

Content Extraction from Webpages Using Machine Learning

Master's Thesis

Hamza Yunis

Bauhaus Universität

26.01.2017

Supervised by:
Prof. Benno Stein
Dr. Andreas Jakoby

Advised by:
Johannes Kiesel

Motivation


SPORT
[Ski World Cup: Kristoffersen blows away rest of field](#)


FIFA gives go-ahead for expanded World Cup

By James Masters and Chris Murphy, CNN

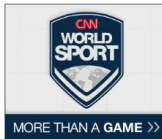
Updated 1617 GMT (0017 HKT) January 10, 2017



Source: CNN

Top stories

[La Liga: Barca title dream over?](#)

[North Korea sends message to Trump amid missile threat](#)


Advertisement

[World Cup: How would FIFA's 48-team plan work? 01:43](#)
Story highlights
[FIFA Council approves expansion plan](#)
[New format set to start in 2026](#)
[Tournament to have 16 more teams](#)

(CNN) — FIFA, soccer's world governing body, has approved a grand plan to revolutionize the World Cup by increasing the number of teams from 32 to 48.

The FIFA Council agreed unanimously to the move Tuesday, with the new format starting in 2026.

[READ: FIFA's 48-team expansion plan explained](#)

 By continuing to browse our site you agree to our use of [cookies](#), [revised Privacy Policy](#) and [Terms of Service](#). More information

 e published later.
e format has stayed



SPORT

Ski World Cup: Kristoffersen blows away rest of field



FIFA gives go-ahead for expanded World Cup

By James Masters and Chris Murphy, CNN

🕒 Updated 1617 GMT (0017 HKT) January 10, 2017



Source: CNN

World Cup: How would FIFA's 48-team plan work? 01:43

Story highlights

FIFA Council approves expansion plan

New format set to start in 2026

Tournament to have 16 more teams

(CNN) — FIFA, soccer's world governing body, has approved a grand plan to revolutionize the World Cup by increasing the number of teams from 32 to 48.

The FIFA Council agreed unanimously to the move Tuesday, with the new format starting in 2026.

[READ: FIFA's 48-team expansion plan explained](#)

Top stories



La Liga: Barca title dream over?



North Korea sends message to Trump amid missile threat



MORE THAN A GAME >>





SPORT

Ski World Cup: Kristoffersen blows away rest of field



FIFA gives go-ahead for expanded World Cup

By James Masters and Chris Murphy, CNN

Updated 1617 GMT (0017 HKT) January 10, 2017



Source: CNN

Top stories



La Liga: Barca title dream over?



North Korea sends message to Trump amid missile threat



Advertisement

World Cup: How would FIFA's 48-team plan work? 01:43

Story highlights

FIFA Council approves expansion plan

New format set to start in 2026

Tournament to have 16 more teams

(CNN) — FIFA, soccer's world governing body, has approved a grand plan to revolutionize the World Cup by increasing the number of teams from 32 to 48.

The FIFA Council agreed unanimously to the move Tuesday, with the new format starting in 2026.

[READ: FIFA's 48-team expansion plan explained](#)

By continuing to browse our site you agree to our use of [cookies](#), [revised Privacy Policy](#) and [Terms of Service](#). More information



e published later.
e format has stayed



SPORT

Ski World Cup: Kristoffersen blows away rest of field



FIFA gives go-ahead for expanded World Cup

By James Masters and Chris Murphy, CNN

Updated 1617 GMT (0017 HKT) January 10, 2017



Source: CNN

World Cup: How would FIFA's 48-team plan work? 01:43

Story highlights

FIFA Council approves expansion plan

New format set to start in 2026

Tournament to have 16 more teams

(CNN) — FIFA, soccer's world governing body, has approved a grand plan to revolutionize the World Cup by increasing the number of teams from 32 to 48.

The FIFA Council agreed unanimously to the move Tuesday, with the new format starting in 2026.

[READ: FIFA's 48-team expansion plan explained](#)

Top stories

La Liga: Barca title dream over?

North Korea sends message to Trump amid missile threat

MORE THAN A GAME >>

Advertisement




SPORT
[Ski World Cup: Kristoffersen blows away rest of field](#)


FIFA gives go-ahead for expanded World Cup

By James Masters and Chris Murphy, CNN

Updated 1617 GMT (0017 HKT) January 10, 2017



Source: CNN

Top stories


[La Liga: Barca title dream over?](#)

[North Korea sends message to Trump amid missile threat](#)


Advertisement

[World Cup: How would FIFA's 48-team plan work? 01:43](#)

Story highlights

[FIFA Council approves expansion plan](#)
[New format set to start in 2026](#)
[Tournament to have 16 more teams](#)

(CNN) — FIFA, soccer's world governing body, has approved a grand plan to revolutionize the World Cup by increasing the number of teams from 32 to 48.

The FIFA Council agreed unanimously to the move Tuesday, with the new format starting in 2026.

[READ: FIFA's 48-team expansion plan explained](#)

 By continuing to browse our site you agree to our use of [cookies](#), [revised Privacy Policy](#) and [Terms of Service](#). More information

 published later.
format has stayed



SPORT

Ski World Cup: Kristoffersen blows away rest of field



FIFA gives go-ahead for expanded World Cup

By James Masters and Chris Murphy, CNN

🕒 Updated 1617 GMT (0017 HKT) January 10, 2017



Source: CNN

World Cup: How would FIFA's 48-team plan work? 01:43

Story highlights

FIFA Council approves expansion plan

New format set to start in 2026

Tournament to have 16 more teams

(CNN) — FIFA, soccer's world governing body, has approved a grand plan to revolutionize the World Cup by increasing the number of teams from 32 to 48.

The FIFA Council agreed unanimously to the move Tuesday, with the new format starting in 2026.

[READ: FIFA's 48-team expansion plan explained](#)

Top stories



La Liga: Barca title dream over?



North Korea sends message to Trump amid missile threat



Advertisement





SPORT

Ski World Cup: Kristoffersen blows away rest of field



FIFA gives go-ahead for expanded World Cup

By James Masters and Chris Murphy, CNN

🕒 Updated 1617 GMT (0017 HKT) January 10, 2017



Source: CNN

World Cup: How would FIFA's 48-team plan work? 01:43

Story highlights

FIFA Council approves expansion plan

New format set to start in 2026

Tournament to have 16 more teams

(CNN) — FIFA, soccer's world governing body, has approved a grand plan to revolutionize the World Cup by increasing the number of teams from 32 to 48.

The FIFA Council agreed unanimously to the move Tuesday, with the new format starting in 2026.

[READ: FIFA's 48-team expansion plan explained](#)

Top stories



La Liga: Barca title dream over?



North Korea sends message to Trump amid missile threat



Advertisement



What is the Main Content?

What is the Main Content?

Definition (i): The main content is what the webpage is *supposed to communicate* according to the publisher.

What is the Main Content?

Definition (i): The main content is what the webpage is *supposed to communicate* according to the publisher.

- We cannot always tell what the webpage publisher wants to communicate.

What is the Main Content?

Definition (i): The main content is what the webpage is *supposed to communicate* according to the publisher.

- We cannot always tell what the webpage publisher wants to communicate.
- A single webpage may have different publishers, each wanting to communicate a different type of information.

What is the Main Content?

Definition (i): The main content is what the webpage is *supposed to communicate* according to the publisher.

- We cannot always tell what the webpage publisher wants to communicate.
- A single webpage may have different publishers, each wanting to communicate a different type of information.

Definition (ii): The main content is what makes the webpage *interesting in to the user*.

What is the Main Content?

Definition (i): The main content is what the webpage is *supposed to communicate* according to the publisher.

- We cannot always tell what the webpage publisher wants to communicate.
- A single webpage may have different publishers, each wanting to communicate a different type of information.

Definition (ii): The main content is what makes the webpage *interesting in to the user*.

- Different users may have different interests in the webpage.

What is the Main Content?

Definition (i): The main content is what the webpage is *supposed to communicate* according to the publisher.

- We cannot always tell what the webpage publisher wants to communicate.
- A single webpage may have different publishers, each wanting to communicate a different type of information.

Definition (ii): The main content is what makes the webpage *interesting in to the user*.

- Different users may have different interests in the webpage.

Definition (iii): The main content of a webpage consists of information that *cannot be found in other webpages*.

What is the Main Content?

Definition (i): The main content is what the webpage is *supposed to communicate* according to the publisher.

- We cannot always tell what the webpage publisher wants to communicate.
- A single webpage may have different publishers, each wanting to communicate a different type of information.








Definition (ii): The main content is what makes the webpage *interesting in to the user*.

- Different users may have different interests in the webpage.

Definition (iii): The main content of a webpage consists of information that *cannot be found in other webpages*.

- Usually used in template recognition.

What is the Main Content?


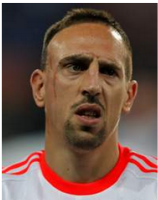
Edition: InternationalRegister / LoginFollow us on

HOMENEWSGOAL 50COMPETITIONSCLUBSLIVETRANSFERSFEATURESBET

FC Bayern München » **Franck Ribéry**

FRANCK RIBÉRY

Club PageNational Team PagePlayer ProfilePlayer NewsTransfer Zone




FC Bayern München

Franck Ribéry



★★★★★3.24

Date of Birth:	Apr 7, 1983 (Age 33)
Place of Birth:	Boulogne-sur-Mer
Nationality:	France
Height:	170 cm.
Weight:	72 Kg.
Position:	Midfielder
Squad Number:	7








GOAL LIVE SCORES APP



SQUAD

Manuel Neuer
Goalkeeper

What is the Main Content?


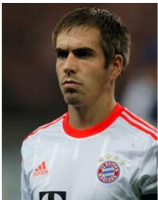
Edition: InternationalRegister / LoginFollow us on

HOME NEWS GOAL 50 COMPETITIONS CLUBS LIVE TRANSFERS FEATURES BET

FC Bayern München » **Philipp Lahm**



PHILIPP LAHM

Club PageNational Team PagePlayer ProfilePlayer NewsTransfer Zone



FC Bayern München

Philipp Lahm★★★★☆ 3.09








Date of Birth:	Nov 11, 1983 (Age 33)
Place of Birth:	München
Nationality:	Germany
Height:	170 cm.
Weight:	62 Kg.
Position:	Defender
Squad Number:	21



SQUAD

Manuel Neuer
Goalkeeper

What is the Main Content?



Edition: InternationalRegister / LoginFollow us on

HOMENEWSGOAL 50COMPETITIONSCLUBSLIVETRANSFERSFEATURESBET

FC Bayern München » Philipp Lahm


PHILIPP LAHM

Club PageNational Team PagePlayer ProfilePlayer NewsTransfer Zone

FC Bayern München



Philipp Lahm★★★★☆ 3.09

Date of Birth:	Nov 11, 1983 (Age 33)
Place of Birth:	München
Nationality:	Germany
Height:	170 cm.
Weight:	62 Kg.
Position:	Defender
Squad Number:	21



Download on the App Store

SQUAD

	Manuel Neuer Goalkeeper	
---	----------------------------	---

What is the Main Content?

The main content is the non-noisy content!

What is the Noisy Content?

What is the Noisy Content?

- Advertisements.

What is the Noisy Content?

- Advertisements.
- Navigation links.

What is the Noisy Content?

- Advertisements.
- Navigation links.
- Links to promoted webpages.

What is the Noisy Content?

- Advertisements.
- Navigation links.
- Links to promoted webpages.
- Legal information.

What is the Noisy Content?

- Advertisements.
- Navigation links.
- Links to promoted webpages.
- Legal information.
- Irrelevant information.

What is the Noisy Content?

- Advertisements.
- Navigation links.
- Links to promoted webpages.
- Legal information.
- Irrelevant information.
- Input elements.

Types of HTML Elements

Types of HTML Elements

- Content elements.
- Inline semantic elements.
- Sectioning elements.

```
<ul>
  <li>List item 1.</li>
  <li>List item 2.</li>
</ul>
<div>
  <p>This is the <span class="important">first</span>
  paragraph.</p>
  <p>This is the <span class="important">second</span>
  paragraph.</p>
</div>
```

Types of HTML Elements

- Content elements.
- Inline semantic elements.
- Sectioning elements.

```
<ul>
  <li>List item 1.</li>
  <li>List item 2.</li>
</ul>
<div>
  <p>This is the <span class="important">first</span>
  paragraph.</p>
  <p>This is the <span class="important">second</span>
  paragraph.</p>
</div>
```

Types of HTML Elements

- Content elements.
- Inline semantic elements.
- Sectioning elements.

```
<ul>
  <li>List item 1.</li>
  <li>List item 2.</li>
</ul>
<div>
  <p>This is the <span class="important">first</span>
  paragraph.</p>
  <p>This is the <span class="important">second</span>
  paragraph.</p>
</div>
```


Types of HTML Elements

- Content elements.
- Inline semantic elements.
- Sectioning elements.

```
<ul>
  <li>List item 1.</li>
  <li>List item 2.</li>
</ul>
<div>
  <p>This is the <span class="important">first</span>
  paragraph.</p>
  <p>This is the <span class="important">second</span>
  paragraph.</p>
</div>
```

Types of HTML Elements

Elements to Be Classified

Elements to Be Classified

- Paragraph elements: `<p>`.

Elements to Be Classified

- Paragraph elements: `<p>`.
- `<div>` elements.

Elements to Be Classified

- Paragraph elements: `<p>`.
- `<div>` elements.
 - If they do not have content element descendants.

Elements to Be Classified

- Paragraph elements: `<p>`.
- `<div>` elements.
 - If they do not have content element descendants.
- Table cell elements: `<th>` and `<td>`.

Elements to Be Classified

- Paragraph elements: `<p>`.
- `<div>` elements.
 - If they do not have content element descendants.
- Table cell elements: `<th>` and `<td>`.
 - If they do not have content element descendants.

Elements to Be Classified

- Paragraph elements: `<p>`.
- `<div>` elements.
 - If they do not have content element descendants.
- Table cell elements: `<th>` and `<td>`.
 - If they do not have content element descendants.
- List item elements: ``.

Elements to Be Classified

- Paragraph elements: `<p>`.
- `<div>` elements.
 - If they do not have content element descendants.
- Table cell elements: `<th>` and `<td>`.
 - If they do not have content element descendants.
- List item elements: ``.
- Header elements: `<h1>`, `<h2>`, `<h3>`, `<h4>`, `<h5>`, and `<h6>`.

Elements to Be Classified

- Paragraph elements: `<p>`.
- `<div>` elements.
 - If they do not have content element descendants.
- Table cell elements: `<th>` and `<td>`.
 - If they do not have content element descendants.
- List item elements: ``.
- Header elements: `<h1>`, `<h2>`, `<h3>`, `<h4>`, `<h5>`, and `<h6>`.
- Image elements: ``.

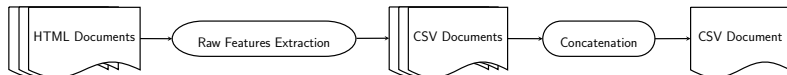
Learning Workflow



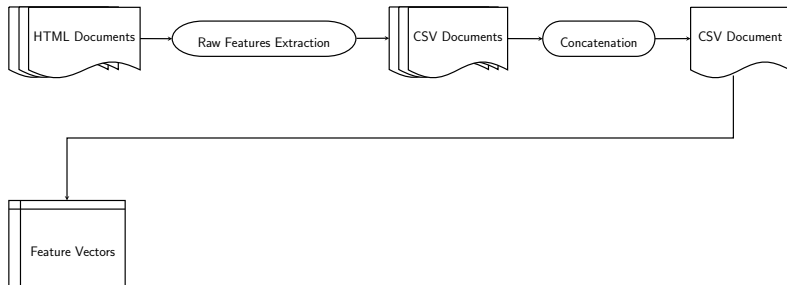
Learning Workflow



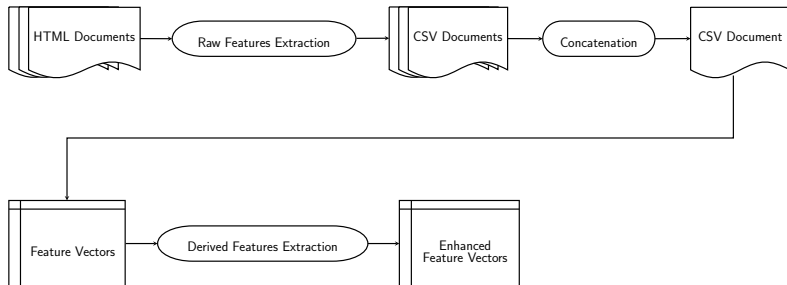
Learning Workflow



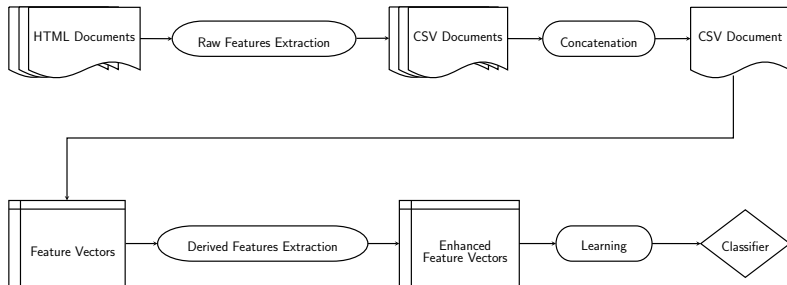
Learning Workflow



Learning Workflow



Learning Workflow



Feature Engineering

Feature Engineering

```
<div id="comments-section">
  <div class="comment">
    <div class="comment-header">
      <ul>
        <li>Author name.</li>
        <li>Comment title.</li>
      </ul>
    </div>
    <div class="comment-content">
      <p>The body of the comment</p>
    </div>
  </div>
  ...
</div>
```

Feature Engineering

```
<div id="comments-section">
  <div class="comment">
    <div class="comment-header">
      <ul>
        <li>Author name.</li>
        <li>Comment title.</li>
      </ul>
    </div>
    <div class="comment-content">
      <p>The body of the comment</p>
    </div>
  </div>
  ...
</div>
```

Feature Engineering

```
<div id="comments-section">
  <div class="comment">
    <div class="comment-header">
      <ul>
        <li>Author name.</li>
        <li>Comment title.</li>
      </ul>
    </div>
  </div>
  ...

```

Raw features:

- ancestor_names="div, div, div, ul"
- ancestor_classes="NO_CLASSES, comment, comment-header, NO_CLASSES"
- inner_text="Author name."

Feature Engineering

```
<div id="comments-section">
  <div class="comment">
    <div class="comment-header">
      <ul>
        <li>Author name.</li>
        <li>Comment title.</li>
      </ul>
    </div>
  </div>
  ...
```

Raw features:

- ancestor_names="div, div, div, ul"
- ancestor_classes="NO_CLASSES, comment, comment-header, NO_CLASSES"
- inner_text="Author name."

Derived features:

- is_desc_comment="1"
- is_desc_cookies="0"
- is_desc_section="1"
- inner_text_length=2

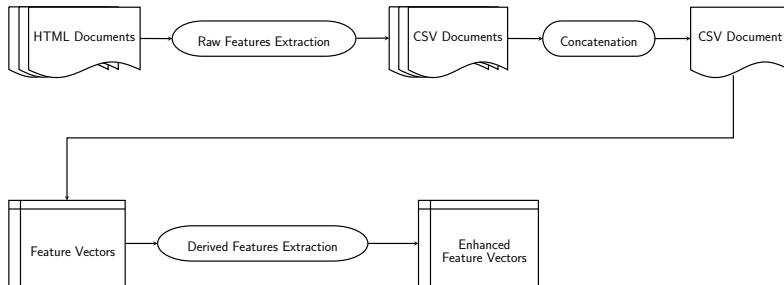
Evaluation

Evaluation Workflow



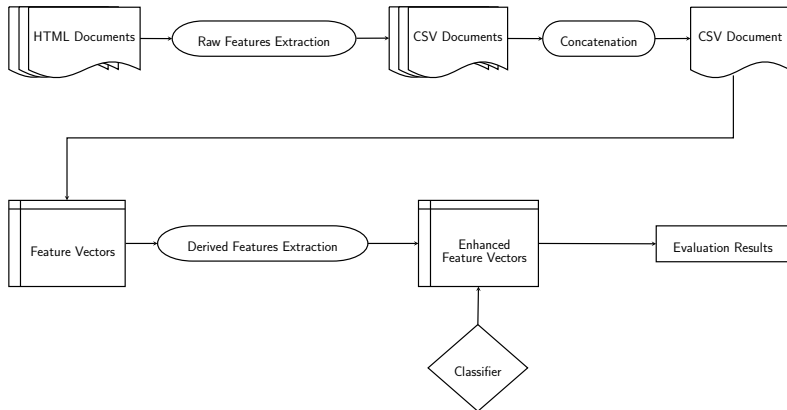
Evaluation

Evaluation Workflow



Evaluation

Evaluation Workflow



Evaluation

Confusion matrix:

Actual Class \ Predicted Class	"Noisy"	"Main"
	"Noisy"	"Main"
"Noisy"	<i>tn</i>	<i>fp</i>
"Main"	<i>fn</i>	<i>tp</i>

Evaluation metrics:

$$\text{precision} = \frac{tp}{tp + fp}$$

$$\text{recall} = \frac{tp}{tp + fn}$$

$$\mathbf{F}_{\beta} = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

Element-based results for textual content elements:

Actual Class \ Predicted Class	"Noisy"	"Main"
"Noisy"	4625	211
"Main"	277	1018

precision = 0.828

recall = 0.786

F_1 = 0.806

Text-based results for textual content elements:

Actual Class \ Predicted Class	"Noisy"	"Main"
	"Noisy"	"Main"
"Noisy"	496921	19618
"Main"	28654	163908

precision = 0.893

recall = 0.851

F_1 = 0.871

Element-based results for small and medium-size images ($\leq 40000\text{px}$):

Actual Class \ Predicted Class	"Noisy"	"Main"
	"Noisy"	"Main"
"Noisy"	900	5
"Main"	97	25

precision = 0.833

recall = 0.205

F_1 = 0.328

Element-based results for large images ($> 40000\text{px}$):

Actual Class \ Predicted Class	"Noisy"	"Main"
	"Noisy"	"Main"
"Noisy"	122	6
"Main"	10	29

precision = 0.828

recall = 0.743

F_1 = 0.783

Thank you for your attention!