

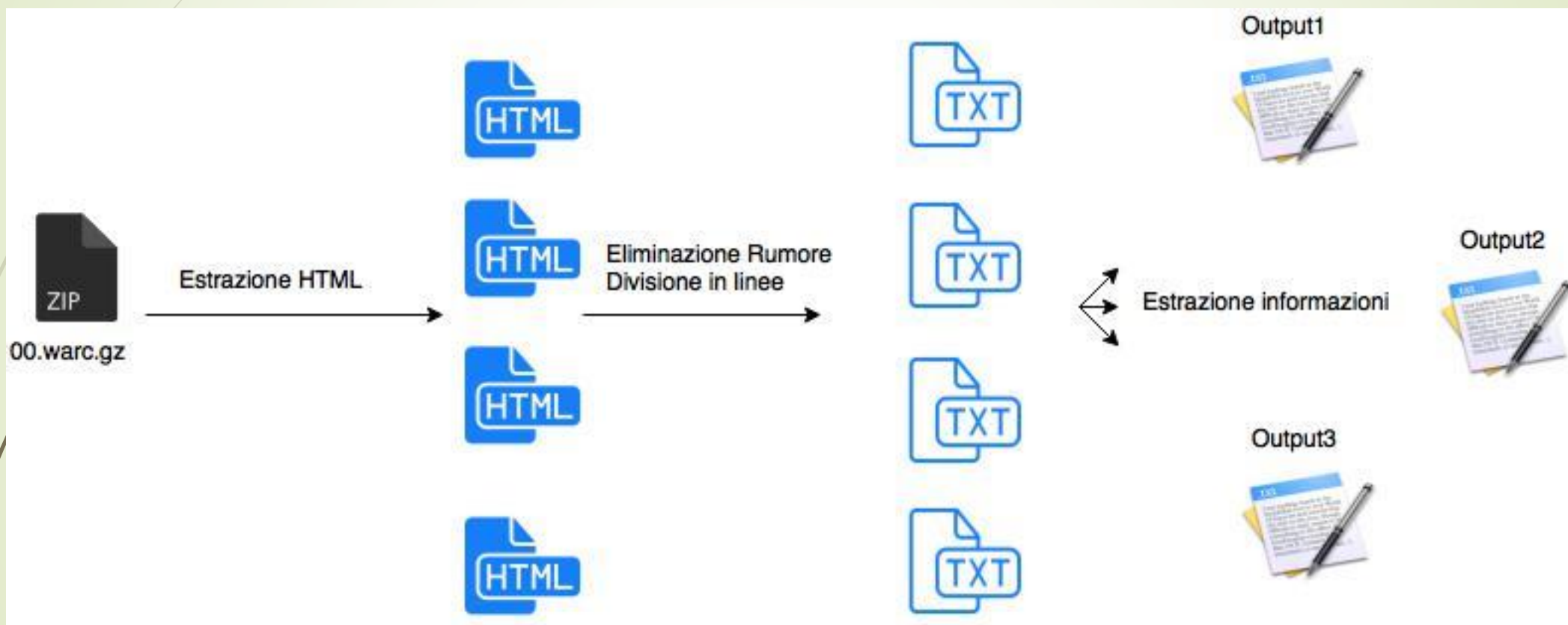
Explore3

Tag Mining

Terzo progetto Analisi e gestione dell'informazione su web 2014/2015

Pietro Coronas, Rudi Veshti, Dorjan Dika

Step



Estrazione record HTML

```
1 <!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
2 <html>
3 <head>
4 <STYLE type="text/css">
5 BODY,P,A {font-family:sans-serif; color:black; background-color:white; font-size:12pt; font-weight:normal;}
6 H1 {font-family:sans-serif; color:black;font-size:16pt;font-weight:bold;}
7 </STYLE>
8 </head>
9 <html>
10 <head>
11 <title>Accessible Aviation Equiz</title>
12 <STYLE type="text/css">BODY,P,A {font-family:<? echo "" ;?>; background-color:<? echo "" ;?>; color:<? echo "" ;?>; font-size:<? echo "" ;?> pt; font-weight:normal;}H1
    {font-family:<? echo "" ;?>; color:<? echo "" ;?>;font-size:<? echo +4 ;?>pt; font-weight:bold;}</STYLE></head>
13
14 <body>
15 | <center>
16 <H1><br>Quiz Selection<br></H1>
17 <P align = "left"><A>You can choose to review either 11 questions or a reduced bank of lash Card Questions </A><A>The lash Card Questions are the questions closest to the
    regulation or concept the FAA is testing.</A> <A>If you review only the lash Card Questions you will be able to pass the test with a good score.</A><A>We recommend
    reviewing all the questions to make you as prepared as possible for the test. </A><A>The FAA ATP test has 80 questions and passing is 70%.<br></A></p>
18 <P><a href='?quiz=1'>All Questions</a></P>
19 <P><a href='?quiz=2'>Flash Card Questions</a></P>
20 | </center>
21 </body>
22 </html>
23
```

- Il record HTML viene estratto dal Warc e salvato in un file
 - Il file avrà nome *TREC_ID.txt*

Eliminazione rumore : JSoup

```
1 Keyword Search Keyword Search To search the board, enter one or more words below. Placing a + sign in front of a word requires the word. Placing a - sign in front of a word  
excludes the word. Placing a "phrase in quotes" matches the entire phrase. Search for: Search topic: All topics on board Public Forum Look in: Subject lines Names of  
authors Text of messages Type of page: Pages containing messages Pages with or without messages
```

- Eliminazione dei tag: #Select, #Button, #Script,
- Metodo `doc.text()` : Estrae il testo dalla pagina HTML e lo riporta in un'unica stringa

Divisione in Linee

```
1 Keyword Search Keyword Search To search the board, enter one or more words below
2 Placing a + sign in front of a word requires the word
3 Placing a - sign in front of a word excludes the word
4 Placing a "phrase in quotes" matches the entire phrase
5 Search for: Search topic: All topics on board Public Forum Look in: Subject lines Names of authors Text of messages
6 Type of page: Pages containing messages Pages with or without messages
7
```

- Inizialmente basata su Linee di lunghezza prefissata
 - Troppa perdita di informazione: nel 90% dei casi divisa tra più linee
- Basata sui caratteri `<point><space>` o `<space><space>`
 - Pro: Minor perdita di informazione , Linee ne troppo lunghe ne troppo corte
 - Contro: In alcuni casi l'informazione viene divisa su più linee o si trova in una linea che verrà successivamente scartata come rumore: (Jan. 14, 2012)

Eliminazione Rumore : *Regex*

```
1 ||Trec_id: 00012 ||String: Home | Contact | Policies Copyright © 1995-2008, zaz Corporation
2
3 ||Trec_id: 00014 ||String: German 3 Home | Contact | Policies Help build the largest human-edited directory on the web. Submit a Site - Open Directory Project
4
5 ||Trec_id: 00029 ||String: Home | Contact | Policies Help build the largest human-edited directory on the web. Submit a Site - Open Directory Project -
6
7 ||Trec_id: 00033 ||String: www.geocities.com E-Mail Password E X P A N D m i n i m i z e
8
9 ||Trec_id: 00035 ||String: \?D? - z/, Chinese Simplified - 00perdomain.com z/, Chinese Simplified, World - Top \?D? \? 46 ?H 103 ^?{\?
10
11 ||Trec_id: 00035 ||String: ? 2,615 ?? 50 ? 11 Home | Contact | Policies Help build the largest human-edited directory on the web. Submit a Site - Open
12
13 ||Trec_id: 00036 ||String: ?U - z/, Chinese Simplified - 00perdomain.com z/, Chinese Simplified, World - Top ?U z/114z/114 ?
14
15 ||Trec_id: 00042 ||String: www.jah.ee w3 - k?ige kiiremad uudisedw3 - k?ige kiiremad uudised Reaalajas uudised paljudelt eesti- ja v?lismaistelt v?ljaannetelt
16
17 ||Trec_id: 00044 ||String: T??indi - F?royskt - 00perdomain.com F?royskt, World - Top T??indi News and Media - Faroe Islands, Europe, Regional
18
19 ||Trec_id: 00044 ||String: 1 Dimmal?ttingDimmal?tting Landsins st?rsta bla?
20
21 ||Trec_id: 00047 ||String: www.mookart.nl E-Mail Password E X P A N D m i n i m i z e
```

- Abbiamo definito una *Regex* per identificare linee contenenti rumore
 - Queste linee vengono scartate e non passano per la fase di match
 - Estratto del file *Cluweb09_Noise.txt*

Estrazione Informazioni

- L'estrazione avviene tramite Regex
 - **Link** : in formato www | http:// | https | sito.com
 - **Data**: Vari formati, tra cui D/M/Y | M D, Y | D-M-Y | D M Y |
 - **Distanza** : Basata su km | kilometer | mi | ft | yd | m | cm | dm | mm | ...
 - **Unità**: Basata su mw | kw | ton | µg | µl | µgml | kg | kilogram | pound | mb | gb | pb....
 - **Numeri**: Telefonici/Fax/Sequenze: (800) 123 123 1234 | (800) 328-3456 | 540-662-9041 | 310.566.7560 | 0114115953 |
 - **Email**: Vari formati: info@1969web.com | mail@legacyadventure.com | tomcat-dev@jakarta.apache.org
 - **Tempo**: 11:23 (AM/PM) | 12:23:45 (AM/PM) | 980s
 - **Dimensioni**: 23x45 cm | 23 cm x 34 cm | ...


Estrazione Informazioni

- L'estrazione avviene tramite Regex
 - **Indirizzi:** 91st Street | 581 Flushing Avenue | 1543 President Street | 4706 - 54Street
 - **Indirizzi IPV4:** 255.255.255.0 | 150.0.5.1 | 127.0.0.1
 - **File Format:** .flac | .mpeg | .avi | .mp3 | .gif |
 - **Money:** \$17.95 | usd 75 million | euro 75 milion | ...
 - **Stati americani e provincie:** New York | California | Washington | ..
 - Testata e funzionante ma non implementata nel progetto finale.
 - #StatiAmericani: 32332



Output

- Clueweb09_StringaDaRimpiazzare(Output1)
 - Contiene solo il Match della Regex
- Clueweb09_FraseSenzaTag(Output2)
 - Contiene la Stringa in cui è presente il Match della Regex
- Clueweb09_FraseConTag(Output3)
 - Contiene la Stringa in cui è presente il Match della Regex con i Tag al posto della Stringa Match
 - Se in una frase sono presenti più Match, riportiamo direttamente un'unica frase invece che più copie della stessa frase con diversi Match
- Clueweb09_Noise
 - Contiene le stringhe considerate rumore
- Clueweb09_Stats
 - Contiene statistiche



Statistiche: *Estratto da Clueweb09_stats.txt*

- 1 Dataset, **35581** Html estratti in **2 minuti e 35 secondi**
- Tempo Impiegato Estrazione Informazioni/Scrittura Output: **circa 57 minuti**
- Stringhe Utili Estratte: **139090**
- Stringhe Scartate: **231676**

Statistiche: *Estratto da Clueweb09_stats.txt*

➤ Tag totali Trovati:

- || Tag: #Time **32913**
- || Tag: #FileFormat **208**
- || Tag: #Money **15472**
- || Tag: #Link **18812**
- || Tag: #Unit **83**
- || Tag: #Address **1169**
- || Tag: #Dimension **161**
- || Tag: #Date **53888**
- || Tag: #Email **3886**
- || Tag: #IPV4Address **161**
- || Tag: #Number **10537**
- || Tag: #Distance **1800**