BIA660 Final Project

Team 8
members:
Bofei Zhang
Siqi Jiang
Tianzhang Qi
Jianhan Wang

Using H-index in the credibility increasing of the paper

What is H-index?

The h-index is an *author-level metric* that attempts to measure both the *productivity* and *citation impact* of the publications of a scientist or scholar.

The h-index correlates with *obvious success indicators* such as winning the Nobel Prize, being accepted for research fellowships and holding positions at top universities. The index is based on the set of the scientist's most cited papers and the number of citations that they have received in other publications.

The index was suggested in 2005 by J. E. Hirsch, a physicist at UC San Diego, as a tool for determining theoretical physicists' relative quality and is sometimes called the *Hirsch index* or *Hirsch number*.

---Wikipedia

Why H-index?

- *Productivity of the paper*
- *Citation impact of the paper*
- *People with high H-index must have obvious success indicators*
- *More intuitive to people*
- *Suitable for measuring the overall achievement of senior scientists who have been engaged in scientific research for many years.*
- *H-index is attracting more and more attention in the global scientific community, and has become an important reference index or hiring basis for professors in some European and American universities.*

The necessity of using H-index

Since our purpose is to improve the credibility of the paper, we believe that the author's credibility to some extent represents the credibility of the paper. The H-index is a relatively accurate reflection of one's academic achievements. The higher a person's H-index, the more influential and valuable his papers are. Moreover, the H-index can also be used not only to evaluate researchers' academic performance in the past, but also to predict their academic achievements in the future, which makes it more convincing for us to use the H-index to evaluate authors and increase their credibility.

Gather information about the authors

Download the SCORE_csv.csv

SCORE_csv

titile_CR **author_first_CR**
author_last_CR DOI_CR
pub_year_CR ISSN_CR
publication_CR pub_short
paper_id source_WOS … …

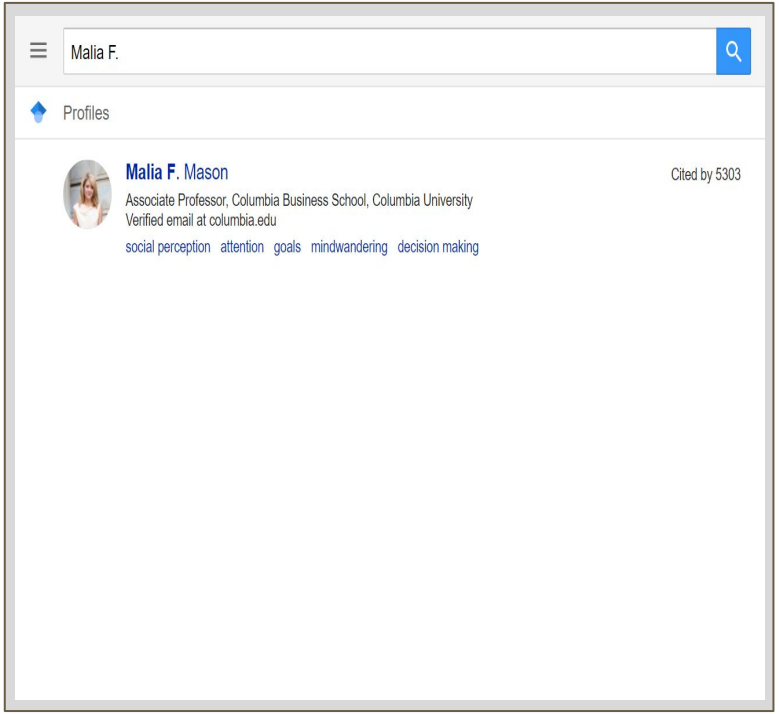
Authors'name_csv

author_first_CR
author_last_CR

Match the author's academic archives

Write a crawler to implement the following steps:

- ➔ Open the “personal academic archive search” page
- ➔ Find the search box
- ➔ Submit the search author name and open the search results page
- ➔ Go to the search author details page
- ➔ Locate the profile link location and save all links to your local computer



Crawling the H-index

Modules: BeautifulSoup, Request, Selenium

- Use Request to download the target pages.
- Use BeautifulSoup to extract the author's profiles and corresponding H-index.
- Use Selenium to realize automatization and crawl all authors' H-indexes.
- Use Pandas to merge the H-indexes we crawled with the DARPA Claims dataset.

Challenges

- It’s hard to collaboratively work due to the current circumstance.
- Not familiar with some web mining tools and libraries.
- Obey the robots.txt file

Conclusion and Future Work

- Enhanced the DARPA SCORE Claims Dataset by determine author's H-index
- Implemented a H-index web crawling system.
- Learned the web crawling techniques
- Improve the current system
- Try other enhancements

Thanks for listening!