Bag of Words and TF-IDF Assignment

Nathan Sawmiller

February 2025

1 Vocabulary

All unique words (lowercased, punctuation removed, no stemming/lemmatization):

`a, ago, from, improve, is, long, more, novel, novels, read, should, the, this, time, to, vic`

Total unique words: 19

2 Task 1: Bag of Words Table

Sentence	a	ago	from	improve	is	long	more	novel	novels	read	should	the	this	time	to
S1	1	1	1	0	1	1	0	1	0	0	0	0	1	1	0
S2	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0
S3	0	0	0	0	0	0	1	0	1	1	1	0	0	0	0
S4	0	0	0	1	1	0	0	0	1	1	0	0	0	0	2

3 Task 2: Document Frequency (DF) and IDF

Word	DF	Appears in	IDF
			$\log_{10}(4/\text{DF})$
a	1	S1	0.602
ago	1	S1	0.602
from	1	S1	0.602
improve	1	S4	0.602
is	3	S1 S2 S4	0.125
long	1	S1	0.602
more	1	S3	0.602
novel	2	S1 S2	0.301
novels	2	S3 S4	0.301
read	2	S3 S4	0.301
should	1	S3	0.602
the	1	S2	0.602
this	1	S1	0.602
time	1	S1	0.602
to	1	S4	0.602
victorian	1	S2	0.602
vocabulary	1	S4	0.602

you	1	S3	0.602
your	1	S4	0.602

4 Task 2: TF-IDF Table

(Rounded to three decimal places)

Sentence	a	ago	from	improve	is	long	more	novel	novels	read	should	the	t
S1	0.602	0.602	0.602	0.000	0.125	0.602	0.000	0.301	0.000	0.000	0.000	0.000	0
S2	0.000	0.000	0.000	0.000	0.125	0.000	0.000	0.301	0.000	0.000	0.000	0.602	0
S3	0.000	0.000	0.000	0.000	0.000	0.000	0.602	0.000	0.301	0.301	0.602	0.000	0
S4	0.000	0.000	0.000	0.602	0.125	0.000	0.000	0.000	0.301	0.301	0.000	0.000	0

5 Notes

- It seems that documents are understood to be sentences too.
- The result of the TF-IDF table seems to suggest a kind of matrix is created between the two features of the sentences or documents.
- I say features because it seems that Document Frequency (DF) and Inverse Document Frequency (IDF) are two different features of documents, where TF-IDF is just the ratio of such documents over the document freq of a word, all as an image of a log in base 10.
- **TF** = raw term frequency (from BoW table)
- **IDF** = $\log_{10}(4/\text{DF})$
- **TF-IDF** = TF × IDF

6 Answers to Questions

6.1 Which word has the highest TF-IDF score in Sentence 4, and why does it have such a high value?

The word within document four which has the highest value is ‘to’. Without looking into why just yet, I know that by the TF-IDF expression TF-IDF = TF × IDF means any large value would be either connected to the Term Frequency of that word or Inverse Document Frequency; looking into the IDF of the word ‘to’ I find that it was on the high end, meaning its Document Frequency was small & only appeared in one document; looking into the table for Bag of Words I find too that though it appeared twice it was only in its sentence rather than distributed like most others were. This would amount to its expression being TF-IDF = $2 \times 0.602 = 1.204$, which can now be seen as having the largest TF while one of the highest IDF; so, I would say that the term within the expression that explains its value is that of its TF or rather the raw cell count in the Bag of Words table; but why is its attribute of having the largest occurrence in a single document rather than a couple arises this property of it?

This property I’m describing above is central function of TF-IDF; it’s to weight or reward words which are repeated a lot in a single document as it’s supposed that such an attribute of a word means it’s thematically important to that document. This is why its raw term frequency is used while being a multiple of its IDF, being a weight of the whole corpus for the rarity of words while penalizing ones more frequent.

6.2 Why does the very common word “is” have a low TF-IDF score across most sentences?

Much like the prior analysis for the word ‘to’ the reason for its low TF-IDF score is due to its distribution across the documents (sentences). As I noted before, the factor for TF-IDF is that of IDF which will penalize terms which appear frequently across the documents of the corpus. This can be seen in this case by $\log_{10} \frac{N}{\text{DF}}$ where $N = 4$ (the documents) & $\text{DF} = 3$ (Doc Freq).

6.3 How does TF-IDF improve upon simple Bag of Words when representing document importance?

Unlike the raw amount of term’s appearance within the given documents given by the Bag of Words algorithm, TF-IDF generates a value representative of its appearance relative to its rarity which given by the IDF of the word.