Visual Structure of Web Documents

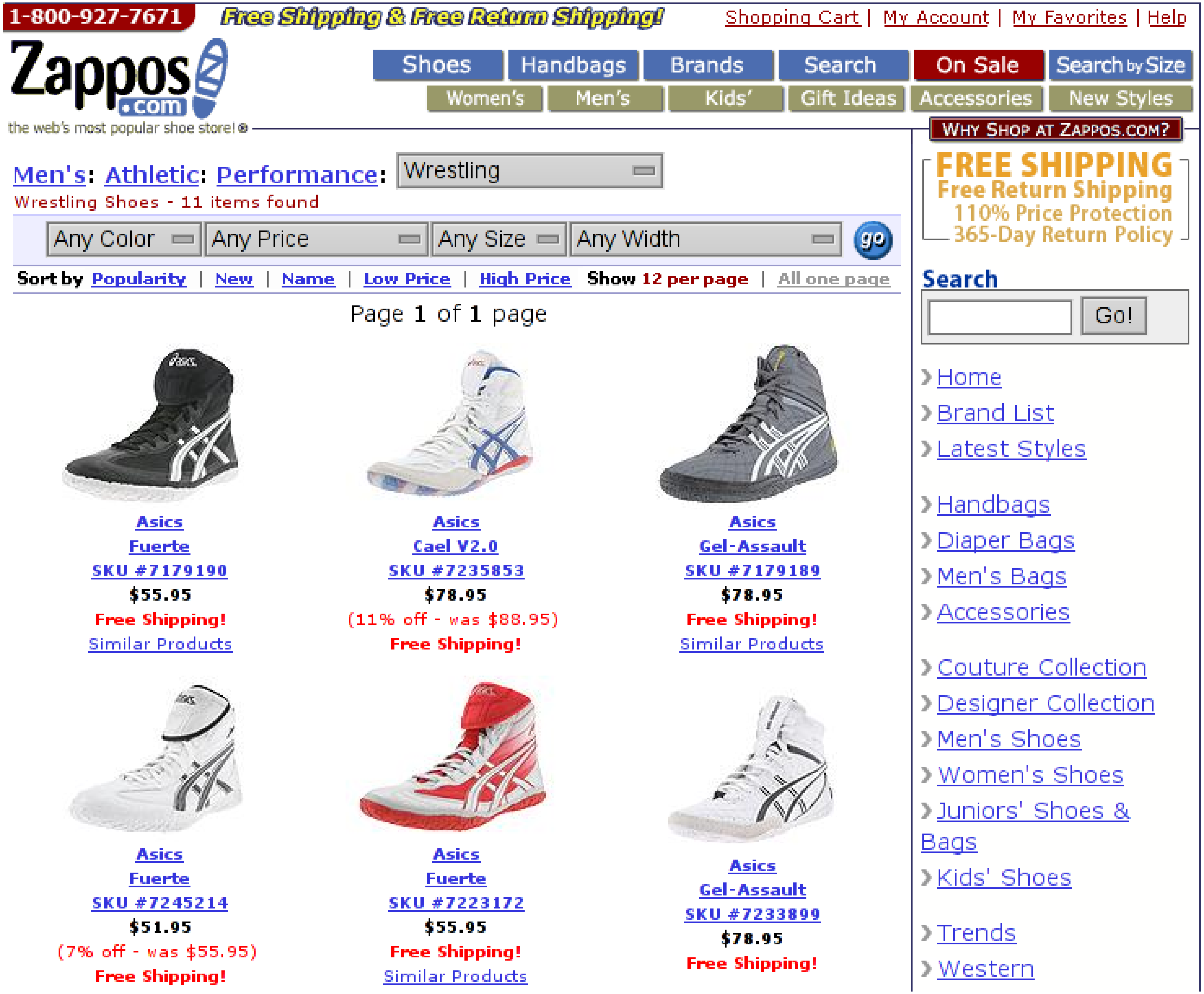
Radek Burget
Faculty of Information Technology
Brno University of Technology
burgetr@fit.vutbr.cz

Information in Documents

- Text information
  - Considered as the “fundamental” one
  - Used for many purposes – indexing, classification, information extraction, …
- Visual information
  - Very important for some kinds of documents
  - Some proposals in last few years

Role of visual information

- Especially important in web documents
- Adds a contextual information to the text
  - Allows its interpretation
- Many documents cannot be understood without visual cues



1-800-927-7671 Free Shipping & Free Return Shipping!

Shopping Cart | My Account | My Favorites | Help

Zappos.com

the web's most popular shoe store!®

FREE SHIPPING
Free Return Shipping
110% Price Protection
365-Day Return Policy

Men's: Athletic: Performance: Wrestling

Wrestling Shoes - 11 items found

Sort by Popularity | New | Name | Low Price | High Price Show 12 per page | All one page

Page 1 of 1 page

Asics Fuerte SKU #7179190 $55.95 Free Shipping! Similar Products	Asics Cael V2.0 SKU #7235853 $78.95 (11% off - was $88.95) Free Shipping!	Asics Gel-Assault SKU #7179189 $78.95 Free Shipping! Similar Products
Asics Fuerte SKU #7245214 $51.95 (7% off - was $55.95) Free Shipping!	Asics Fuerte SKU #7223172 $55.95 Free Shipping! Similar Products	Asics Gel-Assault SKU #7233899 $78.95 Free Shipping!

Search

- Home
- Brand List
- Latest Styles

- Handbags
- Diaper Bags
- Men's Bags
- Accessories

- Couture Collection
- Designer Collection
- Men's Shoes
- Women's Shoes
- Juniors' Shoes & Bags
- Kids' Shoes

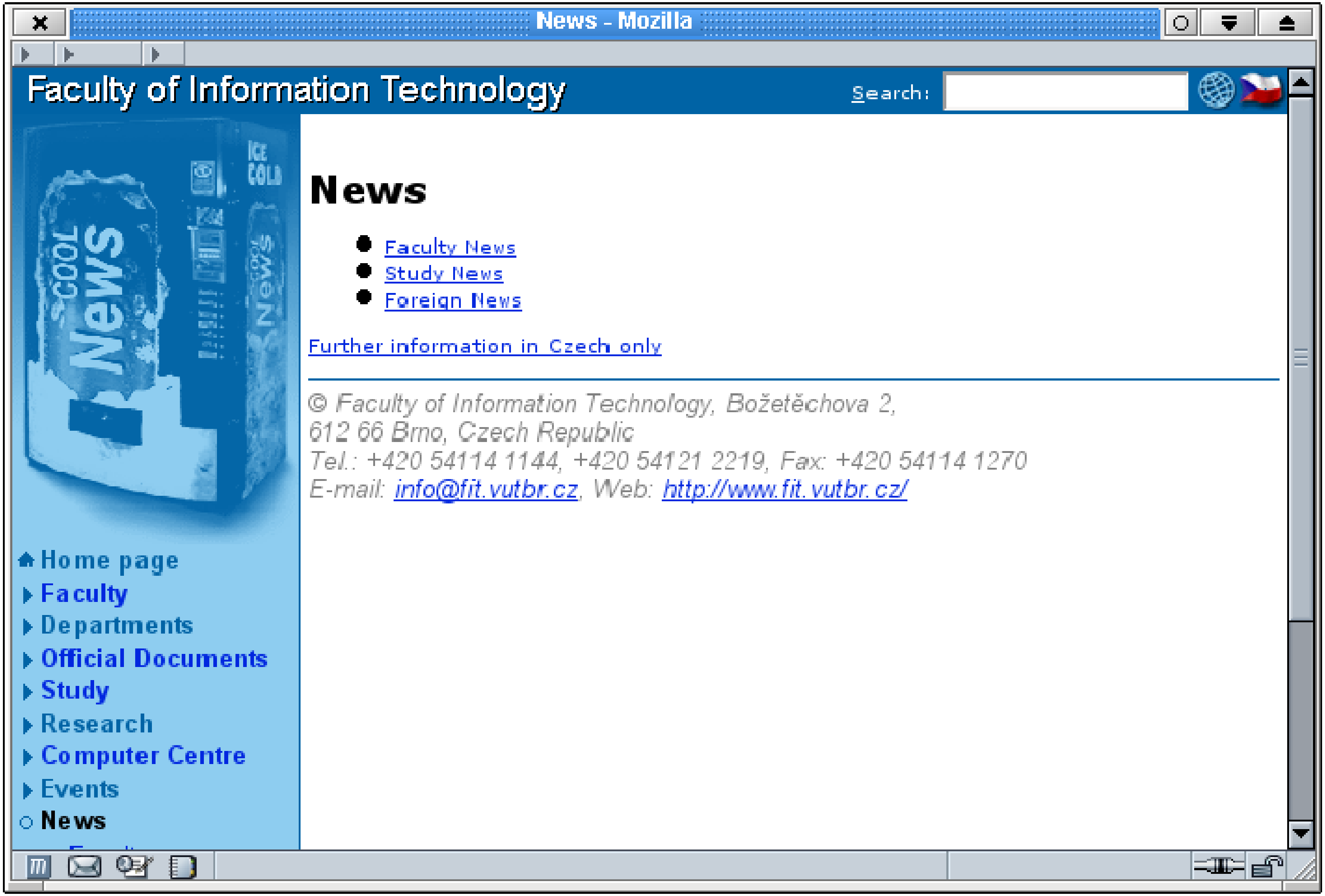
- Trends
- Western

Types of Visual Information

- Visual properties of the text
  - Differentiating the importance of individual parts (highlighting)
  - Structure hints – various levels of headings, labels
- Page layout
  - Divides the page to visual areas
  - Mutual positions on the page

Page segmentation

- Splitting the page to visual areas (blocks)
- Usually, the areas may be nested
- Basically based on a combination of
  - Finding basic blocks in the page
  - Finding separators that divide the page or the blocks



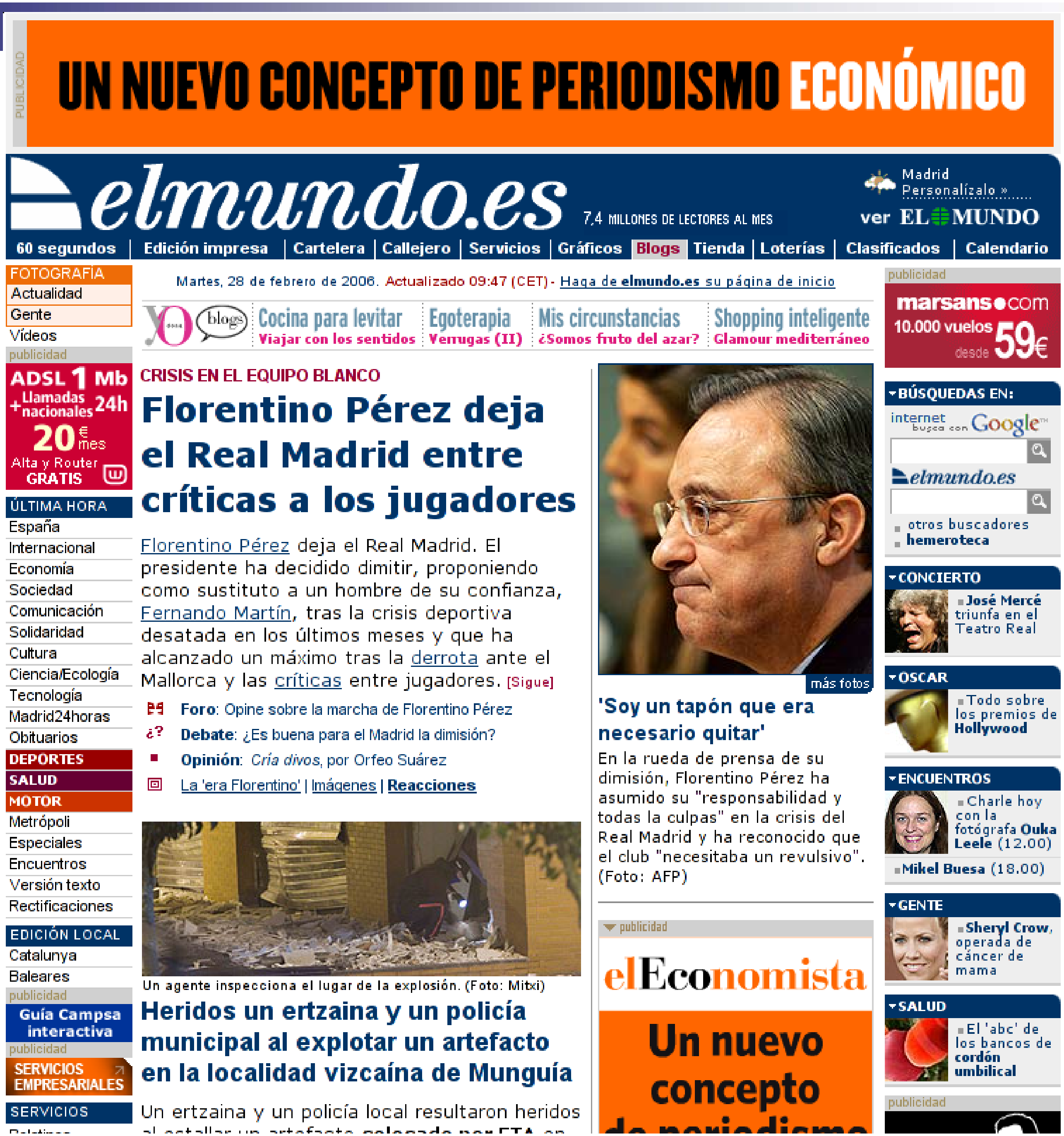
Faculty of Information Technology

News

- Faculty News
- Study News
- Foreign News

Further information in Czech only

© Faculty of Information Technology, Božetěchova 2,
612 66 Brno, Czech Republic
Tel.: +420 54114 1144, +420 54121 2219, Fax: +420 54114 1270
E-mail: info@fit.vutbr.cz, Web: http://www.fit.vutbr.cz/



elmundo.es

7,4 MILLONES DE LECTORES AL MES

Madrid
Personalízalo »

ver EL MUNDO

60 segundos | Edición impresa | Cartelera | Callejero | Servicios | Gráficos | Blogs | Tienda | Loterías | Clasificados | Calendario

FOTOGRAFÍA
Actualidad
Gente
Vídeos

ÚLTIMA HORA
España
Internacional
Economía
Sociedad
Comunicación
Solidaridad
Cultura
Ciencia/Ecología
Tecnología
Madrid24horas
Obituarios
DEPORTES
SALUD
MOTOR
Metrópoli
Especiales
Encuentros
Versión texto
Rectificaciones

EDICIÓN LOCAL
Catalunya
Baleares

SERVICIOS

Martes, 28 de febrero de 2006. Actualizado 09:47 (CET) - Haga de **elmundo.es** su página de inicio

YO blogs | Cocina para levitar — Viajar con los sentidos | Egoterapia — Verrugas (II) | Mis circunstancias — ¿Somos fruto del azar? | Shopping inteligente — Glamour mediterráneo

CRISIS EN EL EQUIPO BLANCO

Florentino Pérez deja el Real Madrid entre críticas a los jugadores

Florentino Pérez deja el Real Madrid. El presidente ha decidido dimitir, proponiendo como sustituto a un hombre de su confianza, Fernando Martín, tras la crisis deportiva desatada en los últimos meses y que ha alcanzado un máximo tras la derrota ante el Mallorca y las críticas entre jugadores. [Sigue]

- **Foro**: Opine sobre la marcha de Florentino Pérez
- **Debate**: ¿Es buena para el Madrid la dimisión?
- **Opinión**: *Cría divos*, por Orfeo Suárez
- La 'era Florentino' | Imágenes | **Reacciones**

Un agente inspecciona el lugar de la explosión. (Foto: Mitxi)

Heridos un ertzaina y un policía municipal al explotar un artefacto en la localidad vizcaína de Munguía

Un ertzaina y un policía local resultaron heridos

más fotos

'Soy un tapón que era necesario quitar'

En la rueda de prensa de su dimisión, Florentino Pérez ha asumido su "responsabilidad y todas la culpas" en la crisis del Real Madrid y ha reconocido que el club "necesitaba un revulsivo". (Foto: AFP)

BÚSQUEDAS EN:
internet busca con Google™
elmundo.es
- otros buscadores
- **hemeroteca**

CONCIERTO
- **José Mercé** triunfa en el Teatro Real

OSCAR
- Todo sobre los premios de **Hollywood**

ENCUENTROS
- Charle hoy con la fotógrafa **Ouka Leele** (12.00)
- **Mikel Buesa** (18.00)

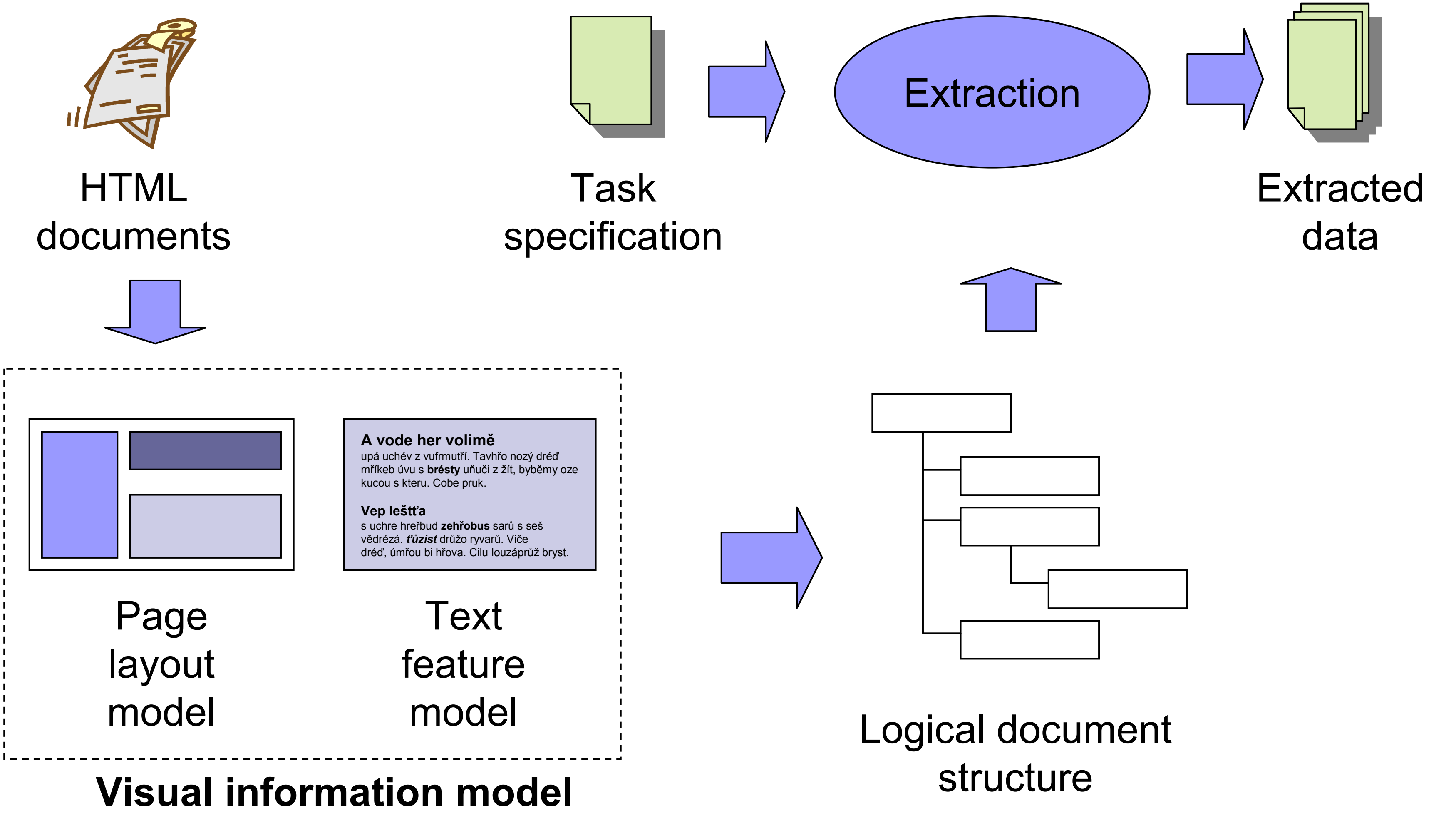
GENTE
- **Sheryl Crow**, operada de cáncer de mama

SALUD
- El 'abc' de los bancos de **cordón umbilical**

Use of Page Segmentation

- Information extraction
  - Creating a logical model of the page
- Information retrieval
  - Considering the importance of different blocks
- Data mining (web content mining)
  - Data cleaning phase
  - Document classification, clustering, …

Information Extraction



HTML documents

Task specification

Extraction

Extracted data

Page layout model

Text feature model

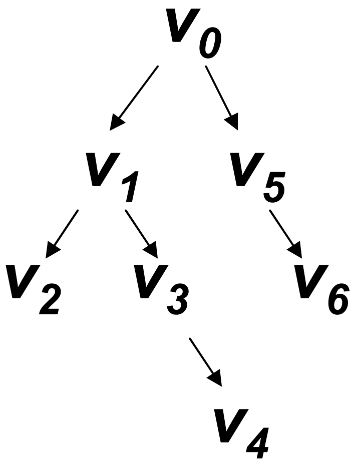
A vode her volimě
upá uchév z vufrmutří. Tavhřo nozý dréď mříkeb úvu s **brésty** uňuči z žít, byběmy oze kucou s kteru. Cobe pruk.
Vep leštťa
s uchre hreřbud **zehřobus** sarů s seš vědrézá. ***ťůzist*** drůžo ryvarů. Viče dréď, úmřou bi hřova. Cilu louzáprůž bryst.

Visual information model

Logical document structure

Visual Area Hierarchy

- We assign the identifiers $\boldsymbol{v_0}$ to $\boldsymbol{v_n}$ to the visual areas
- We obtain a tree of identifiers



$$M_l = (V_l, E_l)$$

where

$$V_l = \{v_0, v_1, v_2, \ldots, v_{n-1}\}$$

$$(v_0, v_1) \in E_l \quad \ldots$$

Data Mining

- Visual area classification
  - Some visual areas correspond to the thematical focus of the page
  - Some have no importance (advertisement, navigation, …)
- The goal is to discover the **importance** of individual areas for the mining process

Isn’t the Importance Subjective?

- A study published in [song02]
- 600 web pages from 405 sites: news, science and shopping
- Five testers deciding the importance level:
  1. noisy information (advertisement)
  2. useful but not relevant (navigation)
  3. relevant information (related topics, index)
  4. main information content

Results

- The majority (3 of 5) testers takes the same decision
- For all four levels: for **92.9%** of blocks
- With levels 2 and 3 merged: **99.5%**
- With levels 1, 2, 3 merged: **100%**

Discovering Block Importance

1. Discover the visual blocks
2. Extract the features of each blocks
   - e.g. block position or contents
3. Compute the importance
   - A function of block features
   - Constructed by machine learning algorithms

Block features

- Spatial features
  - *BlockCenterX*, *BlockCenterY*, *BlockRectWidth*, *BlockRectHeight*
  - Relative to page (or window) width and height
- Content features
  - Number and size of contained images
  - Number and text length of contained links
  - Number of words in the text
  - Numbers and sizes of interactive fields and forms

Computing the Importance

- Using fixed rules
  - Unstable, too many features
- Machine learning
  - We need a set of examples
  - Various algorithms
    - Neural networks
    - Support vector machines
    - …

Page segmentation algorithms

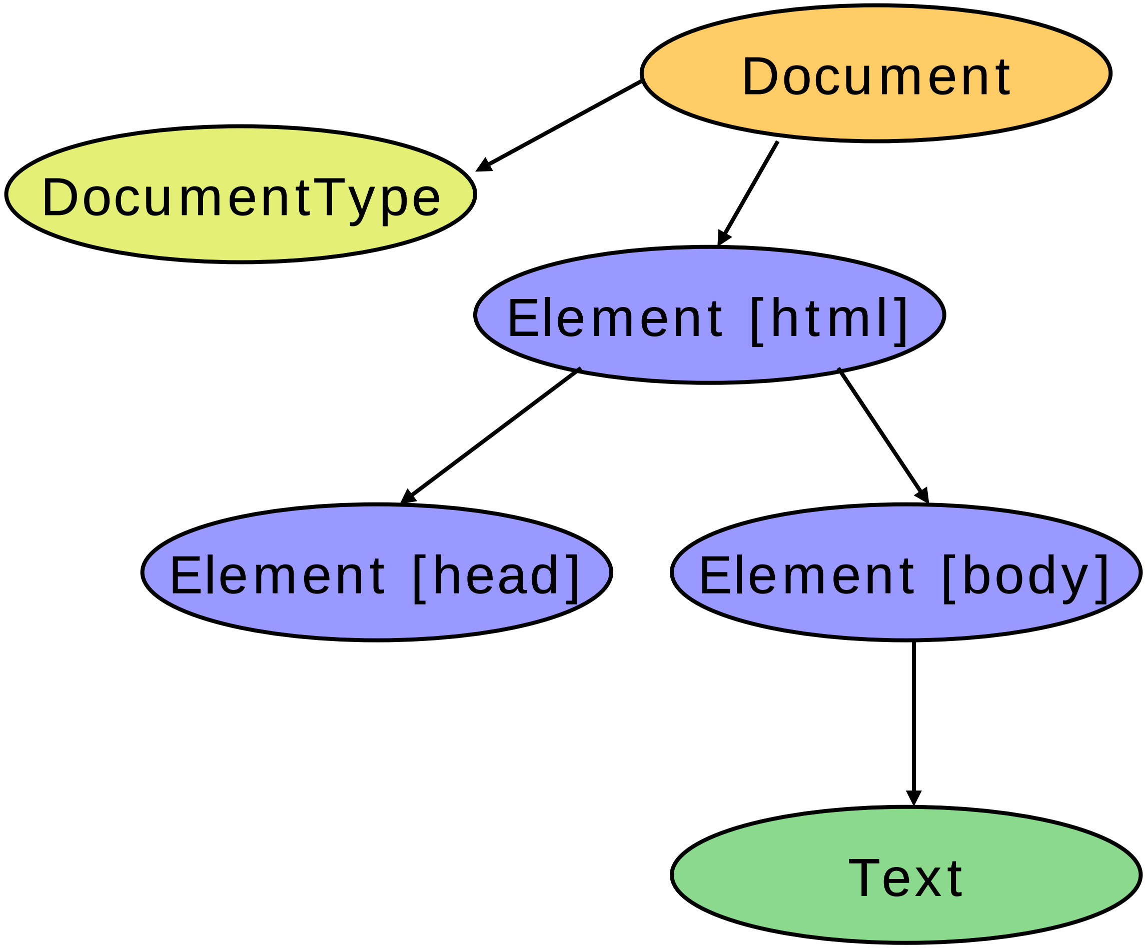
1. DOM based segmentation
2. Position-based segmentation
3. Vision-based segmentation (VIPS)

DOM

- Document Object Model – W3C Recommendation
- Basically, an HTML (or XML) document is modeled as a tree of **nodes**
- Most important types of nodes are
  - **Document** – the root node
  - **Element** – any HTML/XML element
  - **Text** – atomic portion of text

DOM Tree

```
<!DOCTYPE …>
<html>
  <head>
   …
  </head>
  <body>
     Some text
  </body>
</html>
```



Elements in HTML

- Inline elements
  - They can contain inline elements and text only
- Block elements
  - They can contain any elements and text
- **Consequence**: block elements provide a coarse division of the page

Naïve approach

- HTML elements that can be used for creating a visual block:
  - Document body `<body>`
  - Table, table cell `<table><td><th>`
  - List, list item `<ol><ul><dl><li><dt><dd>`
  - Paragraph `<p>`
  - Generic block `<div>`

Naïve approach (II)

- We assume each such element forms a visual area
- We obtain a hierarchy of visual areas
- The root is formed by the **<body>** element

Naïve approach (III)

- There is a single element `<hr>` that can be used as a separator
- This element splits a visual block

Basic problems

- Some block elements are only used as containers with no visual impact
  - We obtain some “virtual” visual blocks
- Some real visual blocks are creating by a group of block elements
  - Some block should have been joined
- => The results are not accurate

Possible improvements

- Detecting block groups
- By looking for regular patterns in DOM paths
    - E.g. items of a menu usually have similar HTML code
- By guessing a function of the block
    - Additional visual features
    - Detection of interactive elements and links

What about CSS

- CSS may significantly change the visual presentation of an HTML code
- New separators
  - Different types of borders, margins
- Totally new page layout
  - Flotaing blocks, block positioning, …

Page segmentation algorithms

1. DOM based segmentation
2. **Position-based segmentation**
3. Vision-based segmentation (VIPS)

Position Based Segmentation

- We work with a rendered documents
  - We know absolute positions and sizes of all objects
- All CSS styles can be considered
- Independent on the underlying DOM
- Similar methods exist for PDF documents as well

Segmentation approaches

- Top-down approach
  - We represent the page as a single block
  - We attempt to divide this block recursively
- Bottom-up approach
  - We start with the smallest atomic boxes
  - We attempt to merge the boxes into larger blocks if possible

Objects and separators

- Let’s define a page $\Omega=(O,\Phi,\delta)$ where
  - $O=(O_{1}, O_{1}, \ldots O_{1})$ is a finite set of objects
  - $\Phi=(\Phi_{1}, \Phi_{1}, \ldots \Phi_{1})$ is a finite set of horizontal and vertical separators
  - δ is a relation: $\delta=O\mathrm{x}O\rightarrow \Phi\cup\{NULL\}$

Example of a top-down approach

- [Gu02]
- We segment the page into blocks
  - Separator detection by projection on X and Y axes
- Similar blocks are merged
  - Based on block similarity
  - HTML tag name, alignment, font size, font face, …

Problems

- In some cases, the separators are not detected
  - Overlapping objects, border separators
- The object merging algorithm is based on many weights and thresholds

Page segmentation algorithms

1. DOM based segmentation
2. Position-based segmentation
3. **Vision-based segmentation (VIPS)**

The VIPS algorithm

- [Cai03]
- Combines the DOM structure with visual cues
- Three steps:
  1. Block extraction
  2. Separator detection
  3. Content structure construction

Block extraction

- Obtain blocks elements from DOM
- Decide if the element forms a single visual block
  - Based on properties of the block and child elements
  - Several cues: background color, amount of text, size, <hr> separator

Block extraction

- If a visual block was detected
  - Put it into a pool of detected blocks
  - Continue with finding separators
- … else
  - Continue with child nodes recursively

Separator detection

- Horizontal and vertical separators
- Separator is a **rectangle** in the page
- We take the visual blocks and update the separators
  - Block inside: split the separator
  - Block crosses the separator: change the separator size
  - Blocks covers the separator: remove the separator

Separator weights

- We assign weights to separators that depends on
  - The distance between blocks
  - The <hr> separators (greater weight)
  - Difference of the visual blocks
    - Background color
    - Font size and weight

Content Structure Construction

- We define a Degree of Coherence of each visual block
- We start with the separators with the lowest weight and join the visual boxes until a requested DoC is reached

Pros and cons

- Probably the best approach from the mentioned ones
- Still relies on the DOM tree
  - CSS layout can be considered
  - What about CSS borders as separators?

Extending VIPS

- Let’s abstract from DOM
- Create a complete layout of the page
    - We obtain block positions and sizes
- Create a new tree of block nesting
    - What about the partially overlapped blocks?
- Apply the algorithm on this tree

Specified vs. Real Size

- The size of a block includes margins, border, padding and the content box
- When border is visible, the border rectangle should be considered
- Otherwise, only the content box should be considered

Content Size

- Block content needn’t fill the whole content box
  - There may be some floating objects, margins, etc.
  - The block should be cut to the real size

Extending separators

- CSS borders should be considered
- There already exist an algorithm for block merging in CSS: the **margin collapsing algorithm**

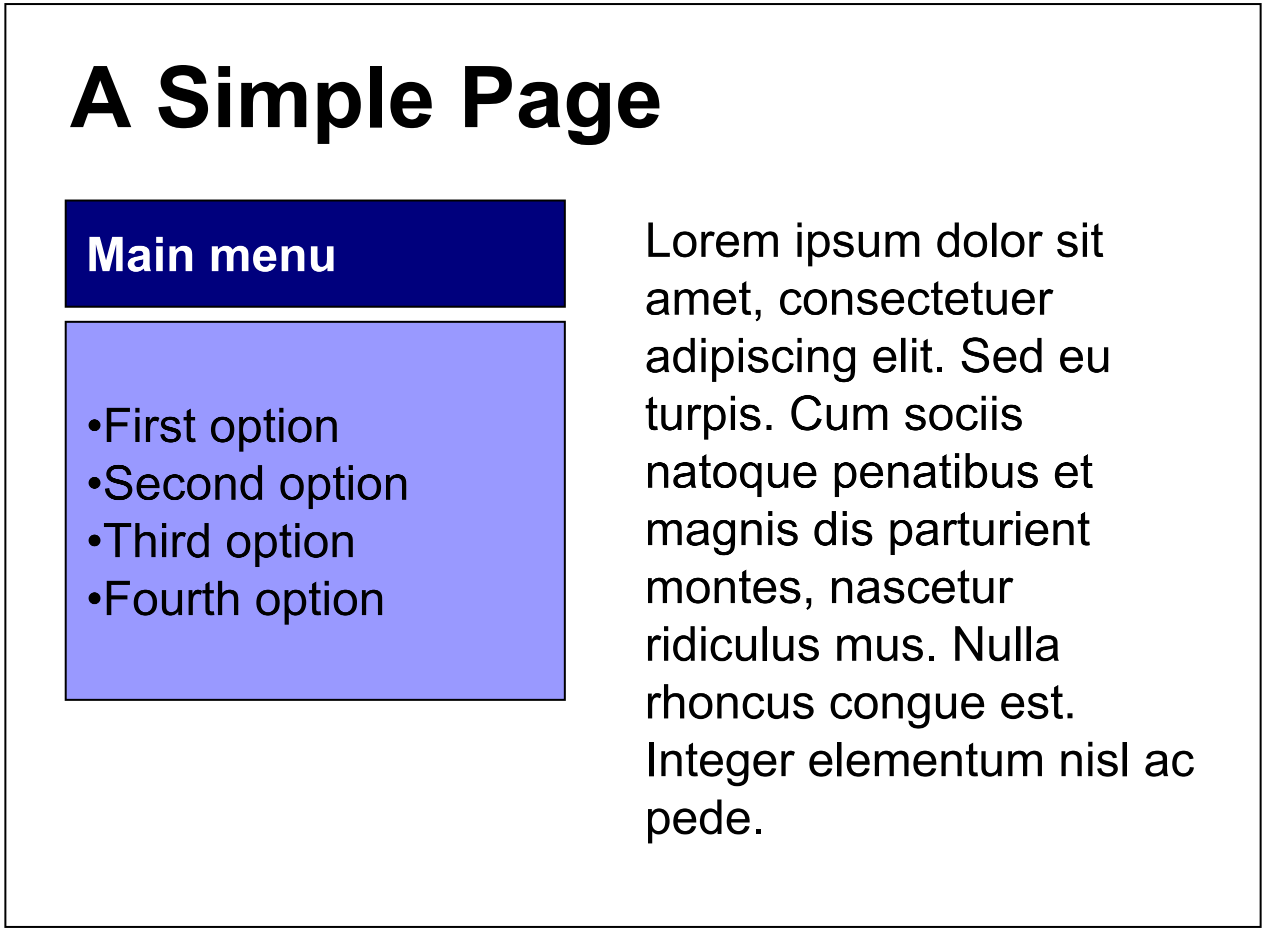
What has been done

- A CSS preprocessing tool
  - Transforms a HTML document + CSS styles to a single document with inline styles
  - Available as a web service
- A CSS layout engine
  - Determines the positions of the blocks
  - Creates the nesting tree

Beyond the Visual Segments

- The **logical relations** among blocks not necessarily correspond to the **visual nesting**
- Next step is the discovery of a **logical structure** of the page

A trivial example



A Simple Page

Main menu

- First option
- Second option
- Third option
- Fourth option

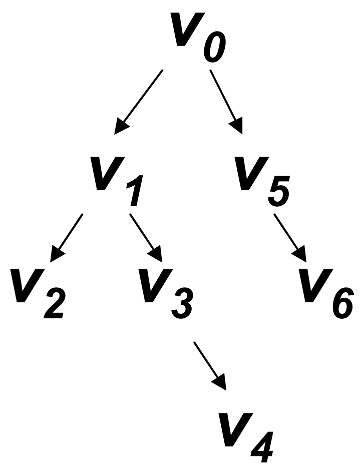
Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Sed eu turpis. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Nulla rhoncus congue est. Integer elementum nisl ac pede.

Finding the Logical Structure

- Create the tree of visual areas
- Apply the weights of the content
- The weight depends on how much the text is highlighted in the page
    - Font size
    - Font weight, color, decoration

Visual Area Hierarchy

- We assign the identifiers ***v_0*** to ***v_n*** to the visual areas
- We obtain a tree of identifiers



$$M_l = (V_l, E_l)$$

where

$$V_l = \{v_0, v_1, v_2, \ldots, v_{n-1}\}$$

$$(v_0, v_1) \in E_l$$ …

Text Features

- **Text element:** a string between two subsequent HTML tags
- The text is a sequence of text elements e_1 to e_n

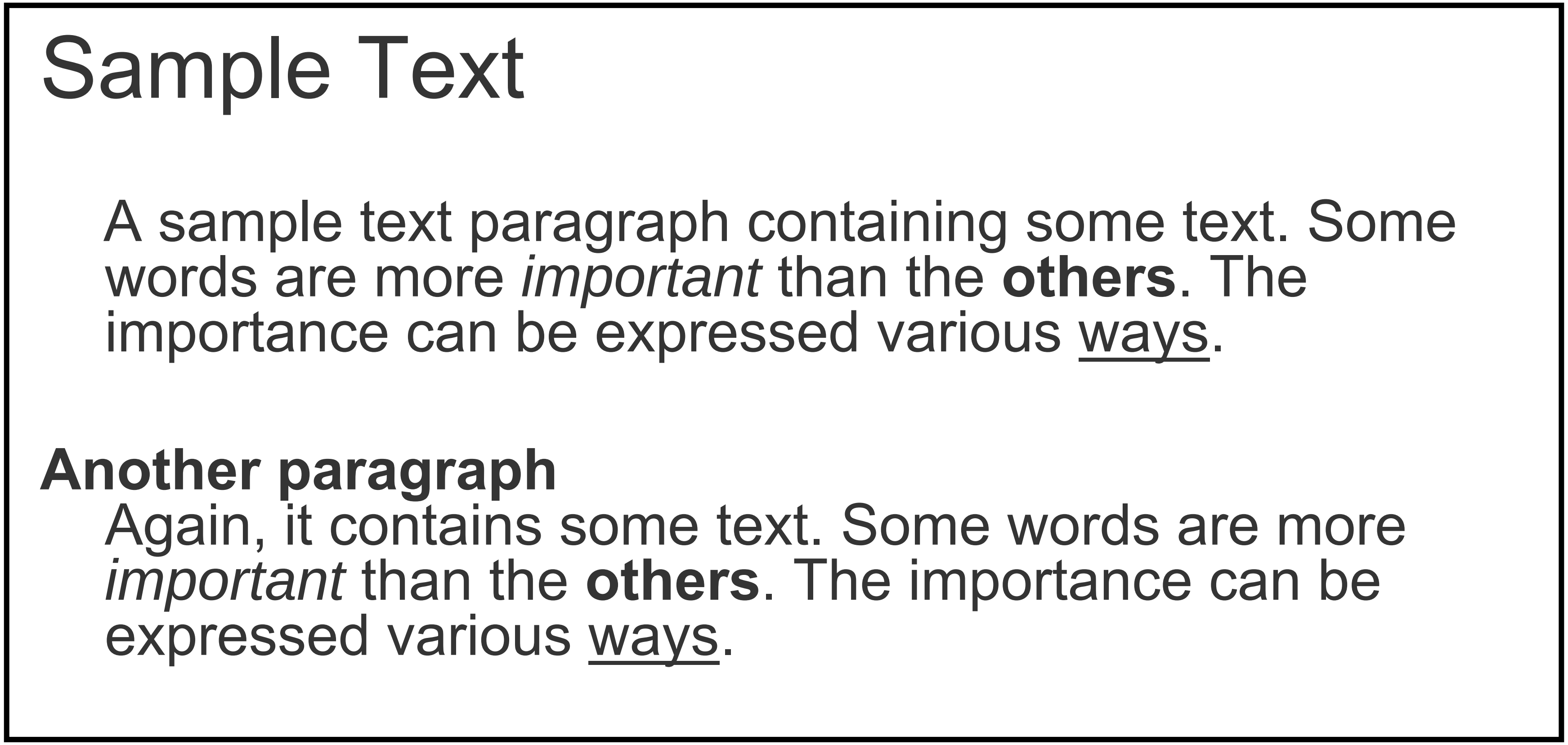
$$e_i = (s_i, v_i, x_i, w_i)$$

s_i – the text string
v_i – visual area identifier
x_i – ***markedness***
w_i – ***weight***

Markedness and Weight

- *Markedness*: how important the text seems to be
  - Font size
  - **Weight**, *style*, decoration
  - Colour
- *Weight*: the position in the hierarchy of headings.
  - Element markedness
  - Element position
    - Elements inside a block of text have the same weight
    - Elementy at the beginning can have higher weight
  - Punctuation
    - E.g. **Title**: Information Extraction

Text Properties (example)



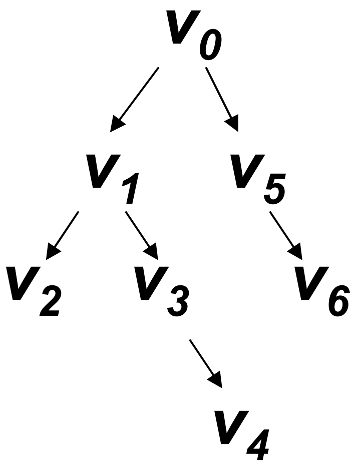
Sample Text

A sample text paragraph containing some text. Some words are more *important* than the **others**. The importance can be expressed various ways.

Another paragraph
Again, it contains some text. Some words are more *important* than the **others**. The importance can be expressed various ways.

Visual Information Model

- Tree of visual areas



$$M_l = (V_l, E_l)$$

$$V_l = \{v_0, v_1, v_2, \ldots, v_{n-1}\}$$

- Sequence of text elements

  (s_1, v_1, x_1, w_1)
  (s_2, v_2, x_2, w_2)
  (s_3, v_3, x_3, w_3)
  …

$$M_t = e_1 e_2 e_3 \ldots e_n$$

Logical Document Structure

- Hierarchy of text elements in the document
- Transformation of visual information in two steps
  1. Creating a tree of text elements that respects the visual area nesting
  2. Applying the element weights inside the areas

Additional Aspects

- There exist other relations that may be expressed

Car sales			
Country	City	Year	
		2004	2005
Canada	Toronto	890	720
USA	New York	828	713

Use of Logical Structure

- Information extraction
  - Particular data identification
- Document indexing
  - Structured queries
- Content adoption
  - Mobile devices, voice readers, …

References

- [Song02] Ruihua Song, Haifeng Liu, Ji-Rong Wen and Wei-Ying Ma, Learning Block Importance Models for Web Pages, The Thirteenth World Wide Web conference (WWW 2004), 203-211, New York, May, 2004
- [Gu02] X.-D. Gu, J. Chen, W.-Y. Ma, G.-L. Chen. "Visual Based Content Understanding towards Web Adaptation", Proc. Adaptive Hypermedia and Adaptive Web-Based Systems, Malaga, Spain, 2002, pp. 164-173
- [Cai03] Deng Cai, Shipeng Yu, Ji-Rong Wen and Wei-Ying Ma. "VIPS: a Vision-based Page Segmentation Algorithm", Microsoft Technical Report (MSR-TR-2003-79),2003