COMMERCIAL AND FREE(OPEN-SOURCE) SOFTWARE PACKAGES FOR DM

By: Damon Quire (Topic 3)

To discuss open source/free software we must first know exactly what that entails. Open source software is software that can be modified, shared, and used by the general public because it's publicly accessible. So it's software that can be downloaded, dissected, and more by anyone in the public domain. Opensource.com defines open source software as software with source code that can be inspected, modified, and enhanced by the public. I personally stream my video gameplay to interact with the community that I reside in depending on the game. There is a TON of open source software for your stream that allow you to automate a lot of the tedious processes that streaming includes. It also allows for a much more enjoyable experience for the viewers by adding things such as event handlers that spark an animation on the screen when a certain situation occurs. For instance when someone follows my channel a little animation pops up on screen and thanks that individual personally for following me.

So open source software is very useful for many reasons. It helps with very tedious things that only individuals in IT would use, but it's also useful for regular everyday people as well. For instance, Microsoft Firefox is actually an open source web browser. It's free to download, you can add on other applications that run with it in order to make it run differently like ad blockers or password savers, and many other things that allow you to customize your Mozilla Firefox Web Browsing experience to your own liking. You don't have to pay for a license, you don't have to purchase it outright, you simply go to the source, and download it for free.

So now that we have a good feel of what open source software is, what software of that nature is available for Data Mining purposes? Thenewstack.com has a list of the top software for Data Mining that is indeed, open source. The first being Rapidminer. This software is written in Java’s

programming language. The interesting thing about it is that it gives you template based data mining tools. Meaning, you don't actually have to write hardly any code. This tool is said to be useful for things other than the actual data mining, for instance, visualization, predictive analysis, statistical modeling, evaluation, and deployment. This tool looks really interesting to me based on the template based experience. Sounds like they already have the code written and all you do is input your data and it mines it according to what template you choose.

Another example would be Weka, which is again is now a java based data mining tool. The New Stack says that it can handle several data mining tools including data processing, clustering, classification, regression, visualization, and feature selection. It apparently doesn't include sequence modeling which diagrams the flow of logic within your system says Indico. It's also free under General Public License which means users can customize it however they see fit.

R-Programming is another example of open source data mining software that The New Stack lists. It supports data mining but stands out because it allows for graphical techniques to be easily used to visually exhibit the results of the data mine. This allows for users to easily see the averages, among other things without having to just look at a set of relevant data. This makes it much more user friendly for figuring out patterns in the data.

You also can use a product called “Orange”. Orange is written in python. This software appears to be a favorite for many unexperienced data miners or programmers. I visited the website of the creators of Orange and will no briefly go over its capabilities listed on the website. These features include interactive data visualization. This allows you to explore statistical distributions

with plots. It also allows for a more in depth statistical monitoring with use of decision trees, hierarchical clustering, and more. It also incorporates visual programming. Their GUI apparently allows you to focus on the exploration of data instead of the backend coding involved. It allows you to use many to choose from widgets in order to do a lot and display a lot of the work for you. They also include great tools for learning. They have widgets designed to teach the user how to data mine. They have many add-ons that allow you to customize your experience and use other helpful tools not originally programmed into the software. At the bottom of their page it shows that a professor from Baylor, a scientist from France, and another professor from the University of Pavia in Italy all use and highly recommend Orange as and easy to use, easy to learn/teach open source data mining software.

The next software listed on the New Stack is that of KNIME. Which allows for all three phases of data processing: extraction, transformation, and loading. It gives you a GUI which helps create the nodes needed for processing. It is easy to extend and allows you to add many plug ins to make it perform with different wanted tweaks. Due to its components that utilizes machine learning, it has caught they eye of many people within business intelligence and financial heads themselves. Given the screenshot of their graphical user interface, it looks relatively user friendly and allows for customization of your data plots to make it easier to follow by the human eye no doubt.

The last software mentioned in this article is called NLTK or natural language toolkit. Which it appears is one of the leaders for language processing tasks. These include data mining, machine learning, data scripting, and sentiment analysis. They claim to be the leader for building python

programs for human language data. As you said in class text mining can be actually pretty tedious and interesting. They claim to be suitable for many professions including linguists, engineers, students, educators, researchers, and more. They have a book called Natural Language Processing with Python. Which they claim provides an introduction to programming for language processing itself. This will guide people thru python coding, using corpora, throwing texts into categories, and analyzing structures of the text.

After researching all of the open source software for data mining that I have discussed, NLTK sounds the most interesting because I think it would be cool to analyze patterns and categories in texts. It would be really neat to actually dive into texts and find certain aspects of it more than just using everyone's favorite CTRL+F. This could even help you find key points without knowing what those points were to begin with by finding key important words within the texts. I'm sure there are much more in depth reasons for data mining thru text but even what I can come up with sounds interesting. So I'm sure if I dove into what the actual deep capabilities are of this technique it would no doubt surprise me and impress me with how much you can do with this tool.

The easiest to use tool without personally using them I'd say would have to be Rapidminer. Mostly because it is a template based software that already gives you the general aspects that you need without the need for much coding from your part. This would allow me to simply give it my data, tell it what I want by choosing the template and starting the process.

WORKS CITED:

https://thenewstack.io/six-of-the-best-open-source-data-mining-tools/

https://orange.biolab.si/

https://indico.io/blog/sequence-modeling-neuralnets-part1/

http://www.nltk.org/

http://www.cs.waikato.ac.nz/ml/weka/