**Mathematical Review**

Term Frequency - Inverse Document Frequency (TF-IDF) is a Machine learning technique used as a weighting factor for features. The weight increases as the word occurs more times in the text but is offset by the frequency of occurrences in the dataset. The offset reduces the importance of common words such as “the”, “a” and “and”, which appear frequently but do not serve any purpose when it comes to text classification.

tf-idf (t, d) = tf (t, dn) \* idf (t, D)

The tf-idf function is a multiplication of term frequency (TF) function and the inverse document frequency (IDF) function. The First function is the term frequency function.

tf (t, dn) = ft| d

The tf in the function stands for term frequency and the function expresses how many times a single term occurs in a document. The f represents frequency, t represents given term and dn represents a specific document in the dataset.

idf (t, d) =

The second function is called IDF function, where the total number of documents (N) is divided by the number of documents that contain given word. The function is logarithmic in order to dampen its effects.

In order to demonstrate how this works, here is example with calculation on short sentences. We have two documents d1 and d2, both documents contain one sentence. Both sentences are the same except for one difference, one sentence contains negative word not, thus reversing the meaning of the sentence. The TF-IDF should identify, which word in this document is important.

|  |  |
| --- | --- |
| **Documents** | **Data** |
| Document d1 | Computer science is easy. |
| Document d2 | Computer science is not easy. |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Terms (t)** | **TF** | | **IDF** | **TF-IDF** | |
| **d1** | **d2** | **d1** | **d2** |
| computer | 1 | 1 |  | 0 | 0 |
| science | 1 | 1 |  | 0 | 0 |
| is | 1 | 1 |  | 0 | 0 |
| easy | 1 | 1 |  | 0 | 0 |
| not | 0 | 1 |  | 0 |  |

The TF column counts the occurrence of all words in each document. We can see, that only difference is that d1 does not contain word ‘not’. In the IDF column we use the IDF function, which means number of documents divided by number of occurrences of the word in the documents. In this case every word was for every word except for not, which was contained in only one of the documents, so the result was . With all the values from TF and IDF, we can calculate the TF-IDF. From the results we can see, that the TF-IDF technique labeled all the words with zero value, except for word not in d2. TF-IDF was contained once in d2 and the IDF value for that word was and after multiplication we end up with a result of . TF-IDF values for other words in both documents are 0, because every other word is contained in both documents thus IDF value is which is equal to 0 and whatever we multiply with the IDF value will end up being 0.

There are also other variations of TF-IDF algorithm. For example, Sublinear Term Frequency (STF), which can become useful if the TF is required to be scaled less then linearly. STF is like TF, but it puts the result on a logarithmic scale (Cambridge UP, 2019).

stf (t, d) = log (ft| d)

Changes can also be made for the IDF. For IDF Smoothing we can add 1 to the inside of the logarithmic expression. This change will never let the IDF value to be a 0 and thus all TF-IDF results will have at least some value (Cambridge UP, 2009).

idf (t, d) =

**References**

Cambridge UP (2009), ‘Scoring, term weighting and the vector space model’, *Introduction to Information Retrieval*, Online edition: Cambridge UP, pp. 109-133

Scikit-learn developers (2019), Scikit-learn documentation, *‘6.2. Feature Extraction’*, [Online]. Available at: <https://scikit-learn.org/stable/modules/feature_extraction.html> [Accessed 26.3.2020]

Aizawa, A., 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, *39*(1), pp.45-65 [Online]. Available at: https://pdf.sciencedirectassets.com/271647/1-s2.0-S0306457300X01616/1-s2.0-S0306457302000213/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEDEaCXVzLWVhc3QtMSJHMEUCIG72I7YlS2enbKHYJ0QVcy08QF4L%2BI0rTsDpIfDyFVPbAiEA0oqyyIWqNn7omadrKdEfRfoLheOqsfATSU2W6p1xOBUqtAMIeRADGgwwNTkwMDM1NDY4NjUiDIqGBCqLZxIw%2B9OPYSqRA2ZpuW7I9ev0yLCFtO5BTM3DyuvLEoElRd6Xre8vV8T7cTYT4zrRcZ0vrsLt0VFuHdzPiLGNngk2aQZr0dcXnsTswNgCyzvhnz%2BcrNwrFjf0OdieE02Go3bBsdDP%2F6nAKc9F1%2BB0P2oE9xT9At3B1tKCoihGpvC5g3IHIm2A7aPyBCR%2Fg38qODtIF8KZO2YcnEmwssb899J%2F7hL7I7ecfN1ZD9X%2BcU4TKZBMuMIyQOtWGS5HHk38j9cca7wzQG7ZvTFY39GKG0R43Azwi1dbhc1wPIdFOVJxxma8TAoLYQ3j4worBI%2FOkAyZWlIyKLI3cgz2tDNmlEchP%2BoyjO%2BcGBdezg9vxHRyNirrhIb9ZOfXzg0uzM7cJwToeOp0xunM1v9JonHGGXk%2Fq2IyqeB2hvXzIM%2BiqMvjU%2BxUQNjteKSgmDx7MOcj%2BqRs1fKOKCfY%2BPyTCWYv8Php9UOX%2B8MXY6k8oOBQZIYDKIKiutlxAPhZKXHl8ySwF5lsGH1HQVhSsSzdHiR2NLVq6HK7%2FalX5F6jMLzf9fUFOusBvdQa0nXFDmru%2Bfrnd860gJWn6fx2WogW9WZ6BfY7qRENMFkk05g762cqVkSxKHkSYTixNWi23dvQB9m0Of2D6gZRVuEBgvZSM3lDwtOUHUqX216ssSzfMzxb7qmsnewstQqv0fji21QptyEPaPlt361VFOZfBkC0krNqlDlVFe5nwgYfYVZDUsqvS2jcOd2%2FZsW7vj98cuIbrct1wEWNGQKik5bNVdk2DGK7BlauuSoh9oNIQc6M1OgQ7sI2AgS%2BsJnWHLbYocq2exY2oLO1mEtywcJUp9kxsc2eRT4GgA5ENzvY4i3S%2FUC%2Bwg%3D%3D&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20200514T172506Z&X-Amz-SignedHeaders=host&X-Amz-Expires=300&X-Amz-Credential=ASIAQ3PHCVTY37HSHNND%2F20200514%2Fus-east-1%2Fs3%2Faws4\_request&X-Amz-Signature=2f2182d9a354968eb6eb3e130eddf8c5072de4229bc8ede0a525a4eb42a44f4e&hash=224508b53985aef402445441e5c77b6709604e86e9bf92b04ce44fb19df79516&host=68042c943591013ac2b2430a89b270f6af2c76d8dfd086a07176afe7c76c2c61&pii=S0306457302000213&tid=spdf-25d95727-52f5-44f9-9fb3-9c03a54dd2ae&sid=a4010be3399de14d3359efa38abf555e9627gxrqb&type=client [Accessed 20.3.2020]