Summary of our research: data collection and NLP calculation

1. We decided to use the available html text instead of converting pdfs to texts due to potential inaccuracies. First, we tried to directly get text from the EBSCO database. But it only has access to a limited number of full-text articles.
2. We use SAGE database as our primary source for collecting full-text articles because it includes JM and JMR journals. To collect an article, three steps are required: First, collect the unique DOI for each article. This is done by using the EBSCO database. Second, generate an URL for each article using its DOI sequence, which is the URL we will later collect article from. Third, use Selenium to scrape the URL and collect what we are interested in (i.e., exclude references, etc.).
3. scraping citation data through: 谷粉学术. It is basically a copy of Google Scholar for Chinese researchers because Google is banned in China. It will return the same result as Google Scholar when you search something. We scrape it using Selenium too. There is a trick that it changes the languages used to represent [citation numbers, , etc. ].