Computational Linguistics

Assignment 1 (2021-10-26)

Winter Semester 2021/22

Student: Uliana Eliseeva

Random Text Generation

In this task, I chose the Jungle Book by Rudyard Kipling. In the preprocessing of data, I preferred to keep original punctuation and case in order to train models with this information as well. A few examples of the generated texts are provided in the end of this report.

It's striking how the quality of the generated text changes with the increase in N from a literal nonsense to an almost fully coherent piece.

The bigram model is impossible to comprehend: meaningful chunks do not extend further than 2 words (sometimes 3, when the model is either "lucky" to pick a word that could be a logical continuation of the previous two, or it had no other than meaningful choices to sample from in its learned probability distribution). However, on a local level that spans 2 adjacent words there were no semantic or syntactic violations observed. I ran the model on a Russian book as well and it turned out even a morphologically rich language is not a problem for it. So, the model is able to capture and output correct POS sequencing and colligation and, hence, produce locally reasonable bigram, which is expected from a bigram model.

The trigram model demonstrates longer sensible chunks in most cases spanning a clause. The peculiar characteristic of this model is that it had a higher chance of producing puns and semantic structures that we would call stylistic devices, for example:

half in and half shut

beat of a prayer

To produce them, the model would have to follow syntactic rules but break semantic rules as almost any stylistic device (like zeugma and metaphor above) are based on unexpected semantic violations. We can call this pattern of behaviour creativity. I'm inclined to believe that the model is more creative in terms of novelty and unpredictable combinations, in contrast to the 4-gram model, since it has a greater degree of freedom, the sequences it's bound to are smaller. But there is a tradeoff between creativity and coherence, which I see as a main difference between the 2 models.

The 4-gram model, as the ultimate among the 3, outputs the best quality in terms of coherence. It attempts discourse elements like clarifications, dialogues, doesn't deviate that wildly from the topic, it feels more like there's a story. At this complexity level, we can actually recognise a genre and (if we compare authors with different styles) writing style (short sentences vs longer ones, particular word choices etc.). It still reads like schizophrenia delusions, but less so than the trigram model. As mentioned above, the 4-gram model has less degree of freedom, but it is also creative in a sense that it doesn't have to repeat full sentences from the original text to generate a clear sequence, so it stays more coherent. Below, there are 2 examples of generated sentences, the difference is that the 4-gram model could "create" a more unique one than the trigram model:

Mowgli, hast thou anything to say?""Nothing. [book]

4-gramm: "But what will our Pack say?""Nothing. [generated]

3-gramm: "--you may do practically ANYTHING with public domain" [same as in book]

This is illustrative of the 4-gram model, because in general it gives out syntactic chunks very similar to the original text but yet more grammatical and various more often than the trigram model.

All models were messy with punctuation, “forgetting” to close or open brackets or quotation marks, with 4-gram one performing slightly better in this sense. Also in every model there were cases of out-of-context pieces of text such as information about the ebook, author, other external links and metadata. This poses a question of better preprocessing as well as warns us that the model learns what we feed it, so we have to be cautious about the content we choose as our training corpus.

Generated texts

bigram model:

1) is Rikki-tikki-tchk! Wahooa!" I right to be dealt now that make better than to a clear night. I ran down it. Sit down by himself. If a landslide, and the regular beat of a prayer to fight at last--that man popped his foot."In a half the trail of the days when something to move so that you do with lumps of their heavy sea is as he had taken him in a Project Gutenberg-tm License(available with a weasel in the tin dipper and filth on his

2) top of exporting a sign with rage. He's nose was rather awful."It is dead mongoose fights, that some time with their prisoner. Let's a refund set alight. That was open, and down?""I went out of the only the wildest tusker of noise all that garden path from one of the life went too soon as it came down Lukannon--because thou wouldst not be affected in paragraph 1.E.8." Now you think of drouth,"there is not uniform and hullaballoo of wild elephants,

trigram model:

1) ," The last words were retained as-is. The science of the camp through the long hissing snorts that Indian cattle give, and with almost no restrictions whatsoever. You're hung up till daylight. What ye will not wake him, kicking bodies beneath him, and then he forgets that he should not have come up to date contact information can be bought at a clumsy trot in the world's end) that I may pay for elephants, except once, when the Pack that love thee and all associated files of various

2) could can. When he returns he swears that he is the fault of this license and intellectual property infringement, a means of obtaining a copy upon request, of whom thou hast avenged Nathoo's death. So he hurried on, where he was, and I saw the great folds of the Jungle." The gun-bullocks got up to Karait with the Lone Wolf lay with her big gray nose poked him under the terms of this agreement shall be thy kill if we hauled out, of course that's a great assembly in his

4-gram model:

1) death it man, the Pack know it, the Pack know it, and told Kamya, one of the Jungle.""But what will our Pack say?""Nothing. I did not know his own strength in the least understand, and about plowing, of which he did not like or understand this kind of life. Volunteers and financial support to provide volunteers with the assistance they need, are critical to reaching Project Gutenberg-tm License. You must have seen old Kerick polishing off a drove. He's done that for thirty years.

2) men and day of the Gods of the Jungles is with him.""Something is coming uphill," said Two Tails, with a grunt."Tabaqui is with him.""H'm!" said the troop-horse, thinking hard."I couldn't stand that. I also have made a promise--a little promise. Only thy coat is lacking before I keep

my word. With the knife--with the knife of the hunter is known by the gloss of his hide. If ye find that the Bullock can toss you, or