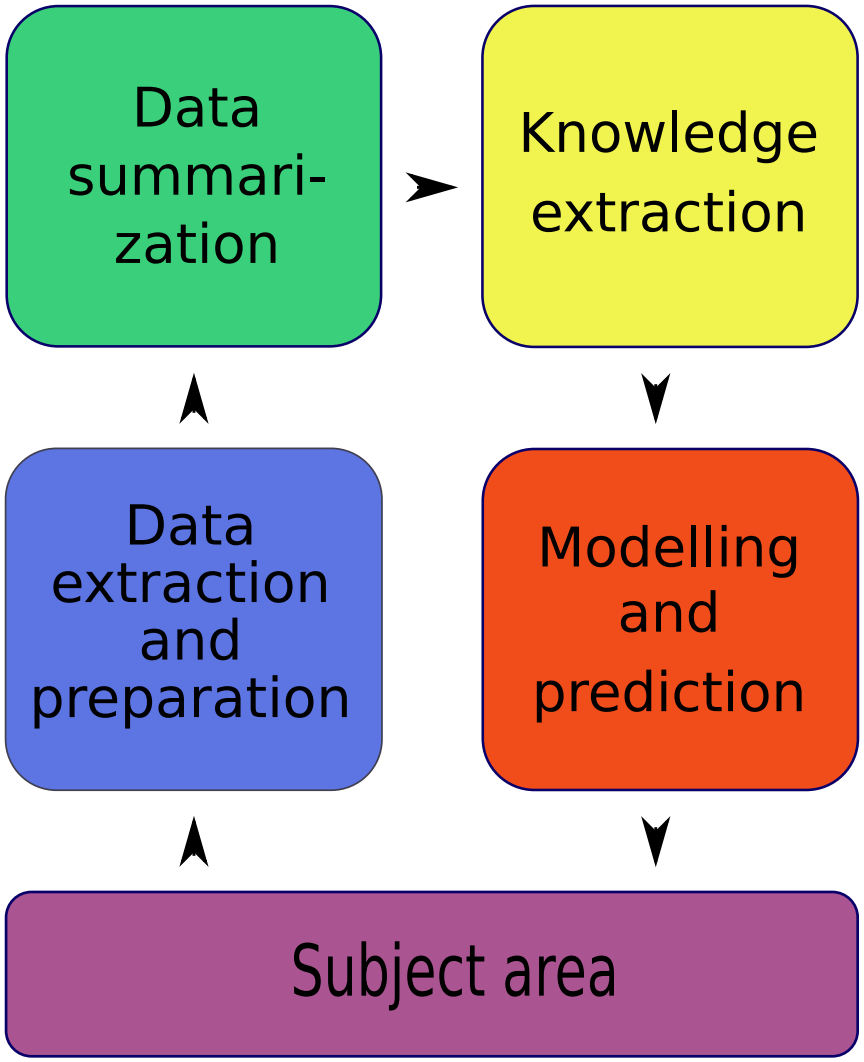
The data analysis cycle



Data Analysis: Text and Network Analysis

Institute of Technology Tallaght

Department of Computing

Multinomial Naïve Bayes for the Prediction of Document Type

- Naïve Bayes can be used with unstructured data
- One example is the *classifiaction of documents*
- Documents are treated as *bags of words*
- As in basic Naïve Bayes, the key to estimating the probability of a document belonging to a class, given the evidence (i.e. the count of different words in it), are the probabilities of documents having a particular combination of words when they are in a particular class (i.e. the probability of evidence given the outcome). This is calculated as:

$$P(E|C) = N! \times \prod_{i=1}^{k} \frac{P_i^{n_i}}{n_i!}$$

where $N = n_1 + n_2 + ... + n_k$ is the number of words in the document being classified, k is the number of unique words in all the documents in the data set (all training documents), $n_1, n_2...$ are the counts of each of these unique words in the document that is being classified

and P_i is the probability of unique word i appearing across all documents of class C in the training set training set.

- The probabilities of membership in different classes are then calculated for a document as:

$$P(C|E) = \frac{P(E|C_1)P(C_1)}{P(E|C_1)P(C_1) + P(E|C_2)P(C_2) + ...{}_P(E|C_o)P(C_o)}$$

where $P(C_i)$ is the *a priori* probability of a document belonging to class C and o is the number of document classes (types).

Example

Let's say there are two words in the vocabulary of a group of documents: **`yellow`** and **`blue`**. A type of document called 'yellowish green', which we will denote with **`Y75_B25`**, has $P(\texttt{yellow}|\texttt{Y75\_B25}) = 0.75$ and $P(\texttt{blue}|\texttt{Y75\_B25}) = 0.25$. The probability of a 'yellowish green' document consisting of the words **`{yellow, yellow, yellow}`** is:

$$P(\texttt{yellow, yellow, yellow}|\texttt{Y75\_B25}) = 3! \times \frac{0.75^3}{3!} \times \frac{0.25^0}{0!} = 0.42$$

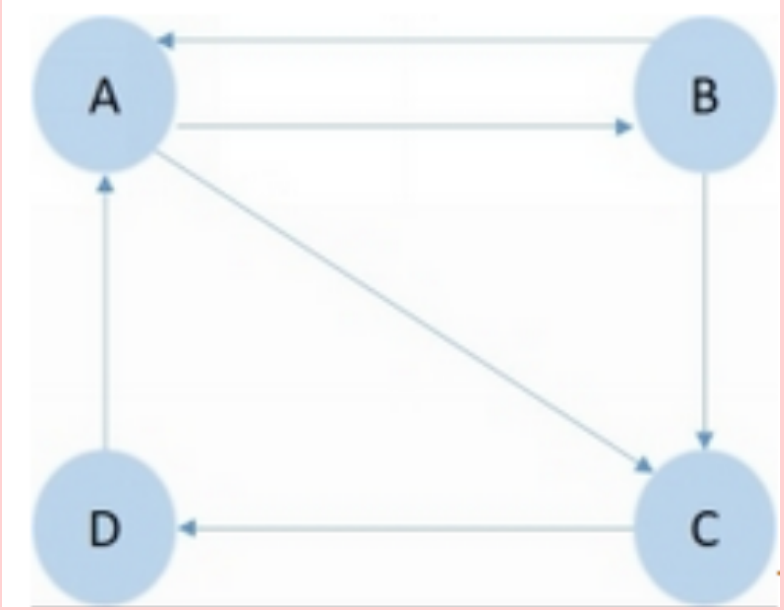
Now let's say that there is only one other class, 'very blue green', which we will denote with **`Y10_B90`** and which has $P(\texttt{yellow}|\texttt{Y10\_B90}) = 0.1$ and $P(\texttt{blue}|\texttt{Y10\_B90}) = 0.9$. The probability of a 'very blue green' document consisting of the words **`{yellow, yellow, yellow}`** is:

$$P(\texttt{yellow, yellow, yellow}|\texttt{Y10\_B90}) = 3! \times \frac{0.1^3}{3!} \times \frac{0.9^0}{0!} = 0.001$$

We can see that the prior probabiliy of class **`Y10_B90`** would have to be about 250 or more times greater than that of class **`Y75_B25`** in order for a classification of the given document as **`Y10_B90`** to happen.

Importance of Nodes and Page Rank

- When network connections are directional, node influence or rank can be calculated
- One way to do this is using a link analysis algorithm, such as the Google Page Rank algorithm. With this The algorithm is based around the idea that the more in-links a node has, the more influential it is considered to be.
- The model will be presented using an example:



For the four nodes in the picture, the values of influence (or rank) are related as follows:

$$Ra = 0.5Rb + Rd$$

$$Rb = 0.5Ra$$

$$Rc = 0.5Ra + 0.5Rb$$

$$Rd = Rc$$

The *influence matrix* is:

	Ra	Rb	Rc	Rd
Ra	0	0.50	0	1.00
Rb	0.50	0	0	0
Rc	0.50	0.50	0	0
Rd	0	0	1.00	0

Before calculation starts, equal rank is assigned to all the nodes, after which they have the following values:

Variable	InitVal
Ra	0.250
Rb	0.250
Rc	0.250
Rd	0.250

Then an iterative process of assigning new rank values is performed, using the values from the previous iteration as the right-hand-side values in the equations.

Variable	InitVal	Iteration1	Iteration2	....	Iteration8
Ra	0.250	0.375	0.313	....	0.333
Rb	0.250	0.125	0.188	....	0.167
Rc	0.250	0.250	0.250	....	0.250
Rd	0.250	0.250	0.250	....	0.250

The process is stopped once the rank values stabilise.

The Google Page Rank algorithm also includes a *damping factor* accounting for the fact that some proportion of page requests do not come from another page:

$$Ra = \frac{1-d}{N} + d \times (0.5Rb + Rd)$$
$$Rb = \frac{1-d}{N} + d \times (0.5Ra)$$
$$Rc = \frac{1-d}{N} + d \times (0.5Ra + 0.5Rb)$$
$$Rd = \frac{1-d}{N} + d \times (Rc)$$

Typically, $d = 0.85$. This number represents the frequency with which the average web-surfer accesses pages from bookmarks rather than by link from another page.

References

[DSB] *Data Mining: Practical Machine Learning Tools and Techniques*, by Ian H. Witten, Eibe Frank, Mark A. Hall, Christopher J. Pal, Kindle Direct Publishing eBook, 2016.

[MSD] *Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*, by Glenn J. Myatt and Wayne P. Johnson, John Wiley & Sons, 2014.