DL4Seq - Assignment 3

Otmazgin, Shon
shon711@gmail.com

Rassin, Royi
isroei5700@gmail.com

May 15th, 2021

1 Challenge

Q: Can the two languages be distinguished using a bag-of-words approach? Explain why.

A: It depends what the features are. in general BOW disregarding grammar and word/characters order. Hence if the features are unigram characters, positive and negative examples with the same multiplicity will have the same BOW representation, but if the features are more sophisticated like 1 if *b* follow by *c* BOW will work (in that case the BOW is sum of 1 feature so it is kind of naive BOW).

Q: Can the two languages be distinguished using a bigram or trigram based approach? Explain why

A: No for bigram and fot trigram. For bigram all the possible features are: $\{DD, Da, aa, aD, Db, bb, bD, Dc, cc, cD, Dd, dd, dD\}$ (D is a digit) for both languages. Therefore there is no bigram feature which can distinguish between the two languages. For trigram we can have feature *bDc* which can indicate for positive example and trigram feature *cDb* which can indicate for negative example, but this is only for examples with 1 and only 1 digit between *b* and *c*, with that restriction we can assume for low accuracy for trigram too.

Q: Can the two languages be distinguished using a convolutional neural network? Explain why

A: Yes. Convolutional neural's filters "search" for "object" all over the input. Particularity in sequences, if we will optimize filter's k value (k is k-gram in the sequence) we will be able to distinguish between the languages. We saw that trigram already have major feature but rare examples, so higher k-gram with the filter's ability to learn the importance of each k-gram in the sequence will cover it. **Note:** It will be hard to distinguish between examples with *b* follow *c* not in the k-gram filter length. But we can assume that these examples are very rare and given a good dataset we can optimize the filter sizes.