**Mining data from Ravelry.com to design the perfect yarn**

**Introduction**

Ravelry is the world’s most popular online network for fibre artists and is colloquially known as ‘the facebook of knitting’. It is used by businesses and individuals worldwide to exhibit and advertise knitting and crocheting products and patterns. Since it was founded in 2007, it has grown to include sizeable databases of a variety of knitting products which can be rated by any of the 6 million users. This proposal refers to the full dataset of yarn products.

For clarity, in the context of knitting and crocheting, a ‘yarn’ is composed of raw fibres. These are interlocked to create a continuous length, colloquially known as a ‘ball of wool’.

**Overview of the dataset**

The yarn dataset was obtained through use of an API key, which is publically available to any Ravelry user with a free account. The key was used to authenticate with Ravelry via OAuth. Then, the ‘show yarn’ method was used, as per the Ravelry API documentation, to retrieve data for all yarns present in the Ravelry yarn database (Forbes, 2017). This method returns a ‘yarn(full)’ object for each yarn which R interprets as a single-row dataframe containing various attributes for the yarn. These objects were combined using rbind to create the final dataset - a dataframe with all yarns in the database.

The dataset contains 109895 entries and 25 attributes in its raw form. However, the number of attributes is likely to change because:

* Not all attributes are independent. For example ‘yarn\_gauge’ and ‘max\_needle\_size’ are inter-dependant.
* Some attributes are in the form of arrays (or lists when loaded into R). The ‘Fiber\_Type’ attribute, for example, contains the percentage of different fibre types in each yarn. This attribute will likely be split into a number of new attributes – one for each fibre type, with each record giving the percentage present in the yarn.
* New attributes may be introduced that are resultant on existing attributes. For example, a Boolean ‘natural\_fiber’ attribute may be created to signify whether the yarn is composed of natural fibres.

With the exception of the 4 attributes which are of type *list,* attributes are of types *Boolean, numeric* and *character*. The list type attributes will be converted into multiple attributes of type *Boolean, numeric* and *character*. The dataset contains many null values which will be carefully considered during modelling.

**Project objective and proposal of data-mining methods**

The overall rating of each yarn will be used as the dependant variable, with the other attributes used to predict this value. The overall rating for each yarn is the product of the average user rating and the number of ratings. The objective of this project is to identify which characteristics are necessary to produce a highly rated yarn.

As discussed above, a degree of data cleaning will be necessary to ensure the attributes used are sensible. Dimensionality reduction (PCA or similar) will be attempted to remove attributes which do not contribute to the overall rating. Since the data has a known class, this dataset is suited for supervised learning methods. At this stage, it is unknown which techniques will perform best but since the size of the dataset is not excessive, a supervised neural network of some type, support vector machine and ensemble methods (eg. random forest) should all be viable. (Law & Weiner, 2002)

**References**

Forbes, C., (2017) Ravelry API documentation, Available from <https://www.ravelry.com/api> [accessed on 16/02/2017]

Liaw, A., Wiener, M., (2002) Classification and regression by randomforest, *R News*  (2/3) 18–22 Available from: http://www.bios.unc.edu/~dzeng/BIOS740/randomforest.pdf