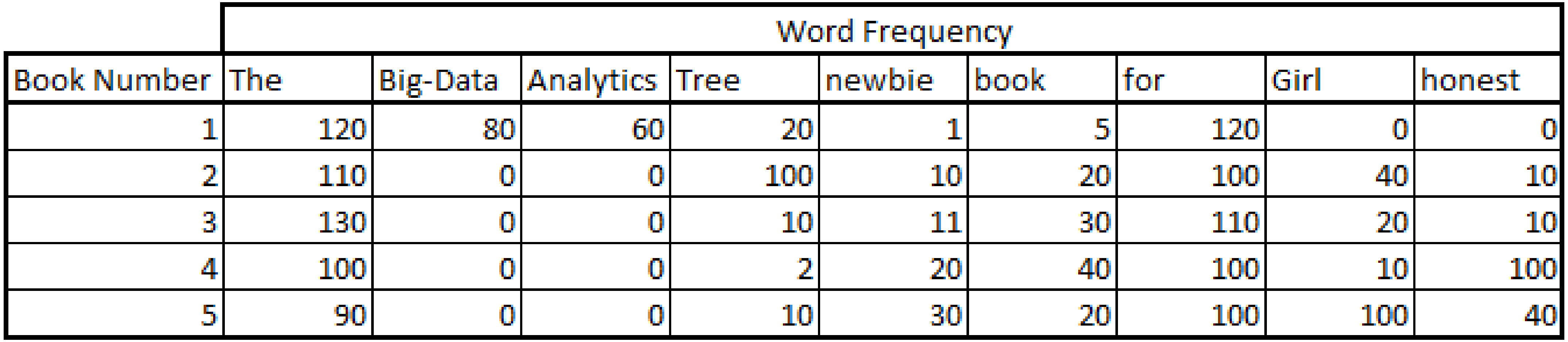
Solving the puzzle using Text Mining

- Term Frequency (TF) Matrix :

This is the most obvious technique to find out the relevance of a word in a document. The more frequent a word is, the more relevance the word holds in the context. Here is a frequency count of a set of words in the 5 books :



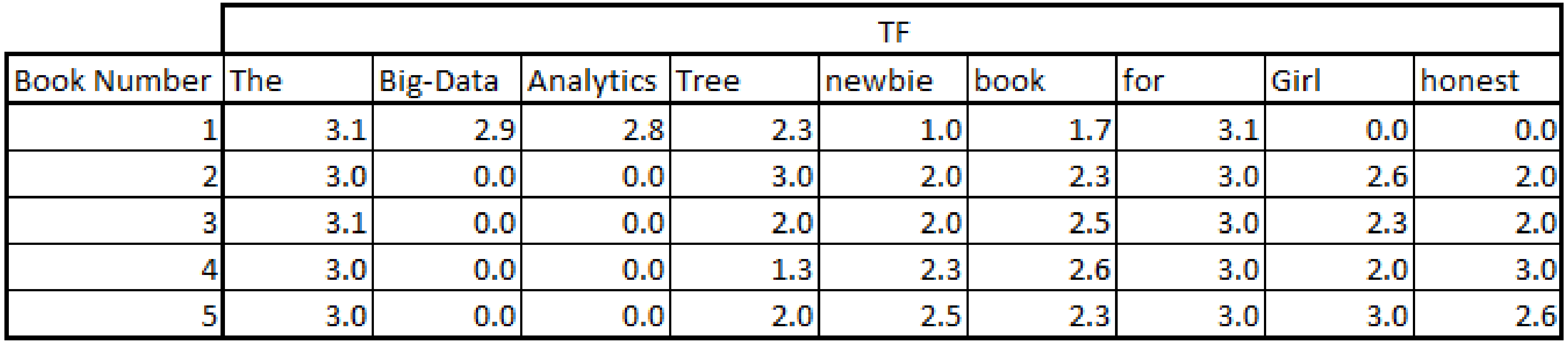
	Word Frequency								
Book Number	The	Big-Data	Analytics	Tree	newbie	book	for	Girl	honest
1	120	80	60	20	1	5	120	0	0
2	110	0	0	100	10	20	100	40	10
3	130	0	0	10	11	30	110	20	10
4	100	0	0	2	20	40	100	10	100
5	90	0	0	10	30	20	100	100	40

Term Frequency (TF)

- One way to check Term Frequency (TF) is to just count the number of occurrence.
- But it has been observed that if a word X occurs in document A 1 time and in B 10 times, its generally not true that the word X is 10 times more relevant in B than in A.
- Hence it is good to apply following transformation on TF :
  -

$$\begin{array}{ll} TF = 1 + \log(TF) & \text{if } TF > 0 \\ TF = 0 & \text{If } TF = 0 \end{array}$$

Calculations:



	TF								
Book Number	The	Big-Data	Analytics	Tree	newbie	book	for	Girl	honest
1	3.1	2.9	2.8	2.3	1.0	1.7	3.1	0.0	0.0
2	3.0	0.0	0.0	3.0	2.0	2.3	3.0	2.6	2.0
3	3.1	0.0	0.0	2.0	2.0	2.5	3.0	2.3	2.0
4	3.0	0.0	0.0	1.3	2.3	2.6	3.0	2.0	3.0
5	3.0	0.0	0.0	2.0	2.5	2.3	3.0	3.0	2.6

- Now to find the relevance of document in the query, you just need to sum up the values of words in the query.
  - Document 1 : 1.7 + 3.1 + 2.8 + 1 = 8.6
  - Document 2 :2.3 + 3.0 + 0 + 2 = 7.3
  - Document 3 : 2.5 + 3.0 + 0 + 2 = 7.5
  - Document 4 : 2.6 + 3.0 + 0 + 2.3 = 7.9
  - Document 5 : 2.3 + 3.0 + 0 + 2.5 = 7.8

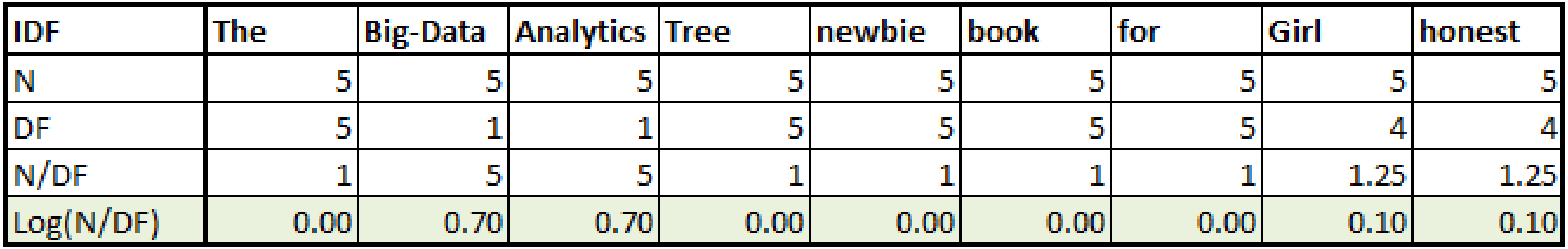
- Document 1 will be more relevant to display for the query. Since, document 4 and 5 are not far away from Document 1. They might turn out to be relevant too. This is because of the stopwords which elevates all the scores with similar magnitude.

Inverse Document Frequency Matrix(IDF) :

- IDF is another parameter which helps us find out the relevance of words.
- It is based on the principle that less frequent words are generally more informative.

- IDF = log (N/DF)
  - N represents the number of documents and DF represents the number of documents in which we see the occurrence of this word.

Calculations

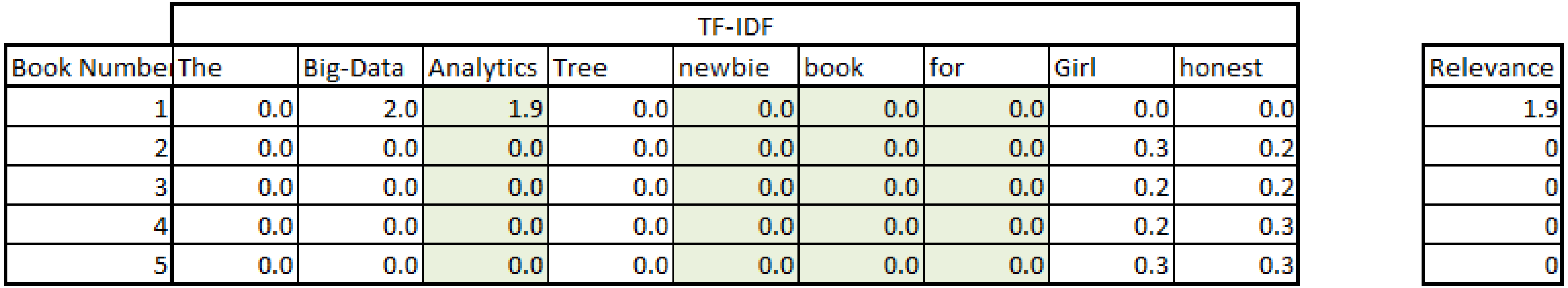


IDF	The	Big-Data	Analytics	Tree	newbie	book	for	Girl	honest
N	5	5	5	5	5	5	5	5	5
DF	5	1	1	5	5	5	5	4	4
N/DF	1	5	5	1	1	1	1	1.25	1.25
Log(N/DF)	0.00	0.70	0.70	0.00	0.00	0.00	0.00	0.10	0.10

- We now can clearly see that the words like “The” “for” etc. are not really relevant as they occur in almost all the document. Whereas, words like honest, Analytics Big-Data are really niche words which should be kept in the analysis.

TF-IDF Matrix :

- As we now know the relevance of words (IDF) and the occurrence of words in the documents (TF), we now can multiply the two. Then, find the subject of the document and thereafter the similarity of query with the document.



	TF-IDF								
Book Numbe	The	Big-Data	Analytics	Tree	newbie	book	for	Girl	honest
1	0.0	2.0	1.9	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.2
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.2
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.3
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.3

Relevance
1.9
0
0
0
0

- Now it clearly shows that Document 1 is most relevant to query!