Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Programa de Ingeniería Electrónica y Telecomunicaciones
ScientoPy v 1.3.3, Installation and User Manual

By: Juan Pablo Ruiz Rosero `jpabloruiz@unicauca.edu.co`

Contents

1 Installation

1. For Windows download and install the Python 3 latest version (for example Python 3.6.5) from: `https://www.python.org/downloads/`.
2. For Debian or Ubuntu run these commands to install Python3:

   ```
   sudo apt-get install python3 python3-tk python3-pip
   ```

3. To use wordCloud in Windows, install Microsoft Visual C++ Redistributable para Visual Studio 2017 according to these instructions: `https://www.scivision.co/python-windows-visual-c++-14-required/`
4. Install the unidecode, numpy, scipy, matplotlib, and wordcloud Python libraries. For Windows, enter in the command line (Windows + R, cmd, and Enter), and run the installation script:

   ```
   python3 -m pip install --user unidecode numpy scipy matplotlib wordcloud
   ```

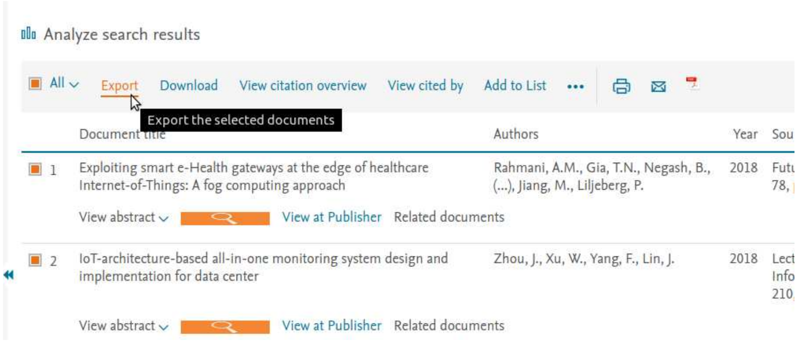
   IMPORTANT NOTE: If you have installed Python 2 and Python 3, run the previous and the following commands with `python3` instead `python`

2 Download the bibliometric dataset

This section describes how to download the proper dataset from Scopus and WoS. Define a search criteria, it will be used for Scopus and WoS. For this guide we are using: "Internet of thing" AND "Gateway"

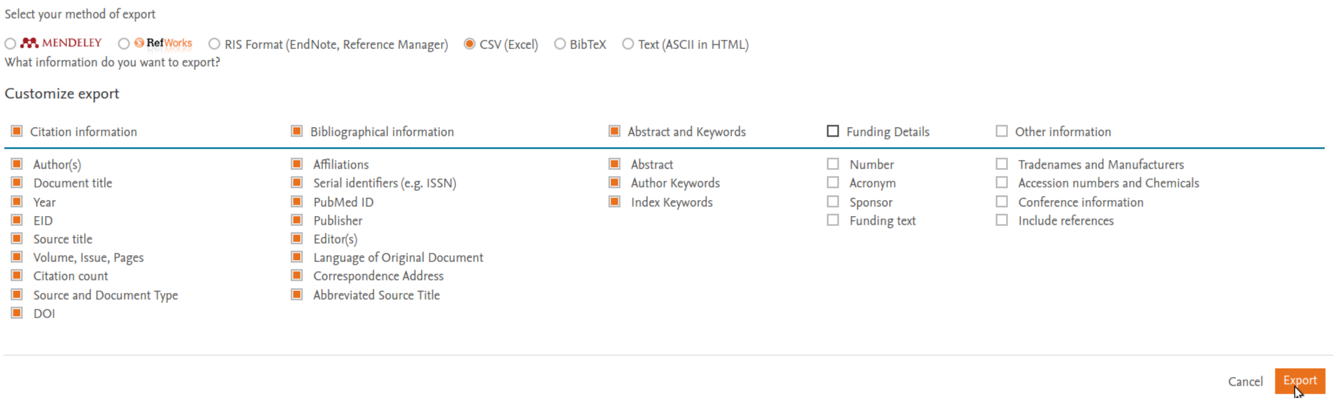
2.1 Download the dataset from Scopus

1. Make your search with the defined search criteria for Article title, Abstract, Keywords.
2. Select all the results and click on Export:



		Document title	Authors	Year	Sou
☑	1	Exploiting smart e-Health gateways at the edge of healthcare Internet-of-Things: A fog computing approach	Rahmani, A.M., Gia, T.N., Negash, B., (...), Jiang, M., Liljeberg, P.	2018	Futu 78,
☑	2	IoT-architecture-based all-in-one monitoring system design and implementation for data center	Zhou, J., Xu, W., Yang, F., Lin, J.	2018	Lect Info 210

3. Select as method of export **CSV (Excel)**, and select the Customize export **Citation information, Bibliographical information, Abstract and Keywords**, then click on Export:



Select your method of export

○ MENDELEY ○ RefWorks ○ RIS Format (EndNote, Reference Manager) ● CSV (Excel) ○ BibTeX ○ Text (ASCII in HTML)

What information do you want to export?

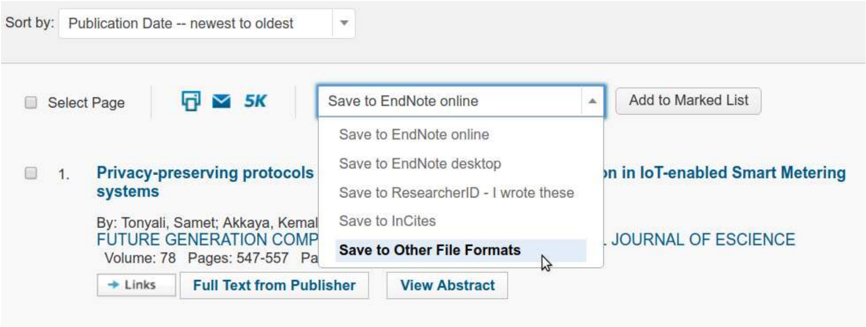
Customize export

☑ Citation information	☑ Bibliographical information	☑ Abstract and Keywords	□ Funding Details	□ Other information
☑ Author(s)	☑ Affiliations	☑ Abstract	□ Number	□ Tradenames and Manufacturers
☑ Document title	☑ Serial identifiers (e.g. ISSN)	☑ Author Keywords	□ Acronym	□ Accession numbers and Chemicals
☑ Year	☑ PubMed ID	☑ Index Keywords	□ Sponsor	□ Conference information
☑ EID	☑ Publisher		□ Funding text	□ Include references
☑ Source title	☑ Editor(s)			
☑ Volume, Issue, Pages	☑ Language of Original Document			
☑ Citation count	☑ Correspondence Address			
☑ Source and Document Type	☑ Abbreviated Source Title			
☑ DOI				

4. Save the file on the folder `/ScientoPy/dataIn`

2.2 Download the dataset from WoS

1. Make your search with the defined search criteria for Topic.
2. Select **Save in Other File Formats**



Sort by: Publication Date -- newest to oldest

□ Select Page

Save to EndNote online

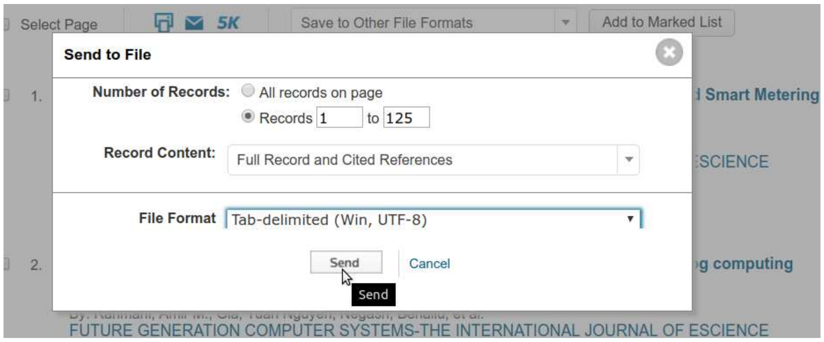
□ 1. Privacy-preserving protocols ...on in IoT-enabled Smart Metering systems

By: Tonyali, Samet; Akkaya, Kemal

FUTURE GENERATION COMP... JOURNAL OF ESCIENCE

Volume: 78 Pages: 547-557 Pa

3. Select the number of records to download, on Record Contented select **Full Record and Cited References**, on File Format select **Tab-delimited (Win, UTF-8)**, and click on Send.



Send to File

Number of Records: ○ All records on page

● Records 1 to 125

Record Content: Full Record and Cited References

File Format Tab-delimited (Win, UTF-8)

FUTURE GENERATION COMPUTER SYSTEMS-THE INTERNATIONAL JOURNAL OF ESCIENCE

4. Save the file on the folder `/ScientoPy/dataIn`

3 Running the ScientoPy scripts

This section describes the ScientoPy scripts to pre-process and process the bibliometric dataset.

3.1 Pre-processing

First we need to pre-process the downloaded dataset. This pre-process joint all the downloaded files from one folder to a single file. Also, this process remove the duplicated files. To pre-process the example dataset (Internet of things papers) run this command inside ScientoPy folder:

```
python3 preProcess.py dataInExample
```

Then, inside the folder `ScientoPy/dataPre` you will find the following files:

- **papersPreprocessed.tsv:** this file contains the information of all papers after the pre-process. This file will be used by the others scripts as the input data.
- **PreprocessedBrief.tsv:** this file briefs the pre-process statics results, such as duplicated papers removed, types of documents, and others.

To find more options of the pre-processing script you can run:

```
python3 preProcess.py -h
```

3.2 Extract the top topics

With this script you can extract the top topics of a selected criterion. The ScientoPy script criterion are described on Table 1:

Table 1: ScientoPy criterion description

Criteron	Description
author	Authors last name and first name initial
sourceTitle	Publication or journal name
subject	Research areas, only from WoS documents
authorKeywords	Author keywords
indexKeywords	Keywords generated by the index, from WoS {Keyword Plus}, and from Scopus {Indexed keywords}
bothKeywords	AuthorKeywords and indexKeywords are used for this search
abstract	Document abstract, for use with pre-defined topics and asterisk wildcard
documentType	Type of document
dataBase	Database where the document was extracted (WoS or Scopus)
country	Country extracted from authors affiliations
institution	Institution extracted from authors affiliations
institutionWithCountry	Institution with country extracted from authors affiliations

For example, to find the top author's keywords you can run this script:

```
python3 scientoPy.py authorKeywords
```

This will generate a list with the top 10 topics on the selected criterion (in this case authorKeywords), with the number of documents per topic, and the h-index associated to each one. Also, this script graphs the evolution of each topic per year, and saves the quantitative results on the folder `ScientoPy/results`.

This script have more options like, save the plot on a file, or increase the number of topic results. For more information you can run:

```
python3 scientoPy.py -h
```

3.3 Analyze pre-defined topics inside a criterion

If you want to make an analysis of pre-defined topics, such as the two selected countries papers evolution, you can use the `scientoPy.py` script, with the option `-t`, to specify the topics:

```
python3 scientoPy.py country -t "United States; Brazil"
```

You can analyze any topic in any criterion. Put the topics on the `-t` argument. Divide the topics with the `;`. Also, you can integrate two or more topics in one, by dividing it with `,`. This is very useful for abbreviations and plural singulars, for example:

```
python3 scientoPy.py authorKeywords -t \
"WSN, Wireless sensor network, Wireless sensor networks; RFID, RADIO FREQUENCY IDENTIFICATION"
```

Note: The command is very long, for that reason the command was divided by "∵ If you have problems in Windows, remove the "änd put the command in one single line.

3.3.1 Asterisk (*) wildcard

You can use the asterisk wildcar to find phrases or words which starts or ends with the letters that you have inserted. For example, if you want to find "device", "devices", and "device integration", enter the following command:

```
python3 scientoPy.py authorKeywords -t "device*"
```

ScientoPy will print the topic found for the previous search using the asterisk wildcard:

```
Topics found for device*:
"devices;device management;Device Interactions;Device objectification;Device;Device integration"
```

You can use this information, to find the statics of each specific topic found, like this:

```
python3 scientoPy.py authorKeywords -t \
"devices;device management;Device Interactions;Device objectification;Device;Device integration"
```

3.3.2 Parametric plot

Also, you can see the results with a parametric graphic (add `--parametirc`). This option plot the accumulative documents, average growth rate (AGR), and h-Index of the selected topic, for example:

```
python3 scientoPy.py authorKeywords -t \
"WSN, Wireless sensor network, Wireless sensor networks; RFID, RADIO FREQUENCY IDENTIFICATION" \
--parametric
```

This script have more options like, save the plot on a file, or others. For more information you can run:

```
python3 scientoPy.py -h
```

3.4 Finding trending topics

This script finds the top trending topics based on the higher average growth rate (AGR) over the others. The AGR is calculated on two years periods, using the following Equation (1):

$$AGR = \frac{\sum\limits_{i=Y_s}^{Y_e} P_i - P_{i-1}}{(Y_e - Y_s) + 1}, \tag{1}$$

where:
AGR = Average growth rate;
Y_s = Start year;
Y_e = End year;
P_i = Number of publications on year i.

To find the top trending topics on author's keywords criterion, you can run the following script:

```
python3 scientoPy.py authorKeywords --trend --windowWidth 2 --startYear 2006 --endYear 2017 \
-l 10 -s 3 --parametric --agrForGraph --graphTitle "Internet of things trending topics"
```

This script will find the top 200 topics, then it calculates the AGR for the last 2 years (–windowWidth 2). Finally, the 200 top topics are sorted from the highest AGR in the last 2 year period to the lower. The first 3 AGR topics are filtered (they correspond to the keyword Internet of thigs), and the next 10 topics are garph in a parametric plot.

3.5 Analysis based on the previous results

ScientoPy generates an output file with all the results documents from the last run script. For example if we run the command:

```
python3 scientoPy.py country -t "Canada" --noPlot
```

ScientoPy will create an output file with all documents that have authors with affiliation in Canada. This output file can be used by ScientoPy to perform an analysis based on the previous results, in that way if we run the following command with the option `-r` or `--previousResults` after the previous one to analyze based on the previous results:

```
python3 scientoPy.py authorKeywords -r --bar --graphTitle "Internet of things in Canada top author keywords"
```

We will obtain the top author keywords from papers where the author affiliation correspond to Canada. Also, we can run the following command to know which are the countries that have more common documents with Canada:

```
python3 scientoPy.py country -r --bar --graphTitle "Internet of things in Canada top colaboration countries"
```

Note: the ScientoPy's documents output file is only generated when the `-r` or `--previousResults` is not used. In that way, if we run many times a ScientoPy command with that option, the documents output file will not overwritten.

3.6 Output files and directories

After run some ScientoPy commands or after run all the commands by executing the `exampleGenerateGraphs.sh` script you will find the following folder and files structure:

```
ScientoPy
├─dataInExample
├─dataPre
│ ├─papersPreprocessed.tsv
│ └─PreprocessedBrief.tsv
├─graphs
├─Manual
└─results
  ├─AuthorKeywords.tsv
  ├─AuthorKeywords_extended.tsv
  └─papersPreprocessed.tsv
```

- **dataInExample:** contains Scopus and WoS example data set for the search criteria "Internet of things" AND "Gateway" downloaded in 27 November 2017. This is the input example for preprocess script.
- **dataPre:** output folder for the preprocess results, and input folder for scientoPy script.
- **papersPreprocessed.tsv:** preprocesed papers data with all input documents integrated, filtered, and duplication removed. This is the input file that scientoPy script uses.
- **PreprocessedBrief.tsv:** preproceses brief table that shows the preprocess results related to total papers found per data base, the omitted papers, the duplicated papers count per data base, and the total number of papers per paper type (Conference paper, article, review...)
- **graphs:** graphs output folder for preprocess and scientoPy scripts
- **Manual:** folder with the pdf manual and example paper with scientoPy commands highlighted used for graph and tables generation.
- **results:** output folder for scientoPy result output files
- **AuthorKeywords.tsv:** scientoPy output file for the selected criterion (in this case authorKeywords) that shows the top topics or the custom topics with the total number of documents, the Average Growth Rate (AGR), the Average Documents per Year (ADY), the h-index, and the documents per each year.
- **AuthorKeywords_extended.tsv:** scientoPy output file for the selected criterion that show the top or custom topics with the documents related to each one.
- **papersPreprocessed.tsv:** inside the results folder, this file contains the output papers from the last scientoPy used script. This is used as an input for scientoPy script when it use the option `-r` or `--previousResults`

4 ScientoPy graph types

ScientoPy has 4 different ways to graph the results described on Table 2.

Below are showed some examples of these graphs types, with the used command.

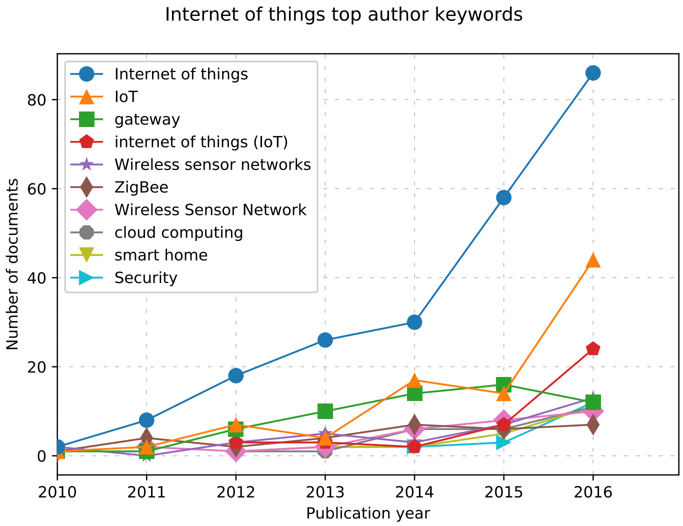
4.1 Time line graph

Command:

```
python3 scientoPy.py authorKeywords --startYear 2010 --endYear 2016 \
--graphTitle "Internet of things top author keywords"
```

Table 2: ScientoPy output graphs types

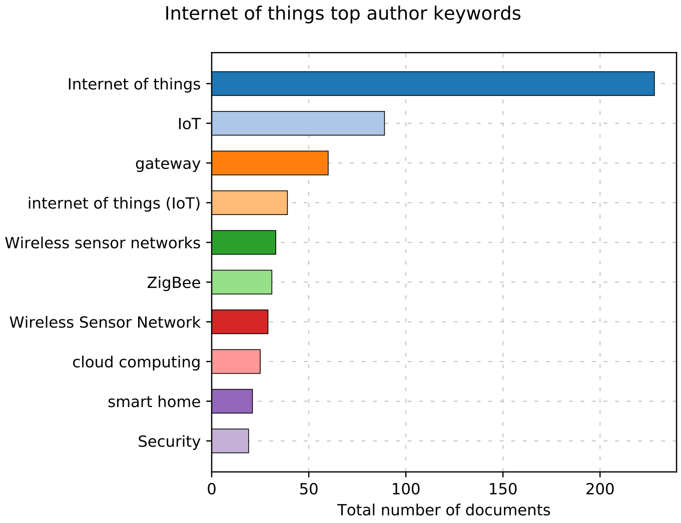
Graph type	Argument	Description
Time line	No argument	Graphs the number of documents of each topic vs the publication year
Horizontal bars	--bar	Graphs the total number of documents of each topic in horizontal bars
Word cloud	--wordCloud	Generate a word cloud based on the topic total number of publications
Parametric	--parametric	Graphs two plots, one with the accumulative number of documents vs the publication year, and other with the average papers per year vs the h-index
Parametric2	--parametric	The total number of documents vs average papers per year



4.2 Horizontal bars graph

Command:

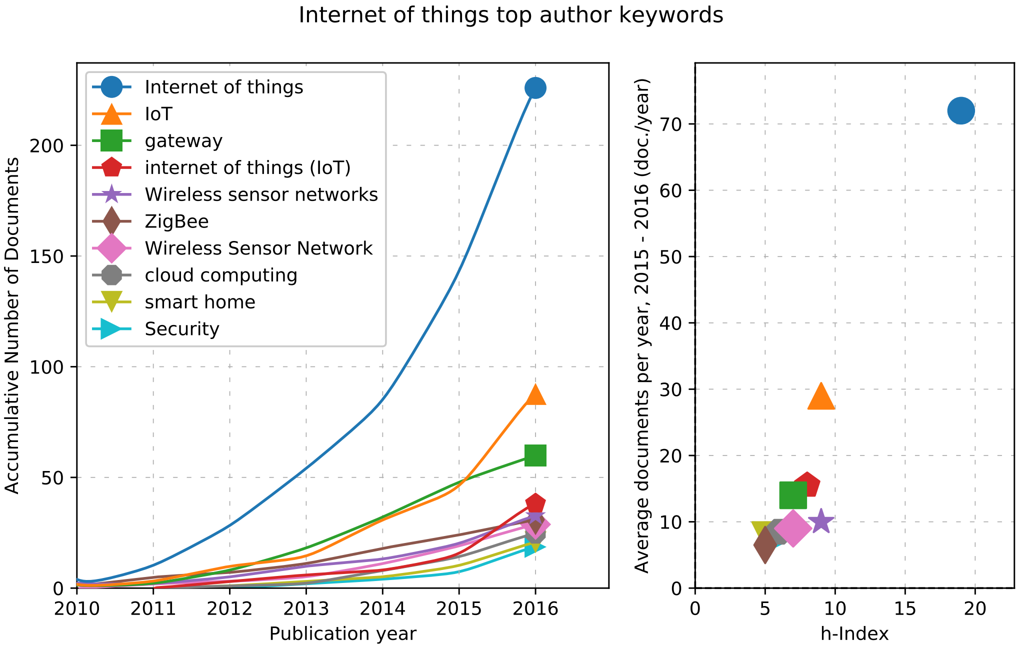
```
python3 scientoPy.py authorKeywords --startYear 2010 --endYear 2016 --bar \
--graphTitle "Internet of things top author keywords"
```



4.3 Parametric graph

Command:

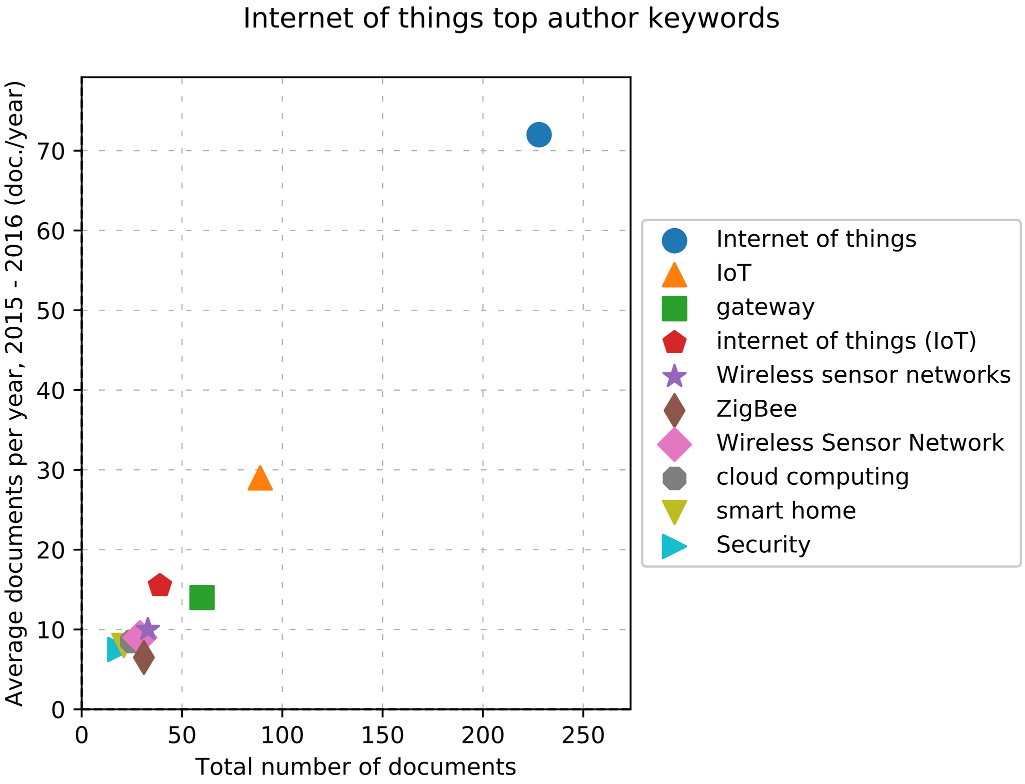
```
python3 scientoPy.py authorKeywords --startYear 2010 --endYear 2016 --parametric \
--graphTitle "Internet of things top author keywords"
```



4.4 Parametric2 graph

Command:

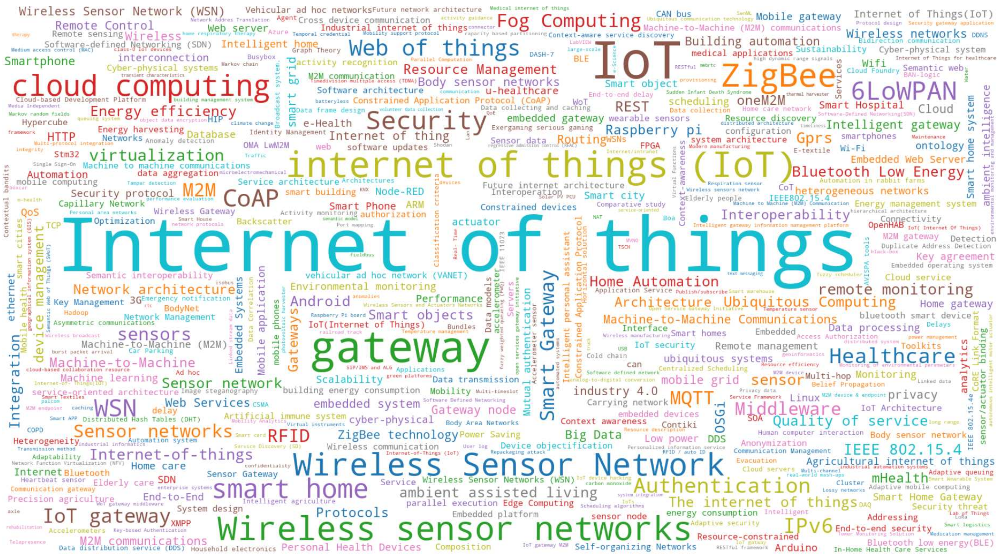
```
python3 scientoPy.py authorKeywords --startYear 2010 --endYear 2016 --parametric2 \
--graphTitle "Internet of things top author keywords"
```



4.5 Word cloud graph

Command:

```
python3 scientoPy.py authorKeywords --startYear 2010 --endYear 2016 -l 500 --wordCloud
```


Internet of things
IoT
ZigBee
internet of things (IoT)
cloud computing
gateway
Wireless Sensor Network
Wireless sensor networks
smart home
Security
Web of things
6LoWPAN
CoAP
M2M
Fog Computing
Resource Management
Energy efficiency
virtualization
Healthcare
Authentication
Middleware
Quality of service
Internet-of-things
Sensor networks
WSN
RFID
MQTT
IPv6
IEEE 802.15.4
Interoperability
Architecture
Ubiquitous Computing
Machine-to-Machine Communications
Home Automation
remote monitoring
Smart Gateway
Bluetooth Low Energy
Raspberry pi
Intelligent gateway
Gprs
IoT gateway
The internet of things
mHealth
Android
Smart objects
sensors
Machine-to-Machine
Network architecture
Sensor network
Web Services
embedded system
Gateway node
ZigBee technology
Big Data
industry 4.0
privacy
sensor
mobile grid
ambient assisted living
Protocols
Remote Control
Wireless Sensor Network (WSN)
Intelligent home
Smartphone
Automation
HTTP
Node-RED
Smart Phone
Building automation
Mobile gateway
Internet of Things(IoT)
Wireless networks
u-healthcare
Body sensor networks
e-Health
Internet of thing
Routing
REST
oneM2M
Smart Hospital
Cloud
Smart city
Integration
device management
Smart homes
Data processing
Personal Health Devices
Arduino