WHO • SAID • THAT • ?

Comparing performance of TF-IDF and fastText to identify authorship of short sentences

Problem

Identifying authorship of text can be done by matching features extracted from text.

Currently mostly performed on large text documents, not on short everyday sentences, which could benefit personal assistents or chatbots.

Compare performance of two popular extraction techniques TF-IDF and fastText by answering the questions:

- Which mistakes does the model make?
- Does performance change as sentence length increases?

TF-IDF measures importance of words by counting occurrences in document

fastText represents words with similar meanings and context the same way

Method

1. Retrieve data

Download transcripts of television series

F·R·I·E·N·D·S

consisting of 236 episodes

→

2. Parse data

Remove side characters
Split lines
Remove duplicate lines
...
+-55.000 lines

→

3. Extract features

Sentence

TF-IDF ⟷ fastText

→

4. Classify features

Using logistic regression

Choosing from 6 characters with 80:20 train:test split

→

5. Evaluate metrics

Which mistakes: confusion matrix

Min. sentence length: accuracy, confidence, cross-entropy loss

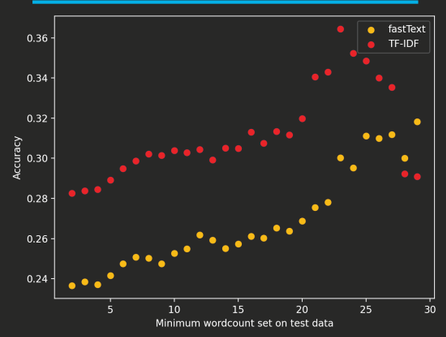
Results

Confusion matrix

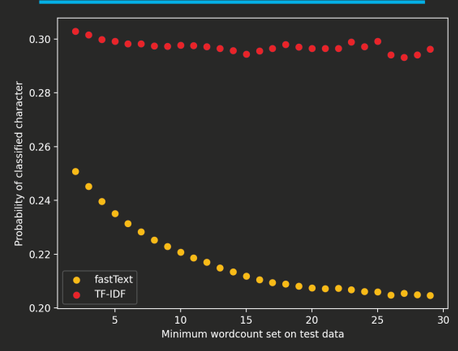
Amount of correct predictions

fastText	332	294	569	673	331	515
TF-IDF	457	408	656	636	433	654

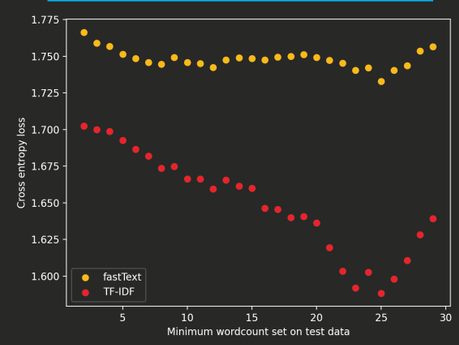
Accuracy



Confidence



Cross-entropy loss



Conclusion

TF-IDF outperforms fastText in every measurement, but its performance is only slightly better than randomly guessing the original character, reaching an accuracy of 28 percent when making a distinction between 6 characters.

Accuracy increases linearly at the same rate for both techniques test set's sentence length increases.

TFIDF's confidence remains constant as this limit is set on either the test or training data, whereas fastText's confidence decreases and increases, respectively.

Cross-entropy loss, however, remains constant for fastText and decreases for TF-IDF as the minimum word count set on the test data increases.

TU Delft

Thomas van Tussenbroek
CSE3000, 25/06/2020
David Tax; Marco Loog; Tom Viering; Arman Naseri Jahfari; Stavros Makrodimitris;