NLP 2019 HW1

Or Stern, Koren Maliniak

1. Results

	en	es	fr	in	it	nl	pt	tl
en	14.9545	17.55731	16.69178	20.44755	18.74125	18.38444	17.60872	19.65232
es	19.80442	14.04512	16.98129	20.31389	17.04775	19.60626	15.61376	20.51082
fr	19.02182	16.74322	14.48761	21.2174	18.43668	19.18305	17.47108	21.31608
in	19.47016	18.06237	18.51041	15.39236	19.03788	19.71923	18.12052	17.65649
it	19.68963	16.39105	17.42927	20.2749	14.41314	20.14743	16.2086	19.77255
nl	18.42385	18.85474	18.26952	20.4413	20.20772	14.82694	19.12487	20.2137
pt	19.54274	15.6655	17.27037	20.21653	17.01865	20.01167	13.75632	19.71976
tl	18.14727	18.21279	18.02511	17.99447	18.4966	19.60879	17.92131	14.69454

2. Some Considerations taken:
    a. We followed and implemented the methods exactly as stated in the assignment (signatures, return values etc…).
    b. We decided to treat the corpus as one big file and not split it. As tweets can have varied length.
    c. We did not exclude emojis because we think that might help the model. Different languages (different cultures) might use emojis differently.
    d. We padded with <e> and <s>. And while counting unigrams we dealt with that in order to avoid bias towards those signs.
3. Code
    a. https://github.com/koren88i/NLP-HW1