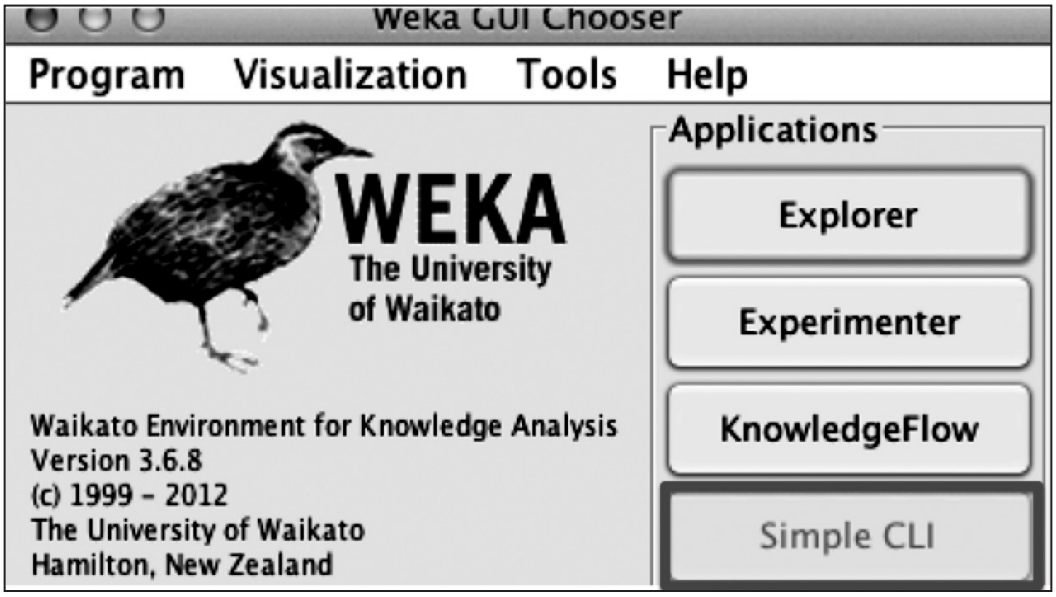
Weka – GUI Way to Learn Machine Learning



In this era of data science where R and Python are ruling the roost, let's take a look at another data science tool called Weka. Weka has been around for quite a while and was developed internally at University of Waikato for research purpose. What makes Weka worthy of try is the easy learning curve.

For someone who hasn't coded for a while, Weka with its GUI provides easiest transition into the world of Data Science. Being written in Java, those with Java experience can call the library into their code as well.

Personally, I had my first shot at Data Science when I took a course at University of Waikato. It was a healthy introduction and gave me a smooth transition into the Data Science. Later when I had to tackle larger problems I moved onto R. So I strongly recommend Weka as a learning tool for those looking to get into the world of data.

Below is the pedagogy of stepwise learning which will help you to understand the concepts in a better & concrete manner:

Step 1: What is Weka and Why to Use It?

According to Wikipedia:,

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

You might want to have a look at this video from Brandon Weinberg. This video will give you considerable insight to this amazing tool. You might not understand everything through this video but will certainly get a hang of things.

Step 2: Setting up Machine

Now, that we are acquainted with Weka, we can proceed to the next stage. To know more about the tool and the people behind its success, you can have a look at this site on Project Weka. Moreover, you can also download the software and get the latest version for your system from this link.

Step 3: Learning the Basics of Weka

The best way of getting started with Weka is using MOOC offered by University of Waikato. Data Mining with Weka is a well reputed course, but it isn't available around the year. Yet, not to worry, in such cases one can access the course videos from this Youtube Channel. The official link of this course can be viewed here. The Data Sets which will be discussed in here can be downloaded from this link. The page has further links to data sets. Weka uses data in ARFF format. In case data is not in ARFF format, you can convert it from CSV to ARFF format by taking help from this video.

Step 4: Data Sets

Having tried our hands at Data sets provided by the course coordinators, we will try our hands on a fresh data set from Kaggle. Since the format would be of .csv, convert it to ARFF

format, so that we can read it onto the Weka interface. After having done these courses, once has attained enough skills to start working and analyzing data sets using Weka GUI . Those who visited the MOOC link would have the seen the course 'More Data mining with Weka'.

Step 5: More Data Mining with Weka

Here, some more advanced features of using the software have been discussed. It builds up the experience from using the previous course, hence it's a prerequisite.

Besides this course you might want to have a look at this YouTube Lecture Series from Rushdi Shams.There are total of 38 lectures. You can skip a few of initial 2-3 lectures if you find the content to be similar to the courses above. This course has been built on various skills which are complementary to those provided by the above series.

There are some interesting discussions happening on Reddit about Weka. It is advisable to go through the mentioned link to gather news on Weka and how it is being used by others. This should give one enough perspective about next possible step after Weka.

Step 6: Weka Command Line

Next Step: As of now we have been relying on using Weka using Weka GUI. As of now both the courses rely on GUI for the purpose , those with experience in JAVA Programming can rely on calling Weka from within JAVA Code. This is useful because when trying or working out with large data sets scripting helps in automating your work. Also, since JAVA is used for Hadoop Framework, Weka can be used for BigData as well. You can read more using Weka in BigData from here.

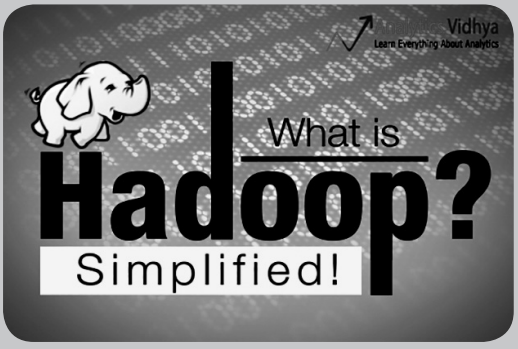
So, those interested in this aspect of Weka can try this lecture series by Dr Noureddin Sadawi. You may like to checkout this Weka API tutorial playlist also. The emphasis is on calling Weka API from within JAVA code, it repeats some of the above concepts but we use Weka using a command line interface.

Step 7: Word2Vec Challenge

Having gained significant insight, we will now have a look at sentiment analysis. There is a small data set with data sets size around 25 MB. So these can be processed using the Weka GUI. For data sets larger than 40 MB, we need to use the command line method. This discussion might be useful.

This path has been contributed by Abhinav Unnam, who interned with us last year. Abhinav is currently undergoing a dual degree course from IIT Roorkee, one of the India's finest Engineering Colleges. He started his machine learning journey through Weka and enjoys participating in several Kaggle competitions today using R and Kaggle.

What is Hadoop?



Scenario 1: Any global bank today has more than 100 Million customers doing billions of transactions every month

Scenario 2: Social network websites or e-Commerce websites track customer behavior on the website and then serve relevant information / product.

Traditional systems find it difficult to cope up with this scale at required pace in cost-efficient manner.

This is where Big data platforms come to help. In this article, we introduce you to the mesmerizing world of Hadoop. Hadoop comes handy when we deal with enormous data. It may not make the process faster, but gives us the capability to use parallel processing capability to handle big data. In short, Hadoop gives us capability to deal with the complexities of high volume, velocity and variety of data (popularly known as 3Vs).

Please note that apart from Hadoop, there are other big data platforms e.g. NoSQL (MongoDB being the most popular), we will take a look at them at a later point.

Introduction to Hadoop:

Hadoop is a complete eco-system of open source projects that provide us the framework to deal with big data. Let's start by brainstorming the possible challenges of dealing with big data (on traditional systems) and then look at the capability of Hadoop solution.

Following are the challenges I can think of in dealing with big data :

1. High capital investment in procuring a server with high processing capacity.

2. Enormous time taken

3. In case of long query, imagine an error happens on the last step. You will waste so much time making these iterations.

4. Difficulty in program query building

Here is how Hadoop solves all of these issues :

1. High capital investment in procuring a server with high processing capacity: Hadoop clusters work on normal commodity hardware and keep multiple copies to ensure reliability of data. A maximum of 4500 machines can be connected together using Hadoop.

2. Enormous time taken: The process is broken down into