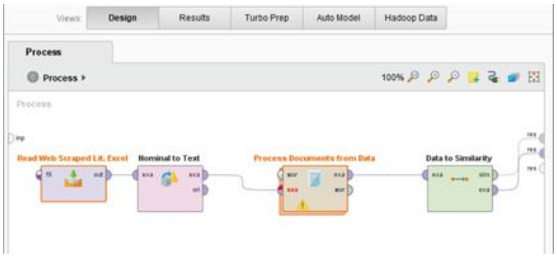

Fig 6(a). Rapidminer process model deriving term similarity from the scraped literature search

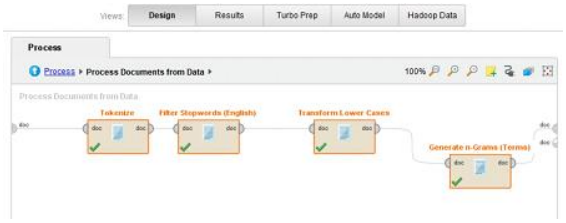

Fig 6(b). text-mining sub-process components (tokenize, filter stop words, transform lower case and generate n-grams).

It could be seen that most regarding the components which can be below GUI-based elements that focus on increasing the simplicity of use as well as time and energy to value through the device.

V. CONCLUSION

So when to scrape websites and when to not and Now the very first thing a scraper have to check when scraping a website is terms of services and the robot.txt file now the robot.txt file is used by websites owners they can specify what website pages that can be scraped or not so if a scraper finds the web page that tries to scrape is forbidden in the robot.txt file In this case scrapers is not allowed to scrape that website.

The second main thing a scraper have to check is does the website have a public API. If yes a scraper have to check if there are some limitations because some websites limit the amount of request. Next a scraper have to check if the API provides all the data it wants now in case of the API is paid it doesn't provide all the data and it has some limitations In this case the only solution you have is to use web scraping on it.

So reasons why scrapers generally use web scraping are data analysis and machine learning so data analysis is like evaluating the data use an analytical and statistical tools now of course in order to do some data analysis a scraper must have large amounts of data or what it is called data sets and the more data a scraper have the more accurate data analysis will be.

In web scraping the other big thing is machine learning now machine learning is like a set of algorithms that allows the computer or the software to be more accurate in predicting outcomes without being explicitly programmed and this also requires huge amounts of data now the more data scraper have the more scraper's system can learn by itself.

VI. FUTURE WORK

Now it comes to the maintenance and the stability of the spider. Now believe it or not the stability of the spider is like a couplet and 100 percent dependent on the website you are going to scrape and that's because if the website changes it's user interface. Scraper will end up with broken XPath expressions and CSS selctors and if scraper first created the spider they didn't use for example javascript to render the content and then after owners of website decided to redesign the website and make it fully dependent on javascript. So in this case scraper can end up with a broken spider from A to Z. so a scraper have to keep itself up to date with web scraping technology to avoid any kind of wastage and a desktop application scrapy is required which can scrape websites in the real time by using that desktop application.

REFERENCES

[1] BuzzSumo.com[Online]Available:https://buzzsumo.com/blog/filtering-the-worlds- content-5-ways-to-stay-ahead/. (Accessed: 28-Sep-2019).

[2] D. Pratiba, A. M.s., A. Dua, G. K. Shanbhag, N. Bhandari, and U. Singh, "Web Scraping And Data Acquisition Using Google Scholar," 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), 2018.

[3] Mooney, S. J., Westreich, D. J., & El-Sayed, A. M. Epidemiology in the era of big data. Epidemiology, 26(3), 390. 2015.

[4] D. D. Prasetya, A. P. Wibawa, and T. Hirashima, "The performance of text similarity algorithms," International Journal of Advances in Intelligent Informatics, vol. 4, no. 1, p. 63, 2018..

[5] V. Bharanipriy and V. Prasad," WEB CONTENT MINING TOOLS: A COMPARATIVE STUDY," International Journal of Information Technology and Knowledge Management, Volume 4, No. 1, pp. 211-215. January-June 2011.

[6] Kushmerick Nicholas; Weld Daniel S.; Doorenbos Robert," Wrapper Induction for Information Extraction," Proceedings of the International Joint Conference on Artificial Intelligence, 1997.

[7] Liu, L.; Pu, C.; Han, W.; , "XWRAP: an XML-enabled wrapper construction system for Web information sources," Data Engineering, 2000. Proceedings. 16th International Conference on , vol., no., pp.611-621, 2000.7.] Zoratti, "MYSQL Security Best Practices," 2006 IET Conference on Crime and Security, London, 2006, pp. 183-198.

[8] Tripathy, A.K.; Joshi, N.; Thomas, S.; Shetty, S.; Thomas, N., "VEDD-a visual wrapper for extraction of data using DOM tree," Communication, Information & Computing Technology (ICCICT), 2012 International Conference on , vol., no., pp.1,6, 19-20 Oct. 2012.

[9] S. M. Al-Ghuribi and S. Alshomrani, "A Comprehensive Survey on Web Content Extraction Algorithms and Techniques," 2013 International Conference on Information Science and Applications (ICISA), 2013.

[10] W. H.gomaa and A. A. Fahmy, "A Survey of Text Similarity Approaches," International Journal of Computer Applications, vol. 68, no. 13, pp. 13–18, 2013.

[11] D. M. Thomas and S. Mathur, "Data Analysis by Web Scraping using Python," 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2019, pp. 450-454, doi: 10.1109/ICECA.2019.8822022.

[12] Glez-Peña, D.; Lourenço, A.; López-Fernández, H.; Reboiro-Jato, M.; Fdez-Riverola, F. Web scraping technologies in an API world. Brief Bioinform. 2013, 15, 788–797.

[13] Berglund, A.; Wg, X.S.L.; Boag, S.; Wg, X.S.L.; Chamberlin, D.; Wg, X.M.L.Q.; Almaden, I.B.M.; Fern, M.F.; Wg, X.M.L.Q.; Kay, M.; et al. XML Path Language (XPath) 2.0; W3C. 2010. Available online: https://www.w3.org/TR/xpath20/ (accessed on 5 December 2019).

[14] Haddaway, R.N. The Use of Web-scraping Software in Searching for Grey Literature. Grey J. 2015, 11, 186–190.

[15] 19. Dwivedi, S. Comprehensive Study of Data Analytics Tools. In Proceedings of the 2016 Symposium on Colossal Data Analysis and Networking (CDAN), Indore, India, 18–19 March 2016.