Conditional Entropy vs Tokenization (mBERT)

Estimated Conditional Entropy vs Number of sub−tokens of a word in test dataset

Language: De, En, Fr, It, Vi, Th, Zh−CN, Yue, Ja, Sv, Kor