In questa fase del mio lavoro mi sono soffermato sullo studio delle parole chiave che appaiono nell’output del mio tokenizer. Il file sul quale ho concentrato la mia analisi è il seguente: **ManyUrls**

<https://docs.google.com/spreadsheets/d/1Ci2vQf2bFicYPd4F6DvDvki8qljtFV9FdQbgQmcjUZw/edit#gid=612594098>

composto da  7292 url.

***Prima fase di lavoro***

Come prima operazione ho scritto del codice che mi producesse un file con tutte le informazioni che mi interessavano. L’output del codice è riportato nel file **testManySites** <https://docs.google.com/spreadsheets/d/1Yc52zAa2Ad5BuTztSmFewlgb2AeBkbwJb8f3OU5PXIk/edit#gid=1885290236>

Sono riportate nel file colonne del tipo: Sito - Sitemap - Google YES/NO - eventuali errori nel codice.

Studiando l’output **1854** siti tra quelli esaminati hanno almeno una sitemap. Ecco l’elenco dei siti:

<https://docs.google.com/spreadsheets/d/1Yc52zAa2Ad5BuTztSmFewlgb2AeBkbwJb8f3OU5PXIk/edit#gid=544069085>

**810** presentano un eventuale errore nel codice. Nel seguente file sono riportati sito e relativo errore.

<https://docs.google.com/spreadsheets/d/1Yc52zAa2Ad5BuTztSmFewlgb2AeBkbwJb8f3OU5PXIk/edit#gid=1841824200>

Ho delle colonne del genere:

|  |  |
| --- | --- |
| <https://1210wpht.radio.com/> | java.io.IOException: Server returned HTTP response code: 403 for URL: https://1210wpht.radio.com/sitemap.xml |
| <http://all-comic.com/> | crawlercommons.sitemaps.UnknownFormatException: Error parsing XML for: http://all-comic.com/sitemap.xml |
| <http://arabadonline.com/> | java.io.FileNotFoundException: http://arabadonline.com/sitemap.php |
| <http://bethlehem.thelehighvalleypress.com/> | org.apache.http.conn.HttpHostConnectException: Connect to bethlehem.thelehighvalleypress.com:80 [bethlehem.thelehighvalleypress.com/209.50.150.237] failed: Connection timed out: connect |

**4628** sono i siti senza sm.

Ecco l’elenco per un’eventuale verifica.

<https://docs.google.com/spreadsheets/d/1Yc52zAa2Ad5BuTztSmFewlgb2AeBkbwJb8f3OU5PXIk/edit#gid=2026396448>

***Seconda fase di lavoro***

Come seconda fase del mio lavoro ho lanciato il codice del tokenizer che mi ha prodotto un file csv con colonne di questo tipo:

|  |  |  |
| --- | --- | --- |
| **KeyWord** | **Occurences** | **TotalSites** |
| **sitemap** | 186527 | 1627 |
| **shop** | 83756 | 4 |
| **post** | 14304 | 290 |
| **section** | 2577 | 38 |
| **tag** | 880 | 85 |
| **video** | 836 | 174 |

Link al file completo: <https://docs.google.com/spreadsheets/d/1Yc52zAa2Ad5BuTztSmFewlgb2AeBkbwJb8f3OU5PXIk/edit#gid=1282692560>

Come si vede dalla tabella alcuni termini sono colorati in modo differente. Questo perché ho considerato alcuni termini da cui posso ricavare informazioni importanti.

Ho identificato 8 macro sezioni di termini da cui posso ricavare informazioni utili:

Legenda:

1. **Author**

I termini che compaiono in possibili sitemap contenenti liste di autori (con i relativi articoli) o info biografiche, sono:

* Author
* Authors

Link di esempio:

* https://24sunnews.com/sitemap-authors.xml
* https://blog.freightview.com/sitemap-authors.xml
* https://www.orlandopinstripedpost.com/sitemaps/authors

1. **Tag**

I termini che compaiono in possibili sitemap contenenti liste di tag (con i relativi link ad articoli attinenti al tag) sono:

* Tag
* Tags

Link di esempio:

* <http://www.pleasurehorse.com/taxonomy_ngg_tag.xml>
* <https://bhamnow.com/post_tag-sitemap1.xml>
* <https://bhamnow.com/job_listing_tag-sitemap.xml>
* https://bhamnow.com/post\_tag-sitemap7.xml

1. **Video**

I termini che compaiono in possibili sitemap contenenti liste di video (con relativi link e informazioni ai video) sono:

* Video
* Videos
* ShowVideo
* googleVideoSitemap
* googleVideos
* allVideos
* altre parole attinenti ma contenenti sempre il termine **video**

Link di esempio:

* <https://3dprint.com/sitemap-posttype-video.xml>
* <https://1889mag.com/video-sitemap-1.xml>
* https://www.lonestarball.com/sitemaps/sitemaps/video/1
* https://www.buildseries.com/custom-sitemaps/sitemap-videos-2017.xml
* https://www.christies.com/AllVideos.xml
* https://www.yogajournal.com/sitemaps/ContentVideo/2017.xml

1. **Gallery**

I termini che compaiono in possibili sitemap contenenti liste di articoli e  relativi link alle immagini dell’articolo, sono:

* gallery
* galleries
* Photogallery
* image
* images
* picture
* altre parole attinenti contenenti sempre i termini **image,gallery,picture**

Link di esempio:

* http://damewine.com/image-sitemap-1.xml
* https://www.broadwayworld.com/sitemaps/gallery350000.xml
* http://southernlitreview.com/image-sitemap-1.xml
* <https://www.itnews.com/seo/sitemap/https/images/index.xml>
* <https://www.wonderwall.com/rss/sitemap_googleimages_number_1.xml>
* https://www.housebeautiful.com/sitemap\_galleries\_1.xml.gz/

1. **Section**

I termini che compaiono in possibili sitemap contenenti liste di sezioni (con relativi link ad articoli attinenti) sono:

* section
* subsections
* category
* altre parole attinenti contenenti i termini category,categories,section

Link di esempio:

* <http://www.virginialiving.com/sitemap_section_0.xml>
* https://24sunnews.com/sitemap-tax-category.xml

1. **Download**

I termini che compaiono in possibili sitemap contenenti liste di download (con relativi link) sono:

* download
* downloads

Link di esempio:

* <https://www.jwtintelligence.com/download_tag-sitemap.xml>
* https://www.komando.com/sitemap-posttype-downloads.201805.xml

1. **Post**

I termini che compaiono in possibili sitemap contenenti liste di post(con relativi link) sono:

* post
* altre parole contenenti il termine post

Link di esempio:

* https://internationalliving.com/post-sitemap1.xml
* https://24sunnews.com/sitemap-pt-post-2018-10.xml
* http://1049maxcountry.com/sitemap-pt-post-2018-10.xml

1. **Products**

I termini che compaiono in possibili sitemap contenenti liste di prodotti sono:

* shop
* shopping
* products
* sitemapproducts
* altre parole contenenti i termini shop e product

Link di esempio:

* <https://allaboutshooting.com/sitemap_products_1.xml?from=3875072903&to=554225827873>
* <https://www.agri-pulse.com/ext/resources/sitemaps/sitemap_products.xml.gz>
* https://www.smartneighbor.com/sitemap\_products\_1.xml?from=8387782151&to=1787442888802
* http://thefreshexchange.com/shop-sitemap.xml

**Tabella di riepilogo**

|  |  |  |
| --- | --- | --- |
| **KEYWORD** | **NUMERO DI SITI DIVERSI** | **Percentuale sul totale siti con sitemap** |
| **Author** | 138 | 7,44% |
| **Tag** | 96 | 5,17% |
| **Video** | 227 | 12,24% |
| **Gallery** | 194 | 10,46% |
| **Section** | 224 | 12,08% |
| **Download** | 5 | 0.27% |
| **Post** | 297 | 16% |
| **Products** | 31 | 1,6% |

Naturalmente queste percentuali sono basate su un campione di 1854 siti con almeno una sitemap presente ma il tutto è relativo perchè su siti dove sono presenti download la percentuale di trovare sitemap di download è molto maggiore rispetto a trovare sitemap di download su siti di prodotti. Questo vale in generale per tutto. La probabilità di successo è condizionata naturalmente dal campo di interesse dei siti che andiamo a considerare.

\* E’ disponibile l’elenco dei siti diversi dove questi termini appaiono per poter fare una verifica più precisa.

***Terza fase di lavoro***

Una volta studiati quante volte appaiono questi termini occorre capire come poterli catturare per poi andarli ad analizzare. Secondo il mio studio sono facilmente catturabili tramite espressioni regolari perché compaiono sempre gli stessi termini.

Ad esempio per autore basta fare una cattura nel filename del termine author. Per video una cattura del termine video e così via.

Ho quindi dotato la mia libreria ClassificatoreSitemap di altri metodi che consentono la cattura di sitemap di questo tipo. Il codice completo si trova nella repo su git, qui riporto solamente uno spezzone:

/\*

\* @return true if @param url is a sitemap of a possible       list of author.

\*/

public boolean isAuthorSitemap(String url) {

url = url.toLowerCase();

String regex = ".\*author.\*";

if (url.matches(regex))

return true;

return false;

}

Adesso il mio classificatore è in grado di catturare una sitemap di autori.

**Come si presentano i link riconducibili agli articoli degli autori?**

Facendo uno studio su più siti ho notato che il formato in cui appaiono i seguenti link è sempre simile, ovvero:

<https://sito_in_questione/author/nomeAutore>

Esempio:

* Autore: rajai --> Link ai suoi articoli: <https://24sunnews.com/author/rajai/>
* Autore: samantha --> Link ai suoi articoli: https://blog.freightview.com/author/samantha/
* Autore: cwilliams --> Link ai suoi articoli: https://blog.freightview.com/author/cwilliams/
* Autore: aldon --> Link ai suoi articoli: https://blog.freightview.com/author/aldon/
* Autore: nphillips --> Link ai suoi articoli: https://blog.freightview.com/author/nphillips/
* Autore: edorsey --> Link ai suoi articoli: https://blog.freightview.com/author/edorsey/

etc.

Sfruttando questo tipo di sintassi posso facilmente accedere al nome dell’autore e ai suoi articoli nel seguente modo: (Un esempio di output è stampato sopra)

ClassificatoreSitemap cs = new ClassificatoreSitemap("http://www.solarwakeup.com/");

ArrayList<String> sitemap = (ArrayList<String>) cs.getNoIndexSitemaps();

for (String s : sitemap)

if (cs.isAuthorSitemap(s)) {

System.out.println(s);

SiteMap sm = (SiteMap) cs.getSiteMapType(new URL(s));

Collection<SiteMapURL> links = sm.getSiteMapUrls();

for (SiteMapURL smu : links) {

URL authorURLsitemap = smu.getUrl();

String filename = smu.getUrl().getFile();

String[] parts = filename.split("/");

String authorName = parts[parts.length-1];

System.out.println("Autore: "+authorName+" --> Link ai suoi articoli: "+authorURLsitemap.toString());

}

}

Faccio uno split in base al char ‘/’ e prendo l’ultima stringa disponibile che coincide con il nome dell’autore.

Ho ripetuto questo studio per tutti i termini chiave cercando di tirar fuori delle regole generali. Ad esempio per quanto riguarda il tag si presenta nello stesso modo di author ovvero:

<http://sito/tag/trading>