Extracting Structured Data from Web Pages

CSCI599: Content Detection and Analysis for Big Data

Soravis Taekasem

Authors

Arvind Arasu

Hector Garcia-Molina

Overview

Unstructured Text

- Information is difficult to query

Structured Data

- Data is organize in a certain layout (schema)
- Generate dynamically from a structure source

Star Wars®: Secrets of the Galaxy Deluxe Box Set Hardcover – September 27, 2016

by Daniel Wallace (Author)

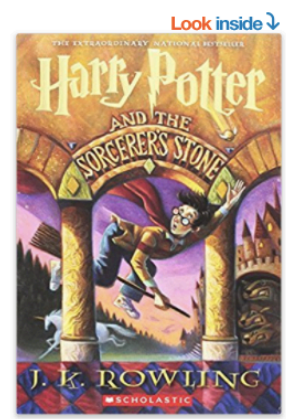
28 customer reviews

See all formats and editions

Hardcover
$62.09

21 Used from $50.17
31 New from $50.18
1 Collectible from $2,862.00

The secrets of the *Star Wars* galaxy have been recorded in a series of handbooks and guides created and kept hidden by the Jedi Order, the Sith, the Bounty Hunters Guild, and the Empire itself. Set in-world, richly illustrated, and annotated by characters such as Luke Skywalker, Leia Organa, Boba Fett, Yoda, and Darth Vader, each volume expands *Star Wars* mythology with details from the inside and deepens readers' experience of the saga. This deluxe edition boxed set collects *The Jedi Path*, *Book of Sith*, *The Bounty Hunter Code*, and *Imperial Handbook* in a handsome and accessible custom slipcase, creating a handy and invaluable library for exploring a galaxy far, far away.



Harry Potter and the Sorcerer's Stone Paperback – September 8, 1999

by J.K. Rowling (Author), Mary GrandPré (Illustrator)

59,828 customer reviews

Book 1 of 8 in the Harry Potter Series

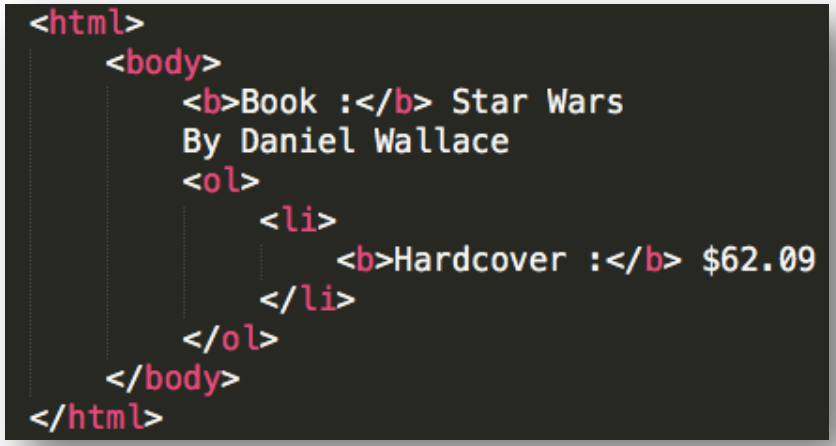
Amazon Charts #4 Most Read

See all 197 formats and editions

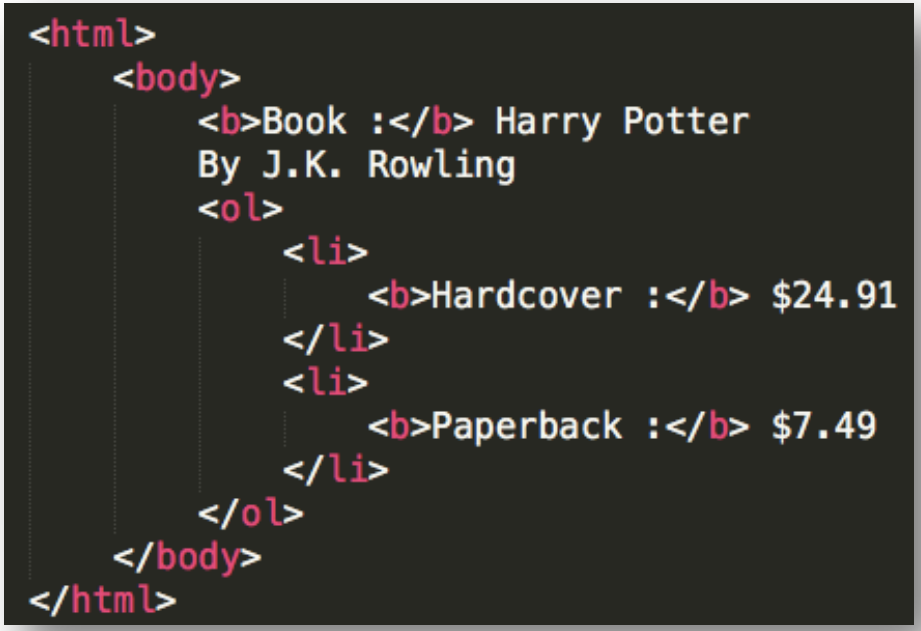
Kindle	Hardcover	Paperback	Audiobook	Mass Market Paperback
$0.00 kindleunlimited	$24.91	$7.49	$0.00	from $1.05
This title and over 1 million more available with Kindle Unlimited $8.99 to buy	15 Used from $20.92 10 New from $24.91 4 Collectible from $60.00	868 Used from $0.25 145 New from $4.00 15 Collectible from $6.89	Free with your Audible trial	132 Used from $1.05 16 New from $5.99 7 Collectible from $4.50

Harry Potter has no idea how famous he is. That's because he's being raised by his miserable aunt and uncle who are terrified Harry will learn that he's really a wizard, just as his parents were. But everything changes when Harry is summoned to attend an infamous school for wizards, and he begins to discover some clues about his illustrious birthright. From the surprising way he is greeted by a lovable giant, to the unique curriculum and colorful faculty at his unusual school, Harry finds himself drawn deep inside a mystical world he never knew existed and closer to his own noble destiny.

Pages Structure



```
<html>
    <body>
        <b>Book :</b> Star Wars
        By Daniel Wallace
        <ol>
            <li>
                <b>Hardcover :</b> $62.09
            </li>
        </ol>
    </body>
</html>
```



```
<html>
    <body>
        <b>Book :</b> Harry Potter
        By J.K. Rowling
        <ol>
            <li>
                <b>Hardcover :</b> $24.91
            </li>
            <li>
                <b>Paperback :</b> $7.49
            </li>
        </ol>
    </body>
</html>
```

Objective

Data Extraction

- Automatically extracting the structure data encoded in the web pages
- Without any learning examples or knowledge of the template
- To pose a complex queries over the data

Avoid Human Input

- Time consuming
- Error-prone
- Difficult to extract semi-structured pages
- Templates change very frequently

Challenges

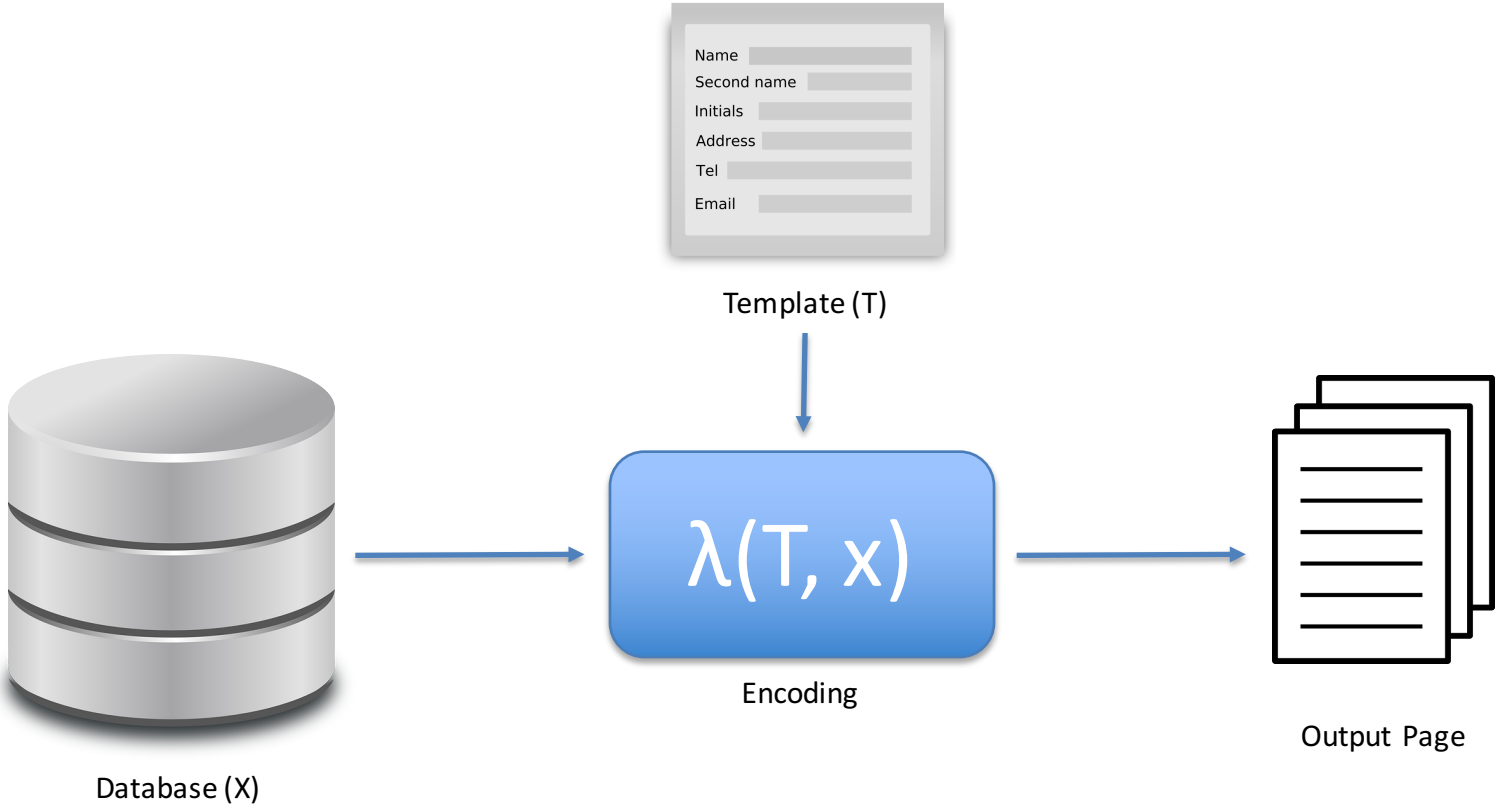
Template vs Data

- There is no obvious way of differentiating between text that is part of template and text that is part of data

Complex Schema

- The schema of data is usually not a simple set of attributes, but is more complex and semi-structured

Model of Page Creation



ExAlg

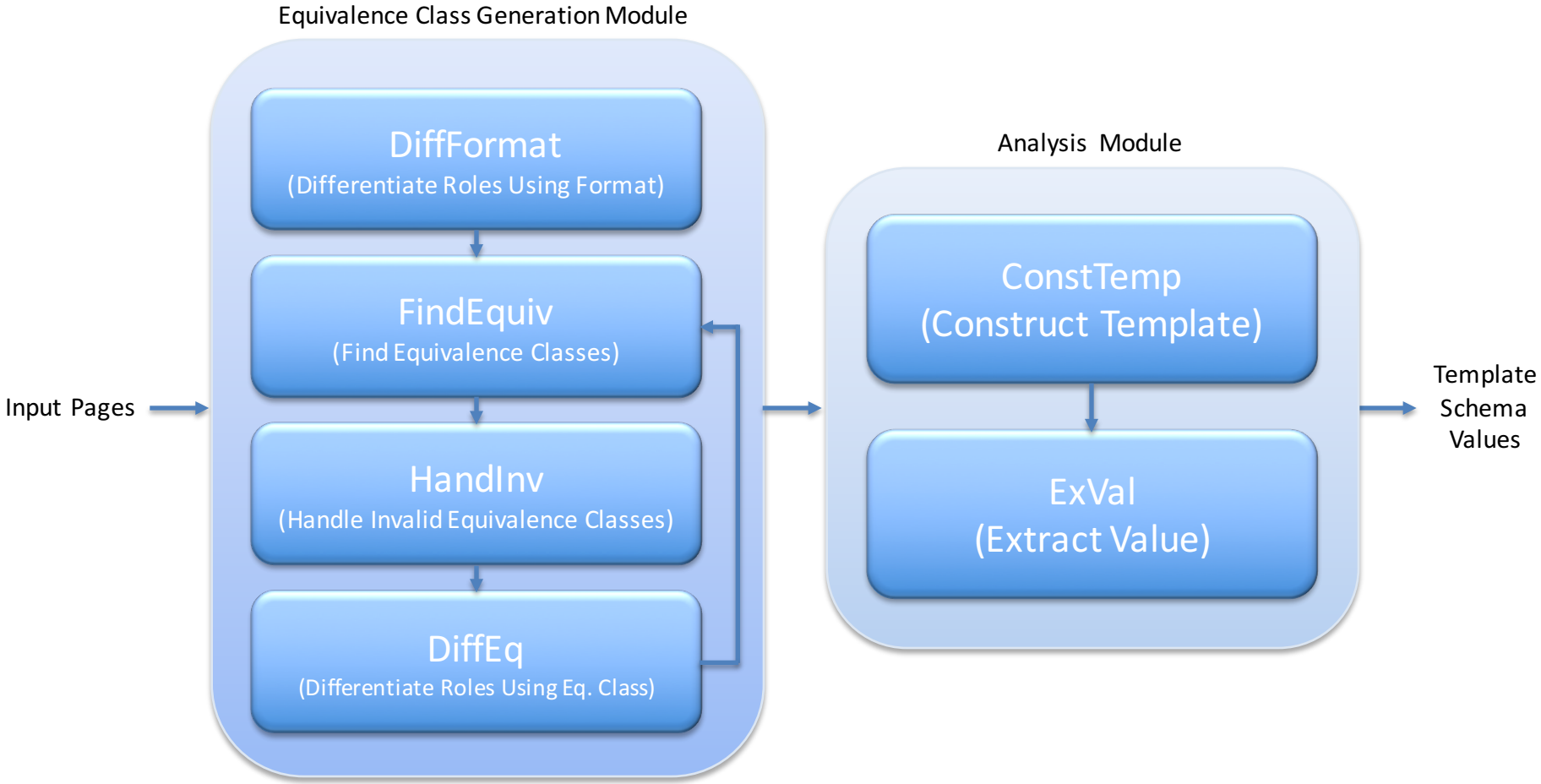
Equivalence Class Generation Stage

- Discovers sets of tokens associated with the same type constructor in the template used to create the input pages

Analysis Stage

- Use the sets of tokens to deduce the template
- Extract value from the page using template

Modules of ExAlg



Equivalence Classes

Definition

- Sets of tokens having the same frequency of occurrence in every input page (same occurrence-vector)

LFEQs

- Large and Frequently occurring EQuivalence classes
- Almost always, LFEQs are formed by tokens associated with the same type constructor in the template used to create the input pages

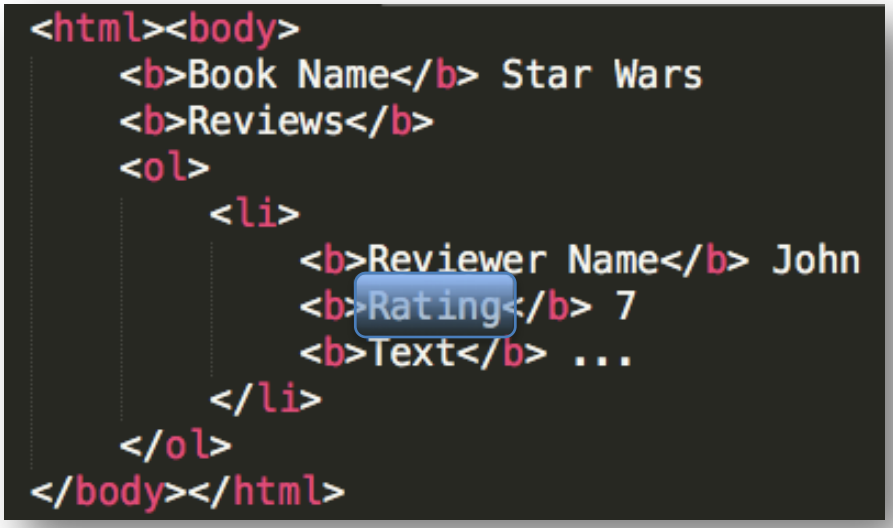
Occurrence-Vector

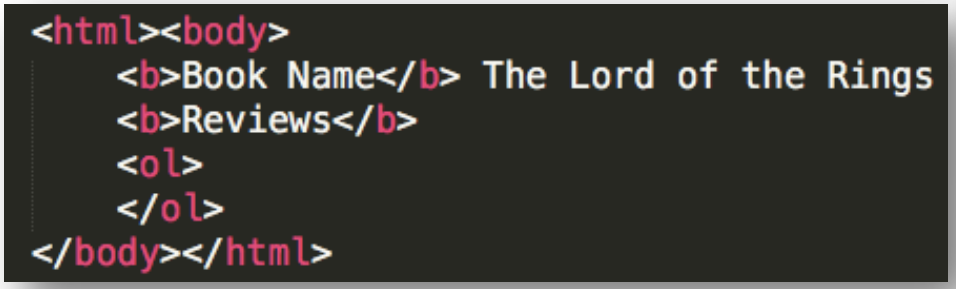
```
<html><body>
    <b>Book Name</b> Star Wars
    <b>Reviews</b>
    <ol>
        <li>
            <b>Reviewer Name</b> John
            <b>Rating</b> 7
            <b>Text</b> ...
        </li>
    </ol>
</body></html>
```

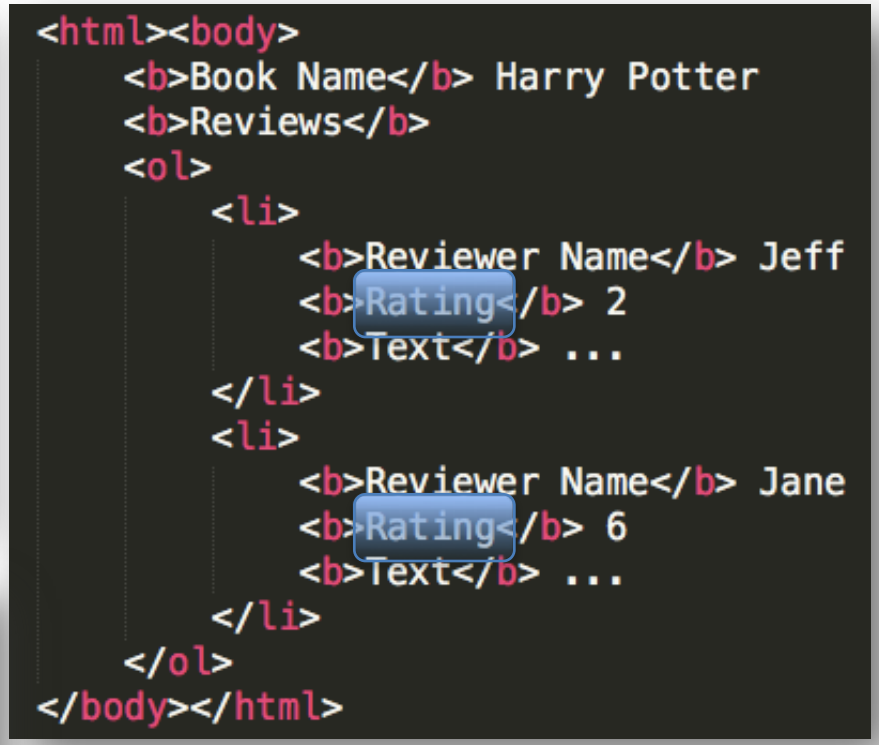
```
<html><body>
    <b>Book Name</b> Harry Potter
    <b>Reviews</b>
    <ol>
        <li>
            <b>Reviewer Name</b> Jeff
            <b>Rating</b> 2
            <b>Text</b> ...
        </li>
        <li>
            <b>Reviewer Name</b> Jane
            <b>Rating</b> 6
            <b>Text</b> ...
        </li>
    </ol>
</body></html>
```

```
<html><body>
    <b>Book Name</b> The Lord of the Rings
    <b>Reviews</b>
    <ol>
    </ol>
</body></html>
```

Occurrence-Vector

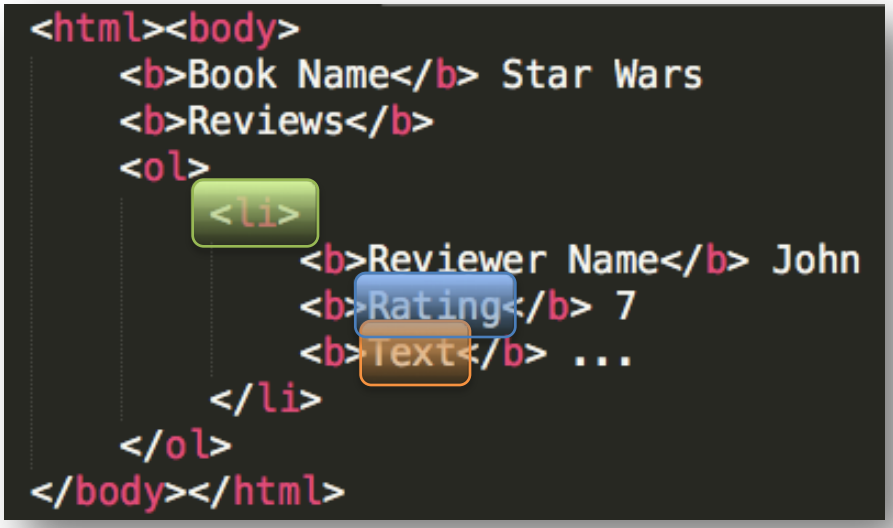


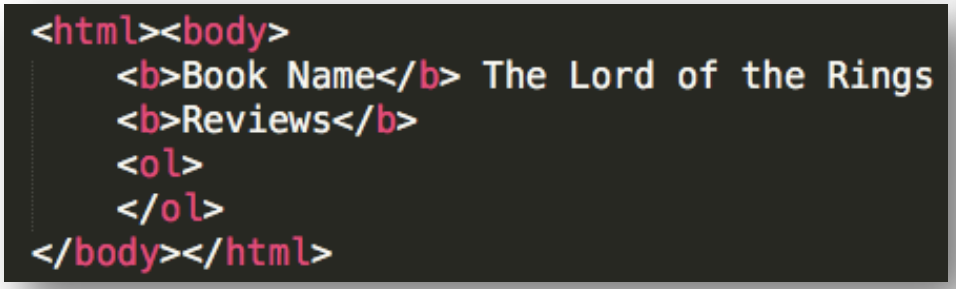


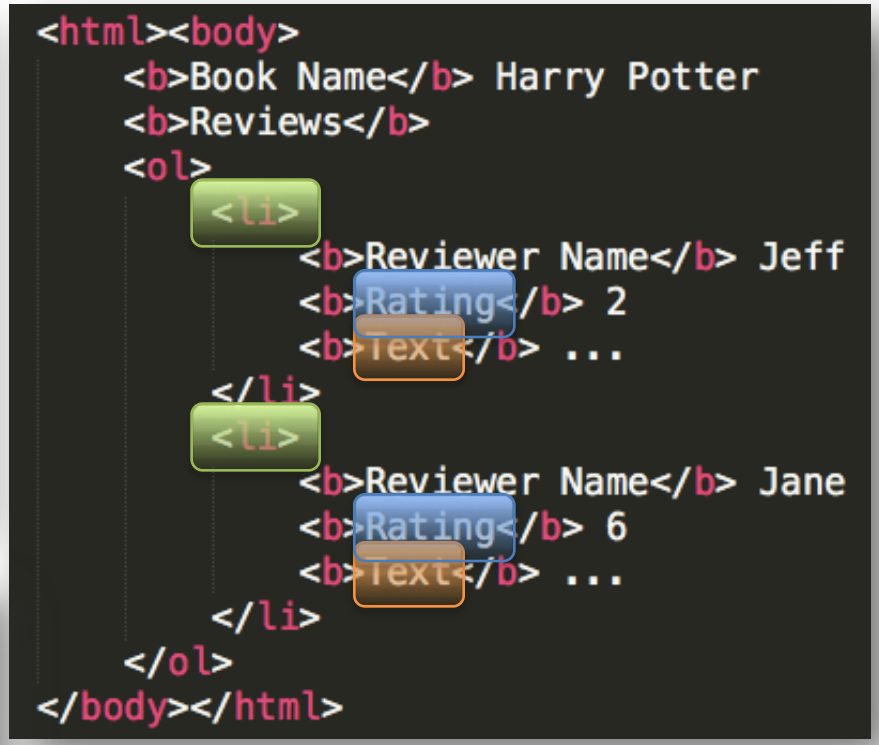


Occurrence-Vector = <1, 2, 0>

Equivalence Classes

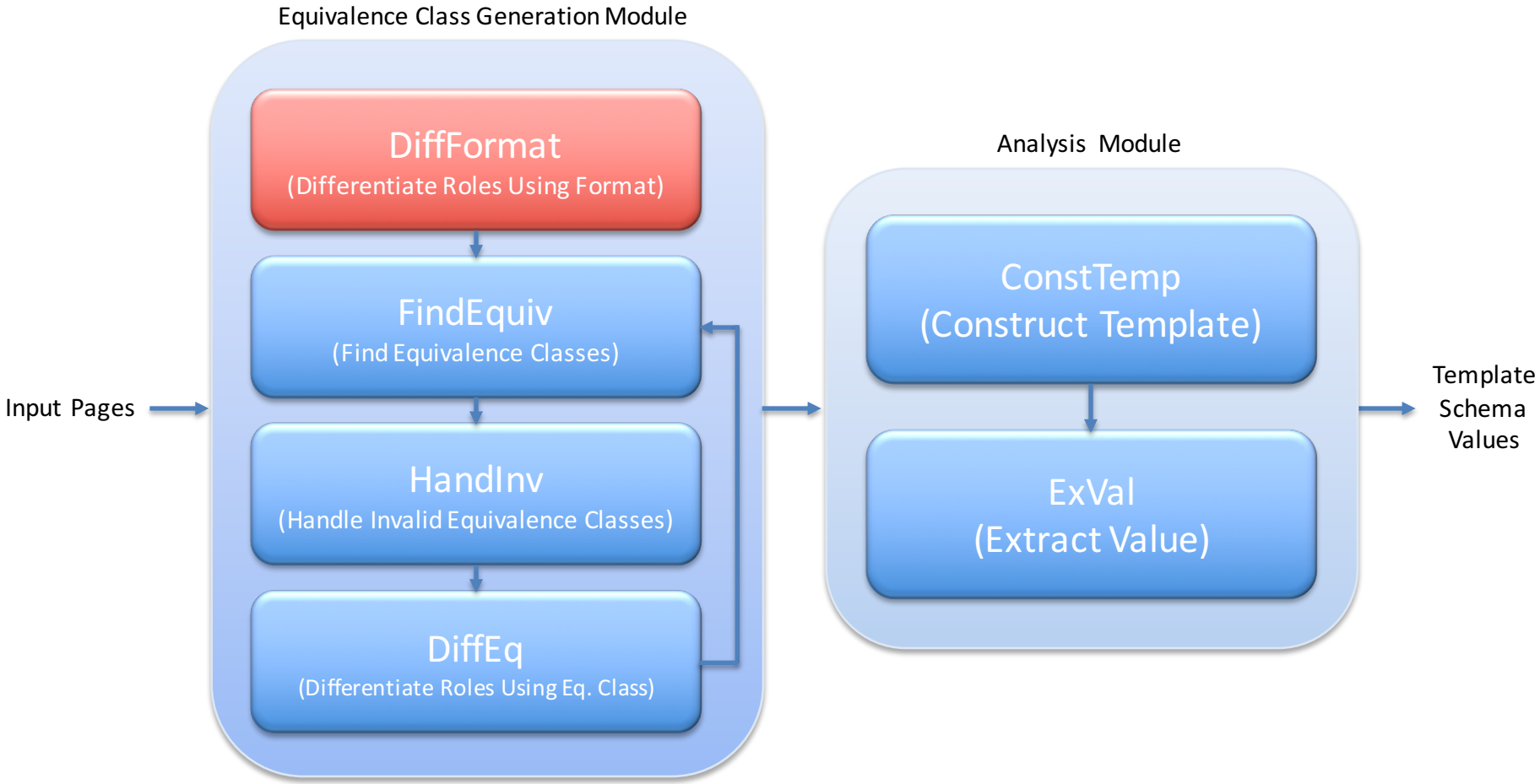






Equivalence Class = {<li>, Rating, Text, ...}

Modules of ExAlg



DiffFormat Sub-module

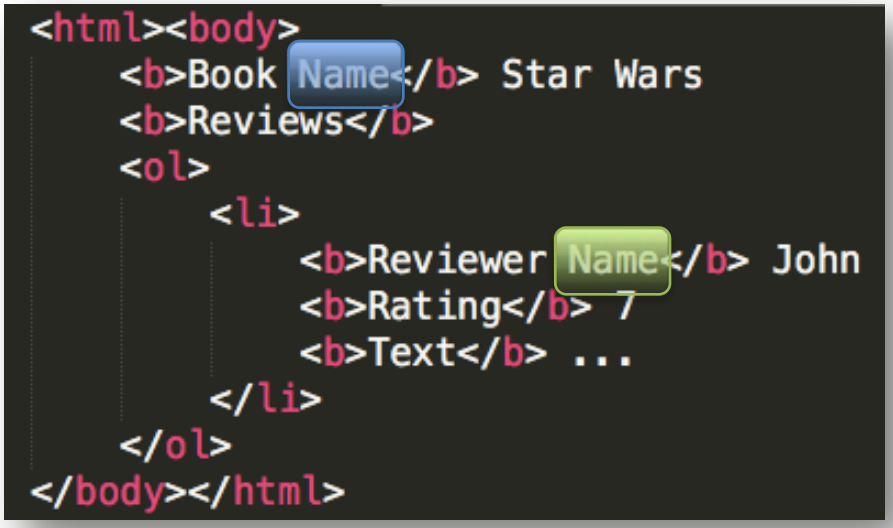
Differentiate roles using Format

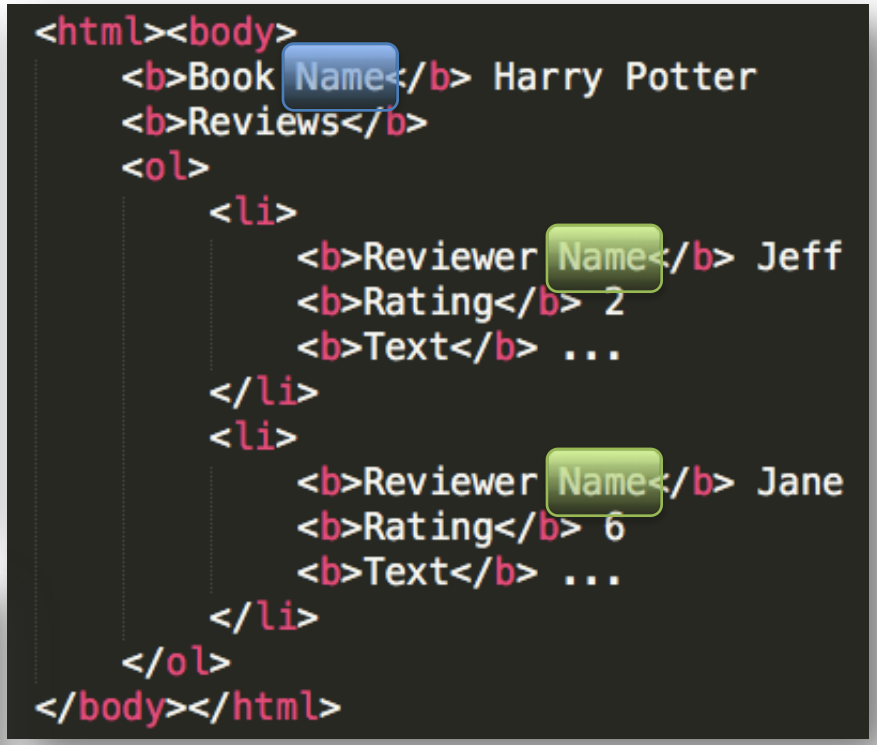
- Use the HTML formatting of input pages to differentiate token roles
- Two tokens with different occurrence path have different roles

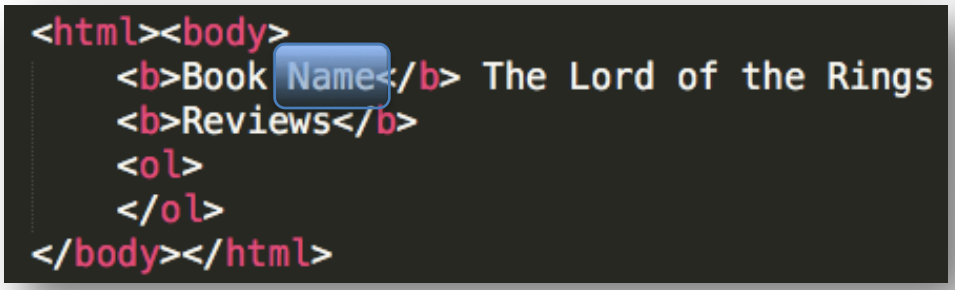
Input & Output

- **Input:** a set of input web pages
- **Output:** input web pages as strings of dtokens

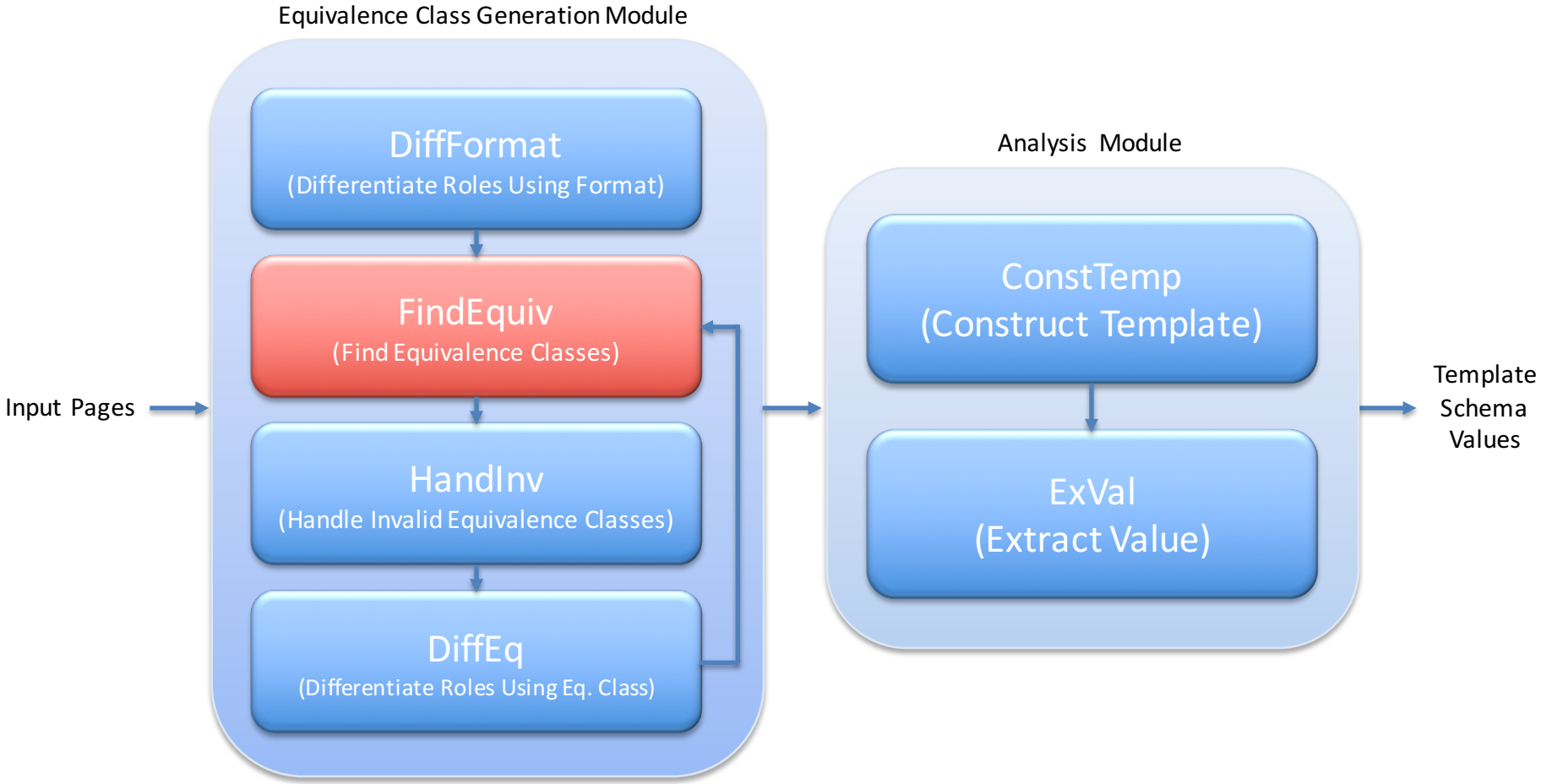
Token Roles







Modules of ExAlg



FindEquiv Sub-module

Find Equivalence classes

- Computes occurrence vectors and equivalence classes of the dtokens in the input pages
- Determines LFEQs from **SizeThres** and **SupThres** parameters

Input & Output

- **Input:** strings of dtokens
- **Output:** a set of LFEQs

LFEQs

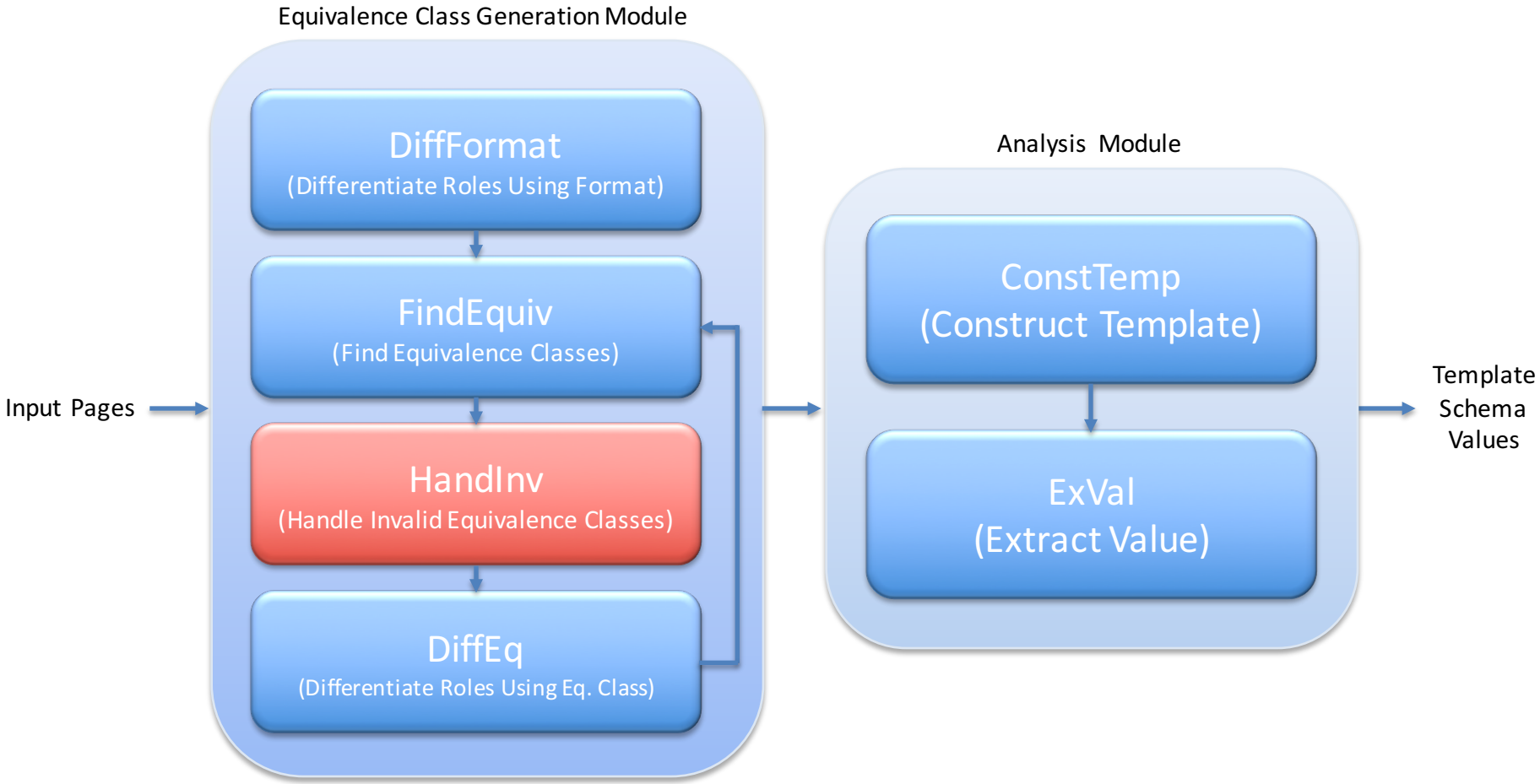
Large and Frequently occurring EQuivalence classes

- **Large:** number of tokens greater than **SizeThres**
- **Frequently:** number of occurring greater than **SupThres**

Intuition

- Almost always, LFEQs are formed by tokens associated with the same type constructor in the template used to create the input pages

Modules of ExAlg



HandInv Sub-module

Handel Invalid equivalence classes

- Detect and identify invalid LFEQs using violations of ordered and nesting properties

Input & Output

- **Input:** a set of LFEQs
- **Output:** an ordered set of nested LFEQs

Ordered & Nesting Properties

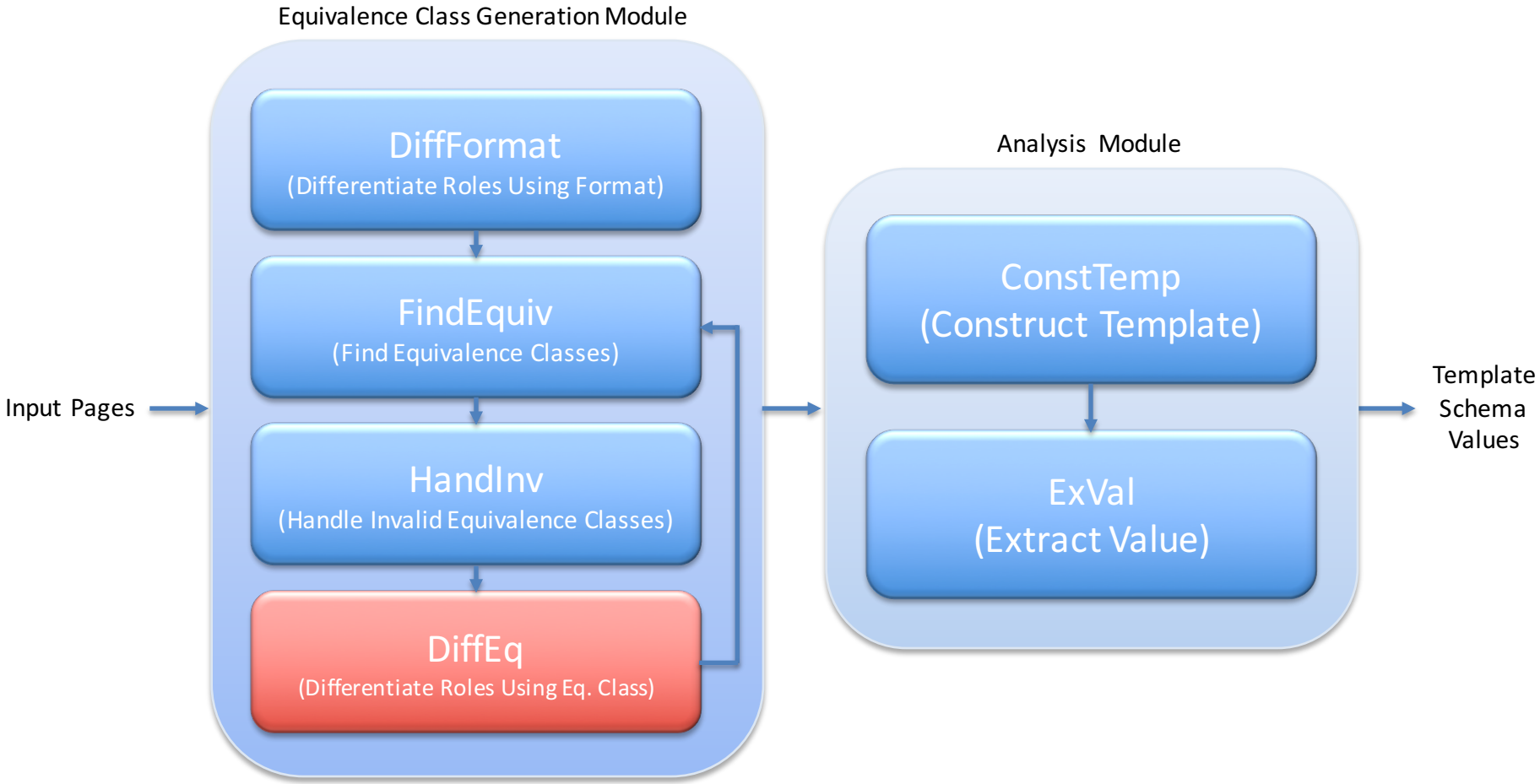
Ordered Properties

- An equivalence class is ordered
- The span of each occurrence of Ɛ is subdivided into a certain (m-1) position
- Token Pos(k) may only occur between t_k and t_{k+1}

Nesting Properties

- A pair of equivalence classes $Ɛ_1$ and $Ɛ_2$ is nested if the span of all occurrences of $Ɛ_2$ is within some position of occurrence of $Ɛ_1$

Modules of ExAlg



DiffEq Sub-module

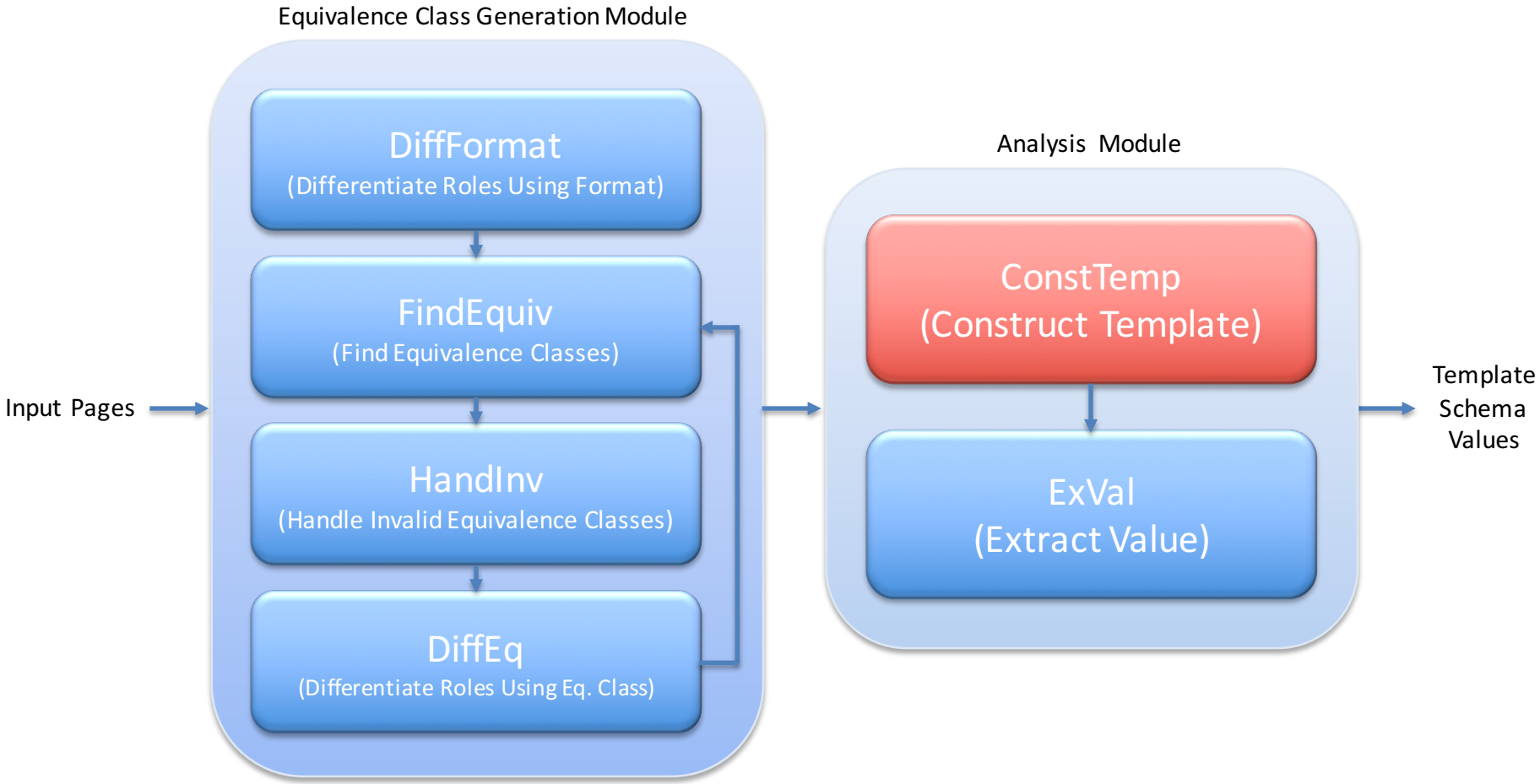
Differentiate roles using Equivalence class

- Use the valid equivalence classes to differentiate token roles
- The role of an occurrence of a token t outside the span of any occurrence of Ɛ, is different from the role of an occurrence within the span of some occurrence of Ɛ
- The role of an occurrence of t within Pos(l) of some occurrence of Ɛ, is different from the role of an occurrence of t within Pos(m) where l ≠ m

Input & Output

- **Input:** an ordered set of nested LFEQs
- **Output:** a strings of dtokens

Modules of ExAlg



ConstTemp Sub-module

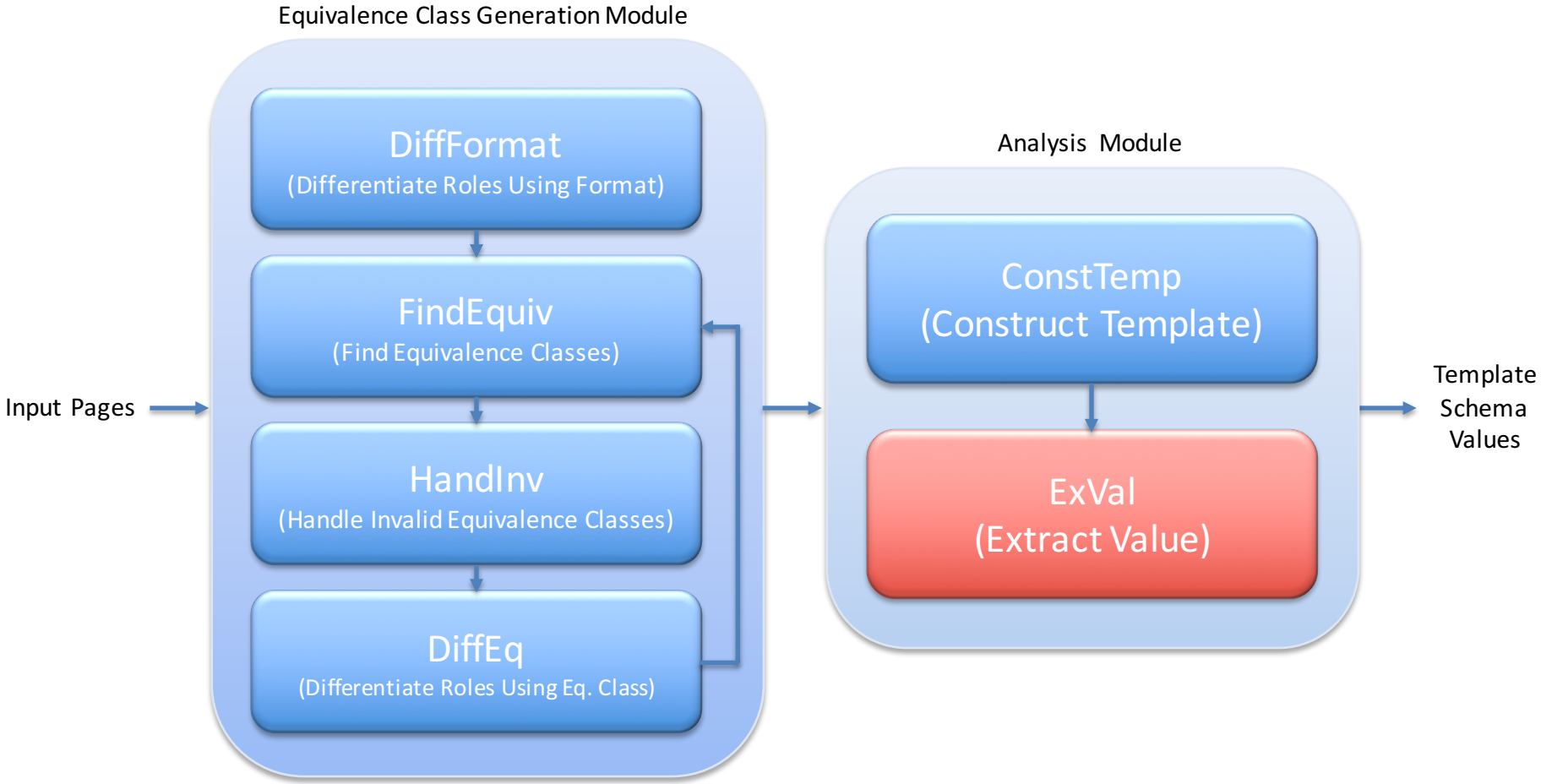
Construct Template

- Consider each set of string corresponding to every occurrence of equivalence class has some recognizable pattern
- Recursively construct the template

Input & Output

- **Input:** a set of LFEQs and strings of dtokens
- **Output:** a template

Modules of ExAlg



ExVal Sub-module

Extract Value

- Data extraction is trivial after we construct the template

Input & Output

- **Input:** a template and input web pages
- **Output:** a set of values

Experiment

Input Collection

- RISE
- RoadRunner
- IEPAD
- Various well-know sites (e.g. eBay, Netflix, DBLP and Google)

Evaluation

- Correct
- Partially correct
- Incorrect

Accuracy

Source	Accuracy	Normalized Accuracy
IEPAD	87.4 %	87.7 %
RISE	76.1 %	67.7 %
RoadRunner	79.8 %	92.5 %
Misc.	76.1 %	71.0 %
Total	**80.0 %**	**82.7 %**

Results

- ExAlg is very effective in extracting the data
- 40% of the input collections were correctly extracted all the attributes
- Other collections were partially correct
- On average, about 80% of the attributes were extracted correctly
- No attributes that were incorrectly extracted
- Partially correct classified attributes can be refined with very small human interaction

Conclusion

ExAlg

- Algorithm for extracting structured data from a collection of web pages generated from a common template
- Discovers the unknown template and then extract the data from input pages
- Use concept of equivalence classes and differentiating roles
- Extremely good in extracting the data from web pages

Pros & Cons

Pros

- Well organized
- Clear explanation
- Good examples

Cons

- Contain a lots of confusing notation and equation

Reference: http://ilpubs.stanford.edu:8090/548/1/2002-40.pdf