Project Summary

In the following analysis, both a structured perceptron, as well as HMM algorithm have been applied to the ner_dataset. In case of the structured perceptron, a model based on the default feature mapper from the skseq package has been run, as well as an extended model with own added features, namely whether the given word:

- ends with ed
- ends with ly
- ends with ing
- ends with or
- starts with up
- contains a full stop
- is “to”
- is “of”
- is “from”
- is “the”
- is “in”

While the first 5 features are helping to differentiate between common word endings, the last added features on the list help to identify common conjunction terms in the English language. The full source code can be found in both python notebooks of the project.

By adding those features, following conclusion can be drawn:

As such, the extended model showed only a slight increase in accuracy for the training set in comparison to the base model (up to 96.1%). However, the performance on the validation set showed up to be similar among both models with 69.6%, while the performance on the testing set was around 3.5% lower with 62.2% for the extended model. A similar conclusion could be drawn by comparing the sentence accuracy metrics. Moreover, by comparing the results of the perceptron with the HMM accuracy, the general conclusion is that the accuracy in this case showed up to be higher (around 97%), as well as very similar across the 3 different data sample sets. Given the obtained results, a main issue is the large number of 0 tags which leads to data imbalance and should be addressed in order to create a more accurate model (in fact in our obtained results a large fraction of the high accuracy can be drawn from the high amount of 0 tags in the dataset).