DATA: 27/10/2018

Nella tabella Studio sitemap sono stati presi in esame 100 url contenuti nel file url-100.csv. Dallo studio risulta che sui 100 siti presi in considerazione 46 siti hanno almeno una sitemap che segue lo schema google tra le loro sitemap.

N.B non sono tutte le sitemap per ogni sito(una volta trovata quella google su alcuni siti per l’elevato numero di sitemap andavo avanti senza riportarle nello schema).

Inoltre sui 100 url presi in questione in 47 non troviamo sitemap. Le cause sono:

* Mancanza di robots.txt. (5 casi)

|  |
| --- |
| <https://www.georgianewsday.com/> |
| http://dailycaller.com/ |
| http://www.945theanswer.com/ |
| http://www.cape-coral-daily-breeze.com/ |
| http://www.westkentuckystar.com/ |
|  |

* Non accessibili dall’europa (3 casi).

|  |
| --- |
| <http://www.nydailynews.com/> |
| <http://www.carrollcountytimes.com/> |
| https://www.ourmidland.com/ |

* Access denied (1 caso).

<http://www.khou.com/>

* Presenza di robots.txt ma mancanza di sitemap. (38 casi).

|  |
| --- |
| <http://www.espn.com/> |
| <http://www.wkow.com/> |
| <http://www.wokv.com/> |
| <http://mynorthwest.com/> |
| <http://www.bostonherald.com/> |
| <http://www.journalgazette.net/> |
| <http://www.arcamax.com/> |
| <http://www.vindy.com/> |
| https://wnyt.com/ |
| <http://www.kwwl.com/> |
| http://www.northcountrypublicradio.org/ |
| https://www.nebraska.tv/ |
| http://www.boston.com/ |
| https://siouxlandnews.com/ |
| http://www.ibtimes.com/ |
| https://www.wpr.org/ |
| http://fox56.com/ |
| https://www.10tv.com/ |
| http://brainerddispatch.com/ |
| https://www.capegazette.com/ |
| https://phys.org/ |
| http://www.bendbulletin.com/ |
| http://www.zerohedge.com/ |
| http://www.contactmusic.com/ |
| http://www.wqow.com/ |
| https://www.mlssoccer.com/ |
| http://www.pinkvilla.com/ |
| <http://www.chattanoogan.com/> |
| https://www.thedailybeast.com/ |
| http://mynewsla.com/ |
| https://www.mediapost.com/ |
| http://www.pireport.org/ |
| http://www.advertisernewsnorth.com/ |
| http://www.pamplinmedia.com/ |
| http://www.advertisernewssouth.com/ |
| http://www.newsoracle.com/ |
| http://wamc.org/ |
| http://www.sunjournal.com/ |

Quindi considerando che siamo partiti da una base di 100 url e di questi 100 in 47 non accediamo a sitemap deduciamo che 53 siti hanno a disposizione sitemap. Inoltre in questi 53 siti 46 hanno almeno una sitemap che segue lo schema di google. Quindi l’**86,8%** dei siti che hanno sitemap ne presenta almeno una google.

Sul campione iniziale invece il **46%** dei siti ha almeno una sitemap Google.

Nel **47%** dei siti non troviamo sitemap. Quando manca sitemap il **10,6%** delle volte è dovuto a mancanza di robots.txt, nel **6,39%** delle volte non riusciamo ad accedere per via delle restrizioni europee, nel **2,1%** delle volte abbiamo l’accesso negato e nel **80%** dei casi c’è il robots.txt ma non c’è nessun link a sitemap.

**Ne consegue che se dei siti hanno sitemap il protocollo google è oramai molto diffuso e adottato.**

**Fase di studio sulle sm Google.**

|  |  |
| --- | --- |
| **URL Sito da esaminare** | **Sitemap** |
| <https://www.delaware1059.com/> | <https://www.delaware1059.com/tncms/sitemap/news.xml> |
| <https://www.starherald.com/> | <https://www.starherald.com/tncms/sitemap/news.xml> |
| <https://www.myrtlebeachonline.com/> | <https://www.myrtlebeachonline.com/sitemap/googlenews/story.xml> |
| <https://nypost.com/> | <https://nypost.com/news-sitemap.xml> |
| <http://triblive.com/> | <https://triblive.com/newssitemap.xml> |
| <https://www.channel3000.com/> | <https://www.channel3000.com/content-syndication-portlet/feedServlet?obfType=GOOGLE_NEWS_SITEMAPS&siteId=400001> |
| <https://www.news8000.com/> | <http://www.news8000.com/content-syndication-portlet/feedServlet?obfType=GOOGLE_NEWS_SITEMAPS&siteId=400002> |
| <https://www.si.com/> | <https://www.si.com/news-sitemap/sitemap_1.xml> |
| <https://abc7news.com/> | <https://abc7news.com/sitemap/news.xml> |
| <https://abc7news.com/> | <https://abc7news.com/sitemap/news/2017.xml> |
| <https://abc7news.com/> | <https://abc7news.com/sitemap/news/2016.xml> |
| <https://abc7news.com/> | <https://abc7news.com/sitemap/news/2015.xml> |
| <https://abc7news.com/> | <https://abc7news.com/sitemap/news/2014.xml> |
| <https://abc7news.com/> | <https://abc7news.com/sitemap/news/2013.xml> |
| <https://www.idahopress.com/> | <https://www.idahopress.com/tncms/sitemap/news.xml> |
| <https://edition.cnn.com/> | <https://www.cnn.com/sitemaps/cnn/news.xml> |
| <https://edition.cnn.com/> | <https://edition.cnn.com/sitemaps/news.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/news-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/news-nfl-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/news-nba-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/news-mlb-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/news-nhl-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/news-boxing-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/news-tennis-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/news-mma-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/news-fantasy-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/news-general-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/news-gametracker-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/video-afl-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/video-boxing-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/video-college-basketball-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/video-college-football-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/video-cricket-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/video-crossfit-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/video-general-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/video-golf-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/video-mlb-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/video-mma-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/video-nba-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/video-nfl-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/video-nhl-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/video-nrl-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/video-nascar-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/video-olympics-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/video-pro-wrestling-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/video-soccer-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/video-tennis-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/video-wwe-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/video-baseball-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/video-basketball-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/video-football-sitemap.xml> |
| <https://www.cbssports.com/> | <https://www.cbssports.com/video-hockey-sitemap.xml> |
| <http://wsls.com/> | <https://www.wsls.com/content-syndication-portlet/feedServlet?obfType=GOOGLE_NEWS_SITEMAPS&siteId=800011> |
| https://www.huffingtonpost.com/ | <http://www.huffingtonpost.com.s3.amazonaws.com/sitemaps/sitemap-google-news.xml> |
| https://www.huffingtonpost.com/ | <https://www.huffingtonpost.com/sitemaps/sitemap-google-video.xml> |
| http://www.philly.com/ | <http://www.philly.com/sitemapgoogle/latest/> |
| http://www.philly.com/ | <http://www.philly.com/sitemapgoogle/partners/> |
| http://www.philly.com/ | <http://www.philly.com/sitemapgoogle/columnists/> |
| http://www.philly.com/ | <http://www.philly.com/sitemapgoogle/projects/> |
| http://www.philly.com/ | <http://www.philly.com/sitemapgoogle/video/> |
| http://www.philly.com/ | <http://www.philly.com/sitemapgoogle/wires/> |
| https://www.themonitor.com/ | <https://www.themonitor.com/tncms/sitemap/news.xml> |
| https://www.hollywoodreporter.com/ | <https://www.hollywoodreporter.com/googlenews.xml> |
| http://www.nbcsports.com/ | <https://www.nbcsports.com/sitemap_video.xml> |
| http://www.nbcsports.com/ | <https://www.nbcsports.com/bayarea/sitemap_video.xml> |
| http://www.nbcsports.com/ | <https://www.nbcsports.com/boston/sitemap_video.xml> |
| http://www.chicoer.com/ | <https://www.chicoer.com/sitemap.xml?yyyy=2018&mm=10&dd=16> |
| http://www.kmov.com/ | <https://www.kmov.com/tncms/sitemap/news.xml> |
| http://www.masslive.com/ | <https://www.masslive.com/sitemap-blog.xml> |
| http://www.masslive.com/ | <https://www.masslive.com/sitemap-secure-blog.xml> |
| http://www.masslive.com/ | <https://www.masslive.com/sitemap.xml> |
| http://www.masslive.com/ | <https://www.masslive.com/sitemap-secure-www.xml> |
| http://www.heraldtribune.com/ | <http://www.heraldtribune.com/section/google-news-sitemap> |
| http://www.heraldtribune.com/ | <http://www.heraldtribune.com/section/google-recent-sitemap> |
| http://www.heraldtribune.com/ | <http://www.heraldtribune.com/section/google-sitemap-index> |
| http://www.heraldtribune.com/ | <http://www.heraldtribune.com/section/googlesitemap_rss_index> |
| http://www.heraldtribune.com/ | <http://www.heraldtribune.com/section/googlevideositemap> |
| http://www.rttnews.com/ | <http://www.rttnews.com/GoogleNews.xml> |
| https://www.today.com/ | <https://www.today.com/sitemap/today/sitemap-news> |
| https://www.today.com/ | <https://www.today.com/sitemap/today/sitemap-2018-10-video.xml> |
| https://www.theepochtimes.com/ | <https://www.theepochtimes.com/assets/uploads/sitemap/sitemap_news.xml.gz> |
| https://www.theepochtimes.com/ | <https://www.theepochtimes.com/assets/uploads/sitemap/sitemap_news_m.xml.gz> |
| https://www.mymotherlode.com/ | <https://www.mymotherlode.com/post_google_news.xml> |
| http://www.washingtonexaminer.com/ | <https://www.washingtonexaminer.com/news-sitemap.xml> |
| http://www.goal.com/en-US/ | <https://www.goal.com/en-us/google-news/page> |
| http://www.thevalleydispatch.com/ | <http://www.thevalleydispatch.com/news-sitemap.xml> |
| http://www.theindychannel.com/ | <http://www.theindychannel.com/content-syndication-portlet/feedServlet?obfType=GOOGLE_NEWS_SITEMAPS&siteId=100004> |
| https://www.charlotteobserver.com/ | <https://www.charlotteobserver.com/sitemap/googlenews/story.xml> |
| http://www.cbr.com/ | <https://www.cbr.com/post_google_news.xml> |
| https://www.aspendailynews.com/ | <https://www.aspendailynews.com/tncms/sitemap/news.xml> |
| http://deadline.com/ | <https://deadline.com/news-sitemap.xml> |
| http://www.montgomerynews.com/ | <http://www.montgomerynews.com/tncms/sitemap/news.xml> |
| http://www.santafenewmexican.com/ | <http://www.santafenewmexican.com/tncms/sitemap/news.xml> |
| http://www.cjonline.com/ | <https://www.cjonline.com/section/google-news-sitemap> |
| http://www.phillytrib.com/ | <http://www.phillytrib.com/tncms/sitemap/news.xml> |
| http://www.uticaod.com/ | <http://www.uticaod.com/section/google-news-sitemap> |
| https://www.clickondetroit.com/ | <https://www.clickondetroit.com/content-syndication-portlet/feedServlet?obfType=GOOGLE_NEWS_SITEMAPS&siteId=800002> |
| http://www.raycomnewsnetwork.com/ | <http://www.raycomnewsnetwork.com/newssitemap.xml.gz> |
| http://www.hotnewhiphop.com/ | <https://www.hotnewhiphop.com/sitemap/sitemap_news.xml> |
| http://www.hotnewhiphop.com/ | <https://www.hotnewhiphop.com/sitemap/sitemap_news_100.xml> |
| https://www.keyt.com/ | <http://www.keyt.com/content-syndication-portlet/feedServlet?obfType=GOOGLE_NEWS_SITEMAPS&siteId=200005> |
| https://www.ktvz.com/ | <http://www.ktvz.com/content-syndication-portlet/feedServlet?obfType=GOOGLE_NEWS_SITEMAPS&siteId=200001> |
| http://www.channel4000.com/ | <http://www.channel4000.com/content-syndication-portlet/feedServlet?obfType=GOOGLE_NEWS_SITEMAPS&siteId=700001> |
| <https://www.cnet.com/> | <https://www.cnet.com/sitemaps/news.xml> |

ANALISI DEI TERMINI RICORRENTI NELLE URL SITEMAP GOOGLE NEWS

Questi sono dei link alle sm di Google. Andiamo ad esaminare i termini ricorrenti:

(tralascio quelle per i video che seguono anche lo schema google)

* **/sitemap/news**
* **googlenews**
* **news-sitemap**
* **newssitemap**
* **GOOGLE\_NEWS\_SITEMAPS**
* **sitemaps/news**
* **sitemaps/.\*/news**
* **news-.\*-sitemap**
* **sitemap-google-news**
* **sitemapgoogle**
* **google-.\***
* **google-news**
* **sitemap\_news**
* **googlesitemap**
* **STUDIO SUI TERMINI PIÙ FREQUENTI NELLE INDEX SITEMAP.**

Esaminando una lista di index sitemap di questo tipo:

|  |
| --- |
| <https://www.delaware1059.com/sitemap.xml> |
| <https://www.starherald.com/sitemap.xml> |
| <https://www.myrtlebeachonline.com/sitemap/archive/story.xml> |
| <https://nypost.com/sitemap-nypost-section.xml> |
| <https://nypost.com/sitemap-nypost-post_tag.xml> |
| <https://nypost.com/sitemap-nypost-authors.xml> |
| <https://nypost.com/sitemap-nypost-pages.xml> |
| <http://www.dailyjournal.net/sitemap.xml> |
| <https://www.channel3000.com/sitemap.xml> |
| <http://www.news8000.com/sitemap.xml> |
| <https://www.si.com/article-sitemap.xml> |
| <https://www.si.com/gallery-sitemap.xml> |

attraverso l’uso del tokenizer siamo riusciti a ricavare i termini principali. Quasi sempre infatti un archivio termina con sitemap.xml. I termini più frequenti sono questi:

|  |  |  |
| --- | --- | --- |
| sitemap | 49 | 39 |
| index | 9 | 9 |
| nypost | 4 | 1 |
| archive | 3 | 2 |
| sitemapindex | 3 | 3 |

e sono riportati nel foglio output indexSM tokenizer.

Queste sono invece alcune occorrenze dei termini che compaiono nelle sitemap google news. Come vediamo il termine news domina subito dopo quello sitemap. Inoltre sono presenti molte occorrenze della parola google

|  |  |
| --- | --- |
| **KeyWord** | **Occurences** |
| sitemap | 85 |
| news | 54 |
| video | 30 |
| google | 18 |
| sitemaps | 13 |
| tncms | 9 |
| portlet | 8 |
| obftype | 8 |
| content | 8 |
| syndication | 8 |
| feedservlet | 8 |
| siteid | 8 |
| section | 7 |
| sitemapgoogle | 6 |
| googlenews | 4 |