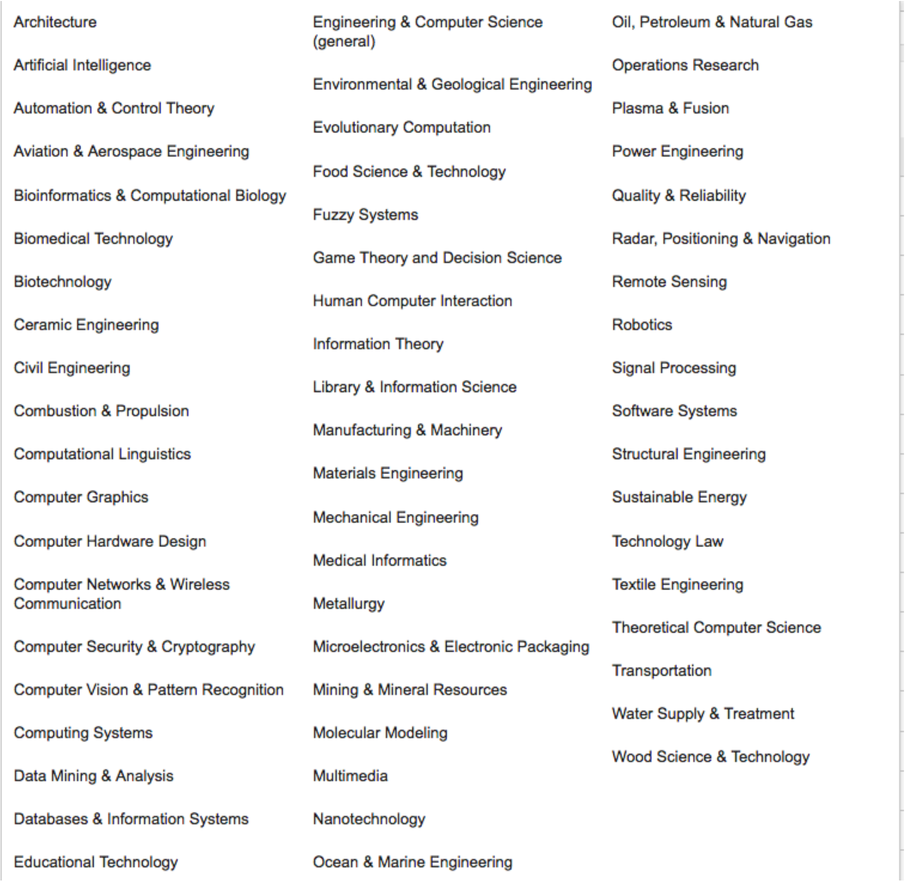
## Project Description (2017)

Objective: create and analyze a research network using Google Scholar API https://github.com/ckreibich/scholar.py

The major limit is the number of allowed daily queries, so you must use multiple accounts or take care not to exceed this limit, by imposing a bound to the number of queries of a given type that you submit. This simply means that you will **not** obtain a full coverage network, but this is expected.

This is your task:
1)In Google Scholar, the following scientific categories are considered for Computer Science:



Architecture

Artificial Intelligence

Automation & Control Theory

Aviation & Aerospace Engineering

Bioinformatics & Computational Biology

Biomedical Technology

Biotechnology

Ceramic Engineering

Civil Engineering

Combustion & Propulsion

Computational Linguistics

Computer Graphics

Computer Hardware Design

Computer Networks & Wireless Communication

Computer Security & Cryptography

Computer Vision & Pattern Recognition

Computing Systems

Data Mining & Analysis

Databases & Information Systems

Educational Technology

Engineering & Computer Science (general)

Environmental & Geological Engineering

Evolutionary Computation

Food Science & Technology

Fuzzy Systems

Game Theory and Decision Science

Human Computer Interaction

Information Theory

Library & Information Science

Manufacturing & Machinery

Materials Engineering

Mechanical Engineering

Medical Informatics

Metallurgy

Microelectronics & Electronic Packaging

Mining & Mineral Resources

Molecular Modeling

Multimedia

Nanotechnology

Ocean & Marine Engineering

Oil, Petroleum & Natural Gas

Operations Research

Plasma & Fusion

Power Engineering

Quality & Reliability

Radar, Positioning & Navigation

Remote Sensing

Robotics

Signal Processing

Software Systems

Structural Engineering

Sustainable Energy

Technology Law

Textile Engineering

Theoretical Computer Science

Transportation

Water Supply & Treatment

Wood Science & Technology

1. You must select **one** among the following categories: Artificial Intelligence, Computer Graphics, Computational Linguistics, Computer Networks and Wireless Communication, Data Mining and Analysis, Evolutionary Computation, Multimedia, Robotics, Computer Security & Cryptography, Computer Vision & Pattern recognition, let's name this category *C1*
2. For *C1*, with 1 query you can select the top 20 venues;

3. For any venue, select 100 articles with h5 index (h index >5). Each "h5" query retrieves 20 articles at a time. In any case, try to select articles with highest number of citations;
4. For any article, extract the authors (store other possible useful metadata, such as the date) and next, generate a query for each author (please note that you have an ambiguity problem here, due to homonymy)
5. For all extracted authors of all articles (let A be this set), obtain their research *keywords* (one query for each author) and their *co-authors* (if they are many, more than one query is required).

After these steps, you have:

a) A tag cloud of keywords associated to the original *C1*
b) A research network with A+K authors (since you also added the co-authors of the initial set A)

Use the generated co-authorship *SN* network to:

- Identify communities
- Identify key players (use any of the algorithms described at lesson) and key separators

**Furthermore**:

1. Select the most cited paper **p**, or (better) a paper that pioneered a new research domain within the considered C1
2. Create the citation network for this paper (this might be possibly quite time consuming in terms of queries, so think of some methodology to limit the queries)
3. Using temporal information on citations, show the influence diffusion network for this research domain during 4-5 years (this means that you should select a not-too-old paper that received many citations in the last 5 years). Hopefully, the nodes of the citation networks (i.e., the researchers who cited the authors in p) should be included in the initial *SN*. These citing nodes become "active" at the time of citation, and the authors of p are the "S" set of initial seeds (see lesson on influence networks). In other terms, on year 1 of publication in the *SN* only the authors of the paper **p** are "infected" with the topic of the paper. On year 2, other nodes (researchers who cited the paper in year 2) become infected, and so on until 2017. You must simply SHOW this viral diffusion of paper **p** on the network, not forecast it (this would be well beyond the scope of a course project)!! But you can observe something, e.g.: to what extent the "virus" (= interest in the topic of the paper) is transmitted trough co-authorship links? Is there any relation between probability of being infected and research keywords of the infected authors? Etc. Comment what you see in your data.

**CREATIVE**: do anything more on these data that you think interesting --and not too complicated. The CREATIVE section MUST be part of the project any how.

HOW TO SUBMIT: You submit both software, documentation, and an (about) 10 pages report to Velardi@di.uniroma1.it and stilo@di.uniroma1,it
You will be evaluated also for the quality of your code.
The project is worth 50% of your final grade, and you can make teams of 2 (NO MORE THAN 2)