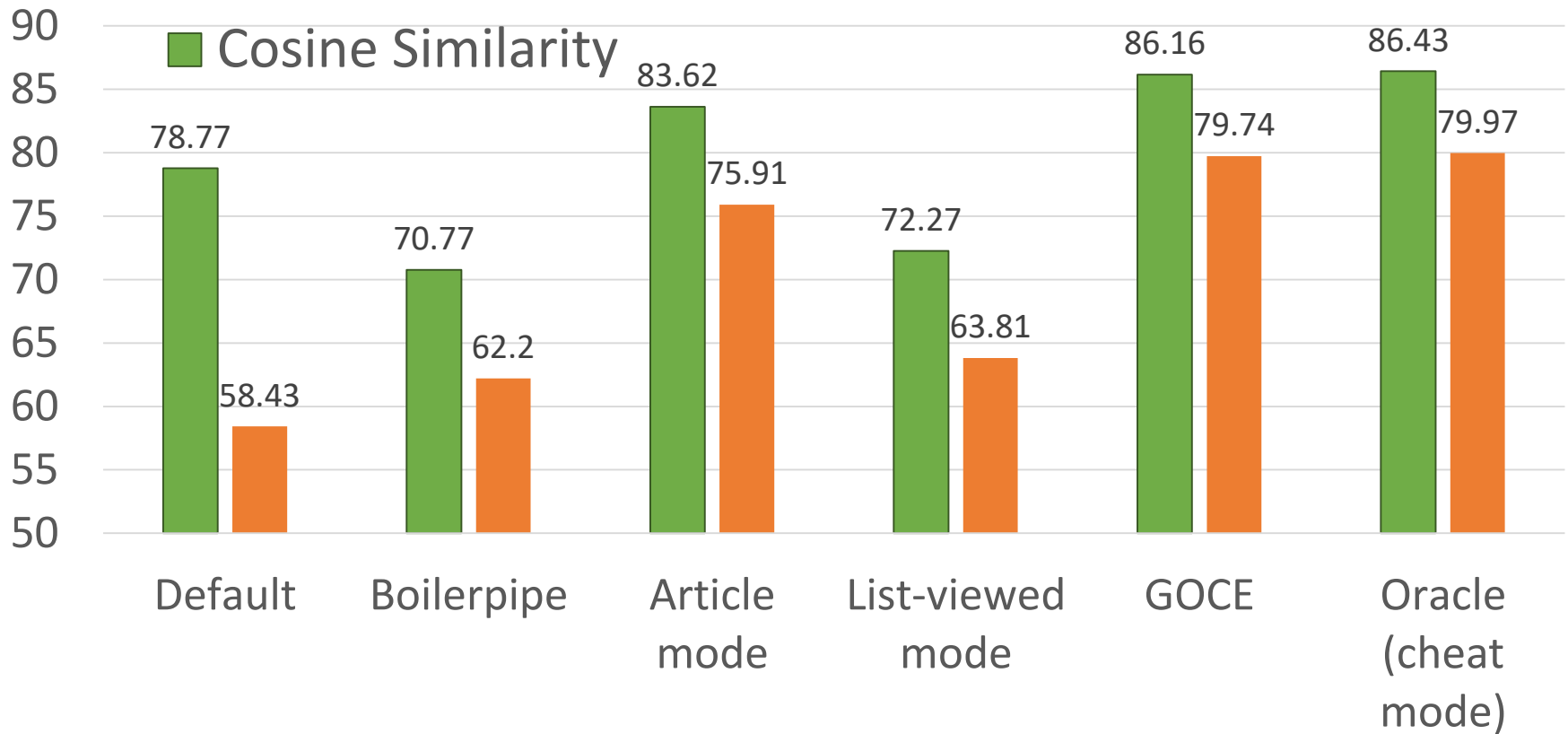
$$r = \frac{\text{length}(\text{gold} \cap \text{extracted sequence})}{\text{length}(\text{gold sequence})}$$

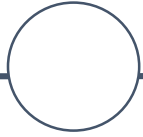
- Compute cosine similarity between 2 bag-of-words vectors of **gold sequence** and **extracted sequence**

$$\cos \theta = \frac{w_i \cdot w_j}{||w_i|| ||w_j||}$$

3

## Evaluation: Content extraction – Full dataset





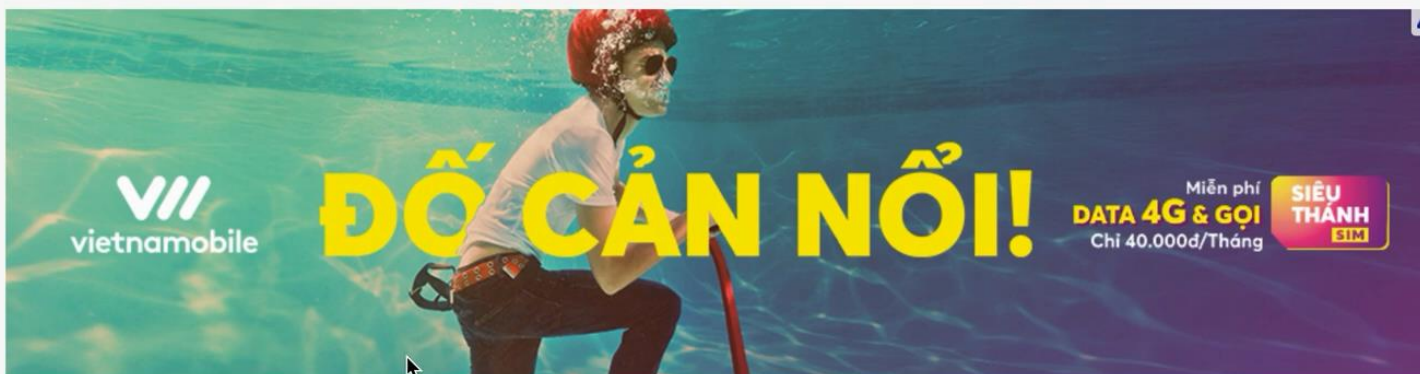
## Summary

### **Main contributions:**

- Proposed content extraction pipeline with CNN
- Improved previous approach to extract article web pages
- Proposed novel method to extract list-viewed web pages
- Built 2 datasets to evaluate the pipeline

### **Future work:**

- Extend web genre
- Improve the pipeline to classify and extract jointly



## Thủ tướng: Đưa tinh thần Park Hang-seo vào công nghiệp hỗ trợ Việt

# Genre-Oriented Web Content Extraction with Deep Convolutional Neural Networks and Statistical Methods

**Bao-Dai Nguyen-Hoang**  
Knorex Vietnam Co., Ltd.  
46 Bach Dang, Tan Binh,  
Ho Chi Minh city, Vietnam  
dai\_nguyen@knorex.com

**Bao-Tran Pham-Hong**  
Knorex Vietnam Co., Ltd  
46 Bach Dang, Tan Binh  
Ho Chi Minh city, Vietnam  
tran\_pham@knorex.com

**Yiping Jin**  
Knorex Pte. Ltd.  
2 Science Park Drive,  
Singapore 118222  
jinyiping@knorex.com

**Phu T. V. Le**  
Knorex Pte. Ltd.  
2 Science Park Drive,  
Singapore 118222  
le.phu@knorex.com

## Abstract

Extracting clean textual content from the Web is the first and an essential step to resolve most of down-stream natural language processing tasks. Previous works in web content extraction focus mainly on web pages with a single main block of textual content, such as news articles and blog posts. They employ techniques that rely largely on the HTML structure of the page to extract the main content.

Little attention has been paid to recognizing different genres of web pages, which can have a tremendous impact on the accu-

## 1 Introduction

Web content extraction is the task of pulling the main textual content out of web pages while removing noise such as clutters (for e.g. HTML tags, scripts, etc.) and boilerplate contents (for e.g. navigation bar, headers, footers, etc.) (Gupta et al., 2003). Humans can easily identify the main content based on the layout and visuals of web pages. However, it is not trivial for a computer program to accurately detect the meaningful content and skip the noise due to the complex and dynamic content layout of web pages.