Analyzing frequency: tf-idf

Suppose we have a text data that contains n document as below.

Document	Texts
1	blah blah
2	blah blah
...	...
n	blah blah

- A document could be a sentence, a paragraph
- A document contains of `terms`. Terms could be a word or a collection of words
- How important is a word/term to a document?

Term frequency (tf)

- The more often the words show up the more important it is
- We could use `Term Frequency`
- TF (Term Frequency) = (Number of times term `t` appears in a document) / (Total number of terms in the document).
- Please notice that the same word/term may have different `tf` in a different document
- For example: the word `love` in document 1 may have different `tf` when compared with that of document 2.

Example

- Calculate the term frequency of the word `cats` for each document in the below text dataset.

	Document	tf
1	I love cats. Cats are renowned for their graceful agility. Cats are awesome!	3/13
2	Cats are furry animals that like to sleep.	1/8
3	Dogs and cats are popular pets that bring joy to many families	1/12
4	Dogs are friendly animals that enjoy companionship.	0

Issues with Term frequency (tf)

- Sometimes: Rare terms are more informative than frequent terms
  - Example: “the”, “is”, “of”…
- We should remove some words such as “the”, “is”, “of” if we use `tf` as a measure of importance
- Or we could create a `weight` for a term so that the rare words would have higher weights.
- Inverse Document Frequency (idf) is such a weight

Inverse Document Frequency (idf)

- IDF (Inverse Document Frequency) = log_e(Total number of documents / Number of documents with term `t` in it).

$$idf(t) = \ln\left(\frac{\text{Total number of documents}}{\text{Number of documents with term t in it}}\right)$$

- The idf of a rare term is high, whereas the idf of a frequent term is likely to be low

- The `idf` of a term is a constant throughout the document. For example the word `love` in Document 1 should have the same `idf` as the word `love` in Document 2.

tf-idf

$$\text{tf-idf}\,(t) = \text{tf}\,(t) \cdot \overset{\text{idf}}{\cancel{\text{tf}}}\,(t)$$

- The `tf-idf` of the same term may have different values in different documents. For example the word `love` in Document 1 may have a different value `tf-idf` compared to the word `love` in Document 2.

Example

- Calculate the idf and tf-idf of the word `cats` for the documents in the below text dataset.

tf		Document	tf-idf
3/13	1	I love pets. Cats are renowned for their graceful agility. Cats are awesome!	.662
1/8	2	Cats are friendly animals that like to sleep.	.036
1/12	3	Dogs and cats are friendly pets that bring joy to many families	.024
0	4	Dogs are friendly animals that enjoy companionship.	0

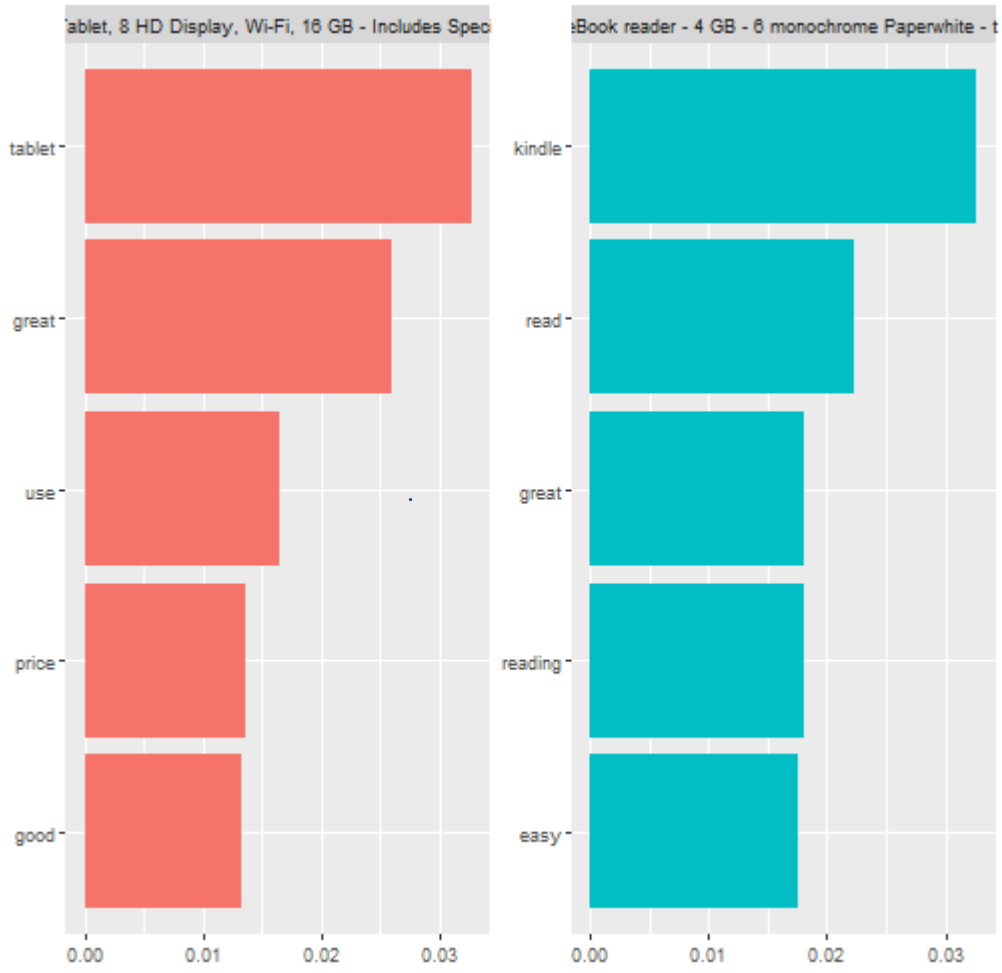
$$idf(cats) = \ln\left[\frac{\#\ \text{documents}}{\#\ \text{document with "cats"}}\right] = \ln\left[\frac{4}{3}\right] \boxed{= .287}$$

$$idf(dogs) = \ln\left[\frac{4}{2}\right] = \boxed{.693}$$

$$tf\ idf = tf \cdot idf$$

Example

Sample Codes



Plot tf-idf

```
## # A tibble: 8,605 × 7
##    document                                   word      n total     tf   idf
##    <chr>                                      <chr> <int> <int>  <dbl> <dbl>
##  1 Amazon Kindle Paperwhite - eBook reade… kind…  1717 52533 0.0327     0
##  2 All-New Fire HD 8 Tablet, 8 HD Display… tabl…  1342 41031 0.0327     0
##  3 Amazon Kindle Paperwhite - eBook reade… read   1169 52533 0.0223     0
##  4 All-New Fire HD 8 Tablet, 8 HD Display… great  1067 41031 0.0260     0
##  5 Amazon Kindle Paperwhite - eBook reade… read…   950 52533 0.0181     0
##  6 Amazon Kindle Paperwhite - eBook reade… great   947 52533 0.0180     0
##  7 Amazon Kindle Paperwhite - eBook reade… easy    925 52533 0.0176     0
##  8 Amazon Kindle Paperwhite - eBook reade… books   778 52533 0.0148     0
##  9 Amazon Kindle Paperwhite - eBook reade… love    772 52533 0.0147     0
## 10 Amazon Kindle Paperwhite - eBook reade… light   686 52533 0.0131     0
## # ℹ 8,595 more rows
```