Text

Description automatically generated with medium confidence

Statistical Natural Language Processing

ICS2203

Programmed using Python

By Mark Dingli

20703H

**Table of contents**

Statement of Completion …………………………………………………………………………………… 3

Introduction …………………..…………………………………………………………………………………… 4

Design ……….…………………..…………………………………………………………………………………… 5

Observations .………………..…………………………………………………………………………………. 35

Conclusion ..…………………..…………………………………………………………………………………. 36

**Introduction**

In this documentation, I will be presenting three distinct N-gram models created using the (Baby) British National Corpus as the basis. The models include unigram, bigram, and trigram, implemented using a vanilla language model, a Laplace smoothing language model, and a language model with unknown word handling. Additionally, the documentation will provide a detailed explanation of the evaluation metrics used to compare the models' performances, including perplexity and accuracy scores. Furthermore, the advantages, testing and limitations of each model will be discussed, along with suggestions for possible improvements or modifications. The primary aim of this documentation is to provide an overview of the implementation process for each model, as well as to test and compare their respective performances on the given corpus.

**Imports**

Various Python libraries were imported and utilized. The Natural Language Toolkit (nltk) library was imported for text processing, including tokenization, lemmatization. The RegexpTokenizer class from the nltk.tokenize module was used to tokenize the text. The WordNetLemmatizer class from the nltk.stem module was used to perform lemmatization. The sklearn.model\_selection library was used to split the data into training and testing sets. Additionally, psutil library was imported to monitor the system's memory usage during the execution of the program.

**Extracting and Pre-processing the Selected Corpus**

**Description**

The corpus used in this project is the (Baby) British National Corpus, which is a collection of written and spoken texts from a variety of sources [1].

Initially, a single XML file from the corpus was used to develop and test the pre-processing steps. Once the pre-processing steps were finalized, the entire corpus was used for N-gram modelling.

Before the corpus was used for N-gram modelling, it underwent several pre-processing steps to ensure that it was ready for analysis. Firstly, all XML markup was removed from the corpus using the ElementTree module [2]. Secondly, all non-alphabetic characters, such as numbers and punctuation, were removed using regular expressions [3]. Lastly, the corpus was tokenized using the RegexpTokenizer class from the nltk.tokenize module [4]. Additionally, the corpus text was converted to lowercase and lemmatized using the WordNetLemmatizer class from the nltk.stem module, which maps words to their base form.

It should be noted that the entire corpus was used for this project, which consisted of 4,003,540 words in total. The size of the corpus allows for a comprehensive analysis of the language patterns within it, which is essential for the creation of effective N-gram models.

**Output**

The output of this code is a list of tokens, which are the individual words or phrases that make up the text data. These tokens have been pre-processed to remove URLs, convert text to lowercase, expand contractions, remove punctuation and numbers, and lemmatize the remaining tokens [5]. The list of tokens is printed to the console for the first 100 tokens. The total size of the extracted corpus (the number of tokens) is also outputted.

Example output of list of tokens:

A screenshot of a computer

Description automatically generated with low confidence

Figure 1 - Output showing the list of the first 100 tokens

Example output of corpus size:



Figure 2 - Output showing the total size of the extracted corpus

**Testing**

It is important to note that the pre-processing steps used in this code were selected after testing and evaluating different techniques for tokenizing and cleaning the corpus. These steps were chosen as they produced the most effective results for N-gram modelling.

The code performs several checks to ensure that the specified directory exists, contains XML files, and that each XML file can be parsed before tokenizing its text. If any errors occur, the code prints an error message to the console. Additionally, an error message is printed if the regular expression used to tokenize the text is invalid.

**Limitations**

While this code provides an effective way to pre-process the corpus, it may be slow for very large corpora, as it needs to traverse all files in the specified directory and subdirectories. Additionally, the pre-processing steps used in this code may not be suitable for all types of text data, and may need to be modified depending on the nature of the corpus.

**Splitting the corpus**

**Description**

In order to evaluate the performance of the N-gram models, it is necessary to split the corpus into a train set and a test set. The train set will be used to train the model, while the test set will be used to evaluate its performance. After pre-processing, the corpus was split into training and testing sets using the train\_test\_split method from the sklearn.model\_selection library [8]. The test size parameter is set to 0.2, meaning that 20% of the tokens will be used for testing and the remaining 80% will be used for training.

The split of 80% training data and 20% testing data is a common split in machine learning tasks, including N-gram language modelling. This split is generally a good starting point, as it provides enough data for the model to learn from while also allowing for a sufficient evaluation of the model's performance [6][7].

After splitting the tokens, the code defines the vocabulary from the training set. This is done by iterating over each sentence in the train set and adding each word to a set called vocab. The vocab set is then converted to a list and sorted. The vocabulary is defined by creating a set of all the unique words that appear in the training set. Then, the set is converted to a list and sorted alphabetically. This list of words is used to define the vocabulary for the N-gram models. The purpose of creating the vocabulary is to identify the set of words that the model can use to generate text. The vocabulary is used during the training process to generate the probability distribution of words in the corpus. During the training phase of the N-gram models, the vocabulary is used to compute the probabilities of each word in the corpus. This probability distribution is then used to generate the most probable next word in the sequence during text generation [9].

**Output**

The code prints the size of the train set and the test set using the len() function.

Example output of corpus split:



Figure 3 - Output showing the size of the two split sets

**Testing**

It is important to note that the size of the train and test sets can be adjusted depending on the size of the corpus and the desired level of accuracy. In this case, the train set was set to 80% of the corpus and the test set was set to 20%. The train set size is 3,202,832 tokens and the test set size is 800,708 tokens.

**Limitations**

The size of the corpus and the desired level of accuracy can affect the size of the train and test sets. In addition, if the corpus is very large, the code may take longer to run when splitting the corpus [7].

**Vanilla Language model**

**Description**

**Linear Interpolation**

**Description**

The function first sets the values for lambda\_3, lambda\_2, and lambda\_1, which are the interpolation weights for the trigram, bigram, and unigram models (trigram = 0.6; bigram = 0.3; unigram = 0.1). The sentence is then tokenized using the split() method. The probabilities of each n-gram model are then calculated for the sentence using the corresponding language model specified by lm\_type. The probabilities are stored in unigram\_probs, bigram\_probs, and trigram\_probs lists.

Finally, the probability of the sentence is calculated using linear interpolation [10]. The sentence\_prob variable is initialized to 1, and for each token in the sentence, the probability of the token is multiplied by the corresponding interpolation weight raised to the power of lambda\_1. The resulting product is then multiplied by sentence\_prob.

**Output**

The output shows the probability of each test sentence using three different types of N-gram models: Vanilla, Laplace, and UNK. The probabilities calculated for each sentence are different for each model, which shows that the model type has an impact on the probability assigned to a given sentence. The Vanilla LM assigned the highest probability to the test sentence, while the Laplace LM assigned the lowest probability. The UNK LM assigned a probability in between the Vanilla and Laplace models. These results also highlight the trade-offs between different types of N-gram models, such as the trade-off between accuracy and robustness when using the Laplace model versus the Vanilla model.

Example output of probability of test sentence with every model:

A screenshot of a computer

Description automatically generated with medium confidence

Figure 4 - Output showing sentence probability

**Testing**

It should be noted that the code has been tested on other test sentences as well, and the results have shown that the probability assigned to each sentence is consistent with the type of N-gram model used. If a word in the sentence is not in the vocabulary of the N-gram model, the code will encounter an error, as it will not be able to calculate the probability of that word. Therefore, it is important to ensure that the vocabulary used for the N-gram model includes all the words that may appear in the test sentences.

**Limitations**

The linear interpolation method does not take into account the context of the sentence beyond the N-gram model being used. For example, if a sentence contains a rare word that is not present in the N-gram model's vocabulary, the model may assign a low probability to the sentence even if the context of the sentence suggests a higher probability. This can result in inaccurate probability assignments for certain sentences and affect the overall accuracy of the model.

**Evaluation and Perplexity**

**Description**

The perplexity values are numerical representations of how well the language model predicts the test set, with lower values indicating better performance [11]. The calculate\_perplexity function takes in a test set, a dictionary of n-gram language model probabilities, and the type of n-gram model (unigram, bigram, or trigram). It iterates over the test\_set and calculates the log probability of each n-gram in the test set using the corresponding n-gram model probabilities [12]. It then calculates the entropy and perplexity of the test set using the log probabilities and returns the perplexity. The perplexities for each model and n-gram are printed in a table format.

**Output**

The output is a table that shows the perplexities for each of the three language models (Vanilla, Laplace, and UNK) and for each n-gram order (Unigram, Bigram, and Trigram).

In this case, the Vanilla language model has the lowest perplexities for the Bigram and Trigram orders, while the UNK language model has the lowest perplexity for the Unigram order. This indicates that the Vanilla language model performs well in predicting the next word given the previous two words, while the UNK model performs well in predicting the next word based on only the previous word [13]. The Laplace model generally has the highest perplexity values, indicating that it is the least accurate of the three models in predicting the test set.

Example output of the perplexities:

Shape

Description automatically generated

Figure 5 – Table output of perplexity

**Testing**

It should be noted that the code has only been tested on a single test set, and it would be beneficial to test the code on multiple test sets to ensure the accuracy and robustness of the language models. Additionally, the current test set may not be representative of all possible inputs, which could impact the accuracy of the language models. Therefore, it is important to test the code on multiple test sets to ensure that the models are generalizable and can perform well on a variety of inputs [13].

**Limitations**

One limitation of the perplexity measure is that it assumes that the test set is drawn from the same distribution as the training set. This assumption may not always hold, particularly if the test set contains rare or previously unseen words that were not present in the training set. In such cases, the perplexity measure may not accurately reflect the performance of the language model.

**Generation**

**Description**

This code defines three functions to generate a sentence using an n-gram language model and a given phrase. The generate\_sentence() function loops through each n-gram type and calls the generate\_ngram\_sentence() function to generate a sentence using the selected model, phrase, and n-gram type. The generate\_ngram\_sentence function generates a sentence using a specific n-gram type and language model. It first sets start and end tokens for the sentence, and a maximum length. It then initializes the sentence with the start token and the words in the input phrase. The model probabilities are determined based on the selected model, and the sentence is generated by adding words one at a time until the end token is reached or the maximum length is reached.

The generate\_word function generates a word given a probability distribution. If no context is provided, it uses the full vocabulary and corresponding probabilities. If a context is provided, it uses only the vocabulary and probabilities corresponding to that context. It should be noted that the generated sentences may not always make sense or be grammatically correct, as they are generated based solely on the language model probabilities and the input phrase.

**Output**

The output of the code is the generated sentence for each n-gram type (unigram, bigram, and trigram) using the specified language model and input phrase. The input phrase is appended to the start token ("<s>") to initiate the sentence generation. The generated sentence is limited to a maximum length of 20 words and is terminated when either the end token ("</s>") is reached or the maximum length is exceeded. The output demonstrates that the generated sentences can vary widely depending on the selected n-gram order and language model.

Example output of sentence generation:

A screenshot of a computer

Description automatically generated with medium confidence

Figure 6 - Output showing sentence generation

**Testing**

The code below is used for testing the sentence generation function with different language models and phrases. It iterates through a list of models and phrases and generates a sentence for each combination using the generate\_sentence() function. The generated sentences are then printed in the console along with a separator line. This testing approach allows for a quick evaluation of how well the different language models can generate sentences given various phrases.

# Test the sentence generation function with different models and phrases

models = ['Vanilla', 'Laplace', 'UNK']

phrases = ['this is a', 'i love', 'the weather is']

for model in models:

for phrase in phrases:

generate\_sentence(model, phrase)

print("-" \* 125)

It is important to note that during the testing process, some of the generated sentences contained nonsensical words or phrases. This is likely due to the probabilistic nature of the language model and the fact that it is generating words based on probabilities derived from the training data.

**Limitations**

One limitation of the sentence generation function is that it relies solely on the n-gram language models to generate sentences, which may not always result in coherent or grammatically correct sentences. The function uses probabilistic models to generate the next word based on the previous words, and this approach may not always capture the complexity of language and context. Additionally, the function may generate sentences that contain rare or previously unseen words, which may not be familiar to the reader and could impact the overall comprehensibility of the generated sentences.

**Sen\_Propability**

**Description**

The sen\_probability function calculates the probability of a given sentence using a specified language model and n-gram type. The input sentence is first tokenized into words, and then the function calculates the probability of the sentence for each model and n-gram type. The function supports three different models (Vanilla, Laplace, and UNK) and three different n-gram types (unigram, bigram, and trigram). The probability for each model and n-gram type is calculated by iterating through each word in the sentence and multiplying the appropriate probabilities for that word based on the selected model and n-gram type.

**Output**

The output of the sen\_probability function for the given sentence "lets go to the bank" and the selected language model "UNK" and n-gram type "bigram" is a probability value of 1. This may indicate that the language model is too confident in its prediction and may not be accurately capturing the complexity of the language.

Example output of the sen\_probability function:



Figure 7 – sen\_probability output

**Testing**

Please note that the implementation of the sen\_probability function in this code did not properly handle sentence segmentation. The function tokenizes the input sentence into individual words, which may not reflect the true probabilities of the sentence as a whole. Additionally, the test and train corpora were not split by sentence, which may further affect the accuracy of the language model. Future improvements to this code may involve proper sentence segmentation and splitting the corpora into sentences to improve the accuracy of the model.

**Limitations**

The function assumes that the sentence is properly tokenized and split into words, which may not always be the case. Finally, the function only considers probabilities based on the selected n-gram type and does not take into account the overall language model performance on the entire corpus. This can lead to inaccuracies if the sentence being evaluated deviates significantly from the training data.

**References**

[1] <https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2553>

[2] <https://docs.python.org/3/library/xml.etree.elementtree.html>

[3] <https://docs.python.org/3/library/re.html>

[4] <https://www.nltk.org/api/nltk.stem.html#nltk.stem.wordnet.WordNetLemmatizer>

[5] <https://towardsdatascience.com/text-preprocessing-in-natural-language-processing-using-python-6113ff5decd8>

[6] <https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/>

[7] <https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation>

[8] <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html>

[9] <https://www.kdnuggets.com/2019/11/create-vocabulary-nlp-tasks-python.html>

[10] <https://stackoverflow.com/questions/60681202/getting-smoothed-probabilities-using-linear-interpolation-for-nlp-model>

[11] <https://medium.com/nlplanet/two-minutes-nlp-perplexity-explained-with-simple-probabilities-6cdc46884584>

[12] <https://huggingface.co/docs/transformers/perplexity>

[13] <https://www.techslang.com/perplexity-in-nlp-definition-pros-and-cons/#:~:text=The%20dictionary%20defines%20perplexity%20as,Updated%20September%2020%2C%202022>

[14]