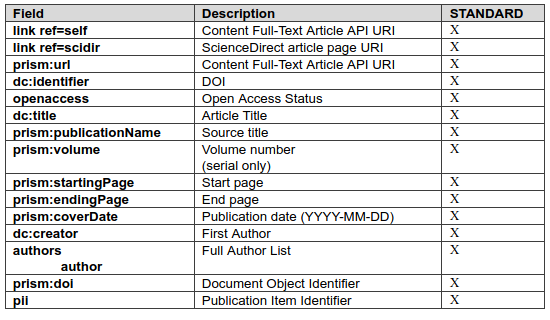
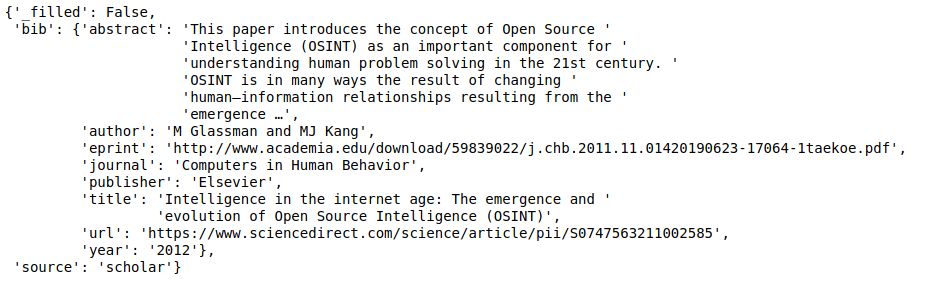
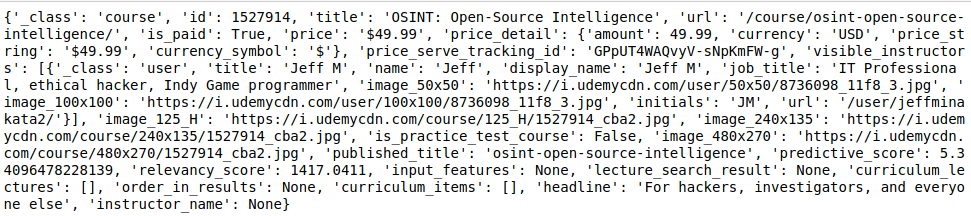
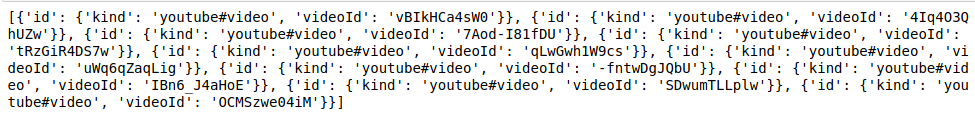
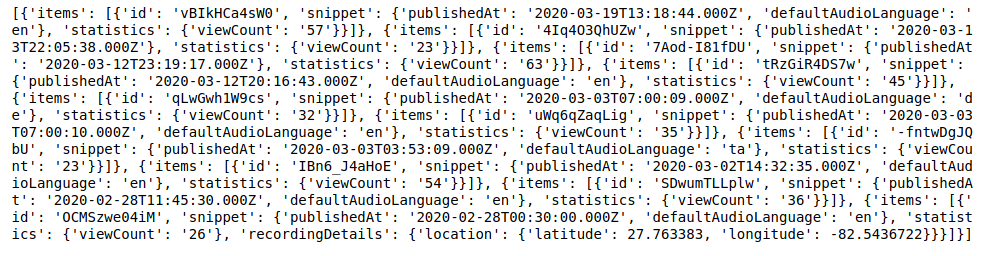
DOCUMENTACIÓN DE LA EXPLORACIÓN DE RECURSOS SOBRE OSINT APLICANDO APIs, WEB SCRAPING Y INTERFACES WEB

1. ARXIV
   * ACERCA DE: Es un servicio de distribución gratuito y un archivo de acceso abierto para 1,673,408 artículos académicos en los campos de física, matemáticas, ciencias de la computación, biología cuantitativa, finanzas cuantitativas, estadística, ingeniería eléctrica y ciencia de sistemas y economía1.
   * API: <http://export.arxiv.org/api>
   * SEGURIDAD: Ninguna
   * FORMATO: Todos los recursos obtenidos por medio de consultas sobre la url estan en formato XML con los siguientes campos
     + id : Código del recurso
     + updated : Fecha de actualización
     + published : Fecha de publicación
     + title : titulo del recurso
     + summary: resumen del recurso
     + author : autores
   * DEBILIDADES: Cuenta con pocos recursos sobre OSINT
   * OPORTUNIDADES: Puedo usarse en la busqueda de otros recursos.
   1. SCIENCEDIRECT
   * ACERCA DE: es un sitio web que proporciona acceso por suscripción a una gran base de datos de investigación científica y médica. Alberga más de 12 millones de contenidos de 3.500 revistas académicas y 34.000 libros electrónicos2.
   * API:<https://api.elsevier.com/>
   * SEGURIDAD: llave suministrada
   * FORMATO: Todos los recursos obtenidos por medio de consultas sobre la url estan en formato XML con los siguientes campos

Tomado de [2]

* + DEBILIDADES: No contiene resumen o palabras claves, el texto completo se enlaza mediante una url lo que dificulta el acceso a terminos ligados a OSINT
  + OPORTUNIDADES: Contiene una gran cantidad de articulos sobre el tema.

1. GOOGLE SCHOLAR
   * ACERCA DE: Google Académico es un buscador de Google enfocado y especializado en la búsqueda de contenido y bibliografía científico-académica
   * API: Ninguna, se debe realizar web scrapping con pyscholar
   * SEGURIDAD: Ninguna
   * FORMATO: Objeto procesado por pyscholar como diccionario
   * DEBILIDADES: Luego de una serie de consultas parece haber un bloqueo de google que impide realizar más peticiones
   * OPORTUNIDADES: Contiene diversa información tanto de libros como articulos, ademas tiene bastante información asociada a OSINT
   * FORTALEZAS: Es uno de los repositorios con más información sobre OSINT.
2. UDEMY
   * ACERCA DE: Es una plataforma de aprendizaje en línea. Está dirigido para adultos profesionales. A diferencia de los programas académicos MOOC conducidos por tradicionales cursos de trabajo creados por el colegio, Udemy utiliza contenido de creadores en línea para vender y así conseguir ganancias.
   * API: <https://www.udemy.com/api-2.0/>
   * SEGURIDAD: Basic Auth, secret y id suministrados por UDEMY
   * FORMATO: La respueta se obtiene en formato Json siguiente el siguiente formato
   * DEBILIDADES: Pocos recursos o resultados
   * OPORTUNIDADES: Puede usarse en la busqueda de otros recursos
   * FORTALEZAS: Es uno de los repositorios con más cursos sobre el tema
3. YouTube API v3
   * ACERCA DE: Es un sitio web para compartir videos. Presenta una variedad de clips de películas, programas de televisión y vídeos musicales, así como contenidos amateurs como videoblogs y YouTube Gaming
   * API: <https://developers.google.com/youtube/v3/>
   * SEGURIDAD: Llave suministrada mediante los proyectos de GOOGLE
   * FORMATO: Respuesta en formato JSON, se aplica filtrado de columnas para reducir uso de unidades para permanecer en lo posible dentro de la cuota, las siguientes imagenes hacen referencia a dos estilos diferentes de búsqueda de la API, búsqueda del listado de videos y búsqueda de los detalles de cada video



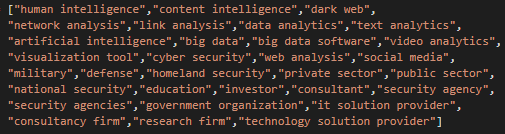
* + DEBILIDADES: La API requiere de llaves cuyas cuotas son limitadas y por tanto también el número de resultados, existen campos que existen únicamente en algunos videos o no fueron especificados. Hay que esperar un día por llave para volver a consultar una vez terminada la cuota
  + OPORTUNIDADES: Con una Key con mayor cuota se podrían obtener más resultados

1. GOOGLE SEACH WEB SCRAPING: Consiste en un script en Python 3 diseñado para recorrer los resultados de google página por página y extraer:
   * Tags utilizados (Aquí se determina si existen recursos tales como imágenes o videos)
   * Tipo de aplicación (ej: PDF)
   * Definir si es Blog ()
   * Definir si es Wiki
   * Agente

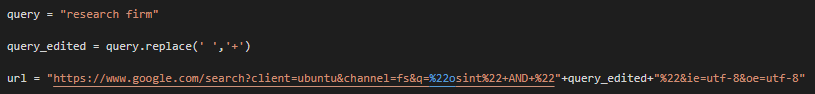
Google tiene bloqueado los robots por defecto por lo que se debe usar un agente en este caso “{"User-Agent": "Mozilla/5.0 (X11; Ubuntu; Linux x86\_64; rv:75.0) Gecko/20100101 Firefox/75.0"}”, incluso en este caso luego de un número determinado de páginas Google determina que es un robot por lo que se deben usar otras estrategias como las VPN.

* + **Librerías**
    - BeautifulSoup: Utilizada para lectura de textos en formato HTML a modo de objeto.
    - Requests: Utilizada para realizar peticiones al motor de búsqueda de Google
    - OS: Manejo de rutas de directorios
    - Json: Escritura y lectura de archivos en formato JSON
    - Sys: Manejo de errores controlados
  + **Términos:**

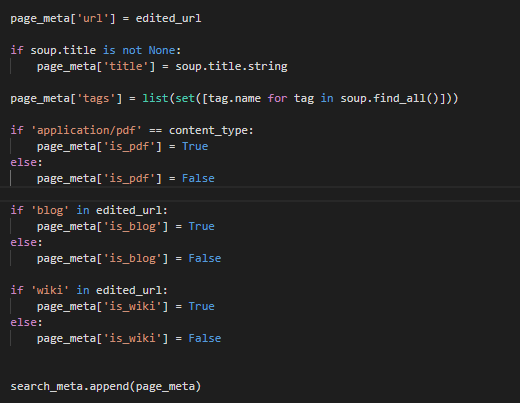
El tipo de búsqueda utilizado es “Termino1” AND “Termino2” donde las comillas obligan al buscador de google buscar la palabra o termino exacto y el AND define que ambos términos deben estar presentes en los resultados, en el caso de este proyecto el término 1 siempre es la palabra **OSINT** mientras que el segundo itera sobre términos predefinidos los cueles pueden observarse a continuación:



* + **Búsqueda**



* + **Extracción de metadata**



* + **Algoritmo**

1. url\_recovered = []
2. print("\033[0;33m-----------------------------------------------NEW WORD-----------------------------------------------")
3. print("URL:"+url)
4. print("WORD: "+query,'\033[m')
5. nextPage = True
6. search\_meta = []
7. count\_page = 0
8. try:
9. while nextPage:
11. count\_page += 1
12. #WE PREPARE THE AGENT TO AVOID FORBIDDEN ON GOOGLE SEARCH
14. response = requests.get(url, agent)
15. if 'Our systems have detected unusual traffic from your computer network' in response.text:
16. print('\033[1;31mSe han dado cuenta --- Somos robots !!!!!!!!!!!!!!! en termino -> '+query,'\033[m')
17. break
18. soup = BeautifulSoup(response.text, "html.parser")
19. print("\033[0;32mGetting results from page:"+str(count\_page)+" - Term:"+query,'\033[m')
20. url\_list = []
21. #WE GET ALL URLS ON CURRENT PAGE
22. for item in soup.find\_all('div',class\_="BNeawe vvjwJb AP7Wnd"):
23. aux = BeautifulSoup(str(item.find\_parents('a')[0]), "html.parser")
24. url\_list.append(str(aux.a['href']))
25. #WE GET NEXT-PAGE-BUTTON URL
26. url\_next = soup.find\_all('div',class\_="ZINbbc xpd O9g5cc uUPGi BmP5tf")[0].a['href']
28. if not url\_next:
29. nextPage = False
30. else:
31. #ASSIGN NEXT URL (nextPage)
32. url = "https://www.google.com"+url\_next
33. for link in url\_list:
34. edited\_url = link.replace('/url?q=','')
35. try:
37. if edited\_url not in url\_recovered:
38. url\_recovered.append(edited\_url)
39. response = requests.get(edited\_url, agent)
40. if response.status\_code == 200:
41. content\_type = response.headers['content-type']
42. soup = BeautifulSoup(response.text, "html.parser")
43. page\_meta = {}
44. page\_meta['url'] = edited\_url
45. if soup.title is not None:
46. page\_meta['title'] = soup.title.string
47. page\_meta['tags'] = list(set([tag.name for tag in soup.find\_all()]))
48. if 'application/pdf' == content\_type:
49. page\_meta['is\_pdf'] = True
50. else:
51. page\_meta['is\_pdf'] = False
52. if 'blog' in edited\_url:
53. page\_meta['is\_blog'] = True
54. else:
55. page\_meta['is\_blog'] = False
56. if 'wiki' in edited\_url:
57. page\_meta['is\_wiki'] = True
58. else:
59. page\_meta['is\_wiki'] = False
60. search\_meta.append(page\_meta)
62. except Exception as e:
63. print("\033[1;31mError 2 ---> "+str(e)+"      line -------> "+str(sys.exc\_info()[-1].tb\_lineno)),'\033[m'
64. pass
65. except Exception as e:
66. print("\033[1;31mError 1 ---> "+str(e)+"      line -------> "+str(sys.exc\_info()[-1].tb\_lineno),'\033[m')
67. pass
68. with open('./results/'+query.replace(' ','\_')+'.json', 'w+') as fp:
69. json.dump(search\_meta, fp, sort\_keys=True, indent=4)
70. REPOSITORIOS CONSULTADOS MANUALMENTE: Algunos repositorios por medio de su buscador e interfaz web muestran la metada pertinente para el caso, por lo que no es necesario obtener la API, en otros casos la API no está disponible por lo que la extracción manual es requerida. Los repositorios a continuación se pueden acceder desde la base de datos de la Universidad Distrital Francisco José de Caldas. A continuación, se comparte el enlace de búsqueda y la descripción de los resultados
    * BOOKLICK
      + Enlace : <https://booklick.udistrital.edu.co/busqueda/osint>
      + Descripción: Con resultados (230) pero sin API y ninguna clasificación
    * NEWSBANK:
      + Enlace: <https://infoweb-newsbank-com.bdigital.udistrital.edu.co/>
      + Descripción: Metadata suministrada a nivel de buscador, No se encontró API
        - Metadata, número de recursos por:
          * Año
          * Fuente
          * Nombre de Fuente
          * País
          * Idioma
    * PASALAPAGINA:
      + Enlace: <http://pasalapagina.com.bdigital.udistrital.edu.co:8080/home-udistrital.html>
      + Descripción: Contenido no relacionado con la academia
    * OXFORD UNIVERSITY PRESS:
      + Enlace : <https://academic.oup.com/journals/search-results?page=1&q=osint&fl_SiteID=5567&SearchSourceType=1&allJournals=1>
      + Descripción: 17 Resultados, metadata suministrada a nivel de buscador, no es necesario obtener la API KEY
        - Metadata, número de recursos por:
          * Tipo
          * Nombre de revista
    * TAYLOR AND FRANCIS GROUP:
      + Enlace: <https://www-tandfonline-com.bdigital.udistrital.edu.co/action/doSearch?AllField=osint>
      + Descripción: 288 Resultados, metadata suministrada a nivel de buscador, no se encontró API
        - Metadata, número de recursos por:
          * Tema
          * Nombre de revista
    * SPRINGER NATURE:
      + Enlace: <https://www.nature.com/>
      + Descripción: Sin Resultados
    * SAGE JOURNALS:
      + Enlace: <https://journals.sagepub.com/action/doSearch?filterOption=allJournal&AllField=osint>
      + Descripción: 21 Resultados, metadata suministrada a nivel de buscador, no es necesario obtener la API KEY (Funciona por medio de Crossref)
        - Metadata, número de recursos por:
          * Tipo
          * Tema
    * SAGE KNOWLEDGE:
      + Enlace: <https://sk.sagepub.com/Search/Results?id=home-search-form&role=search>
      + Descripción: 26 Resultados, metadata suministrada a nivel de buscador, no es necesario obtener la API KEY (Funciona por medio de Crossref)
        - Metadata, número de recursos por:
          * Tipo Contenido
          * Tema
    * IEEE EXPLORER:
      + Enlace: <https://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=OSINT>
      + Descripción: Se pidió la KEY de la API pero nunca se entregó, 62 Resultados, metadata suministrada a nivel de buscador, no es necesario obtener la API KEY
        - Metadata, número de recursos por:
          * Título de conferencia
          * Localización (Ciudad, País)
          * Términos/Temas
    * MATHSCINET:
      + Enlace: <https://mathscinet-ams-org.bdigital.udistrital.edu.co/mathscinet/index.html>
      + Descripción: Sin resultados
    * PROQUEST:
      + Enlace: <https://search-proquest-com.bdigital.udistrital.edu.co/results/FD7C315498654E1BPQ/1?accountid=34687>
      + Descripción: 14 Resultados, metadata suministrada a nivel de buscador, no es necesario obtener la API KEY
        - Metadata, número de recursos por:
          * Tipo de fuente
          * Tema
          * Tipo de documento
    * SPRINGER LINK:
      + Enlace: <https://link-springer-com.bdigital.udistrital.edu.co/search?query=osint>
      + Descripción: 380 resultados, metadata suministrada a nivel de buscador, no es necesario obtener la API KEY
        - Metadata, número de recursos por:
          * Tipo de contenido
          * Tema
          * Subdisciplina
          * Idioma
    * SCOPUS:
      + Enlace: <https://www-scopus-com.bdigital.udistrital.edu.co/results/results.uri?sort=plf-f&src=s&st1=osint&nlo=&nlr=&nls=&sid=3c869ea4a5639ef1245ebc31c2781dec&sot=b&sdt=cl&cluster=scopubyr%2c%222020%22%2ct%2c%222019%22%2ct%2c%222018%22%2ct%2c%222017%22%2ct%2c%222016%22%2ct%2c%222015%22%2ct%2c%222014%22%2ct%2c%222013%22%2ct%2c%222012%22%2ct%2c%222011%22%2ct%2c%222010%22%2ct&sl=20&s=TITLE-ABS-KEY%28osint%29&origin=resultslist&zone=leftSideBar&editSaveSearch=&txGid=42a2a8c320105e942b8756caf40952d7>
      + Descripción: 166 Resultados, metadata suministrada a nivel de buscador, no es necesario obtener la API KEY
        - Metadata, número de recursos por:
          * Año
          * Tema
          * Tipo Documento
          * Tipo Fuente
          * Idioma
          * Pais
    * ZBMATH :
      + Enlace: <https://zbmath.org/?q=osint>
      + Descripción: Sin resultados
    * Britannica Academic Edition:
      + Enlace: <https://academic-eb-com.bdigital.udistrital.edu.co/levels/collegiate/search/articles?query=osint>
      + Descripción: Resultados no relacionados
    * THOMSON REUTERS:
      + Enlace: <http://apps.webofknowledge.com.bdigital.udistrital.edu.co:8080/Search.do?product=WOS&SID=5AYet8GJSoyjjgeWezI&search_mode=GeneralSearch&prID=f6af675c-3256-43c6-b5d5-2b22c3923541>
      + Descripción: 32 Resultados, metadata suministrada a nivel de buscador, no es necesario buscar en la API
        - Metadata, número de recursos por:
          * Año
          * Categoría
          * Tipo Documento
    * EBSCO HOST
      + Enlace: <http://web.b.ebscohost.com.bdigital.udistrital.edu.co:8080/ehost/results?vid=2&sid=409a3bf7-b455-4d6b-a0de-dc8da09b3a75%40pdc-v-sessmgr06&bquery=osint&bdata=JmRiPWE5aCZsYW5nPWVzJnR5cGU9MCZzZWFyY2hNb2RlPVN0YW5kYXJkJnNpdGU9ZWhvc3QtbGl2ZQ%3d%3d>
      + Descripción: 189 Resultados, metadata suministrada a nivel de buscador, no es necesario buscar en la API
        - Metadata, número de recursos por:
          * Tipo de fuente
          * Editor
          * Idioma
          * Pais
71. PAGINAS WEBS SOBRE MOOCS EXPLORADAS
    * The Open University : <http://search.open.ac.uk/public/search/results?q=osint>, Sin resultados
    * Iversity: <https://iversity.org/>, Sin resultados
    * Alison: <https://bookboon.com/blog/2013/12/moocs-list-8-best-massive-open-online-courses-worldwide/>, Resultados no relacionados
    * Open learning: <https://www.openlearning.com/>, Sin resultados.
    * Coursera: <https://es.coursera.org/search?query=osint&>, Resultados no relacionados, Solo 1 resultado buscando por medio del Motor de Google
    * edX: <https://www.edx.org/course?search_query=osint>, Resultados no relacionados
    * Udemy: Se utilizó una API, 83 Resulados
      + Metadata, recursos por:
        - Pagos y no Pagos
        - Costo
        - Titulo
    * SANS: <https://www.sans.org/webcasts/archive/2016>, Tiene Resultados pero no entrega metadata alguna, y se debe verificar uno por uno, tampoco tiene API
72. OTROS REPOSITORIOS:
    * + <https://oyc.yale.edu/> : Sin resultados
      + <https://www.cmu.edu/> : Solo un resultado relacionado
      + <https://www.jisc.ac.uk/> : Sin resultados
      + <https://open.umich.edu/> : Propone repositorios en los que no se tiene ningún resultado relacionado o ya fueron explorados (Twitter API)
      + <https://cedec.intef.es/recursos/>: Problemas de acceso a la pagina
      + <https://emtic.educarex.es/proyectocrea>: Resultados no relacionados
      + <https://www.iadb.org/es>: Resultados no relacionados
      + <http://eduteka.icesi.edu.co/> : Resultados no relacionados
      + <https://ocw.mit.edu/index.htm> : Solo 1 resultado
      + <http://ocw.upm.es/>: Sin resultados
      + <http://www.europana.eu/> : El dominio no está en funcionamiento

REFERENCIAS

[1] <https://arxiv.org/>

[2] <https://dev.elsevier.com/>