TAGmining

Gruppo MASC – Andrea Cremisini e Matteo Sabatini

Che significa TAG mining?

Finanza e Mercati ▸ In primo piano

Borse caute in attesa della Fed. Oggi la Bce decide sui soldi alle banche greche

17 giugno 2015

Tweet 7 Consiglia 7 g+1 1 My24

I mercati europei procedono prudenti nel giorno della Federal Reserve. Mentre gli investitori restano in attesa di vedere come evolverà la trattativa tra la Grecia e i suoi creditori, oggi si concluderà il Fomc, il comitato di politica monetaria della Fed. Non è atteso al momento alcun intervento sui tassi, tuttavia, ci si aspetta una revisione delle stime sulla crescita statunitense e qualche indicazione relativa all'atteso aumento del costo del denaro entro l'anno, il primo da 10 anni. In leggero rialzo anche le altre piazze europee, con l'eccezione di Parigi.

A Piazza Affari il Ftse Mib procede a cavallo della parità. Spunto di Telecom Italia (+3,2%), all'indomani di nuove indiscrezioni sulla volontà di Vivendi di salire nell'azionariato del gruppo di tlc e Mps (+3%).

Estrarre le informazioni significative da un testo attribuendogli dei TAG per classificarli

Esempio di TAG Mining

ARTICOLI CORRELATI

- Tokyo debole in attesa della Fed (-0,2%). Delude l'export, cala l'import dall'Italia

La Bce decide sui soldi alle banche greche

Il Consiglio Bce si pronuncerà oggi in merito a un'ulteriore estensione dellEla alle banche greche. Sebbene si sia parlato della possibilità che la Bce adotti con la Grecia un atteggiamento analogo a quello che obbligò Cipro ad accordarsi con l'Eurogruppo, il discorso pronunciato da Draghi davanti al parlamento europeo lunedì ha segnalato che la Bce non vuole ancora forzare l'esito dei negoziati togliendo il residuo spazio di manovra al governo greco: nel testo preparato, infatti, Draghi era tornato a collegare la decisione sull'Ela a un elemento tecnico, la solvibilità della banche greche, e aveva rimarcato la natura politica dell'eventuale soluzione della crisi. Perciò è probabile che l'aumento dell'Ela dagli attuali 83 miliardi venga autorizzato, magari per un importo limitato, e che non venga neppure imposto un aumento dei margini di garanzia - almeno fino al prossimo consiglio europeo del 25-26 giugno.

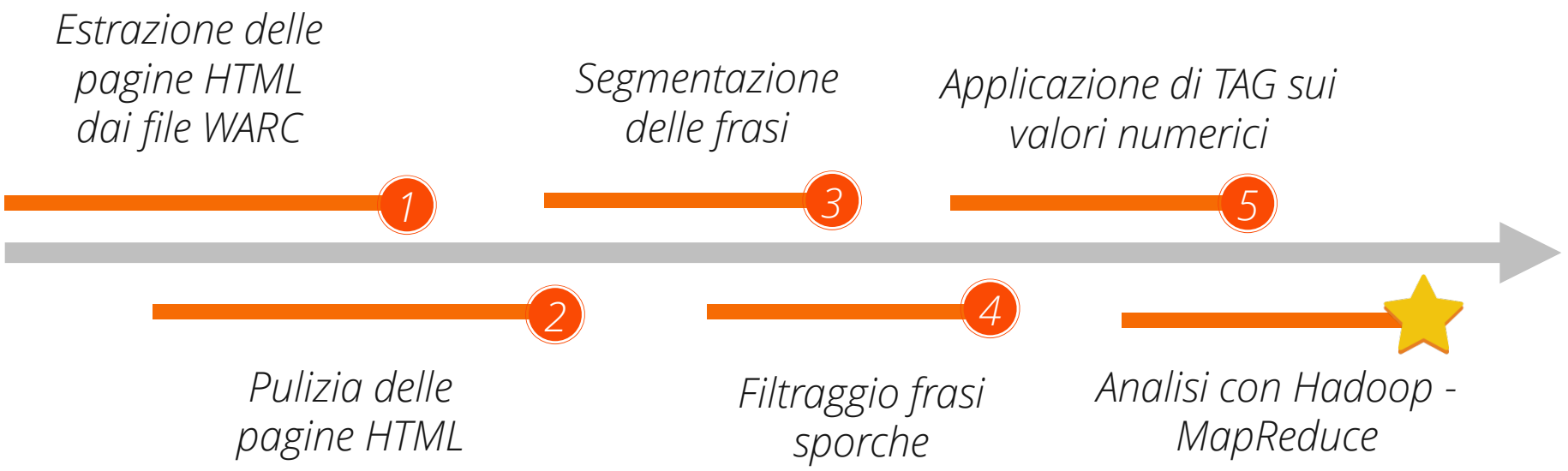
Sul fronte dei cambi, la moneta unica passa di mano a 1,1259 dollari (1,1230 ieri sera), e 139,10 yen (139,61), il biglietto verde vale 123,55 yen (123,43). Il Wti sale dello 0,18% a 60,08 dollari al barile.

Input e Obiettivo

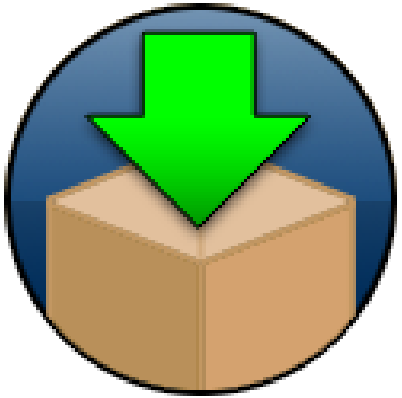
- *Input: ClueWeb09/00warc*

- *Output: tre tabelle:*
  - *out1—> trec-id; stringa_da_rimpiazzare; tag*
  - *out2—> trec-id; frase_estratta_senza_tag*
  - *out3—> trec-id; frase_estratta_con_tag*

Roadmap seguita



Estrazione file HTML



- *Estrazione dei file attraverso GZipped*
- *Pulizia ed estrazione delle informazioni utili dal Warc, in particolare del TREC-ID*

Pulizia della pagine HTML

- *Eliminazione delle parti inutili dell'HTML*
- *Pulizia dalle informazioni non importanti dal body*
- *Tool usato: Jsoup*

Elementi eliminate dall'HTML

- *head*
- *script*
- *title*
- *table (n° parole < 12)*
- *header*
- *meta*
- *nav*
- *footer*
- *sidebar*
- *menu*
- *img*
- *legal*

Pulizia della pagine HTML

PRIMA

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
        "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
        <!-- blojsom-68 -->
        <title>
                Hawk Cast News
                                                                        - Hawk Cast News - Music Edition
                        </title>
                <link rel="stylesheet" href="http://cfpms.ucfsd.org/blojsom_resources/includes/base_layout.css" />
                        <link rel="stylesheet" href="http://cfpms.ucfsd.org/blojsom_resources/stylesheets/v2_layout.css" />
                <link rel="SHORTCUT ICON" href="http://cfpms.ucfsd.org/favicon.ico" />
        <link rel="alternate" type="application/rss+xml" title="RSS" href="feed://cfpms.ucfsd.org/weblog/jwalsh/?flavor=rss" />
        <link rel="alternate" type="application/atom+xml" title="Atom" href="feed://cfpms.ucfsd.org/weblog/jwalsh/?flavor=atom" />
        <link rel="EditURL" type="application/rsd+xml" title="RSD" href="feed://cfpms.ucfsd.org/weblog/jwalsh/?flavor=rsd" />
        <meta name="MSSmartTagsPreventParsing" content="true" />
        <script type="text/javascript" src="http://cfpms.ucfsd.org/blojsom_resources/includes/base_scripts.js"></script>
        <script type="text/javascript" src="http://cfpms.ucfsd.org/blojsom_resources/includes/win32ImageWorkaround.js"></script>
                <script type="text/javascript" src="http://cfpms.ucfsd.org/blojsom_resources/stylesheets/v2_layout.js"></script>
                </head>
<body onload="tryFocusOnEditField();showConfirmMessage();customStartup()">


<div class="span_header" id="header_span_header"></div>
<div class="span_body" id="header_span_body">
        <table id="header_table">
                <tr>
                        <td id="header_description_cell">
                                <h2 id="title"><a href="http://cfpms.ucfsd.org/weblog/jwalsh/">Hawk Cast News</a></h2>
                                <p id="subtitle">A Weekly Audio Podcast from HawkTV</p>
                                                </td>
                        <td id="header_logo_cell"><a href="http://cfpms.ucfsd.org/"><div id="header_logo_img"></div></a></td>
                </tr>
```

DOPO

```
In an idea insipred by a fellow named Domingo, this Special Edition of HCN features the Vocal Ensemble. Look for future editions of Hawk Cast "Music" coming
```

Segmentazione delle frasi

- *Tool usato: Apache OpenNLP*
- *Riconoscimento delle pagine in 4 lingue: inglese, olandese, tedesco e portoghese*

Filtraggio frasi sporche

- *Analisi della lunghezza delle frasi (prese in considerazione solo le frasi lunghe da 3 a 40 parole)*
- *Gran parte delle frasi sporche sono state già eliminate al momento della pulizia delle pagine*

Applicazione di TAG sui valori significativi

Gli elementi significativi sono stati riconosciuti attraverso l'uso di espressioni regolari

TAG considerati

- #NUM (numeri)
- #ORD (numeri ordinali)
- #DIST (distanze)
- #URL
- #MONEY
- #PERC (percentuali)
- #MAIL
- #DATE

Applicazione di TAG sui valori significativi

00196 Fee for any one session $75 Note: Workshops fill early.

00196 Fee for any one session #MONEY Note: Workshops fill early.

00173 All rights reserved This document was last modified 29 Feb 2008

00173 All rights reserved This document was last modified #DATE

Analisi con Hadoop - MapReduce

- Tool usato: *Apache Hadoop 2.6*
- Analisi degli output delle tre tabelle di output ottenute

Analisi con Hadoop - MapReduce

- *URL più puntati all'interno dell'intero archivio Warc*
- *Media del numero di URL per ogni pagina HTML*
- *Numero di tag all'interno dell'archivio*

Analisi con Hadoop – MapReduce: risultati

1

```
avg 2.342072185311546
```

2

```
http://bama.sbc.edu.    812
http://www.jambase.com/Fans/Free    140
 www.ccmattress.com 140
 www.flickr.com 139
 www.cartoonnetwork.2.0 98
http://www.youtube.com/watch    86
http://www.mediafire.com/   82
http://arkansas.indymedia.org   73
http://backend.deviantart.com/embed/view.swf    71
 www.nrcan.gc.ca    67
 www.dadabhagwan.org    64
 www.tudiscoverkit.cl   63
 www.kamamalar.blogspot.com 63
http://drama.yale.edu/  60
http://www.emseg.de/icon/fractal    59
 www.tudiscovirikid.com 58
http://www2.shoutmix.com/   57
 www.pornohud.com   55
 www.beastserver.com    55
 www.barebacked.com 55
 www.ministeriodotrabalho.gov.br    53
 www.discoberkids.com   52
 www.comedi.ro  50
 www.trvmelike.tr   47
 www.finalistlogon.com  46
http://www.mocasting.com/main/wp    44
 www.tamilnatham.com    42
 www.pornohub.com   40
http://www.hackingcough.com/cgi 38
 www.disneilatino.com   37
```

3

```
#DATE    6927
#DIST    879
#MAIL    5262
#MONEY   14853
#NUM     228848
#ORD     7721
#PERC    6587
#URL     13043
TOTAL_KEY    284120
```