Pubmed Analysis
User Documentation

Table of Contents

Description

Pubmed Analysis program was designed to compare trends and applications of two given search terms in Bioinformatics context.
The program queries Entrez database and generates:

- 6 .png files (8 plots) portraying trends and tendencies for each of the given search terms
- 2 .xlxs files containing keywords related to search terms portraying the areas of application of given terms in Bioinformatics

Prerequsites

- Python v. 2.7+ (Installation Guide)
- Biopython (Installation Guide)
- NumPy, Matplotlib (Installation Guide)
- Cygwin (Windows machines only – Installation Guide)

Usage

The program is designed to be launched from the command line by running **python3 src/pubmed.py** or **python src/pubmed.py** command from the project directory.

All relevant parameters are supposed to be set up in **src/settings.py** file prior to the program's execution.

The following parameters can be set:

Name	Default value	Description
ENTREZ_EMAIL	'example@student.uj.edu.pl'	your email (used to query Entrez)
SEARCH_TERM_1	'deep learning'	first term to query Entrez for
MAX_RETURNED_TERM_1	30	maximum amount of returned records for first term
SEARCH_TERM_2	'artificial intelligence'	second term to query Entrez for
MAX_RETURNED_TERM_2	60	maximum amount of returned records for second term
OUTPUT_PATH	'output'	path to store .png plots and .xlxs keywords files in

Tip: Add **[MeSH Terms]** to the end of the term to have Entrez find aliases for given term and search for those as well - i.e. **'artificial intelligence[MeSH Terms]'**.

NOTE: Due to the nature of PubMed database, some search results may not include data regarding references of articles that are returned as a result of a term search. Displaying crucial graphs and plots will not be available if that is the case, and the program will exit with an error logged to a log file.
It is therefore highly suggested to set max MAX_RETURNED values for both terms to at least **30 records**.

Sample Outputs – Interpreting the Results

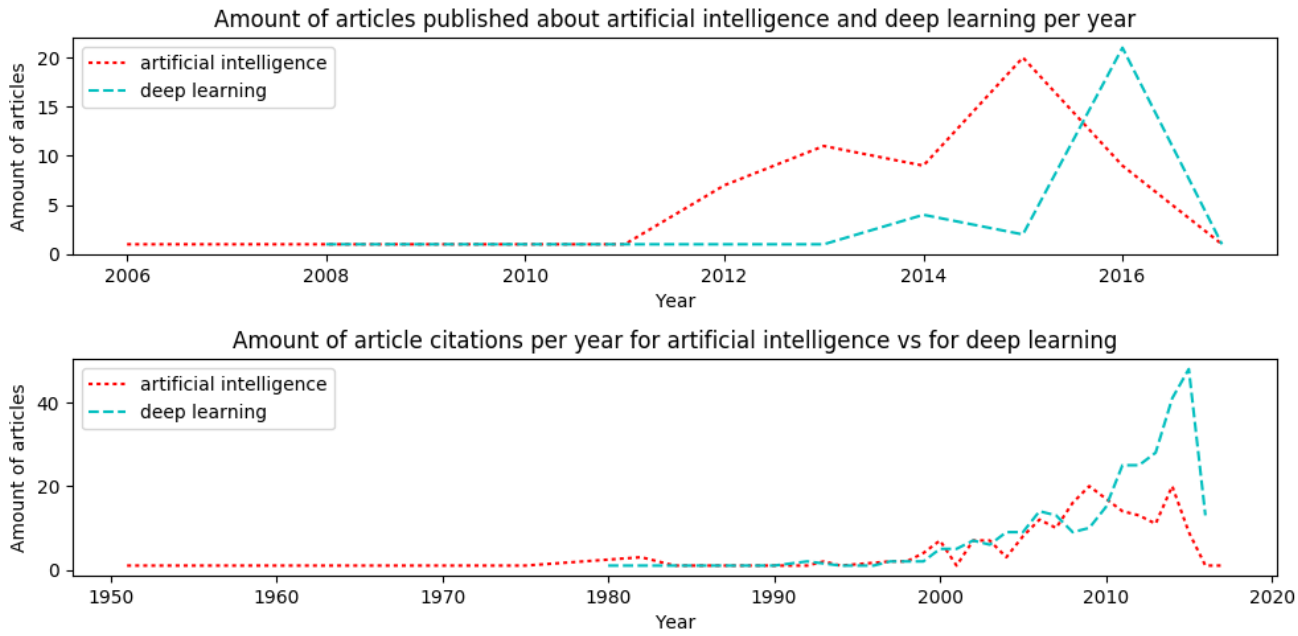
The following plots are created as a result of program execution:

1. Amount of articles published per year (term_1 vs term_2)
2. Amount of article citations per year (term_1 vs term_2)
3. Average age of cited works per year (two plots – one for each term)
4. Percentage of article citations per year (two plots – one for each term)
5. Growth trend for term (two plots – one for each term)

The first graph presents the general idea for the domain and demonstrates a comparison between the popularity of the two terms in question. It becomes apparent, which term has received more attention throughout the years, as well as which one of the two is of greater interest to researchers and scientists at the moment.

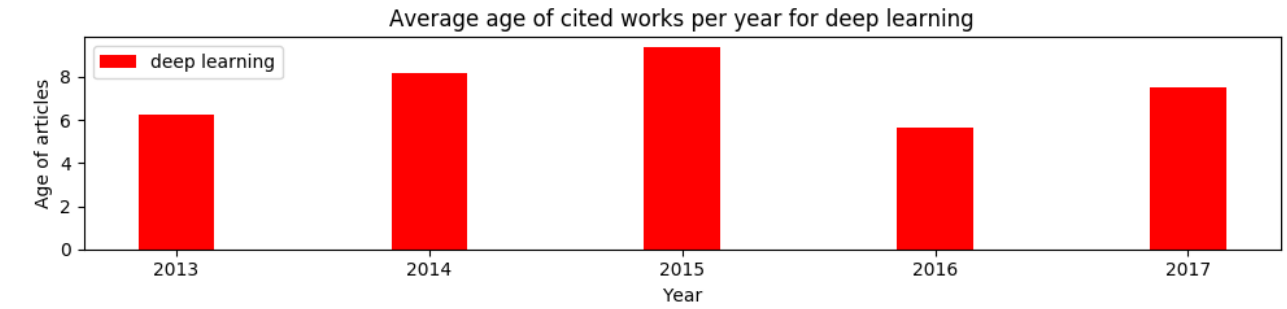
The second graph shows the amount of articles cited each year by the articles from the first graph – this may give us some idea regarding the speed at which the domain is developing.

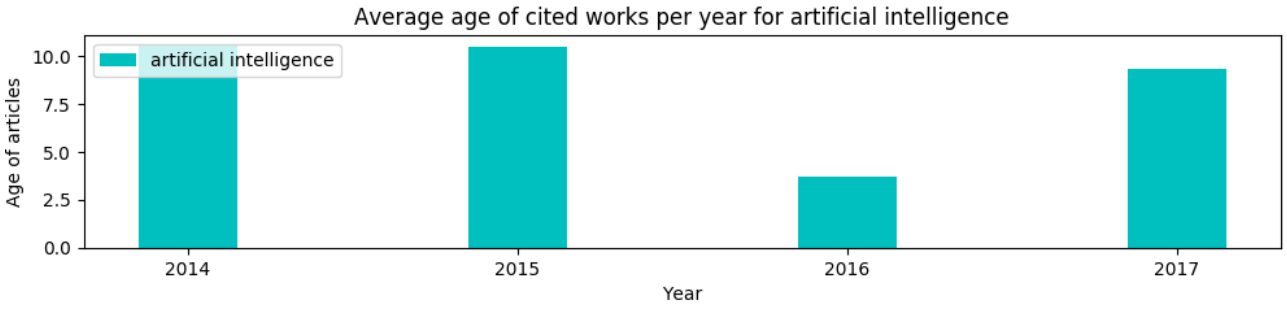
Let us look at the following output graphs for default parameters:



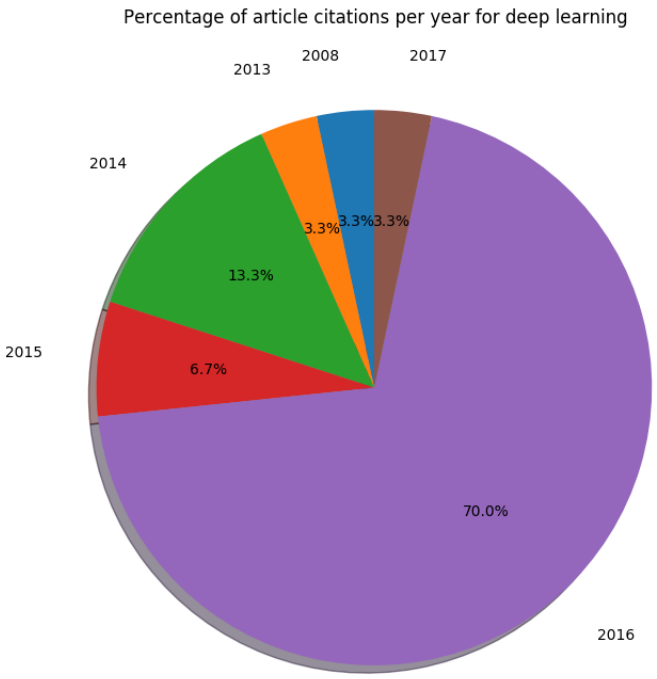
The image presents first and second output graphs.
Based on what is displayed on the graphs, we can note that Artificial Intelligence has been studied for a longer period of time (works from 1950s have been cited by articles related to AI, as opposed to works related to Deep Learning, which cite papers starting from 1980s).
Additionally, we see two other trends: firstly, more works have been written about Deep Learning recently, whereas the number of works written about AI has been dropping since 2015, and secondly, more recent works are being cited by papers connected to Deep Learning as opposed to papers connected to Artificial Intelligence, which suggests that domain is evolving more rapidly.

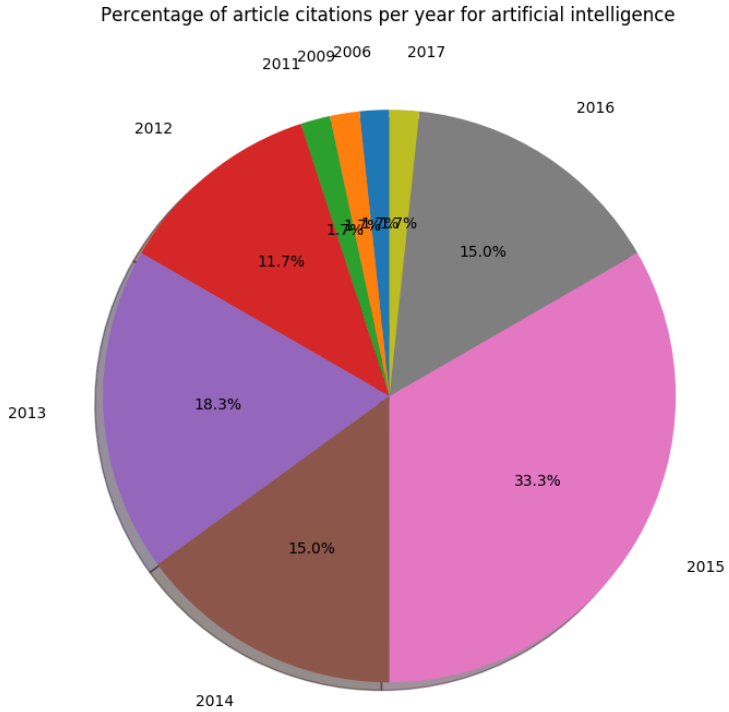
Going forward, the next two plots present the average age of articles referenced for each of the recent years.
Again, based on the below graph we can see that on average papers connected to AI reference papers two years older than the ones referenced by articles connected to Deep Learning.
Since it is the average of all ages, it might be useful to take a look at minimums and maximums for each year – those would give us a better idea regarding the trends of domain development. I can speculate that while more and more newer articles appear regarding both of the terms, the domain for both of them is so old that older works keep being referenced alongside newer ones – which results in rather high average age of cited works for both terms.



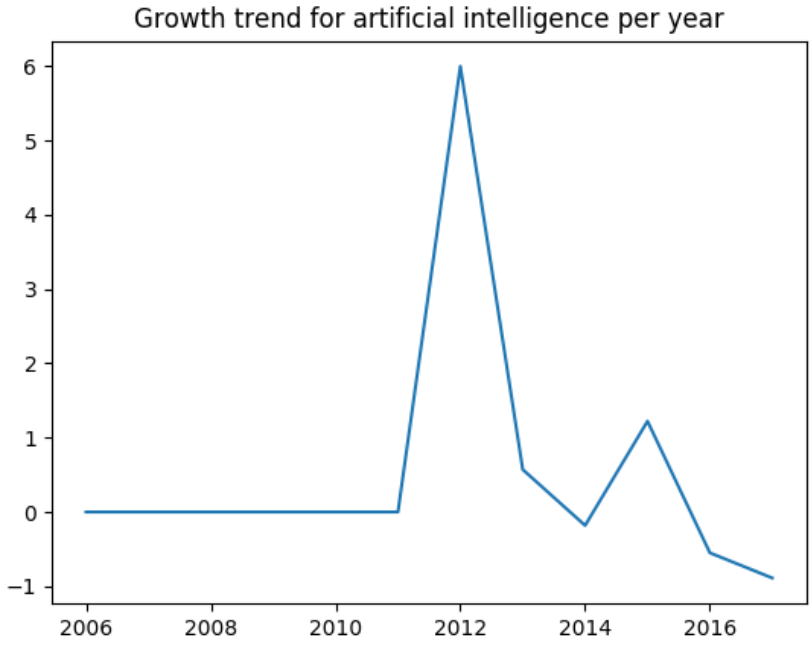


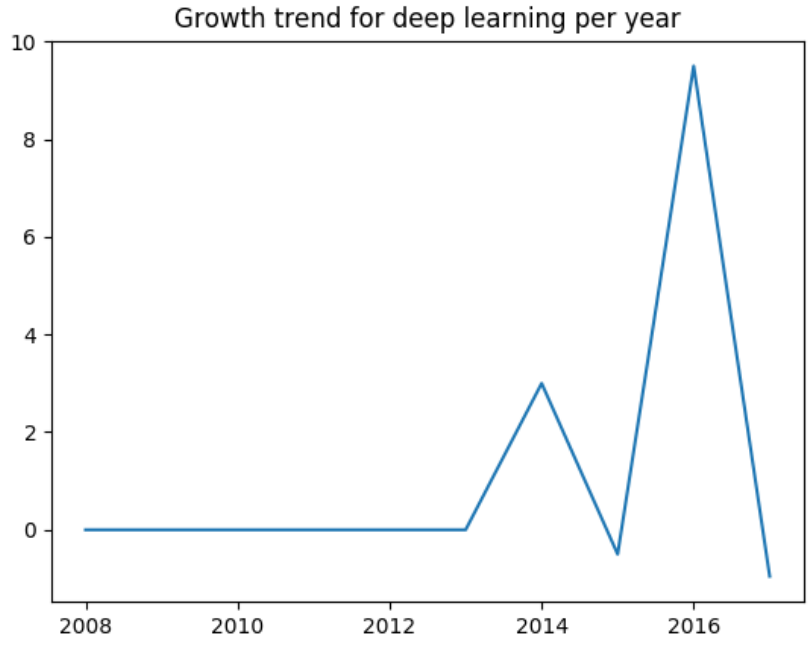
The two graphs that follow demonstrate the percentage of articles created about Deep Learning and Artificial Intelligence for the past couple of years. It is clearly apparent that although AI dominated the field in 2015 and before that, the majority of the articles connected to Deep Learning appeared in 2016. So far all of the graphs, plots and charts indicate that Deep Learning has become a topic of much greater interest in the past year.


Percentage of article citations per year for deep learning
2013
2008
2017
2014
3.3%
3.3%
3.3%
13.3%
2015
6.7%
70.0%
2016


Percentage of article citations per year for artificial intelligence
2011
2009
2006
2017
2016
2012
15.0%
11.7%
18.3%
2013
33.3%
15.0%
2015
2014

The last two graphs are rather minimalistic – they represent the trend of popularity for the terms given. Those graphs are a direct representation of the conclusions regarding the search terms we have made so far.





As we can see, the growth trend for AI has been gradually decreasing in the past couple of years, whereas the growth trend for Deep Learning has peaked in 2016

– the sudden decrease after that is connected to the fact that 2017 has only just begun, and there is insufficient data as of yet to make assumptions about future trends.

Among other outputs are Excel files containing keywords connected to each of the search terms – those may give us some idea regarding the areas of application of given terms.

For instance the first six relevant terms in Excel doc regarding AI are:

crohn's disease
health communication
machine learning
participatory design
synthetic biology
acute aquatic toxicity

Whereas the terms connected to Deep Learning include:

drug discovery
deep approach
problem-based learning
students' approaches to learning (sal)
surface approach
amyotrophic lateral sclerosis

Both of the Excel files contain over 80 entries – those should give a clearer idea of what the usual applications of search terms are.

NOTE: Most accurate results are achieved with a large chunk of data – the bigger the MAX_RETURNED value for each term, the more information the program has to analyze and present particular data. Please, be aware that processing larger data will take more time and will require more RAM.