* 1. Understanding the Challenge
* Can the two languages be distinguished using a **bag-of-words** approach? Explain why.

The two languages **can’t be distinguished** because **bag-of-words** encodes the sequence by **summing** the words vector. Then **we lose the order** of the sequence when we sum all the words vectors.  In fact, two sequences with exactly the same words but not the same order will get the same BOW representation. (e.g., “1 a 1 b 1 c 1 d 1” and “1 a 1 c 1 b 1 d 1” will get the same BOW representation but they each belong to a different language)

* Can the two languages be distinguished using a **bigram or trigram based approach**? Explain why.
  + When using **bigram,** the two languages **can’t be distinguished**.   
    Indeed, we take all the bigram of the sequences and sums it. So, we keep the only a local order of length of two. In fact, bigram can’t inform us if “c” was after or before “b”, because “b” and “c” will never appear in the same bigram because there is always a sequence of digit between them.
  + When using **trigram**, the two languages **can’t be distinguished**.   
    The only difference between bigram and trigram in this task is that of the sequence of digit between “b” and “c” (or “c” and “b”) is of length 1, so the trigram can distinguished if the whole sequence belong to the first language or to the second.  
    But if the sequence of digit between “b” and “c” (or “c” and “b”) is bigger or equal to 2, the trigram can’t inform us if “b” appears before “c” or after. Because no one trigram will contain and “b” and “c”.

In the same way we can extend to all the n-gram method because between “b” and “c” we can get theoretically a sequence of digit of any length so if we have n+1 digit between “b” and “c” we can’t get local order between “b” and “c” with n-gram

* Can the two languages be distinguished using a convolutional neural net-work? Explain why.

No, because CNN is n-gram based , each convolutional is of fix size, but as I said above we can have between “b” and “c” a sequence of digit bigger than each convolutional fixed size. So, CNN can inform only on local size of limit size, but we need to keep information of global ordering and then CNN can’t distinguished between the two language.