


# Analysis of Semi-structured Data Based on Area of Interest

Aleksandr Kozhenkov  
ITMO University

# Different structure of each site

Турция, Сиде ☀️ +25 от 267 \$

ОТЕЛЬ Inside 4★



8 мая, Чт.	на 7 ночей	AI	267\$	Подробнее
27 апреля, Вс.	на 4 ночи	AI	294\$	Подробнее
7 мая, Ср.	на 7 ночей	AI	468\$	Подробнее

Горящий тур в Турцию из Москвы от 22 апреля в 12:00 Для некоторых авиакомпаний в цену не включен топливный сбор.

Edellinen	1	2	3	4	5	...	7184	Seuraava
Matkakohde ⇅	Lähtö ▾		Kumppani ⇅		Hinta ⇅			
<b>RODOS</b> , Kreikka Pelkät lennot	Kuopio ke 23.4.2014				<b>420€</b> 7 päivää	<b>VARAA</b>		
<b>ANTALYA</b> , Turkki Pelkät lennot	Helsinki ke 23.4.2014				<b>215€</b> 8 päivää	<b>VARAA</b>		
<b>ANTALYA</b> , Turkki Määrittelemätön hotelli	Helsinki ke 23.4.2014				<b>275€</b> 8 päivää	<b>VARAA</b>		
<b>ANTALYA</b> , Turkki Määrittelemätön hotelli	Helsinki ke 23.4.2014				<b>255€</b> 8 päivää	<b>VARAA</b>		

# Data search on the page

1) Average length of text

2) Count of blocks

3) Similarity

...

# Pagination handling

- 1) Search for button
- 2) Emulation of keystroke
- 3) Re-search data

# Extraction of essential parts

Algorithm TEX

1) Search for patterns

2) Split by similarity

# Analysis of found data

Using ontologies

1) Date

2) Price

3) Destination

...