

# CS7.401: Introduction to NLP | Assignment 3 report

Aaditya sharma,2019113009

## Q1)

### 1.1)

Created a neural network model using LSTM.

Steps ->

- I read the data and tokenized it.
- Appended "<S>" at the start of every sentence and "<E>" at the end of every sentence.
- Created a list of unique tokens.
- Then I created a word to index and index to word dictionary for the given dataset.
- Then I initialized my model using ->  
`model = Sequential()`  
Took input\_dim of Embedding layer = `len(unique_tokens)`  
Output\_dim = 25  
Input\_length = 4  
Compiled using adam compiler
- Then I trained my model using `model.fit()` and sent data in batches of 5000.
- In `model.fit()` parameters are x,y,epochs , where x is context(n-1 gram) , y is obtained by doing the one-hot encoding of the target(last term of n gram) and epochs = 10
- I saved my model in a directory called modelno1 and I load the model every time to give the probability of any input sentence using the file `language_model.py`

### 1.2)

To get the perplexity score I used my model (`model.predict(x)`) which took each sentence as a parameter and gave the probability for each sentence.

Then from the probabilities get the perplexity score  $(1/p^{1/n})$

For train data avg perplexity = 276.8357364991011 ->

```
276.8357364991011
Resumption of the session 593.89840316838024
I declare resumed the session of the European Parliament adjourned on Friday 17 December 1999, and I would like once again to wish you a happy new year in the h
Although, as you will have seen, the dreaded 'millennium bug' failed to materialise, still the people in a number of countries suffered a series of natural dise
You have requested a debate on this subject in the course of the next few days, during this part-session. 141.2721167516055
In the meantime, I should like to observe a minute' s silence, as a number of Members have requested, on behalf of all the victims concerned, particularly those
Please rise, then, for this minute' s silence. 97.9457366153805
(The House rose and observed a minute' s silence) 354.02213201938116
Madam President, on a point of order. 54.215576011987
You will be aware from the press and television that there have been a number of bomb explosions and killings in Sri Lanka. 372.9859702909767
One of the people assassinated very recently in Sri Lanka was Mr Kumar Ponnambalam, who had visited the European Parliament just a few months ago. 71.136646970
Would it be appropriate for you, Madam President, to write a letter to the Sri Lankan President expressing Parliament's regret at his and the other violent deat
Yes, Mr Evans, I feel an initiative of the type you have just suggested would be entirely appropriate. 340.6864923727093
If the House agrees, I shall do as Mr Evans has suggested. 496.0087291787746
Madam President, on a point of order. 335.61475354250393
I would like your advice about Rule 143 concerning inadmissibility. 363.8313165784278
My question relates to something that will come up on Thursday and which I will then raise again. 93.94466496834914
The Cunha report on multiannual guidance programmes comes before Parliament on Thursday and contains a proposal in paragraph 6 that a form of quota penalties sh
It says that this should be done despite the principle of relative stability. 272.54004411527626
I believe that the principle of relative stability is a fundamental legal principle of the common fisheries policy and a proposal to subvert it would be legally
I want to know whether one can raise an objection of that kind to what is merely a report, not a legislative proposal, and whether that is something I can compe
That is precisely the time when you may, if you wish, raise this question, i.e. on Thursday prior to the start of the presentation of the report. 488.672932972
Madam President, coinciding with this year' s first part-session of the European Parliament, a date has been set, unfortunately for next Thursday, in Texas in A
At the request of a French Member, Mr Zimeray, a petition has already been presented, which many people signed, including myself. 456.0505695089222
However, I would ask you, in accordance with the line which is now constantly followed by the European Parliament and by the whole of the European Community, to
This is all in accordance with the principles that we have always upheld. 499.7161304215323
Thank you, Mr Segni, I shall do so gladly. 137.29779469780482
Indeed, it is quite in keeping with the positions this House has always adopted. 79.30853487822831
Madam President, I should like to draw your attention to a case in which this Parliament has consistently shown an interest. 151.16886428588168
```

For test data avg perplexity = 360.8359099722957->

```
360.8359099722957
When used preventively, it saves the state and the economy a great deal of money. 411.3522971944419
I have completely failed to understand in this debate why a reasonable set of rules was not adopted back in 1993, especially as the Commission and Parliament did
Seven million workers were affected and specific sectors, such as the mobile worker sector, have been subject to ruinous competition over recent years, especiall
It is therefore also a social problem and it is not enough, Mr Crowley, to use tachographs or other technical aids. 432.01078775140877
One does not exclude the other. 143.28506529890075
We also need a framework directive, because the employees who are affected have been working in a grey zone for a long time now. 150.64616457455696
They had no rules, they were not covered by a collective agreement, they were exploited and some also engaged in self-exploitation. 226.44625766146726
We know that this overload sometimes gave rise to alcohol-related problems. 585.0050615485089
Parliament has shown sufficient flexibility; Mrs Smet highlighted our legislative maturity. 323.00287127179524
I think, Mrs Smet, you have proven that we can also fight. 377.3445131085299
The results are acceptable. 573.3564300524033
I am trying to muster support, even if we have not achieved everything we wanted to. 210.08521649508035
But the transition period and the graduated plan are the maximum we have allowed the Council. 516.7943041726865
I hope that no government will use up the full period of time and am counting on constructive competition between the Member States to see which State will be th
Directive 93/104 was already very restrictive as far as worker production is concerned. 561.3344626879476
It does not harmonise social legislation upwards, quite the opposite. 364.0442914052975
It establishes a European framework which falls a long way short of workers' expectations: an 11-hour daily rest period, a 48-hour working week, a 24-hour weekly
What is more, a number of categories were excluded from its field of competence. 443.44644919618435
We find the compromise, which Parliament is to vote on today following the meetings of the Conciliation Committee, unsatisfactory in terms of both the health and
It makes flexibility more widespread, particularly for sea fishermen, establishing the systematic annualisation of reference periods. 424.950218095984
It will permit further exemptions to the already excessive legislation of 48 hours per week. 519.1955877188984
Finally, it will take nine years, in the best possible case, for the working week of doctors in training to be reduced from 58 hours to 48. 441.44957211737386
In fact, the proposed working time organisation institutes social deregulation at the very time when, in France, doctors in training are campaigning for decent w
Proper organisation of working time would entail, we reiterate, a real reduction in working hours accompanied by moves to create a sufficient number of jobs. 19
We cannot support this report as it stands. 297.81792436422734
Mr President, this proposal will bring an additional seven million workers in Europe under the protection of the Working Time Directive, allowing them to have re
However, to get a 48-hour week, junior doctors must wait nine years minimum and possibly twelve years. 209.0970817360806
I will not pretend to be happy with this situation. 441.65099909819804
It is, however, the best that we could achieve if we were to end once and for all the Council's delaying tactics. 381.9434663071523
It has been obvious since 1993 that those excluded from the directive at that time would eventually come under its protection. 245.46944903893376
```

Q2)

2.1)

Created a sequence to sequence model.

Steps ->

- I read the data (both the french and the english dataset) and tokenized it.

- Created a list of unique tokens for both datasets.
- I calculated the max length of the English dataset and then made each sentence of the dataset equal to that length and did a similar thing for the french dataset.
- Then I created a word to index and index to word dictionary for both the datasets.
- Then I initialized my model using ->  
`model = Sequential()`  
 Took input\_dim of Embedding layer = `len(unique_tokens of english dataset)`  
Output\_dim = 15  
Input\_length = max\_len\_of\_english  
Mask\_zero = True  
RepeatVector parameter = max\_len\_of\_french
- Compiled using adam compiler.
- Then trained my model with `model.fit(x,y,epochs)` with batch size = 100.
- Parameter x is the english sequence , y is obtained by doing one-hot encoding and epochs = 10.
- I saved my model in a directory called model2 and I load the model every time (with word to index and index to word dict) to get the translation of any input sentence using the file `machine_translation.py`

## 2.2)

Then load the model and send the data line by line and get the translation of each line . After getting the translation of each line use - `nltk.translate.bleu_score.sentence_bleu` to get the bleu score of each sentence and `nltk.translate.bleu_score.corpus_bleu` to get the corpus bleu score of all the sentences.

Do it for both M1 model and M2 model.

Got M2 by using weights from the model trained on the English for the Encoder and French task for the Decoder and using this fine-tuned the Encoder and Decoder on the parallel corpus for translation.

**For M1 model ->**

**For train dataset** corpus-level BLEU score = 5.31245843232297e-168->

5.31245843232297e-168  
[ 'BF:', ':', ':', ':', 'que', 'de', 'de' ] 0  
[ 'Et', 'nous', 'que', 'de', 'de', 'de', 'de', 'de', 'de', 'de', 'de', 'de', '?' ] 9.418382295637229e-232  
[ 'Nous', 'nous', 'nous', 'n'a-t-il', 'de', 'de', 'de', 'et', 'et', 'que', 'que', 'que', 'de', 'de', 'de', 'Göutez', 'sang', '!' ] 1.0182922  
[ 'Mais', 'y', 'a', 'de', 'de', 'de', 'de', 'de', 'de', 'de', 'de', 'de', 'de', 'de', 'hanche.', '<title>Maira', 'Dunham-Jones', ':', 'R' ]  
[ 'Et', 'ne', 'que', 'pas', 'que', 'que', 'pas', 'pas', 'de', 'de', 'de' ] 0  
[ '<title>Andreas', 'Raptopoulos', ':', ':', ':', ':', 'de', 'de', 'de', 'ouvrères' ] 6.034940380417626e-232  
[ '<title>Mark', 'Shaw', ':', 'Renover', 'nait.', 'impermeable<title>' ] 0  
[ '<title>Maira', 'Kalmán', 'Kitflyer', 's', 'Kitflyer', 's' ] 0  
[ 'Et', 'il', 'a', 'de', 'de', 'de', 'que', 'que', 'que', 'article', 'article', 'Journal', 'Journal', 'Médecine', 'Médecine', 'Et', 'une', 'la' ]  
[ '<title>Jennifer', 'Killingsworth', 'Lee', 'l'ingéniosité', 'de', 'de', 'de', 'de', 'la' ] 3.865846993487796e-155  
[ 'Les', 'Sissay', 'des', 'le', 'de' ] 5.9764165035645784e-232  
[ '<title>Nate', 'Iwasaki', ':', 'JR', 'veau', 'veau', 'Prix', 'Prix', 'Prix', 'de', 'et', 'et', 'fois...', 'fois...', 'hypothèse<title>' ] 8.432010639666965e-232  
[ 'Les', 'avons', 'de', 'de', 'de', 'de', 'de', 'de', 'de', 'de', 'de' ] 0  
[ 'Et', 'nous', 'a', 'de', 'de', 'de', 'de', 'de', 'et', 'et', 'et', 'et', 'et', 'et', 'et', 'et', 'de', 'de' ] 3.3864995705940004e-155  
[ 'Mais', 'Hypponen:', ':', 'vous', 'vous', ':', ':', ':', ':', ':', ':', ':', 'que', 'que', 'de', 'de', '<speaker>Nic', 'Marks</speaker>' ] 1.2508498911928379  
[ 'Nous', 'avons', 'les', 'annuel', 'de', 'de', 'de', 'de', 'de', 'C02?'] 0  
[ 'Nous', 'Besser', 'les', 'de', 'de', 'de', 'de', 'de', 'de', 'de', 'de', 'et', 'et', 'et', 'et', 'et', 'de', 'analysés', 'analysés' ] 9.48832896503462e-  
[ 'Et', ':', 'a', 'a', 'de', 'de', 'de', 'et', 'et', 'et', 'et', 'et', 'a', 'et', 'et', 'et', 'la' ] 9.997801362989555e-232  
[ 'Mais', 'y', 'a', 'soucie', 'de', 'de', 'de', 'de', 'de', 'de', 'de', 'de', 'de', 'de', 'de', 'Tosca', 'Tosca', 'de', 'Nous', 'avons',  
[ 'Beijing' ] 0  
[ 'Alors', 'la', 'souhaitons', 'de', 'de', 'de', 'de', 'primitifs', 'de', 'de', 'de' ] 6.8949282624564e-232  
[ 'Et', 'la', 'de', 'de', 'de', 'de', 'de', 'de', 'de', 'l'humanité<title>' ] 5.622528097505664e-232  
[ 'Comment', 'la', 'les', 'de', 'de', 'société?' ] 8.416851712392762e-232  
[ '<title>David', 'Dunham-Jones', ':', 'Renover', 'le', 'banlieues<title>' ] 0  
[ 'Et', 'la', 'de', 'de', 'de', 'de', 'de', 'de' ] 6.744160953836975e-232  
[ 'Comment', 'étape:', 'des', 'le', 'nait.', 'sacrifier?'] 8.34076112986429e-232  
[ 'Comment', 'que' ] 0

**For test dataset** corpus-level BLEU score = 3.283495847334756e-220->

[illegible]

**For M2 model ->**

**For train dataset** corpus-level BLEU score = 2.155768947364586e-121->

```
p.155768947364586e-121
['David', 'Gallo', 'Voici', 'Bill', 'Lange', 'Je', 'suis', 'Dave', 'Gallo'] 0
['Nous', 'allons', 'vous', 'raconter', 'quelques', 'histoires', 'de', 'la', 'mer', 'en', 'vid', 'o'] 8.038258993350412e-232
['Nous', 'avons', 'des', 'vid', 'os', 'du', 'Titanic', 'parmi', 'les', 'plus', 'spectaculaires', 'jamais', 'vues', 'et', 'nous', 'n', 'allons', 'pas', 'vous', 'la', 'v', 'rit', 'est', 'que', 'le', 'Titanic', 'm', 'me', 's', 'il', 'continue', 'de', 'battre', 'toutes', 'les', 'records', 'de', 'recettes', 'n', 'est', 'p', 'le', 'probl', 'me', 'je', 'crois', 'est', 'qu', 'on', 'tient', 'l', 'oc', 'an', 'pour', 'acquis'] 0
['Quand', 'vous', 'y', 'pensez', 'les', 'oc', 'ans', 'repr', 'sentent', '75', 'de', 'la', 'plan', 'te'] 8.396161215621529e-232
['La', 'plus', 'grande', 'partie', 'de', 'la', 'plan', 'te', 'est', 'd', 'eau'] 0
['La', 'profondeur', 'moyenne', 'est', 'environ', '3', '2', 'km'] 4.94660716462899e-232
['Une', 'partie', 'du', 'probl', 'me', 'je', 'pense', 'est', 'qu', 'en', 'tant', 'sur', 'la', 'plage', 'ou', 'en', 'regardant', 'des', 'images', 'de', 'l', 'oc']
['Il', 'y', 'existe', 'les', 'cha', 'nes', 'de', 'montagnes', 'les', 'plus', 'longues', 'de', 'la', 'plan', 'te'] 6.034940380417626e-232
['La', 'plupart', 'des', 'animaux', 'se', 'trouvent', 'dans', 'les', 'oc', 'ans'] 0
['La', 'plupart', 'des', 'tremblements', 'de', 'terre', 'et', 'de', 'volcans', 'se', 'produisent', 'dans', 'la', 'mer', 'au', 'fond', 'de', 'la', 'mer'] 7.5671
['La', 'biodiversit', 'et', 'la', 'bi densit', 'marines', 'sont', 'plus', 'lev', 'es', 'que', 'dans', 'les', 'for', 'ts', 'tropicales'] 4.752869108205246e-232
['C', 'est', 'pour', 'la', 'plupart', 'inexplor', 'et', 'pourtant', 'il', 'y', 'a', 'de', 'belles', 'vues', 'comme', 'celles', 'ci', 'qui', 'nous', 'captivent', 'Mais', 'quand', 'vous', 'tes', 'la', 'plage', 'imaginez', 'que', 'vous', 'tes', 'au', 'pied', 'd', 'un', 'monde', 'tout', 'fait', 'inconnu'] 7.06030186810811
['Il', 'nous', 'faut', 'disposer', 'd', 'une', 'technologie', 'sp', 'ciale', 'pour', 'p', 'n', 'trrer', 'dans', 'ce', 'monde', 'inconnu'] 7.463640991159612e-232
['Nous', 'utilisons', 'le', 'sous', 'marin', 'Alvin', 'et', 'des', 'cam', 'ras', 'les', 'cam', 'ras', 'ont', 't', 'cr', 'es', 'par', 'Bill', 'Lange', 'avec', 'l', 'Marcel', 'Proust', 'a', 'dit', 'le', 'v', 'ritable', 'voyage', 'de', 'd', 'couverte', 'ne', 'consiste', 'pas', 'chercher', 'de', 'nouveaux', 'paysages', 'mais', 'Nos', 'partenaires', 'nous', 'ont', 'donn', 'de', 'nouveaux', 'yeux', 'non', 'seulement', 'sur', 'ce', 'qui', 'existe', 'les', 'nouveaux', 'paysages', 'au', 'Voici', 'une', 'm', 'duse'] 0
['C', 'est', 'une', 'de', 'mes', 'pr', 'f', 'r', 'es', 'car', 'elle', 'a', 'toutes', 'sortes', 'de', 'parties', 'mobiles'] 8.434560652451766e-232
['Il', 'se', 'trouve', 'que', 'celle', 'ci', 'est', 'l', 'animal', 'plus', 'grand', 'de', 'l', 'oc', 'an'] 6.068009947958691e-232
['Elle', 'peut', 'atteindre', 'jusqu', '45', 'm', 'tres', 'de', 'long'] 0
['Mais', 'voyez', 'vous', 'tous', 'ces', 'bras', 'en', 'mouvement'] 4.952072620509839e-232
['Ils', 'ont', 'ces', 'leurres', 'de', 'p', 'che', 'au', 'dessus', 'Ils', 'montent', 'et', 'descendent'] 6.325072941044999e-232
['Ils', 'ont', 'des', 'tentacules', 'ballants', 'tourbillonnant', 'comme', 'a'] 0
```

For test dataset corpus-level BLEU score =  
4.194058776859423e-215->

```
4.194058776859423e-215
['Quand', 'j', 'avais', 'la', 'vingtaine', 'j', 'ai', 'vu', 'mes', 'tout', 'premiers', 'clients', 'comme', 'psychoth', 'rapeute'] 0
['J', 'tais', 'tudiante', 'en', 'th', 'se', 'en', 'psychologie', 'clinique', 'Berkeley'] 8.81787932857899e-232
['Elle', 'c', 'tait', 'une', 'femme', 'de', '26', 'ans', 'appel', 'e', 'Alex'] 0
['Lorsqu', 'Alex', 'est', 'entr', 'e', 'pour', 'sa', 'premi', 're', 's', 'ance', 'elle', 'portait', 'un', 'jean', 'et', 'un', 'grand', 'top', 'trop', 'large', 'Lorsque', 'j', 'ai', 'entendu', 'a', 'j', 'ai', 't', 'si', 'soulag', 'e'] 8.319100378795605e-232
['Ma', 'camarade', 'de', 'classe', 'avait', 'eu', 'un', 'pyromane', 'comme', 'premier', 'patient'] 1.2183324802375697e-231
['Et', 'moi', 'j', 'avais', 'une', 'fille', 'de', '28', 'ans', 'et', 'quelques', 'qui', 'voulait', 'parler', 'des', 'gar', 'ons'] 0
['Je', 'pensais', 'pouvoir', 'g', 'rer', 'a'] 4.580373270253951e-232
['Mais', 'je', 'ne', 'l', 'ai', 'pas', 'g', 'r'] 1.0120710421309996e-231
['Avec', 'les', 'histoires', 'amusantes', 'qu', 'Alex', 'ramenait', 'durant', 'les', 'sessions', 'c', 'tait', 'facile', 'pour', 'moi', 'de', 'simplement', 'hoc', 'La', 'trentaine', 'c', 'est', 'la', 'nouvelle', 'vingtaine', 'disait', 'Alex', 'et', 'pour', 'ce', 'que', 'j', 'en', 'savais', 'elle', 'avait', 'raison'] 0
['Le', 'travail', 'arrive', 'plus', 'tard', 'le', 'mariage', 'arrive', 'plus', 'tard', 'les', 'enfants', 'arrivent', 'plus', 'tard', 'm', 'me', 'la', 'mort', 'Les', 'jeunes', 'dans', 'la', 'vingtaine', 'comme', 'Alex', 'et', 'moi', 'avons', 'toute', 'la', 'vie', 'devant', 'nous'] 8.560055379685596e-233
['Mais', 'peu', 'de', 'temps', 'apr', 's', 'mon', 'directeur', 'de', 'th', 'se', 'm', 'a', 'pouss', 'questionner', 'Alex', 'sur', 'sa', 'vie', 'amoureuse'] 0
['J', 'ai', 'protest'] 9.788429383461836e-232
['J', 'ai', 'dit', 'Oui', 'elle', 'sort', 'avec', 'des', 'idiots', 'elle', 'couche', 'avec', 'un', 'cr', 'tin', 'mais', 'ce', 'n', 'est', 'pas', 'qu', 'elle', 'Alors', 'mon', 'directeur', 'a', 'r', 'pondu', 'Pas', 'encore', 'mais', 'elle', 'pourrait', 'pouser', 'le', 'prochain'] 8.726094729337945e-232
['De', 'plus', 'le', 'meilleur', 'moment', 'pour', 'pr', 'parer', 'le', 'mariage', 'd', 'Alex', 'c', 'est', 'avant', 'qu', 'elle', 'ne', 'le', 'fasse'] 9.2573
['C', 'est', 'ce', 'que', 'les', 'psychologues', 'appellent', 'l', 'instant', 'de', 'r', 'v', 'lution'] 0
['C', 'est', 'le', 'moment', 'o', 'j', 'ai', 'r', 'alis', 'que', 'la', 'trentaine', 'n', 'est', 'pas', 'la', 'nouvelle', 'vingtaine'] 0
['Oui', 'les', 'gens', 'se', 'casent', 'plus', 'tard', 'qu', 'auparavant', 'mais', 'a', 'ne', 'fait', 'pas', 'de', 'la', 'vingtaine', 'd', 'Alex', 'un', 'temps', 'a', 'fait', 'de', 'la', 'vingtaine', 'd', 'Alex', 'le', 'meilleur', 'moment', 'pour', 'son', 'd', 'veloppement', 'et', 'nous', 'tions', 'assises', 'l', 'le', 'C', 'est', 'ce', 'moment', 'que', 'j', 'ai', 'compris', 'que', 'cette', 'n', 'gligence', 'anodine', 'tait', 'un', 'vrai', 'probl', 'me', 'et', 'qu', 'il', 'a', 'Il', 'y', 'a', 'actuellement', '50', 'millions', 'de', 'personnes', 'dans', 'la', 'vingtaine', 'aux', 'Etats', 'Unis'] 2.6401949804733037e-232
['C', 'est', 'dire', 'environ', '15', 'pour', 'cent', 'de', 'la', 'population', 'ou', '100', 'pour', 'cent', 'si', 'on', 'consid', 're', 'que', 'personne', 'ne', 'Levez', 'la', 'main', 'si', 'vous', 'avez', 'la', 'vingtaine'] 0
['Je', 'voudrais', 'vraiment', 'en', 'voir', 'quelques', 'uns', 'ici'] 2.975482570578e-233
['Oh', 'ouais', 'Vous', 'tes', 'g', 'niaux'] 0
```

We got a very small bleu score for most of the sentences and the reason behind this is that due to computational barriers we can not train our model on a lot of epochs and to get considerable it will take a **LOT** of time. So the model which I am submitting is not very good but if trained for a long time will do its job nicely.