Omer Zucker 200876548
Omer Wolf 307965988

Understanding the Challenge

Question: Can the two languages be distinguished using a bag-of-words approach?

Answer: In our language the order of the letters 'b' and 'c' matters but amount of them is not, as well is the number of digits. Within "Bag of Words" approach the order of the characters is irrelevant, so it's only keeps number of occurrences for each word. Therefore, in our case "Bag of Words" approach will not help us defining whether we got positive example or negative one which attributed by order of sequences.

Question: Can the two languages be distinguished using a bigram or trigram based approach?

Answer: Using bigram or trigram means that we predefine our window before training the data so we should know the size of the data of which being trained as well as the sequences format before we run it. In our case, we are using random function so the size of the sequence is changeable such that the window of bigram/trigram potentially won't be sufficient to contain both 'b's and 'c's sequences. Therefore, using predefine window format such as bigram/trigram might won't be helpful solving our challenge.

Question: Can the two languages be distinguished using a convolutional neural network?

Answer: Convolutional neural networks (CNN) relies both on sequences of words as well as on their relation (extremely useful for image prediction) such that the learning process relies on them. In our case we have some restrictions and predefine format with regard to the words but not to the sequences of which being constructed while running the train. Therefore, CNN won't be a helpful tool for our language.