Lab Work nº 1

Universidade de Aveiro
Departamento de Eletrónica, Telecomunicações e Informática

Algorithmic Information Theory

Ilaria Stefanizzi, Ivo Felix, Vasco Sousa

02/11/2021

1. Introduction

This work was proposed by the Algorithmic Theory of information subject's professor, who presented to the group the finite context model, which is at the base of almost all compression techniques. In this work, the main goal of the finite context model is not to compress data but to collect statistical information from one or more texts so that a text generator can use this trained model to predict the next character of a given context. A context is a sequence of n characters, with n being a positive integer starting at 1.

This report is divided into four chapters; this one introduces the report, the second one describes the whole decision-making process, the third one presents the results achieved. Finally, the last one deepens the group conclusions, describing the solution limitations and potential improvements.

2. Decisions

The developed program is structured following object-oriented programming principles. Therefore, the class Fcm is responsible for collecting statistical information about texts, using finite-context models. In particular, for calculating the entropy of the text, using probabilities of symbols and contexts.

Afterwards, the class Generator is responsible for automatic text generation based on a finite-context model learned beforehand. Furthermore, a class called Main is the entry point to the program. It is responsible for reading the command line arguments and parsing them. It then instantiates both Fcm and Generator using the corrected arguments previously provided. For each of those classes, there are unit tests to assure the robustness of the code. A more detailed description of each of these components will now be presented.

2.1. FCM

FCM is the class responsible for collecting all the statistical data about the text. The two FCM's constructor parameters are the smoothing parameter and the context size. The smoothing parameter is a float variable with values between zero and one (excluding zero). Its purpose is to avoid the problem of assigning zero probability to events after a given context that have not been seen during the construction of the model, which would be problematic as the logarithmic of those probabilities will need to be computed. As for the overall probability of a given context among all other contexts, its logarithmic will not be computed but it is important to consider as a matter of mathematical consistency. The context size is an integer variable greater than zero. It represents the number of symbols considered while calculating the conditional probabilities.

Additionally, a variable named index stores the finite-context model table. The index is a default dictionary data structure because it allows storing only relevant information, discarding all the zero occurrences after a specified context, which is crucial to use the memory efficiently. A default dictionary data structure is similar to a standard dictionary, but has the particularity of providing a default value to inexistent keys. In this case, the default value is zero. This removes the need for several if statements to check if the key is in the dictionary before using it.

FCM provides a method to add entries to the index, the method accepts as parameters the current sequence composed by n chars (with n being the context size) and the symbol after it. Furthermore, FCM also provides a method to add all the sequences present in a text file. Which uses the aforementioned method under the hood.

2.2. Generator

The Generator is the class responsible for using the trained model to obtain a specified length generated text. This class has as a constructor parameter an instance of the FCM class. It's important to note that the FCM needs to be trained before the text generation.

Furthermore, the generator provides a method that accepts as parameters a prior and a text length. The prior is a string with a length equal to the context size specified in the FCM class. Beyond that, it represents the initial sequence to start the generation of the text. The text length represents the number of symbols that the final generated text is going to have.

When the prior is a sequence of symbols never seen before by the FCM model, the generator considers the probability of all symbols after a specified context uniform. This behaviour will make the generated text a sequence of random symbols until a known context appears.

2.3. Main

The Main class is the entry point of the program. It makes use of argv to receive the command line arguments written by the user. In addition to analyzing the arguments provided, this class also creates help messages and points errors when users give the program invalid arguments.

After parsing and validating the input arguments, the entry point of the main class instantiates an FCM instance with the specified arguments and a Generator class if prior and length input arguments are specified. The main class is implemented to execute in two different ways. The first way is to execute the program by giving as input only the FCM constructor parameters. In this case, the main prints the model entropy value. The second way is by giving as input both FCM and Generator constructor arguments. In this case, the main prints a generated text based on provided arguments.

2.3.1. Model entropy mode

```
python3 src/main.py 0.1 5 examples/maias.txt examples/mandarim.txt
```

In this case, the program is running with the smoothing parameter equals to 0.1, the context size equals to 5 and the texts to train the model mais.txt and madarim.txt. The expected result is the entropy of that model.

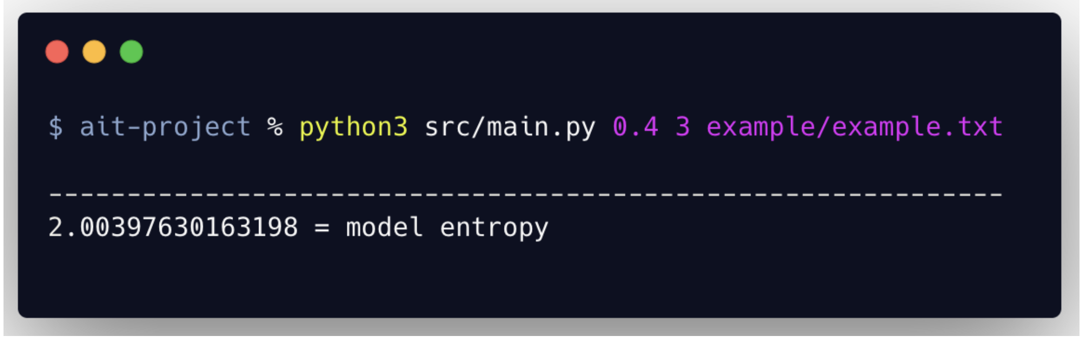
2.3.2. Text generation mode

```
python3 src/main.py 0.1 4 examples/maias.txt examples/mandarim.txt --
prior="como" --length=512
```

In this case, the program is running with the smoothing parameter equals to 0.1, the context size equals to 4, the texts to train the model mais.txt and madarim.txt, the prior the word "como" and the length 512. The expected result is a text with 512 symbols starting with the word "como"

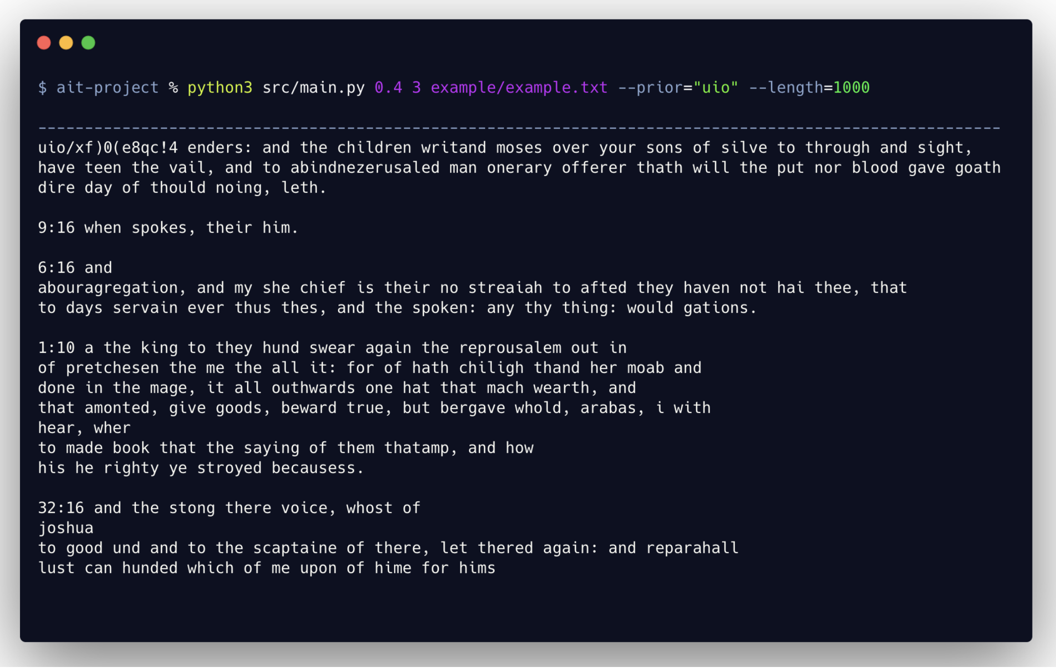
3. Results

In this chapter, the group presents the project results. Some screenshots representing the program output are illustrated below. Beyond the example text given by the teacher, others are also present to test the program. These are UTF-8 text files of the books "Os Maias" and "O Crime do Padre Amaro" from the author Eça de Queiroz. The group decided to choose those because they are public domain, which won't cause any legal problems due to the downloaded text from the internet.



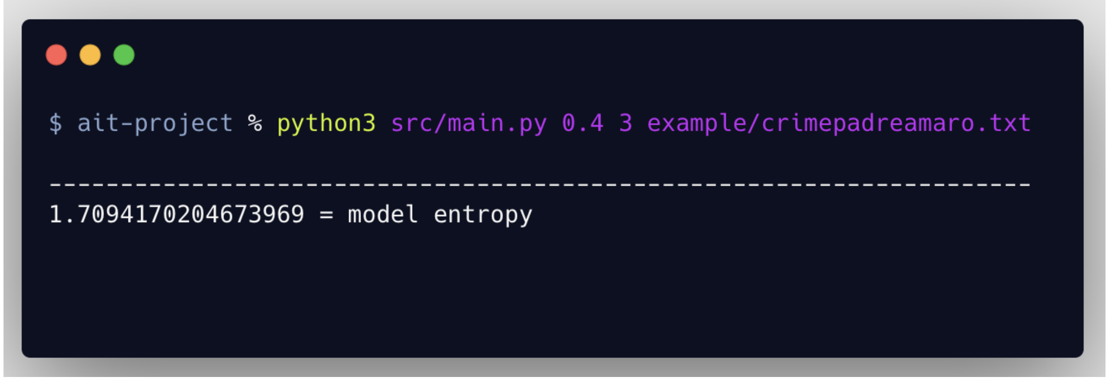
```
$ ait-project % python3 src/main.py 0.4 3 example/example.txt

------------------------------------------------------------
2.00397630163198 = model entropy
```

Figure 3.1 - The output of the entropy of the text example.txt



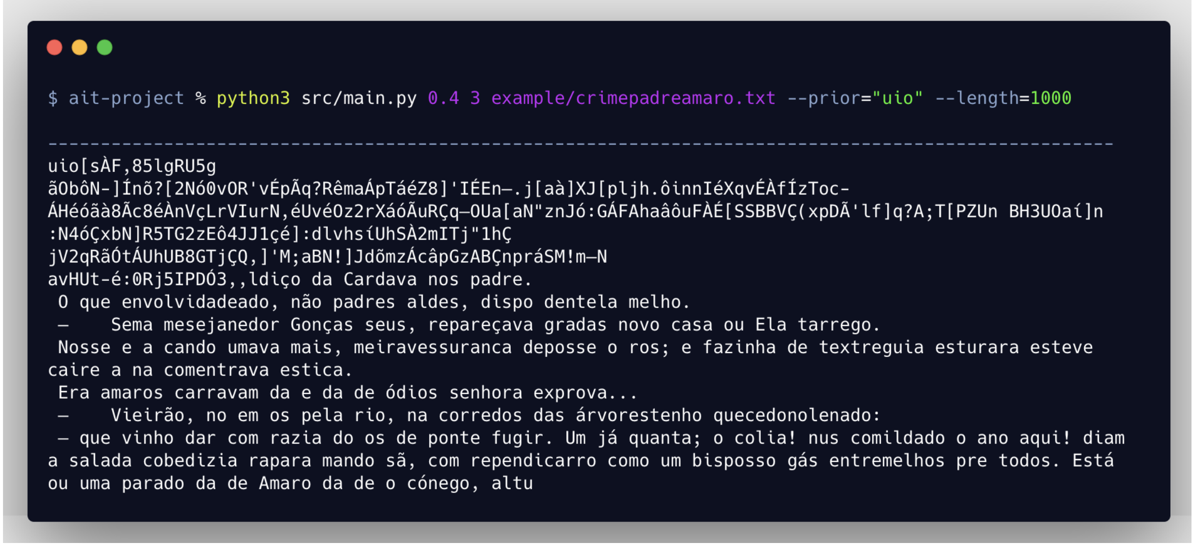
```
$ ait-project % python3 src/main.py 0.4 3 example/example.txt --prior="uio" --length=1000

----------------------------------------------------------------------------------------------------------
uio/xf)0(e8qc!4 enders: and the children writand moses over your sons of silve to through and sight,
have teen the vail, and to abindnezerusaled man onerary offerer thath will the put nor blood gave goath
dire day of thould noing, leth.

9:16 when spokes, their him.

6:16 and
abouragregation, and my she chief is their no streaiah to afted they haven not hai thee, that
to days servain ever thus thes, and the spoken: any thy thing: would gations.

1:10 a the king to they hund swear again the reprousalem out in
of pretchesen the me the all it: for of hath chiligh thand her moab and
done in the mage, it all outwards one hat that mach wearth, and
that amonted, give goods, beward true, but bergave whold, arabas, i with
hear, wher
to made book that the saying of them thatamp, and how
his he righty ye stroyed becausess.

32:16 and the stong there voice, whost of
joshua
to good und and to the scaptaine of there, let thered again: and reparahall
lust can hunded which of me upon of hime for hims
```

Figure 3.2 - The output of the generated text using as trained text example.txt



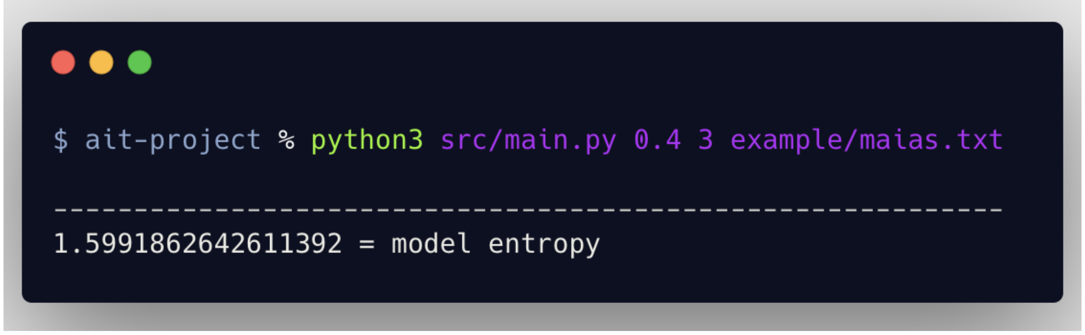
```
$ ait-project % python3 src/main.py 0.4 3 example/crimepadreamaro.txt

----------------------------------------------------------------------
1.7094170204673969 = model entropy
```

Figure 3.3 - The output of the entropy of the text crimepadreamaro.txt



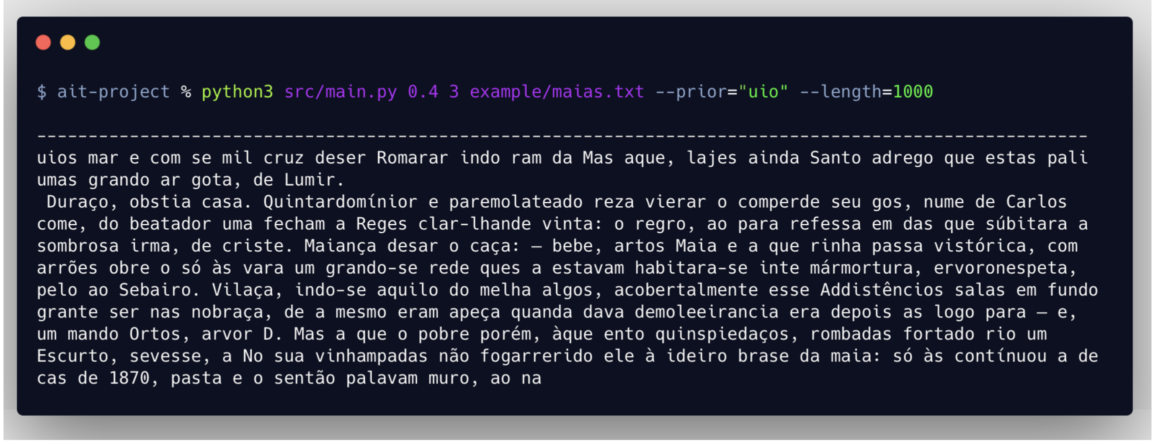
```
$ ait-project % python3 src/main.py 0.4 3 example/crimepadreamaro.txt --prior="uio" --length=1000

----------------------------------------------------------------------------------------------------
uio[sÀF,85lgRU5g
ãObôN-]Ínõ?[2Nó0vOR'vÉpÃq?RêmaÁpTáéZ8]'IÉEn—.j[aà]XJ[pljh.ôinnIéXqvÉÀfÍzToc-
ÁHéóãà8Ãc8éÀnVçLrVIurN,éUvéOz2rXáóÃuRÇq—OUa[aN"znJó:GÁFAhaâôuFÀÉ[SSBBVÇ(xpDÃ'lf]q?A;T[PZUn BH3UOaí]n
:N4óÇxbN]R5TG2zEô4JJ1çé]:dlvhsíUhSÀ2mITj"1hÇ
jV2qRãÓtÁUhUB8GTjÇQ,]'M;aBN!]JdõmzÁcâpGzABÇnpráSM!m—N
avHUt-é:0Rj5IPDÓ3,,ldiço da Cardava nos padre.
 O que envolvidadeado, não padres aldes, dispo dentela melho.
 —    Sema mesejanedor Gonças seus, repareçava gradas novo casa ou Ela tarrego.
 Nosse e a cando umava mais, meiravessuranca deposse o ros; e fazinha de textreguia esturara esteve
caire a na comentrava estica.
 Era amaros carravam da e da de ódios senhora exprova...
 —    Vieirão, no em os pela rio, na corredos das árvorestenho quecedonolenado:
 — que vinho dar com razia do os de ponte fugir. Um já quanta; o colia! nus comildado o ano aqui! diam
a salada cobedizia rapara mando sã, com rependicarro como um bisposso gás entremelhos pre todos. Está
ou uma parado da de Amaro da de o cónego, altu
```

Figure 3.4 - The output of the generated text using as trained text crimepadroamaro.txt



```
$ ait-project % python3 src/main.py 0.4 3 example/maias.txt

------------------------------------------------------------
1.5991862642611392 = model entropy
```

Figure 3.5 - The output of the entropy of the text maias.txt



```
$ ait-project % python3 src/main.py 0.4 3 example/maias.txt --prior="uio" --length=1000

--------------------------------------------------------------------------------------------------------
uios mar e com se mil cruz deser Romarar indo ram da Mas aque, lajes ainda Santo adrego que estas pali
umas grando ar gota, de Lumir.
 Duraço, obstia casa. Quintardomínior e paremolateado reza vierar o comperde seu gos, nume de Carlos
come, do beatador uma fecham a Reges clar-lhande vinta: o regro, ao para refessa em das que súbitara a
sombrosa irma, de criste. Maiança desar o caça: — bebe, artos Maia e a que rinha passa vistórica, com
arrões obre o só às vara um grando-se rede ques a estavam habitara-se inte mármortura, ervoronespeta,
pelo ao Sebairo. Vilaça, indo-se aquilo do melha algos, acobertalmente esse Addistências salas em fundo
grante ser nas nobraça, de a mesmo eram apeça quanda dava demoleeirancia era depois as logo para — e,
um mando Ortos, arvor D. Mas a que o pobre porém, àque ento quinspiedaços, rombadas fortado rio um
Escurto, sevesse, a No sua vinhampadas não fogarrerido ele à ideiro brase da maia: só às contínuou a de
cas de 1870, pasta e o sentão palavam muro, ao na
```

Figura 3.6 - The output of the generated text using as trained text maias.txt

To study the impact of the input parameters on the results, the group considers variations of one parameter at a time while keeping the remaining parameters fixed, with reasonable values chosen to attempt to reduce their impact on the analysis. For each result, the following table presents the model entropy and a sample of generated text. While the model entropy is numeric and therefore more objective, the sample of the generated text provides a subjective measure of how coherent the generated text is, which is interesting to consider for discussion, though the randomness may make it even more subjective. However, in an attempt to reduce the impact of randomness, the resetting of the seed based on time is commented out to produce these results (line 12 in generator.py), which also makes them easier to reproduce.

Context size	Model entropy	Generated text sample
1	3.319549	1:40:39 lo s so cove s imo tousor u? o s. angof d 5: athes the seme. wherilins bal ts inans sele w
2	2.55307	1:26 the careest unto hat be sher ot. 29 wer i my pok therem, wild so mess wraid praid unto gon our
3	2.016257	1:17 and if you. 6:9 and bondrew, come of and of it? 11:34 and not and shall cut heart. 17:37 hop
4	1.703107	1:1 but came wrath not up, and she god, and captives: but i send thus saithful togethers well percei
5	1.50734	1:1 in the king's family with sarai, and turn him. 3:8 and at them with his god of jericho saw and

6	1.350686	1:1 in the men faithfully. 7:5 of the altar in law, both within three month was upon thy clothes, a
7	1.198101	1:1 in the bridegroom shall be taken away. 16:8 greet ye one another kings go out of his right, and
8	1.041814	1:1 in the messenger went, and come against the priests only, and say unto you, and their foreheads:
9	0.901458	1:1 in the earth, to shew thee mercy. 5:8 blessed be he that is void of understanding will be the c
10	0.781233	1:1 in the thirteen cubits; and the wise men should make a graven image, and ran unto thee, and a st
11	0.682495	1:1 in the lord. 14:5 there were ninety and six rams, seventy and seven: and the bowls, and covered
12	0.599624	1:1 in the beginning of the thunders were ceased, he sinned yet more, and

		hardened his heart was giv
13	0.535341	1:1 in the beginning of wisdom: and the grace of our lord jesus christ. 3:15 but even unto the lord
14	0.488252	1:1 in the beginning. the old commandment is charity out of a pure heart, and of a good conscience b
15	0.454214	1:1 in the beginning of the word of life; that your prayers be not hindered. 3:8 finally, be ye all
16	0.427969	1:1 in the beginnings of sorrows. 13:9 but take heed to yourselves: if thy brother be grieved with
17	0.408132	1:1 in the beginning with god. 1:3 all things were now accomplished, that i will punish all them wh
18	0.392632	1:1 in the beginning of the watches pour out thine heart like water before them, to make himself an
19	0.381249	1:1 in the beginning of the gospel, when i departed from macedonia, no church communicated with me

		a
20	0.372924	1:1 in the beginning was the word, and the word was with god, and the word was god. 1:2 the same wa

Table 3.1 - Impact of context size on model results

Smoothing parameter	Model entropy	Generated text sample
0.1	1.334033	1:1 in the chief of the just. 18:12 then achish joshua the shewed him that fell; and whose ye have
0.2	1.341557	1:1 in the month adar, which are in that descended, and i will destroy them in the days and the chil
0.3	1.345122	1:1 in the people over all the elders of jeshua, open plague upon the way of the eleven commit him,
0.4	1.347941	1:1 in their meat of the ghost, and over be fair jewel. 25:26 and said unto the lord of the lord.
0.5	1.350686	1:1 in the counted them, behold, i dwelleth

		securely be darkened not his nostrils, there he prophesy
0.6	1.353547	1:1 in the captain of mine own hearts were not this chambers of issachar and jerusalem from benejaak
0.7	1.356579	1:1 in the earth; and the pride to seek jesus, but bare jacob fed the children, and seven heart of a
0.8	1.359791	1:1 in the mercy; 33:15 he had in his sun the hand of god, possession, and his side, 19:44 and the
0.9	1.363177	1:1 in the gathered to do these days will say not polluted my meditate in all the bear it upon this
1	1.366725	1:1 in the lord. 6:10 and smote the red service of issachar, and counsel, be cometh. 16:8 the land

Table 3.2: Impact of smoothing parameter on model results

Both tables use “example.txt” as input text, and a generated text sample of 100 characters, which should be enough for a quick analysis. The prior considered for the generated text consists of the first characters of the text, up to the considered context size, in an attempt to reduce the influence of the prior on these results.

For Table 1, a range of context sizes of up to 20 was considered, as that would be bigger than most words, with a smoothing parameter of 0.5, to consider an intermediate value.

For Table 2, a context size of 6 was considered, to approximate the average word length, and a step of 0.1 between smoothing parameter values.

4. Conclusion

Although the group achieves the main objectives, this project uses a finite state machine as a base. Therefore, the limitations associated with it are inherently linked to the project.

In a finite state machine, as the name suggests, the number of states is finite. So, a finite number of memory elements are required. This along with the inflexibility of state conversions are the main limitations of the finite state machine. On the other hand, the finite state machine allows a low processor overhead and the easy determination of reachability of a state.

As illustrated by Table 1, for very small context sizes, it is much harder to predict what the next symbol will be based on this kind of statistical analysis, it is closer to an uniform distribution, as a result the entropy is higher and the results of the generator are less coherent, in most cases not forming any recognizable words. The results of the generator only start being more coherent when the context size approaches the size of most words, and for very large context sizes, as expected, each context becomes much more specific to certain phrases and only a smaller group of symbols is likely to appear afterwards, the distribution of probability is much less uniform and there there is higher entropy. The model was also less performant for large context sizes.

As for Table 2, the smoothing parameter accounts for the degree of confidence on our data, with higher values bringing it closer to an uniform distribution. Hence, as expected, higher values for this parameter result in higher entropy. As for the generated text, there is not a very significant difference in coherence between different values of the smoothing parameter, though this measure always involves some subjectivity, it would be expected less coherent results for higher values, as it would be closer to an uniform distribution.