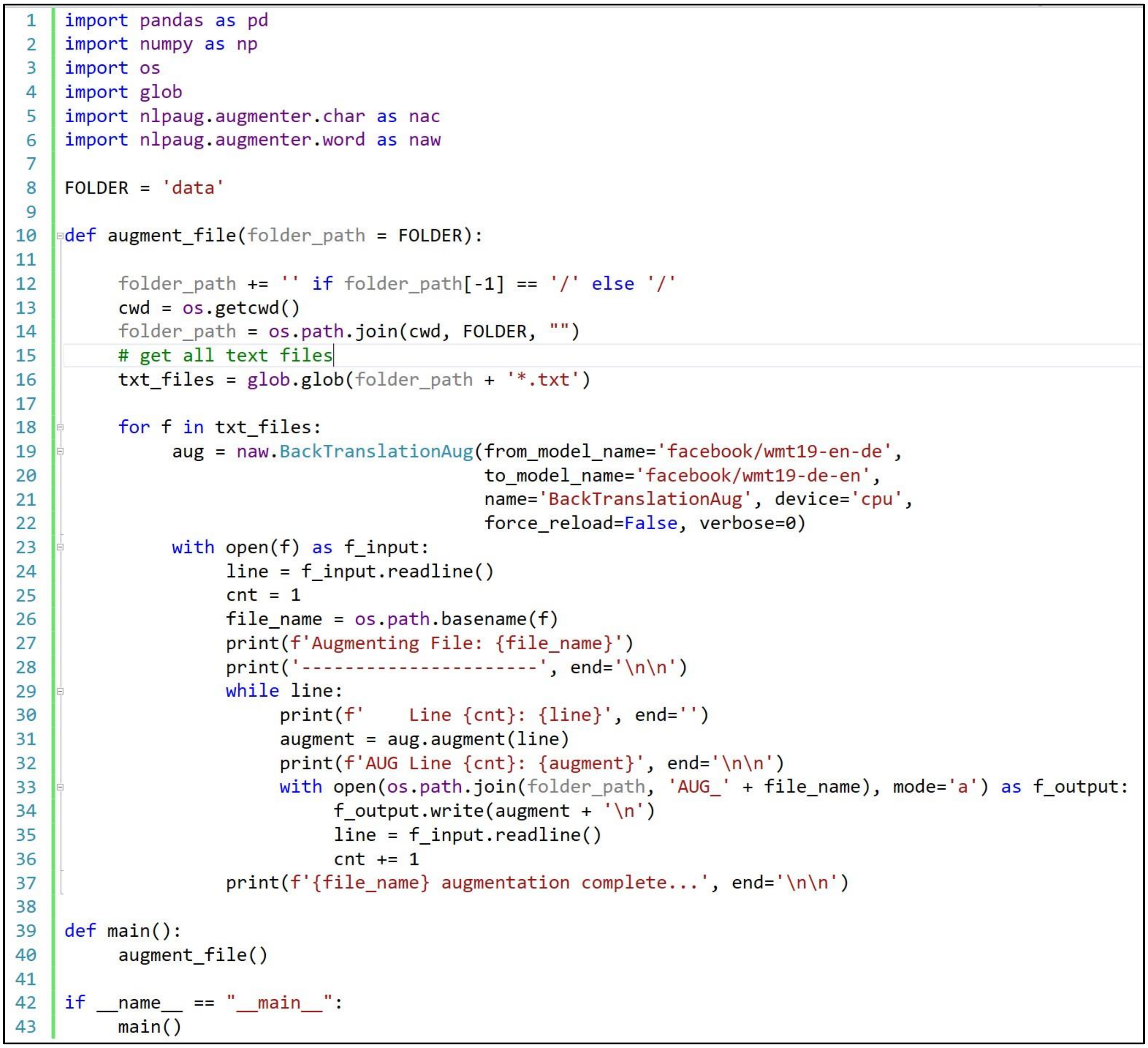
OPTION #1: Text Dataset Augmentation

Scott Miner

Colorado State University – Global Campus

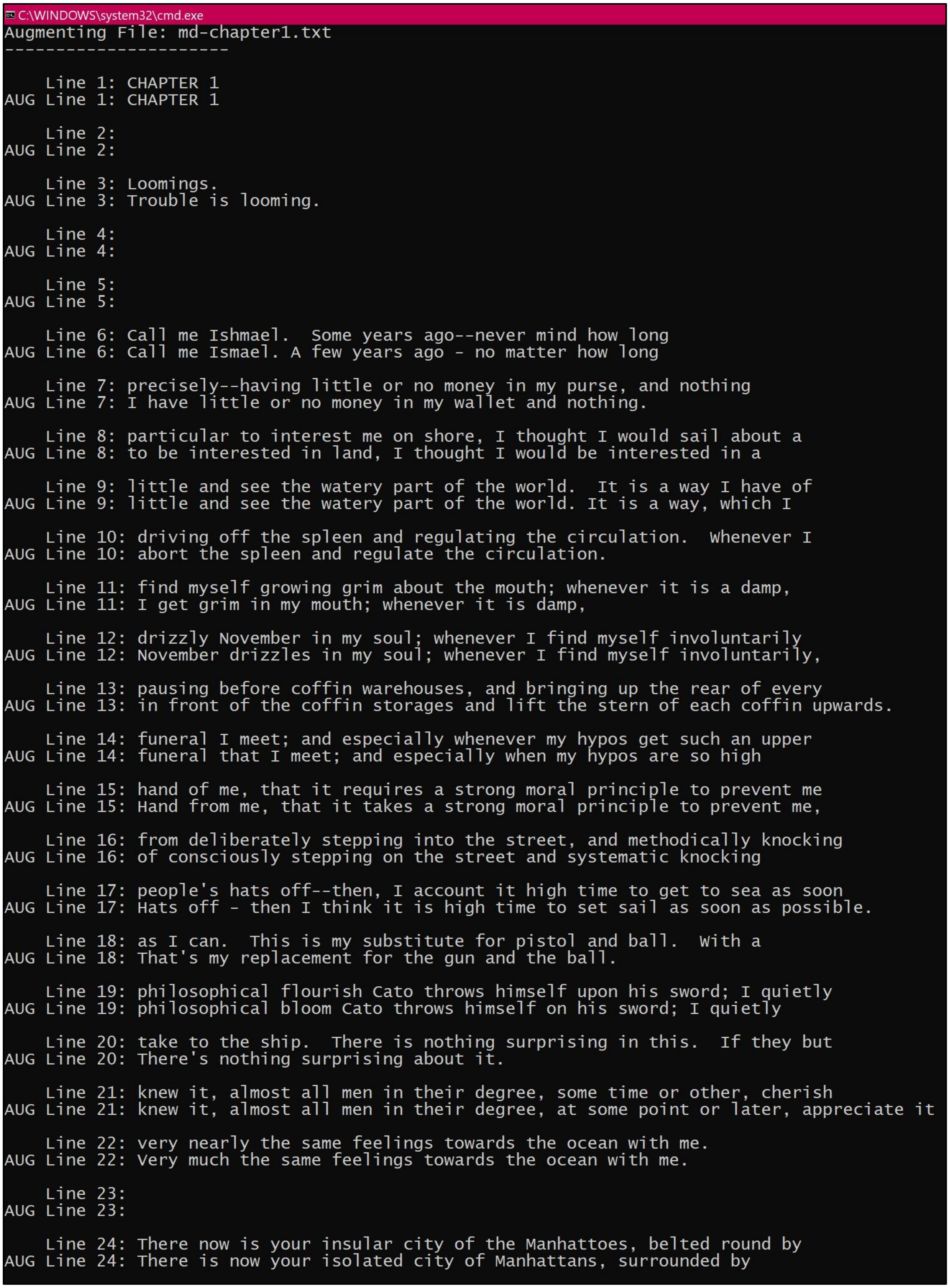
Abstract

This paper presents a Python program that employs round-trip translation (RTT) for text augmentation, expanding the dataset for natural language processing (NLP) tasks. The program reads and augments each line of TXT files in the "data/" folder using the NLPAug Python library's back-translation augmenter. We explore various text augmentation techniques, including synonym augmentation, semantic similarity augmentation, and RTT. The effectiveness of the program is demonstrated by augmenting the classic text, Moby Dick, and comparing the original and back-translated versions side by side. The paper concludes that the RTT technique, while not perfect, can be potentially useful for augmenting training data for NLP tasks, such as training chatbots to recognize variations in user input.



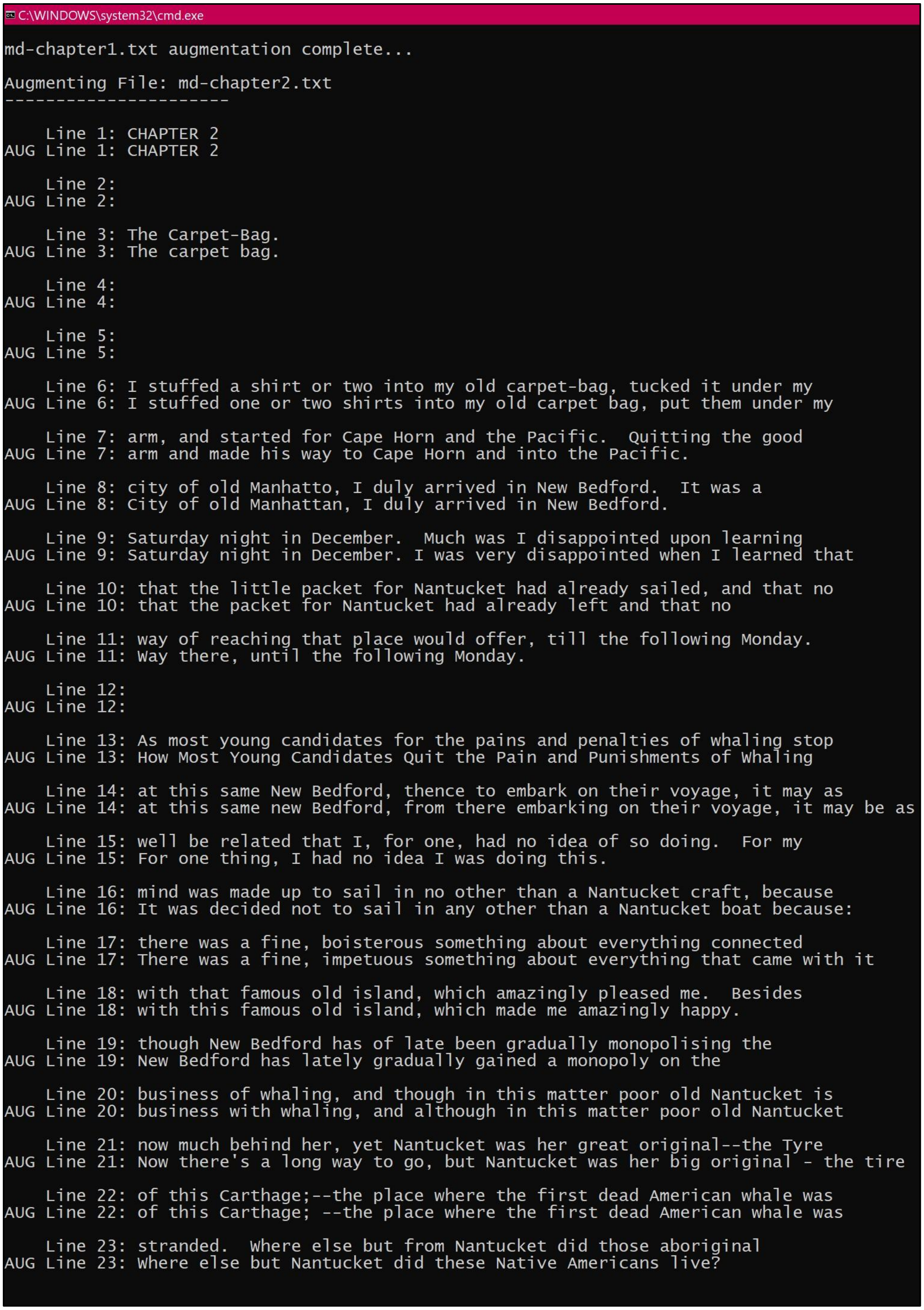
```python
import pandas as pd
import numpy as np
import os
import glob
import nlpaug.augmenter.char as nac
import nlpaug.augmenter.word as naw

FOLDER = 'data'

def augment_file(folder_path = FOLDER):

    folder_path += '' if folder_path[-1] == '/' else '/'
    cwd = os.getcwd()
    folder_path = os.path.join(cwd, FOLDER, "")
    # get all text files
    txt_files = glob.glob(folder_path + '*.txt')

    for f in txt_files:
        aug = naw.BackTranslationAug(from_model_name='facebook/wmt19-en-de',
                                     to_model_name='facebook/wmt19-de-en',
                                     name='BackTranslationAug', device='cpu',
                                     force_reload=False, verbose=0)
        with open(f) as f_input:
            line = f_input.readline()
            cnt = 1
            file_name = os.path.basename(f)
            print(f'Augmenting File: {file_name}')
            print('----------------------', end='\n\n')
            while line:
                print(f'    Line {cnt}: {line}', end='')
                augment = aug.augment(line)
                print(f'AUG Line {cnt}: {augment}', end='\n\n')
                with open(os.path.join(folder_path, 'AUG_' + file_name), mode='a') as f_output:
                    f_output.write(augment + '\n')
                    line = f_input.readline()
                    cnt += 1
            print(f'{file_name} augmentation complete...', end='\n\n')

def main():
    augment_file()

if __name__ == "__main__":
    main()
```

Figure 1. Screenshot of the program code, featuring the implementation of the 'augment_file' function



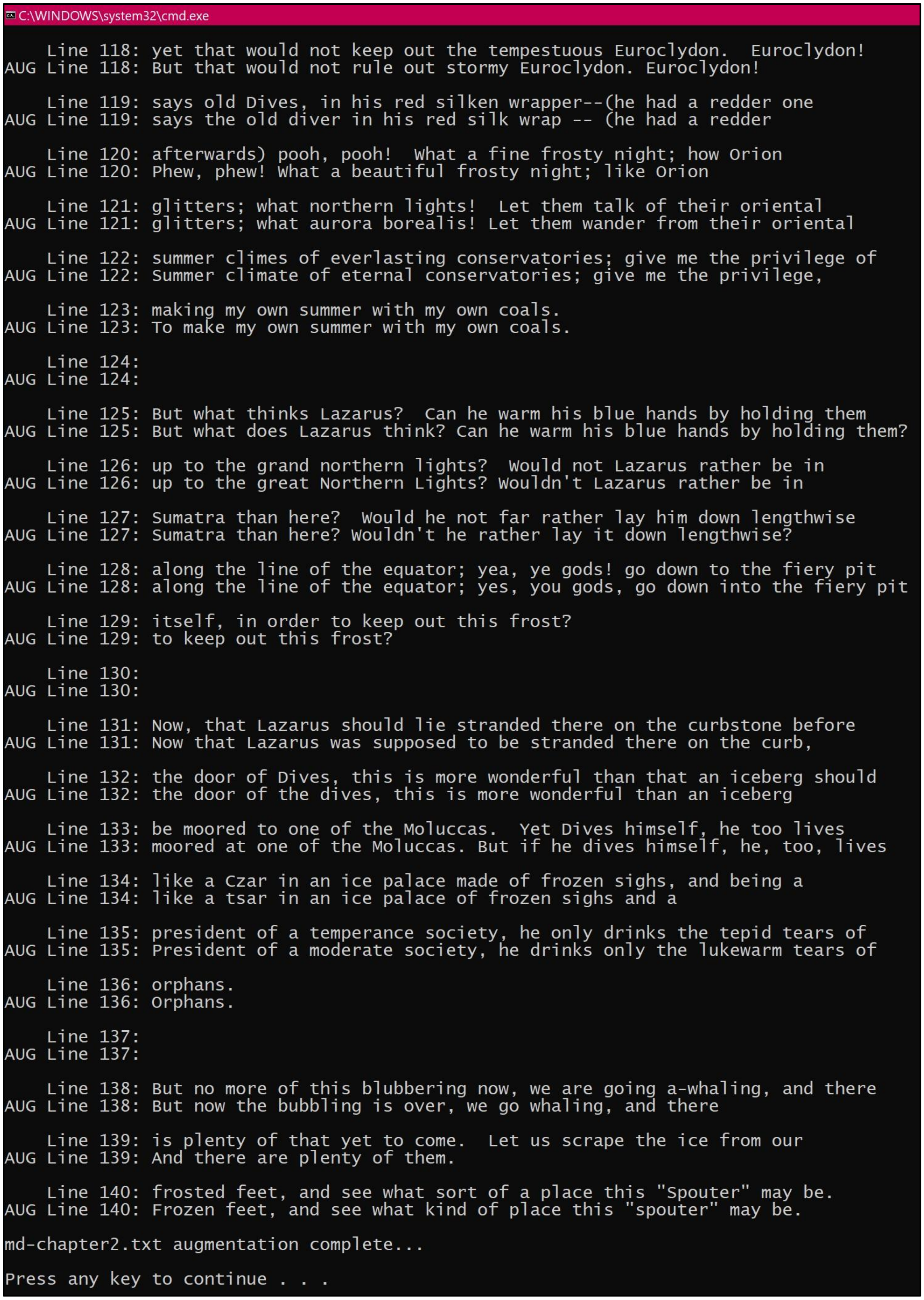
```
Augmenting File: md-chapter1.txt
----------------------

    Line 1: CHAPTER 1
AUG Line 1: CHAPTER 1

    Line 2:
AUG Line 2:

    Line 3: Loomings.
AUG Line 3: Trouble is looming.

    Line 4:
AUG Line 4:

    Line 5:
AUG Line 5:

    Line 6: Call me Ishmael.  Some years ago--never mind how long
AUG Line 6: Call me Ismael. A few years ago - no matter how long

    Line 7: precisely--having little or no money in my purse, and nothing
AUG Line 7: I have little or no money in my wallet and nothing.

    Line 8: particular to interest me on shore, I thought I would sail about a
AUG Line 8: to be interested in land, I thought I would be interested in a

    Line 9: little and see the watery part of the world.  It is a way I have of
AUG Line 9: little and see the watery part of the world. It is a way, which I

    Line 10: driving off the spleen and regulating the circulation.  Whenever I
AUG Line 10: abort the spleen and regulate the circulation.

    Line 11: find myself growing grim about the mouth; whenever it is a damp,
AUG Line 11: I get grim in my mouth; whenever it is damp,

    Line 12: drizzly November in my soul; whenever I find myself involuntarily
AUG Line 12: November drizzles in my soul; whenever I find myself involuntarily,

    Line 13: pausing before coffin warehouses, and bringing up the rear of every
AUG Line 13: in front of the coffin storages and lift the stern of each coffin upwards.

    Line 14: funeral I meet; and especially whenever my hypos get such an upper
AUG Line 14: funeral that I meet; and especially when my hypos are so high

    Line 15: hand of me, that it requires a strong moral principle to prevent me
AUG Line 15: Hand from me, that it takes a strong moral principle to prevent me,

    Line 16: from deliberately stepping into the street, and methodically knocking
AUG Line 16: of consciously stepping on the street and systematic knocking

    Line 17: people's hats off--then, I account it high time to get to sea as soon
AUG Line 17: Hats off - then I think it is high time to set sail as soon as possible.

    Line 18: as I can.  This is my substitute for pistol and ball.  With a
AUG Line 18: That's my replacement for the gun and the ball.

    Line 19: philosophical flourish Cato throws himself upon his sword; I quietly
AUG Line 19: philosophical bloom Cato throws himself on his sword; I quietly

    Line 20: take to the ship.  There is nothing surprising in this.  If they but
AUG Line 20: There's nothing surprising about it.

    Line 21: knew it, almost all men in their degree, some time or other, cherish
AUG Line 21: knew it, almost all men in their degree, at some point or later, appreciate it

    Line 22: very nearly the same feelings towards the ocean with me.
AUG Line 22: Very much the same feelings towards the ocean with me.

    Line 23:
AUG Line 23:

    Line 24: There now is your insular city of the Manhattoes, belted round by
AUG Line 24: There is now your isolated city of Manhattans, surrounded by
```

Figure 2. Screenshot of program output displaying the augmented text of Chapter 1 from Moby Dick, with each line shown sequentially in the console



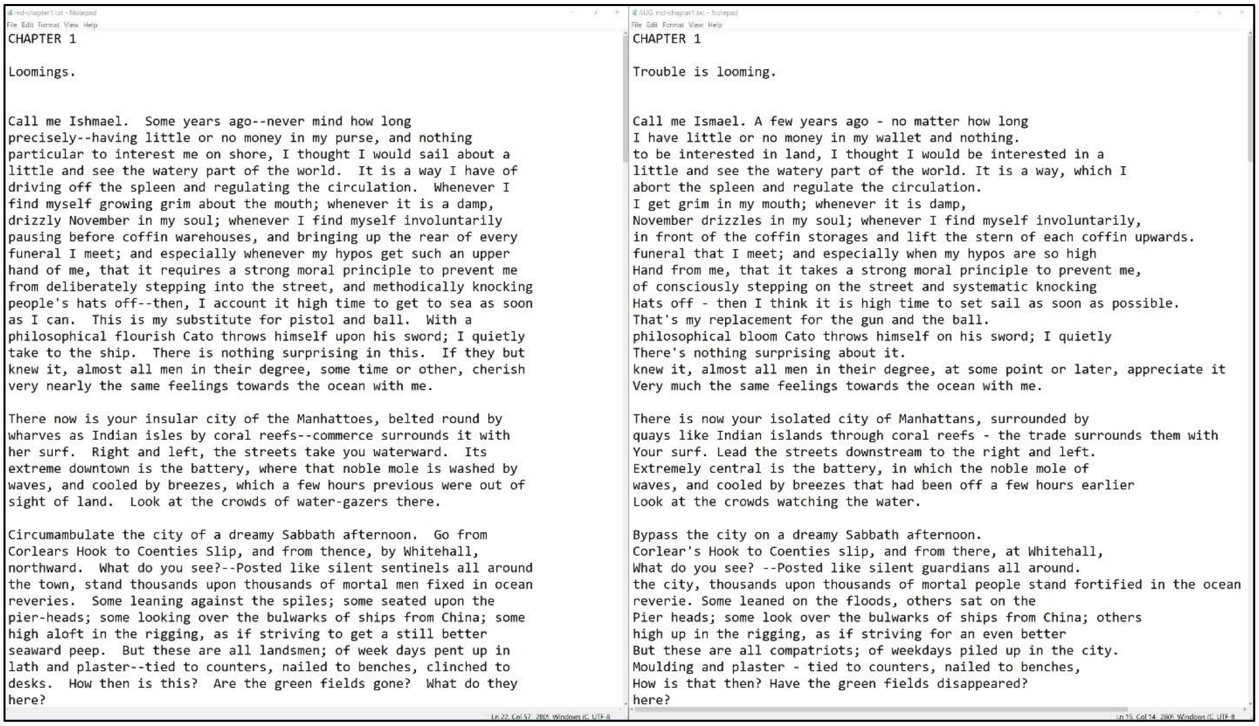
```
md-chapter1.txt augmentation complete...

Augmenting File: md-chapter2.txt
----------------------

    Line 1: CHAPTER 2
AUG Line 1: CHAPTER 2

    Line 2:
AUG Line 2:

    Line 3: The Carpet-Bag.
AUG Line 3: The carpet bag.

    Line 4:
AUG Line 4:

    Line 5:
AUG Line 5:

    Line 6: I stuffed a shirt or two into my old carpet-bag, tucked it under my
AUG Line 6: I stuffed one or two shirts into my old carpet bag, put them under my

    Line 7: arm, and started for Cape Horn and the Pacific.  Quitting the good
AUG Line 7: arm and made his way to Cape Horn and into the Pacific.

    Line 8: city of old Manhatto, I duly arrived in New Bedford.  It was a
AUG Line 8: City of old Manhattan, I duly arrived in New Bedford.

    Line 9: Saturday night in December.  Much was I disappointed upon learning
AUG Line 9: Saturday night in December. I was very disappointed when I learned that

    Line 10: that the little packet for Nantucket had already sailed, and that no
AUG Line 10: that the packet for Nantucket had already left and that no

    Line 11: way of reaching that place would offer, till the following Monday.
AUG Line 11: Way there, until the following Monday.

    Line 12:
AUG Line 12:

    Line 13: As most young candidates for the pains and penalties of whaling stop
AUG Line 13: How Most Young Candidates Quit the Pain and Punishments of Whaling

    Line 14: at this same New Bedford, thence to embark on their voyage, it may as
AUG Line 14: at this same new Bedford, from there embarking on their voyage, it may be as

    Line 15: well be related that I, for one, had no idea of so doing.  For my
AUG Line 15: For one thing, I had no idea I was doing this.

    Line 16: mind was made up to sail in no other than a Nantucket craft, because
AUG Line 16: It was decided not to sail in any other than a Nantucket boat because:

    Line 17: there was a fine, boisterous something about everything connected
AUG Line 17: There was a fine, impetuous something about everything that came with it

    Line 18: with that famous old island, which amazingly pleased me.  Besides
AUG Line 18: with this famous old island, which made me amazingly happy.

    Line 19: though New Bedford has of late been gradually monopolising the
AUG Line 19: New Bedford has lately gradually gained a monopoly on the

    Line 20: business of whaling, and though in this matter poor old Nantucket is
AUG Line 20: business with whaling, and although in this matter poor old Nantucket

    Line 21: now much behind her, yet Nantucket was her great original--the Tyre
AUG Line 21: Now there's a long way to go, but Nantucket was her big original - the tire

    Line 22: of this Carthage;--the place where the first dead American whale was
AUG Line 22: of this Carthage; --the place where the first dead American whale was

    Line 23: stranded.  Where else but from Nantucket did those aboriginal
AUG Line 23: Where else but Nantucket did these Native Americans live?
```

Figure 3. Screenshot of program output displaying the augmented text of Chapter 2 from Moby Dick, with each line shown sequentially in the console.



```
    Line 118: yet that would not keep out the tempestuous Euroclydon.  Euroclydon!
AUG Line 118: But that would not rule out stormy Euroclydon. Euroclydon!

    Line 119: says old Dives, in his red silken wrapper--(he had a redder one
AUG Line 119: says the old diver in his red silk wrap -- (he had a redder

    Line 120: afterwards) pooh, pooh!  What a fine frosty night; how Orion
AUG Line 120: Phew, phew! What a beautiful frosty night; like Orion

    Line 121: glitters; what northern lights!  Let them talk of their oriental
AUG Line 121: glitters; what aurora borealis! Let them wander from their oriental

    Line 122: summer climes of everlasting conservatories; give me the privilege of
AUG Line 122: Summer climate of eternal conservatories; give me the privilege,

    Line 123: making my own summer with my own coals.
AUG Line 123: To make my own summer with my own coals.

    Line 124:
AUG Line 124:

    Line 125: But what thinks Lazarus?  Can he warm his blue hands by holding them
AUG Line 125: But what does Lazarus think? Can he warm his blue hands by holding them?

    Line 126: up to the grand northern lights?  Would not Lazarus rather be in
AUG Line 126: up to the great Northern Lights? Wouldn't Lazarus rather be in

    Line 127: Sumatra than here?  Would he not far rather lay him down lengthwise
AUG Line 127: Sumatra than here? Wouldn't he rather lay it down lengthwise?

    Line 128: along the line of the equator; yea, ye gods! go down to the fiery pit
AUG Line 128: along the line of the equator; yes, you gods, go down into the fiery pit

    Line 129: itself, in order to keep out this frost?
AUG Line 129: to keep out this frost?

    Line 130:
AUG Line 130:

    Line 131: Now, that Lazarus should lie stranded there on the curbstone before
AUG Line 131: Now that Lazarus was supposed to be stranded there on the curb,

    Line 132: the door of Dives, this is more wonderful than that an iceberg should
AUG Line 132: the door of the dives, this is more wonderful than an iceberg

    Line 133: be moored to one of the Moluccas.  Yet Dives himself, he too lives
AUG Line 133: moored at one of the Moluccas. But if he dives himself, he, too, lives

    Line 134: like a Czar in an ice palace made of frozen sighs, and being a
AUG Line 134: like a tsar in an ice palace of frozen sighs and a

    Line 135: president of a temperance society, he only drinks the tepid tears of
AUG Line 135: President of a moderate society, he drinks only the lukewarm tears of

    Line 136: orphans.
AUG Line 136: Orphans.

    Line 137:
AUG Line 137:

    Line 138: But no more of this blubbering now, we are going a-whaling, and there
AUG Line 138: But now the bubbling is over, we go whaling, and there

    Line 139: is plenty of that yet to come.  Let us scrape the ice from our
AUG Line 139: And there are plenty of them.

    Line 140: frosted feet, and see what sort of a place this "Spouter" may be.
AUG Line 140: Frozen feet, and see what kind of place this "spouter" may be.

md-chapter2.txt augmentation complete...

Press any key to continue . . .
```

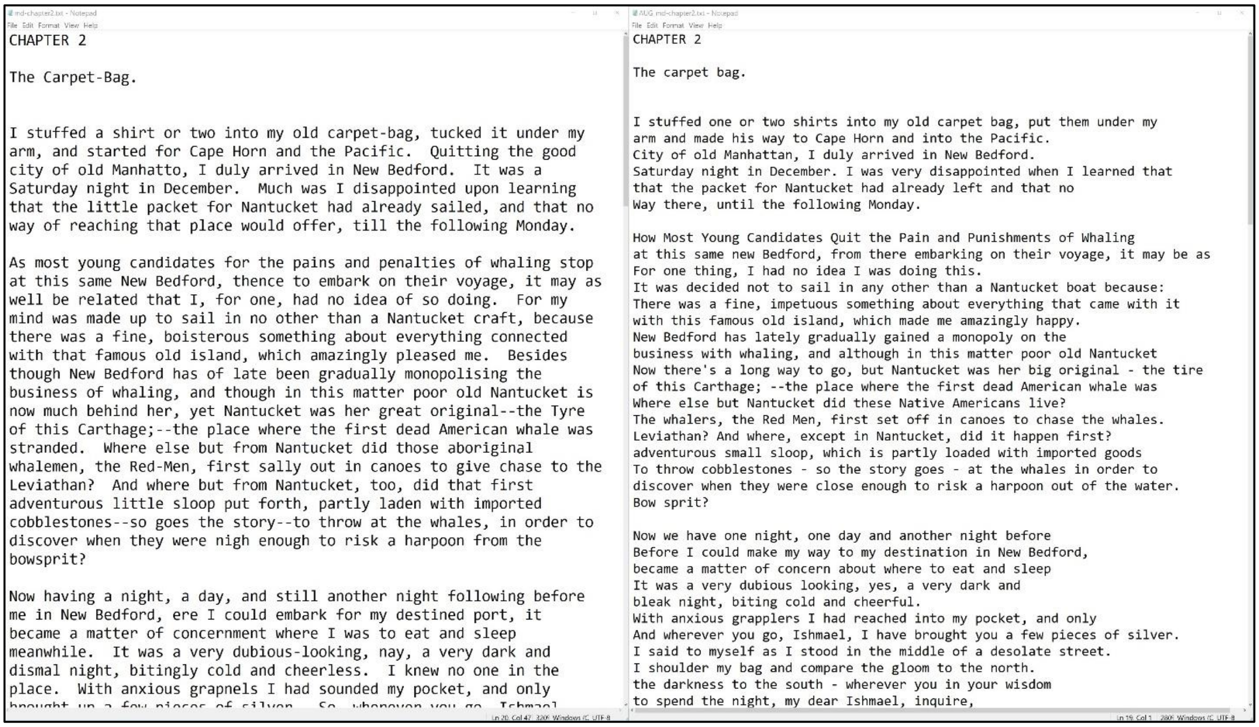
Figure 3. Screenshot of program output displaying the augmented text of Chapter 2 from Moby Dick, with each line shown sequentially in the console.


md-chapter1.txt - Notepad
File Edit Format View Help
Ln 22, Col 57 280% Windows (C UTF-8
AUG md-chapter1.txt - Notepad
File Edit Format View Help
Ln 15, Col 14 280% Windows (C UTF-8

```
CHAPTER 1

Loomings.


Call me Ishmael.  Some years ago--never mind how long
precisely--having little or no money in my purse, and nothing
particular to interest me on shore, I thought I would sail about a
little and see the watery part of the world.  It is a way I have of
driving off the spleen and regulating the circulation.  Whenever I
find myself growing grim about the mouth; whenever it is a damp,
drizzly November in my soul; whenever I find myself involuntarily
pausing before coffin warehouses, and bringing up the rear of every
funeral I meet; and especially whenever my hypos get such an upper
hand of me, that it requires a strong moral principle to prevent me
from deliberately stepping into the street, and methodically knocking
people's hats off--then, I account it high time to get to sea as soon
as I can.  This is my substitute for pistol and ball.  With a
philosophical flourish Cato throws himself upon his sword; I quietly
take to the ship.  There is nothing surprising in this.  If they but
knew it, almost all men in their degree, some time or other, cherish
very nearly the same feelings towards the ocean with me.

There now is your insular city of the Manhattoes, belted round by
wharves as Indian isles by coral reefs--commerce surrounds it with
her surf.  Right and left, the streets take you waterward.  Its
extreme downtown is the battery, where that noble mole is washed by
waves, and cooled by breezes, which a few hours previous were out of
sight of land.  Look at the crowds of water-gazers there.

Circumambulate the city of a dreamy Sabbath afternoon.  Go from
Corlears Hook to Coenties Slip, and from thence, by Whitehall,
northward.  What do you see?--Posted like silent sentinels all around
the town, stand thousands upon thousands of mortal men fixed in ocean
reveries.  Some leaning against the spiles; some seated upon the
pier-heads; some looking over the bulwarks of ships from China; some
high aloft in the rigging, as if striving to get a still better
seaward peep.  But these are all landsmen; of week days pent up in
lath and plaster--tied to counters, nailed to benches, clinched to
desks.  How then is this?  Are the green fields gone?  What do they
here?
```

```
CHAPTER 1

Trouble is looming.


Call me Ismael. A few years ago - no matter how long
I have little or no money in my wallet and nothing.
to be interested in land, I thought I would be interested in a
little and see the watery part of the world. It is a way, which I
abort the spleen and regulate the circulation.
I get grim in my mouth; whenever it is damp,
November drizzles in my soul; whenever I find myself involuntarily,
in front of the coffin storages and lift the stern of each coffin upwards.
funeral that I meet; and especially when my hypos are so high
Hand from me, that it takes a strong moral principle to prevent me,
of consciously stepping on the street and systematic knocking
Hats off - then I think it is high time to set sail as soon as possible.
That's my replacement for the gun and the ball.
philosophical bloom Cato throws himself on his sword; I quietly
There's nothing surprising about it.
knew it, almost all men in their degree, at some point or later, appreciate it
Very much the same feelings towards the ocean with me.

There is now your isolated city of Manhattans, surrounded by
quays like Indian islands through coral reefs - the trade surrounds them with
Your surf. Lead the streets downstream to the right and left.
Extremely central is the battery, in which the noble mole of
waves, and cooled by breezes that had been off a few hours earlier
Look at the crowds watching the water.

Bypass the city on a dreamy Sabbath afternoon.
Corlear's Hook to Coenties slip, and from there, at Whitehall,
What do you see? --Posted like silent guardians all around.
the city, thousands upon thousands of mortal people stand fortified in the ocean
reverie. Some leaned on the floods, others sat on the
Pier heads; some look over the bulwarks of ships from China; others
high up in the rigging, as if striving for an even better
But these are all compatriots; of weekdays piled up in the city.
Moulding and plaster - tied to counters, nailed to benches,
How is that then? Have the green fields disappeared?
here?
```

Figure 4. A side-by-side comparison of Moby Dick's Chapter 1: original text (left) and back-translated text (right) using the round-trip translation augmentation technique.


md-chapter2.txt - Notepad
File Edit Format View Help
Ln 20, Col 47 320% Windows (C UTF-8
AUG md-chapter2.txt - Notepad
File Edit Format View Help
Ln 19, Col 1 280% Windows (C UTF-8

```
CHAPTER 2

The Carpet-Bag.


I stuffed a shirt or two into my old carpet-bag, tucked it under my
arm, and started for Cape Horn and the Pacific.  Quitting the good
city of old Manhatto, I duly arrived in New Bedford.  It was a
Saturday night in December.  Much was I disappointed upon learning
that the little packet for Nantucket had already sailed, and that no
way of reaching that place would offer, till the following Monday.

As most young candidates for the pains and penalties of whaling stop
at this same New Bedford, thence to embark on their voyage, it may as
well be related that I, for one, had no idea of so doing.  For my
mind was made up to sail in no other than a Nantucket craft, because
there was a fine, boisterous something about everything connected
with that famous old island, which amazingly pleased me.  Besides
though New Bedford has of late been gradually monopolising the
business of whaling, and though in this matter poor old Nantucket is
now much behind her, yet Nantucket was her great original--the Tyre
of this Carthage;--the place where the first dead American whale was
stranded.  Where else but from Nantucket did those aboriginal
whalemen, the Red-Men, first sally out in canoes to give chase to the
Leviathan?  And where but from Nantucket, too, did that first
adventurous little sloop put forth, partly laden with imported
cobblestones--so goes the story--to throw at the whales, in order to
discover when they were nigh enough to risk a harpoon from the
bowsprit?

Now having a night, a day, and still another night following before
me in New Bedford, ere I could embark for my destined port, it
became a matter of concernment where I was to eat and sleep
meanwhile.  It was a very dubious-looking, nay, a very dark and
dismal night, bitingly cold and cheerless.  I knew no one in the
place.  With anxious grapnels I had sounded my pocket, and only
brought up a few pieces of silver.  So, wherever you go, Ishmael
```

```
CHAPTER 2

The carpet bag.


I stuffed one or two shirts into my old carpet bag, put them under my
arm and made his way to Cape Horn and into the Pacific.
City of old Manhattan, I duly arrived in New Bedford.
Saturday night in December. I was very disappointed when I learned that
that the packet for Nantucket had already left and that no
Way there, until the following Monday.

How Most Young Candidates Quit the Pain and Punishments of Whaling
at this same new Bedford, from there embarking on their voyage, it may be as
For one thing, I had no idea I was doing this.
It was decided not to sail in any other than a Nantucket boat because:
There was a fine, impetuous something about everything that came with it
with this famous old island, which made me amazingly happy.
New Bedford has lately gradually gained a monopoly on the
business with whaling, and although in this matter poor old Nantucket
Now there's a long way to go, but Nantucket was her big original - the tire
of this Carthage; --the place where the first dead American whale was
Where else but Nantucket did these Native Americans live?
The whalers, the Red Men, first set off in canoes to chase the whales.
Leviathan? And where, except in Nantucket, did it happen first?
adventurous small sloop, which is partly loaded with imported goods
To throw cobblestones - so the story goes - at the whales in order to
discover when they were close enough to risk a harpoon out of the water.
Bow sprit?

Now we have one night, one day and another night before
Before I could make my way to my destination in New Bedford,
became a matter of concern about where to eat and sleep
It was a very dubious looking, yes, a very dark and
bleak night, biting cold and cheerful.
With anxious grapplers I had reached into my pocket, and only
And wherever you go, Ishmael, I have brought you a few pieces of silver.
I said to myself as I stood in the middle of a desolate street.
I shoulder my bag and compare the gloom to the north.
the darkness to the south - wherever you in your wisdom
to spend the night, my dear Ishmael, inquire,
```

Figure 5. A side-by-side comparison of Moby Dick's Chapter 2: original text (left) and back-translated text (right) using the round-trip translation augmentation technique.

Table of Contents

OPTION #1: Text Dataset Augmentation

This paper presents a Python script designed to augment text datasets, effectively expanding the available data for training. Marivate and Sefara (2020) note that numerous image classification tasks have significantly benefited from data augmentation techniques. By altering an image's structure, these methods increase the number of training samples accessible to an algorithm, enhancing the model's resilience. However, applying similar augmentation techniques to text datasets has been challenging, as some approaches demand a deeper understanding of the language in question.

Text Dataset Augmentation for Natural Language Processing (NLP)

To address this issue, this paper explores various text augmentation techniques and presents a Python script that employs round-trip translation (RTT) as an effective means of augmenting text data. By applying this script to any text file within the "data/" folder, the user can successfully expand their dataset while maintaining the integrity of the original information. This improved augmentation approach has the potential to make a significant impact on the development of more robust natural language processing models.

Synonym Augmentation

Marivate and Sefara (2020) classify text augmentation techniques into two primary categories: those that operate on the text source and those that focus on text representation. Ideally, textual data should be augmented by linguistic experts who can manually rephrase sentences using language modeling rules. However, this approach can be costly and time-consuming. An alternative method involves replacing words with synonyms, which falls under the category of augmenting a text's source.

WordNet, an open-source lexical database, is a valuable tool for implementing synonym-based augmentation. It organizes English nouns, verbs, and adjectives into synonym sets or *synsets*, representing distinct lexical concepts, while also documenting the relationships between these groups (Miller et al., 1990). Zhang et al. (2015) successfully employed WordNet to augment training data, enabling a convolutional network to better understand textual input at the character level.

Semantic Similarity Augmentation

Another approach highlighted by Marivate and Sefara (2020) is *semantic similarity augmentation*, which focuses on modifying a text's representation rather than its source. Unlike synonym augmentation, this technique does not rely on dictionaries or thesauri. Instead, it necessitates the use of pre-trained word embeddings or the creation of custom word embeddings.

Word embeddings are dense vector representations of words, generated through neural network-inspired training methods (Levy & Goldberg, 2014). This method distinguishes itself from the previous approach by utilizing distributed word representations to substitute words with those found in similar contexts. In contrast, synonym augmentation depends on manually maintained lexical databases, such as WordNet (Marivate & Sefara, 2020).

Back-Translation Augmentation

Marivate and Sefara (2020) also discuss *round-trip translation* (RTT), sometimes referred to as bi-directional, back-and-forth, or recursive translation. This text augmentation technique expands training data by translating a text into a foreign language and then translating it back into the original language. For example, an English text might be translated into Spanish and then translated back into English.

Somers (2005) explains that RTT is frequently employed to assess machine translation (MT) systems. According to Anon (2003, as cited in Somers, 2005), the original English text should ideally be identical to the back-translated English text. As a result, the round-trip process through a back translation engine can be used to evaluate the accuracy of the translated text and the underlying model (Anon, 2005, as cited in Somers, 2005).

The NLPAug Python Library for Dataset Augmentation

The program presented in this paper employs the *NLPAug* Python library to augment any text dataset stored in the "data/" directory. Nithilaau (2021) explains that NLPAug offers three levels of augmentation: (a) character level, (b) word level, and (c) flow/sentence level.

Character Level Augmentation

Character level augmentation operates on individual characters within the text, making it useful for applications like chatbot training data. Since user input to chatbots often contains typos despite auto-correct features, the NLPAug library provides a "keyboard" augmenter that simulates typographical errors by replacing characters with nearby ones on a keyboard.

Word and Flow Level Augmentation

Word level augmentations in the NLPAug library encompass various methods, including "synonym," "antonym," "random," "spelling," and "split" augmenters. These augmenters manipulate words within the text to create variations in the dataset. Lastly, flow level augmenters such as "sequential" and "sometimes" target the overall structure of sentences or the order of text elements.

The Back Translation Augmenter Program

The "back translation" augmenter is a word-level augmenter that leverages two translation models for augmentation (*Nlpaug.Augmenter.Word.Back_translation — Nlpaug 1.2.0dev Documentation*, n.d.). In this paper's implementation, English text datasets are translated into German and then back to English. Among the various techniques discussed earlier, round-trip translation (RTT) generates augmented text that closely resembles the original content. RTT does not introduce spelling errors, random words, or antonyms into the text. Even “synonym” augmenters sometimes produced unexpected alterations in straightforward sentences. Consequently, RTT was chosen because it provided the most coherent and consistent augmentation results upon initial examination.

Program Overview

The provided function, spanning just under 30 lines of code, utilizes RTT to augment text datasets in a designated folder. As depicted in Figure 1, lines 12-16 compile all TXT files within the “data/” directory into a Python list object. Line 19 generates the “back translation” augmenter object. The program then opens the first text file in the “data/” directory, iterating through each line. It displays the original line and its back-translated version on the console for comparison and writes the augmented version to a new file with an “AUG_” prefix added to the original file name.

Results and Evaluation

To evaluate the program, the first and second chapters of Herman Melville’s classic novel Moby Dick were placed in the “data/” directory. Figures 2 – 4 exhibit the program’s output on the terminal as it processes each line in the text files. Upon completion of each file, the program informs the user and proceeds to the next file. The augmented text is written to new files with the “AUG_” prefix. Figures 5 – 6 display the original Moby Dick chapters alongside their augmented versions. For example, the first sentence in Figure 5 shows the word “Looming” altered to “Trouble is looming.” Although the “back translation” augmenter does not deliver flawless translations, it offers potential value for enhancing training data in natural language processing (NLP) tasks, such as enabling a chatbot to identify user input variations. To further improve the algorithm, the dataset could be augmented sentence-by-sentence instead of line-by-line, since some sentences span multiple lines.

Conclusion

This paper presented a Python program designed to augment any TXT file using back-translation. The program iterates through all files with TXT extensions in its “data/” folder, processing each line-by-line and augmenting the content before displaying it on the console and saving it to a new text file. The paper also explored a range of text augmentation techniques, including synonym augmentation, semantic similarity augmentation, and round-trip translation (RTT). Furthermore, the NLPAug Python library and its diverse applications were discussed. Ultimately, the provided Python script successfully transformed any text dataset (i.e., TXT file) placed in its “data/” folder into an augmented version using RTT and the capabilities of the NLPAug library.

References

Levy, O., & Goldberg, Y. (2014). Dependency-Based Word Embeddings. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 302–308. https://doi.org/10.3115/v1/P14-2050

Marivate, V., & Sefara, T. (2020). Improving short text classification through global augmentation methods. *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 385–399.

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., & Edunov, S. (2019). Facebook FAIR"s WMT19 News Translation Task Submission. *ArXiv Preprint ArXiv:1907.06616*.

Nithilaau. (2021, August 25). NLPAUG - A Python library to Augment Your Text Data. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2021/08/nlpaug-a-python-library-to-augment-your-text-data/

nlpaug.augmenter.word.back_translation—Nlpaug 1.2.0dev documentation. (n.d.). Retrieved September 19, 2021, from https://nlpaug.readthedocs.io/en/latest/augmenter/word/back_translation.html

Somers, H. (2005). Round-trip translation: What is it good for? *Proceedings of the Australasian Language Technology Workshop 2005*, 127–133.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, *28*, 649–657.