Previous model: KNN clusterer that checks for grouped documents in [2,n-1] clusters, assuming that if documents are often in the same cluster they are likely to be from the same author. (No automatic evaluation, manual sampling shows 0% accuracy)

Previous model: Neural network informed by characters, comparing every document to all others (94% accuracy but equals baseline)

Current model: Neural network informed by character embeddings/skipgrams (75-85% accuracy, baseline 50-50)

Future model: Meanshift clustering model that optimizes the number of clusters to work with itself and can be evaluated.

TODO: add features to dataflow

TODO: combine meanshift clustering with character embeddings NN for more informed models.

Dataflow:

1. For every of the 18 problems, all their documents are cut up in characters and normalized
2. Every character is encoded into a number (total differs per language, usually 40-50)
3. The vocabulary size is used for a sampling table, calculating the probabilities of that number begin used.
4. Skipgrams are created based on probabilities, wordpairs next to each other get a 1, further away get a 0 (50/50 result)
5. NN reads all information and builds a model for predicting context of characters(75%)

TODO: Inform model with gold standard labels, not sure what is actually predicted now