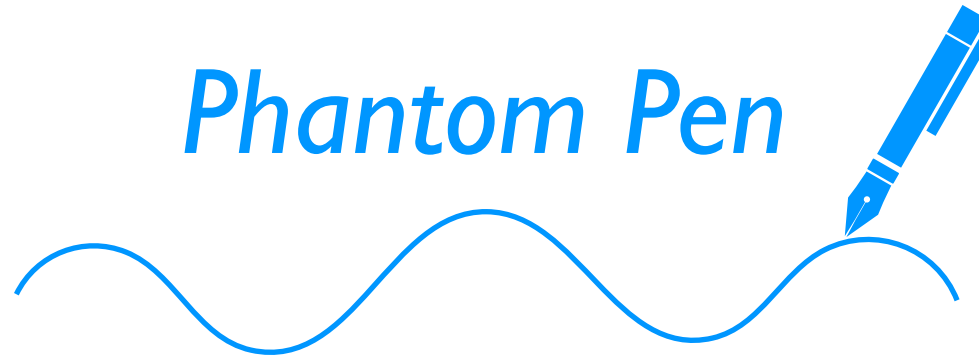


Phantom Pen



Your digital ghost writer

Mobolaji Williams — Insight Data Science

Demo Video

<https://youtu.be/W-UbsOGW3KQ>

Relevant Code

https://github.com/mowillia/phantom_pen

Responding to Writing Needs with Machine Learning

Need #1

Move past a mental block and generate new ideas for writing

Need #2

Integrate the styles of many different writers




Answering the Needs

A program that writes in response to a prompt using the combined styles of many different writers

Author's prompt:



*Created from a compilation of the styles of many writers



How was this built?
(Detailed Version)

Designing Phantom Pen

Phantom Pen



Objective

Build a multifunctional writing program that can not only **generate** new text but can also **classify** user text, **extract** the important sentences and use both of these latter features in generation.

Aspects of the Design

1. Front End

1. User Interface Structure
2. Usage Tips
3. Background
4. Corpora
5. Ethics

2. Back End

1. Language Model
2. Corpora and Aspects of Training
3. Classification Model
4. Extractive Summary

Designing Phantom Pen: Front End

Aspects of the Design

I. Front End

1. User Interface Structure
2. Usage Tips
3. Background
4. Corpora
5. Ethics

PHANTOM PEN

[Home](#)

Usage Tips

Background

Corpora

Ethics

Github Repo

Slides



PHANTOM PEN
YOUR DIGITAL GHOSTWRITER

Phantom Pen is a program that generates *original* writing using retrained versions of Open AI's [Generalized Pre-trained Transformer-2](#). The program has been re-trained on +2000 essays from *The Atlantic*, +160 short stories and 20 works of literature (see [Corpora](#) for a more specific description of the texts used).

There are four choices that enter into the "simple generate" function of the program:

GENERATOR CHOICES

- **CREATIVITY:** The amount of liberty the algorithm will take in generating new text; "1" being the most creative, and "0" leading to repetitive direct copying from corpus.
- **LENGTH:** The number of "tokens" the text generator outputs. A proxy for the number of generated words.
- **CLASS:** The corpus (see [Corpora](#)) that the text generator will use as a basis for its writing style and content.
- **117M/345M MODEL:** The chosen model for text generation. The 345M model is more powerful and produces more coherent text, but takes more than twice as long to run.

GPU STATUS:

SIMPLE GENERATE | CLASSIFY AND GENERATE

CLASSIFY AND GENERATE

CLASSIFY, EXTRACT, AND GENERATE

It is a rare skill to stand at the foot of the mountain and be awed, and yet undaunted; to peer upwards at a summit hidden in the clouds and to respect its terrible beauty without succumbing to terror itself.

CREATIVITY

0.95

LENGTH

100

CLASS

Business

117M Model

EXECUTE

(Enter text and press execute to get an original essay or story!)

PHANTOM PEN

Home

Usage Tips

Background

Corpora

Ethics

Github Repo

Slides

PHANTOM PEN

YOUR DIGITAL GHOSTWRITER

Phantom Pen is a program that generates *original* writing using retrained versions of Open AI's [Generalized Pre-trained Transformer-2](#). The program has been re-trained on +2000 essays from *The Atlantic*, +160 short stories and 20 works of literature (see [Corpora](#) for a more specific description of the texts used).

There are four choices that enter into the "simple generate" function of the program:

GENERATOR CHOICES

- **CREATIVITY:** The amount of liberty the algorithm will take in generating new text; "1" being the most creative, and "0" leading to repetitive direct copying from corpus.
- **LENGTH:** The number of "tokens" the text generator outputs. A proxy for the number of generated words.
- **CLASS:** The corpus (see [Corpora](#)) that the text generator will use as a basis for its writing style and content.
- **117M/345M MODEL:** The chosen model for text generation. The

SIMPLE GENERATE

CLASSIFY AND GENERATE

CLASSIFY, EXTRACT, AND GENERATE

It is a rare skill to stand at the foot of the mountain and be awed, and yet undaunted; to peer upwards at a summit hidden in the clouds and to respect its terrible beauty without succumbing to terror itself.

CREATIVITY

0.95

LENGTH

100

CLASS

Business

117M Model

EXECUTE

17713451 MODEL: the chosen model for text generation. The 345M model is more powerful and produces more coherent text, but takes more than twice as long to run.

GPU STATUS:

(Enter text and press execute to get an original essay or story!)

Creating User Interface

Quick description of what program does

It was important to use size and color to emphasize the text-box and highlight the “EXECUTE” button for “impatient” users

The screenshot shows the PHANTOM PEN web application. On the left is a blue sidebar with the title 'PHANTOM PEN' and navigation links: Home, Usage Tips, Background, Corpora, and Ethics. Below these are buttons for 'Github Repo' and 'Slides'. The main content area has a header with a close button (X) and the title 'PHANTOM PEN' with the subtitle 'YOUR DIGITAL GHOSTWRITER'. A paragraph describes the program as using retrained OpenAI GPT-2 models on a corpus of essays and stories. Below this is a section for 'GENERATOR CHOICES' with three bullet points: CREATIVITY (0.95), LENGTH (100), and CLASS (Business). A fourth choice, MODEL (117M Model), is shown below. A large text box contains a sample generated text: 'It is a rare skill to stand at the foot of the mountain and be awed, and yet undaunted; to peer upwards at a summit hidden in the clouds and to respect its terrible beauty without succumbing to terror itself.' At the bottom right is a large blue 'EXECUTE' button. A footer note says '(Enter text and press execute to get an original essay or story!)'. Annotations with orange lines point to the description paragraph, the text box, and the EXECUTE button.

PHANTOM PEN

Home
Usage Tips
Background
Corpora
Ethics

Github Repo
Slides

PHANTOM PEN
YOUR DIGITAL GHOSTWRITER

Phantom Pen is a program that generates *original* writing using retrained versions of Open AI's [Generalized Pre-trained Transformer-2](#). The program has been re-trained on +2000 essays from *The Atlantic*, +160 short stories and 20 works of literature (see [Corpora](#) for a more specific description of the texts used).

There are four choices that enter into the "simple generate" function of the program:

GENERATOR CHOICES

- **CREATIVITY:** The amount of liberty the algorithm will take in generating new text; "1" being the most creative, and "0" leading to repetitive direct copying from corpus.
- **LENGTH:** The number of "tokens" the text generator outputs. A proxy for the number of generated words.
- **CLASS:** The corpus (see [Corpora](#)) that the text generator will use as a basis for its writing style and content.
- **117M/345M MODEL:** The chosen model for text generation. The 345M model is more powerful and produces more coherent text, but takes more than twice as long to run.

GPU STATUS:

SIMPLE GENERATE | CLASSIFY AND GENERATE
CLASSIFY, EXTRACT, AND GENERATE

It is a rare skill to stand at the foot of the mountain and be awed, and yet undaunted; to peer upwards at a summit hidden in the clouds and to respect its terrible beauty without succumbing to terror itself.

CREATIVITY 0.95 **LENGTH** 100 **CLASS** Business

117M Model

EXECUTE

(Enter text and press execute to get an original essay or story!)

Creating User Interface

Quick description of what program does

It was important to use size and color to emphasize the text-box and highlight the “EXECUTE” button for “impatient” users

The screenshot shows the PHANTOM PEN web application. On the left is a blue sidebar with the title 'PHANTOM PEN' and navigation links: Home, Usage Tips, Background, Corpora, and Ethics. Below these are buttons for 'Github Repo' and 'Slides'. The main content area has a header with a close button (X) and the title 'PHANTOM PEN' with the subtitle 'YOUR DIGITAL GHOSTWRITER'. A text box describes the program: 'Phantom Pen is a program that generates original writing using retrained versions of Open AI's Generalized Pre-trained Transformer-2. The program has been re-trained on +2000 essays from The Atlantic, +160 short stories and 20 works of literature (see Corpora for a more specific description of the texts used). There are four choices that enter into the "simple generate" function of the program:'. Below this is a section titled 'GENERATOR CHOICES' with a bulleted list: 'CREATIVITY: The amount of liberty the algorithm will take in generating new text; "1" being the most creative, and "0" leading to repetitive direct copying from corpus.', 'LENGTH: The number of "tokens" the text generator outputs. A proxy for the number of generated words.', 'CLASS: The corpus (see Corpora) that the text generator will use as a basis for its writing style and content.', and '117M/345M MODEL: The chosen model for text generation. The 345M model is more powerful and produces more coherent text, but takes more than twice as long to run.' At the bottom left of the main area is a 'GPU STATUS:' button. On the right, there are tabs for 'SIMPLE GENERATE' (selected), 'CLASSIFY AND GENERATE', and 'CLASSIFY, EXTRACT, AND GENERATE'. Below the tabs is a large text input box containing the sample text: 'It is a rare skill to stand at the foot of the mountain and be awed, and yet undaunted; to peer upwards at a summit hidden in the clouds and to respect its terrible beauty without succumbing to terror itself.' Below the text box are three input fields: 'CREATIVITY' (0.95), 'LENGTH' (100), and 'CLASS' (Business). Below these is a '117M Model' dropdown menu and a large blue 'EXECUTE' button. At the bottom right, a note says '(Enter text and press execute to get an original essay or story!)'. Orange lines and boxes highlight the description text, the text input box, the EXECUTE button, and the parameter/model choices.

PHANTOM PEN
YOUR DIGITAL GHOSTWRITER

Phantom Pen is a program that generates *original* writing using retrained versions of Open AI's [Generalized Pre-trained Transformer-2](#). The program has been re-trained on +2000 essays from *The Atlantic*, +160 short stories and 20 works of literature (see [Corpora](#) for a more specific description of the texts used).

There are four choices that enter into the "simple generate" function of the program:

GENERATOR CHOICES

- **CREATIVITY:** The amount of liberty the algorithm will take in generating new text; "1" being the most creative, and "0" leading to repetitive direct copying from corpus.
- **LENGTH:** The number of "tokens" the text generator outputs. A proxy for the number of generated words.
- **CLASS:** The corpus (see [Corpora](#)) that the text generator will use as a basis for its writing style and content.
- **117M/345M MODEL:** The chosen model for text generation. The 345M model is more powerful and produces more coherent text, but takes more than twice as long to run.

GPU STATUS:

PHANTOM PEN

SIMPLE GENERATE | CLASSIFY AND GENERATE | CLASSIFY, EXTRACT, AND GENERATE

It is a rare skill to stand at the foot of the mountain and be awed, and yet undaunted; to peer upwards at a summit hidden in the clouds and to respect its terrible beauty without succumbing to terror itself.

CREATIVITY 0.95 **LENGTH** 100 **CLASS** Business

117M Model

EXECUTE

(Enter text and press execute to get an original essay or story!)

Next to the EXECUTE button are parameter/model choices which are preset. Explanations nearby if interested.

Creating User Interface

Intuitive button to allow for minimizing side-plane

Quick description of what program does

It was important to use size and color to emphasize the text-box and highlight the “EXECUTE” button for “impatient” users

The screenshot shows the PHANTOM PEN web application. On the left is a blue sidebar with the title 'PHANTOM PEN' and navigation links: Home, Usage Tips, Background, Corpora, and Ethics. Below these are buttons for 'Github Repo' and 'Slides'. The main content area has a header with a close button (an 'X' in a blue square) and the title 'PHANTOM PEN' with the subtitle 'YOUR DIGITAL GHOSTWRITER'. Below this is a paragraph describing the program: 'Phantom Pen is a program that generates original writing using retrained versions of Open AI's Generalized Pre-trained Transformer-2. The program has been re-trained on +2000 essays from The Atlantic, +160 short stories and 20 works of literature (see Corpora for a more specific description of the texts used). There are four choices that enter into the "simple generate" function of the program:'. This paragraph is highlighted with an orange box. Below it is a section titled 'GENERATOR CHOICES' with a bulleted list: 'CREATIVITY: The amount of liberty the algorithm will take in generating new text; "1" being the most creative, and "0" leading to repetitive direct copying from corpus.', 'LENGTH: The number of "tokens" the text generator outputs. A proxy for the number of generated words.', 'CLASS: The corpus (see Corpora) that the text generator will use as a basis for its writing style and content.', and '117M/345M MODEL: The chosen model for text generation. The 345M model is more powerful and produces more coherent text, but takes more than twice as long to run.' Below the list is a blue button labeled 'GPU STATUS:'. On the right side of the main content area, there are three tabs: 'SIMPLE GENERATE', 'CLASSIFY AND GENERATE', and 'CLASSIFY, EXTRACT, AND GENERATE'. The 'SIMPLE GENERATE' tab is active. Below the tabs is a large text box containing the generated text: 'It is a rare skill to stand at the foot of the mountain and be awed, and yet undaunted; to peer upwards at a summit hidden in the clouds and to respect its terrible beauty without succumbing to terror itself.' This text box is highlighted with an orange box. Below the text box are three input fields: 'CREATIVITY' with a value of 0.95, 'LENGTH' with a value of 100, and 'CLASS' with a dropdown menu showing 'Business'. Below these fields is a dropdown menu for the model, showing '117M Model'. To the right of the model dropdown is a large blue button labeled 'EXECUTE', which is highlighted with an orange box. Below the 'EXECUTE' button is a note: '(Enter text and press execute to get an original essay or story!)'. An orange line connects the 'EXECUTE' button to the text annotation on the right.

PHANTOM PEN

Home

Usage Tips

Background

Corpora

Ethics

Github Repo

Slides

PHANTOM PEN

YOUR DIGITAL GHOSTWRITER

Phantom Pen is a program that generates original writing using retrained versions of Open AI's Generalized Pre-trained Transformer-2. The program has been re-trained on +2000 essays from *The Atlantic*, +160 short stories and 20 works of literature (see [Corpora](#) for a more specific description of the texts used). There are four choices that enter into the "simple generate" function of the program:

GENERATOR CHOICES

- **CREATIVITY:** The amount of liberty the algorithm will take in generating new text; "1" being the most creative, and "0" leading to repetitive direct copying from corpus.
- **LENGTH:** The number of "tokens" the text generator outputs. A proxy for the number of generated words.
- **CLASS:** The corpus (see [Corpora](#)) that the text generator will use as a basis for its writing style and content.
- **117M/345M MODEL:** The chosen model for text generation. The 345M model is more powerful and produces more coherent text, but takes more than twice as long to run.

GPU STATUS:

SIMPLE GENERATE | CLASSIFY AND GENERATE | CLASSIFY, EXTRACT, AND GENERATE

It is a rare skill to stand at the foot of the mountain and be awed, and yet undaunted; to peer upwards at a summit hidden in the clouds and to respect its terrible beauty without succumbing to terror itself.

CREATIVITY LENGTH CLASS

0.95 100 Business

117M Model

EXECUTE

(Enter text and press execute to get an original essay or story!)

Next to the EXECUTE button are parameter/model choices which are preset. Explanations nearby if interested.

Creating User Interface

Intuitive button to allow for minimizing side-plane

Quick description of what program does

It was important to use size and color to emphasize the text-box and highlight the “EXECUTE” button for “impatient” users

The screenshot shows the PHANTOM PEN web application. On the left is a blue sidebar with the title 'PHANTOM PEN' and navigation links: Home, Usage Tips, Background, Corpora, and Ethics. Below these are buttons for 'Github Repo' and 'Slides'. The main content area has a header with a close button (X) and the title 'PHANTOM PEN' with the subtitle 'YOUR DIGITAL GHOSTWRITER'. A paragraph describes the program as using retrained OpenAI GPT-2 models on a corpus of essays and stories. Below this is a section for 'GENERATOR CHOICES' with parameters for Creativity, Length, Class, and Model. A large text box contains generated text. At the bottom are input fields for Creativity, Length, and Class, a dropdown for the Model, and a prominent blue 'EXECUTE' button. A note at the bottom says '(Enter text and press execute to get an original essay or story!)'. Annotations with orange lines point to the close button, the description paragraph, the generated text box, the EXECUTE button, and the sidebar.

PHANTOM PEN

Home

Usage Tips

Background

Corpora

Ethics

Github Repo

Slides

PHANTOM PEN

YOUR DIGITAL GHOSTWRITER

Phantom Pen is a program that generates *original* writing using retrained versions of Open AI's [Generalized Pre-trained Transformer-2](#). The program has been re-trained on +2000 essays from *The Atlantic*, +160 short stories and 20 works of literature (see [Corpora](#) for a more specific description of the texts used).

There are four choices that enter into the "simple generate" function of the program:

GENERATOR CHOICES

- **CREATIVITY:** The amount of liberty the algorithm will take in generating new text; "1" being the most creative, and "0" leading to repetitive direct copying from corpus.
- **LENGTH:** The number of "tokens" the text generator outputs. A proxy for the number of generated words.
- **CLASS:** The corpus (see [Corpora](#)) that the text generator will use as a basis for its writing style and content.
- **117M/345M MODEL:** The chosen model for text generation. The 345M model is more powerful and produces more coherent text, but takes more than twice as long to run.

GPU STATUS:

SIMPLE GENERATE | CLASSIFY AND GENERATE

CLASSIFY, EXTRACT, AND GENERATE

It is a rare skill to stand at the foot of the mountain and be awed, and yet undaunted; to peer upwards at a summit hidden in the clouds and to respect its terrible beauty without succumbing to terror itself.

CREATIVITY 0.95 **LENGTH** 100 **CLASS** Business

117M Model

EXECUTE

(Enter text and press execute to get an original essay or story!)

The usage tips and background info are out of the way on a different page

Next to the EXECUTE button are parameter/model choices which are preset. Explanations nearby if interested.

Creating User Interface

Intuitive button to allow for minimizing side-plane

Quick description of what program does

It was important to use size and color to emphasize the text-box and highlight the “EXECUTE” button for “impatient” users

The screenshot shows the PHANTOM PEN web application. On the left is a blue sidebar with the title 'PHANTOM PEN' and navigation links: Home, Usage Tips, Background, Corpora, and Ethics. Below these are buttons for 'Github Repo' and 'Slides'. The main content area has a header with a close button (X) and the title 'PHANTOM PEN' with the subtitle 'YOUR DIGITAL GHOSTWRITER'. A paragraph describes the program as using retrained OpenAI GPT-2 models on a corpus of essays and stories. Below this is a section for 'GENERATOR CHOICES' with parameters for Creativity, Length, Class, and Model. A large text box contains a generated paragraph. At the bottom are input fields for Creativity (0.95), Length (100), and Class (Business), a dropdown for the Model (117M Model), and a prominent blue 'EXECUTE' button. A 'GPU STATUS' button is also visible. Annotations with orange lines point to various UI elements: the close button, the description paragraph, the generated text box, the EXECUTE button, the parameter choices, the GPU status button, the sidebar, and the overall layout.

PHANTOM PEN
YOUR DIGITAL GHOSTWRITER

Phantom Pen is a program that generates *original* writing using retrained versions of Open AI's [Generalized Pre-trained Transformer-2](#). The program has been re-trained on +2000 essays from *The Atlantic*, +160 short stories and 20 works of literature (see [Corpora](#) for a more specific description of the texts used).

There are four choices that enter into the "simple generate" function of the program:

GENERATOR CHOICES

- **CREATIVITY:** The amount of liberty the algorithm will take in generating new text; "1" being the most creative, and "0" leading to repetitive direct copying from corpus.
- **LENGTH:** The number of "tokens" the text generator outputs. A proxy for the number of generated words.
- **CLASS:** The corpus (see [Corpora](#)) that the text generator will use as a basis for its writing style and content.
- **117M/345M MODEL:** The chosen model for text generation. The 345M model is more powerful and produces more coherent text, but takes more than twice as long to run.

GPU STATUS:

CREATIVITY **LENGTH** **CLASS**

0.95 100 Business

117M Model

EXECUTE

(Enter text and press execute to get an original essay or story!)

It is a rare skill to stand at the foot of the mountain and be awed, and yet undaunted; to peer upwards at a summit hidden in the clouds and to respect its terrible beauty without succumbing to terror itself.

Home
Usage Tips
Background
Corpora
Ethics
Github Repo
Slides

The usage tips and background info are out of the way on a different page

GPU for faster text-generation

Next to the EXECUTE button are parameter/model choices which are preset. Explanations nearby if interested.

Creating User Interface

Intuitive button to allow for minimizing side-plane

Quick description of what program does

It was important to use size and color to emphasize the text-box and highlight the "EXECUTE" button for "impatient" users

The screenshot shows the PHANTOM PEN web application. On the left is a blue sidebar with the title 'PHANTOM PEN' and navigation links: Home, Usage Tips, Background, Corpora, Ethics, Github Repo, and Slides. The main content area has a header with a close button (X) and the title 'PHANTOM PEN' with the subtitle 'YOUR DIGITAL GHOSTWRITER'. Below this is a description of the program and its generator choices. On the right is a large text input box containing a sample paragraph, followed by controls for Creativity, Length, and Class, a model selector, and a large blue 'EXECUTE' button. A GPU status indicator is at the bottom left.

PHANTOM PEN
YOUR DIGITAL GHOSTWRITER

Phantom Pen is a program that generates *original* writing using retrained versions of Open AI's [Generalized Pre-trained Transformer-2](#). The program has been re-trained on +2000 essays from *The Atlantic*, +160 short stories and 20 works of literature (see [Corpora](#) for a more specific description of the texts used).

There are four choices that enter into the "simple generate" function of the program:

GENERATOR CHOICES

- **CREATIVITY:** The amount of liberty the algorithm will take in generating new text; "1" being the most creative, and "0" leading to repetitive direct copying from corpus.
- **LENGTH:** The number of "tokens" the text generator outputs. A proxy for the number of generated words.
- **CLASS:** The corpus (see [Corpora](#)) that the text generator will use as a basis for its writing style and content.
- **117M/345M MODEL:** The chosen model for text generation. The 345M model is more powerful and produces more coherent text, but takes more than twice as long to run.

GPU STATUS:

CREATIVITY 0.95 **LENGTH** 100 **CLASS** Business

117M Model

EXECUTE

(Enter text and press execute to get an original essay or story!)

The usage tips and background info are out of the way on a different page

For developers

GPU for faster text-generation

Next to the EXECUTE button are parameter/model choices which are preset. Explanations nearby if interested.

Designing Phantom Pen: Front End

Aspects of the Design

I. Front End

1. User Interface Structure
2. Usage Tips
3. Background
4. Corpora
5. Ethics

Motivation for Including Additional Sections

PHANTOM PEN

Home

Usage Tips

Background

Corpora

Ethics

Usage Tips:

From my own experience creating and testing the application, I learned useful generation strategies that would help a fresh-user (and would have helped me) obtain better outputs from the program.

Background:

Most people have seen text generators before, and most of these generators are unimpressive. Thus as background I thought it would be useful to explain how the algorithm at the background of this application is different from the others

Corpora:

The most important section from a text generation perspective. The output of the generator is highly dependent on the corpora it has been trained on so the user should be able to know what exactly constitutes that corpora

Ethics:

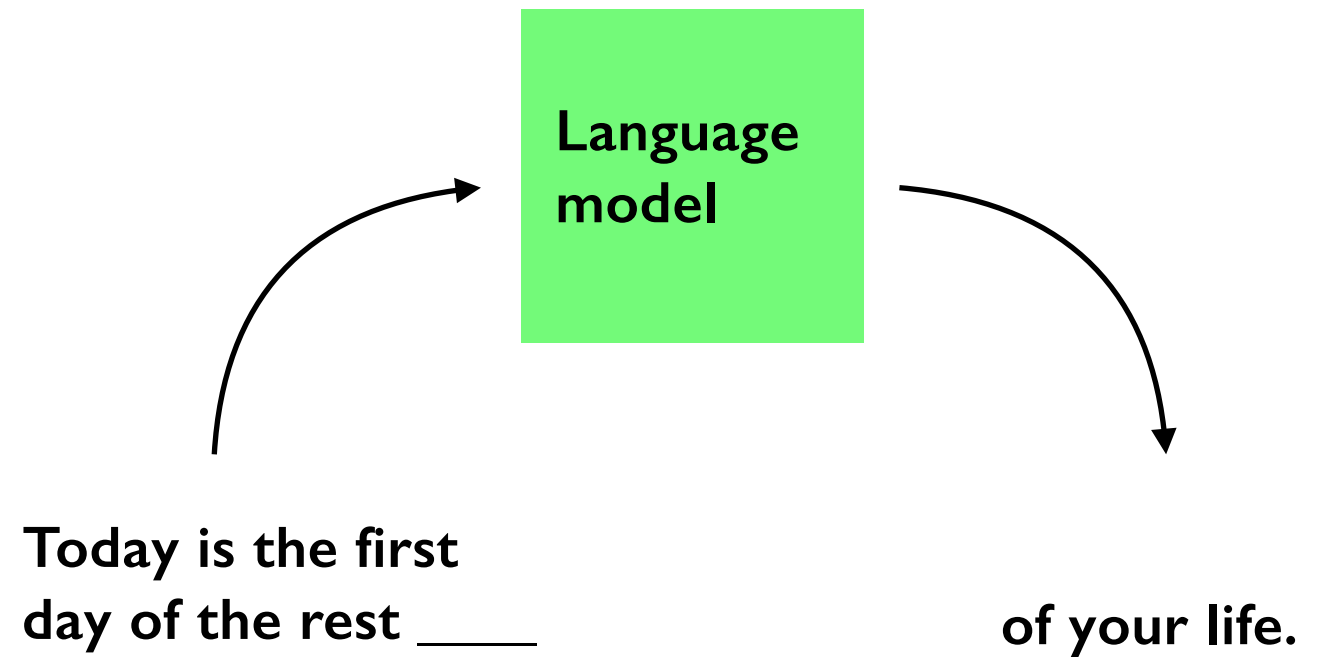
Besides “How do you validate this?” The most common question I received about this program is “How would you deal with plagiarism?” So I wanted to discuss this ethical question and some larger ones surrounding the application.

Designing Phantom Pen: Back End

Aspects of the Design

2. Back End

1. Language Model
2. Corpora and Aspects of Training
3. Classification Model
4. Extractive Summary



Language Model - Choices

Possibilities (besides Transformer Model)

1. Markov Chain Monte Carlo
2. Recurrent Neural Network

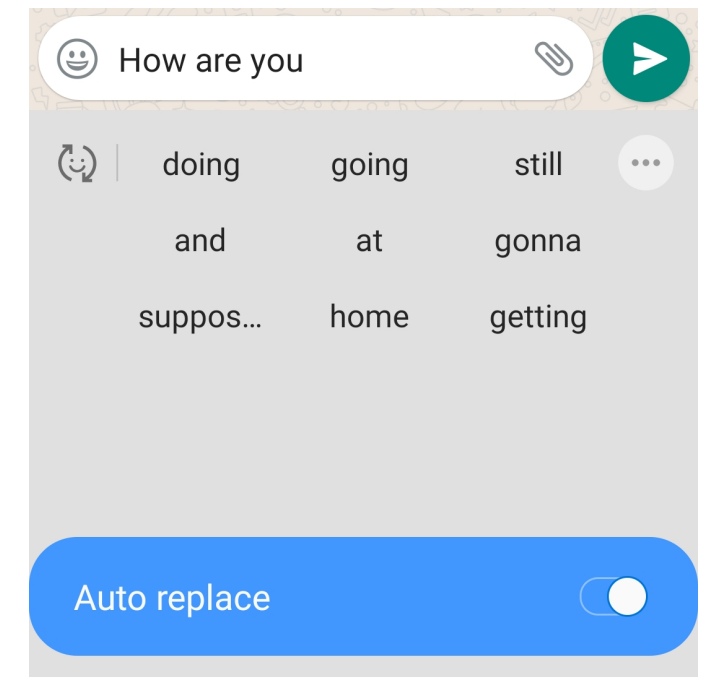
Markov model cannot model context

Possibilities (besides Transformer Model)

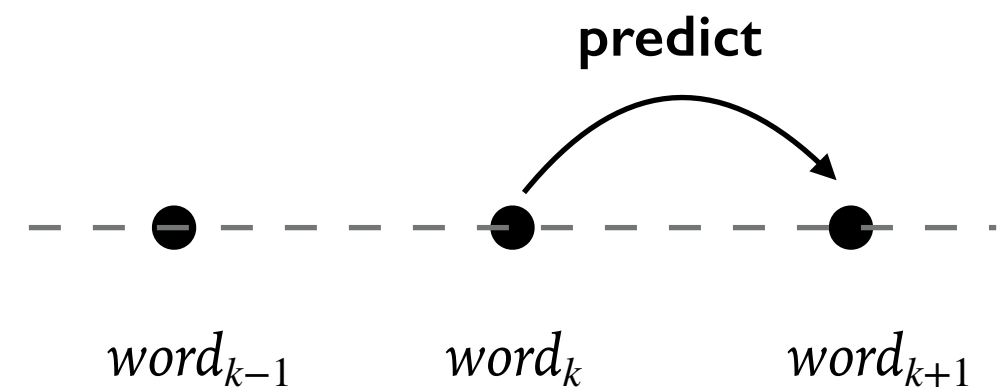
1. Markov Chain Monte Carlo
2. Recurrent Neural Network

Comments about Model

- Conceptually the simplest model to explain
- But, because of its context independence it clearly does not work for generating human-like text.
- Instead, it was used as a baseline model to practice text generation



(WhatsApp)



Markov Chain Monte Carlo

Recurrent Neural Network: Better but not excellent

Possibilities (besides Transformer Model)

1. Markov Chain Monte Carlo
2. Recurrent Neural Network

Comments about Model

- With the Long Short-Term Memory (LSTM) node architecture, RNN's were the SOTA in language model generation
- Clearly much better than Markov Model but still doesn't "sound" like something a human would write (even with LSTM)
- Was the model I was going to use, before I discovered the transformer model

Dinner next week

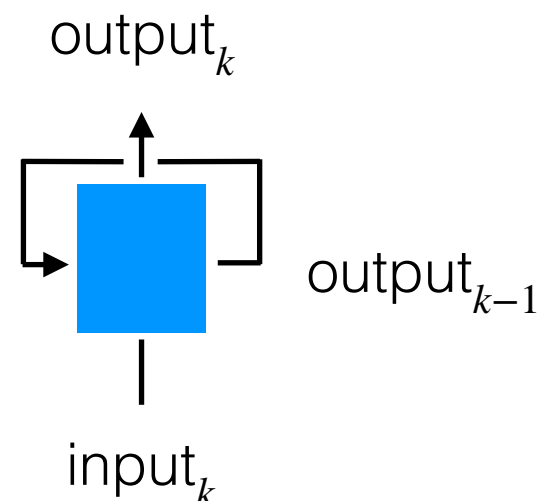
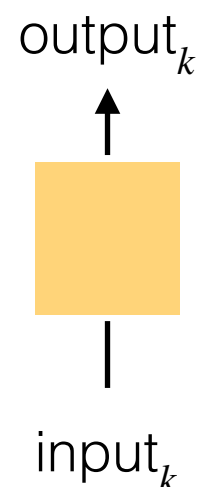
Evan Brown, Maalika Patel

Dinner next week

Hey Evan and Maalika,

Haven't seen you in a while!

(Gmail Smart Compose)



Vanilla Neural Network

Recurrent Neural Network

Output after Training on *Atlantic: Education Corpora*

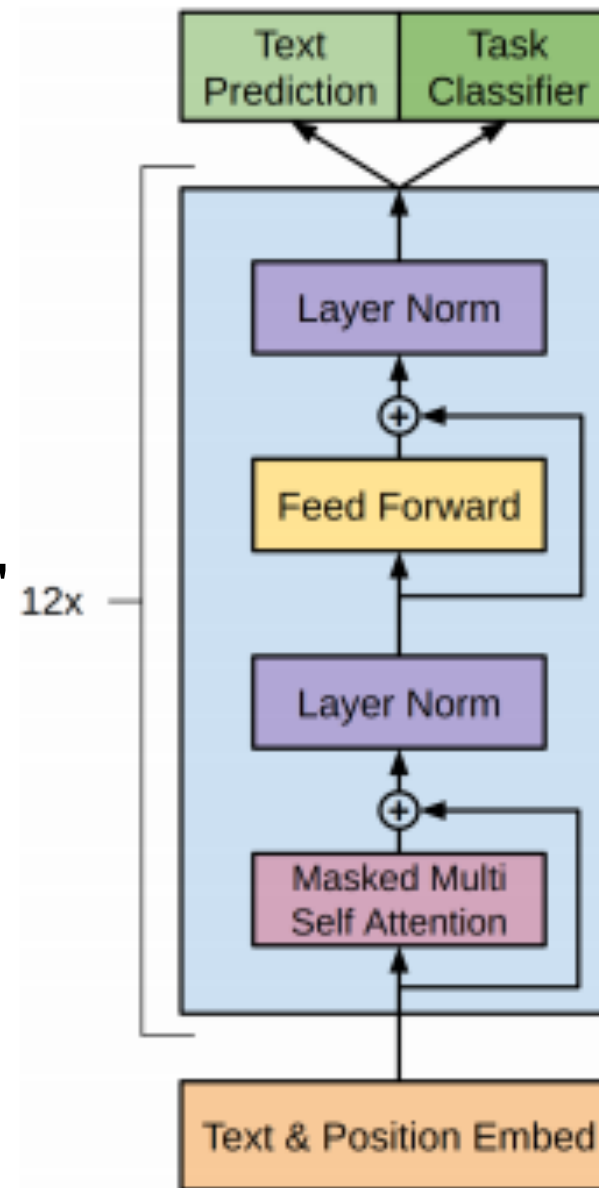
As they develop admissions discussion about health on the college in the four-year college are a letter who were shown about an extremely policy off the feeling about their present.

Transformer Model

Transformer Model

Comments about Model

- Introduced by Google Brain in 2017, and quickly thereafter became SOTA
- “**Attention**” is the key feature of the model
- Open AI released, "small" and "medium" parameter versions of their Transformer Language Model Called Generalized Pre Trained Transformer (GPT)
- Does not propagate hidden states from input to input as in an RNN



Generalized Pre-Trained Transformer (GPT) Architecture

Network Architecture

117M Hparams (Small)

- 12 layers
- 12 Attention Heads
- 768 dimensional embedding
- 1024 length context vector (for position embedding)
- 50257 vocabulary size

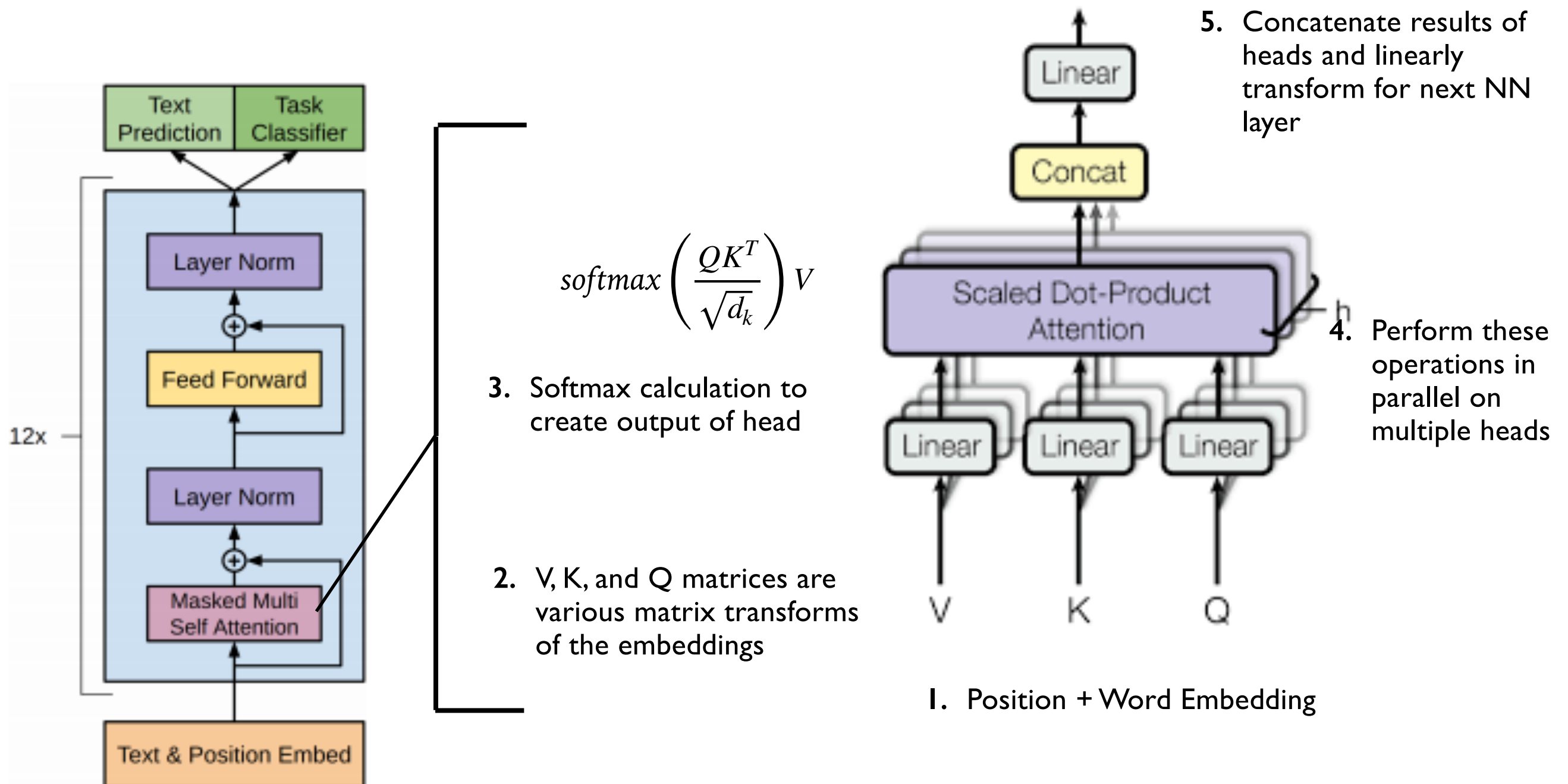
345M Hparams (Medium)

- 24 layers
- 16 Attention Heads
- 1024 dimensional embedding
- 1024 length context vector (for position embedding)
- 50257 vocabulary size (for word embedding)

Transformer Model - Multi headed Attention

Transformer Model

- Multi headed attention

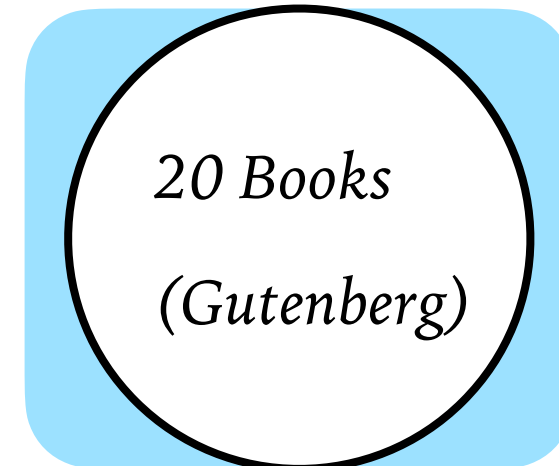
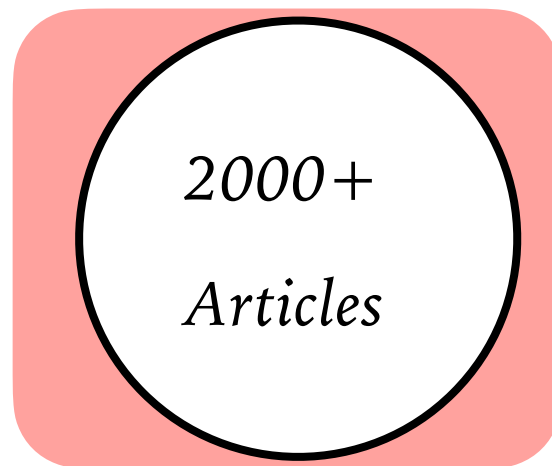


Designing Phantom Pen: Back End

Aspects of the Design

2. Back End

1. Language Model
2. Corpora and Aspects of Training
3. Classification Model
4. Extractive Summary

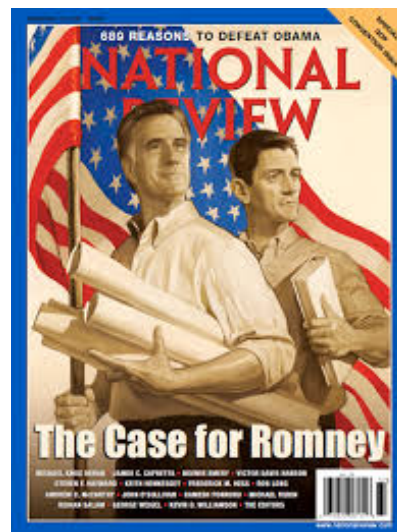


Corpora and Aspects of Training

The initial plan was to use writing from *SLATE*, *The Atlantic*, *National Review*, *SLATE*, and *NY POST* as the various corpora



The Atlantic



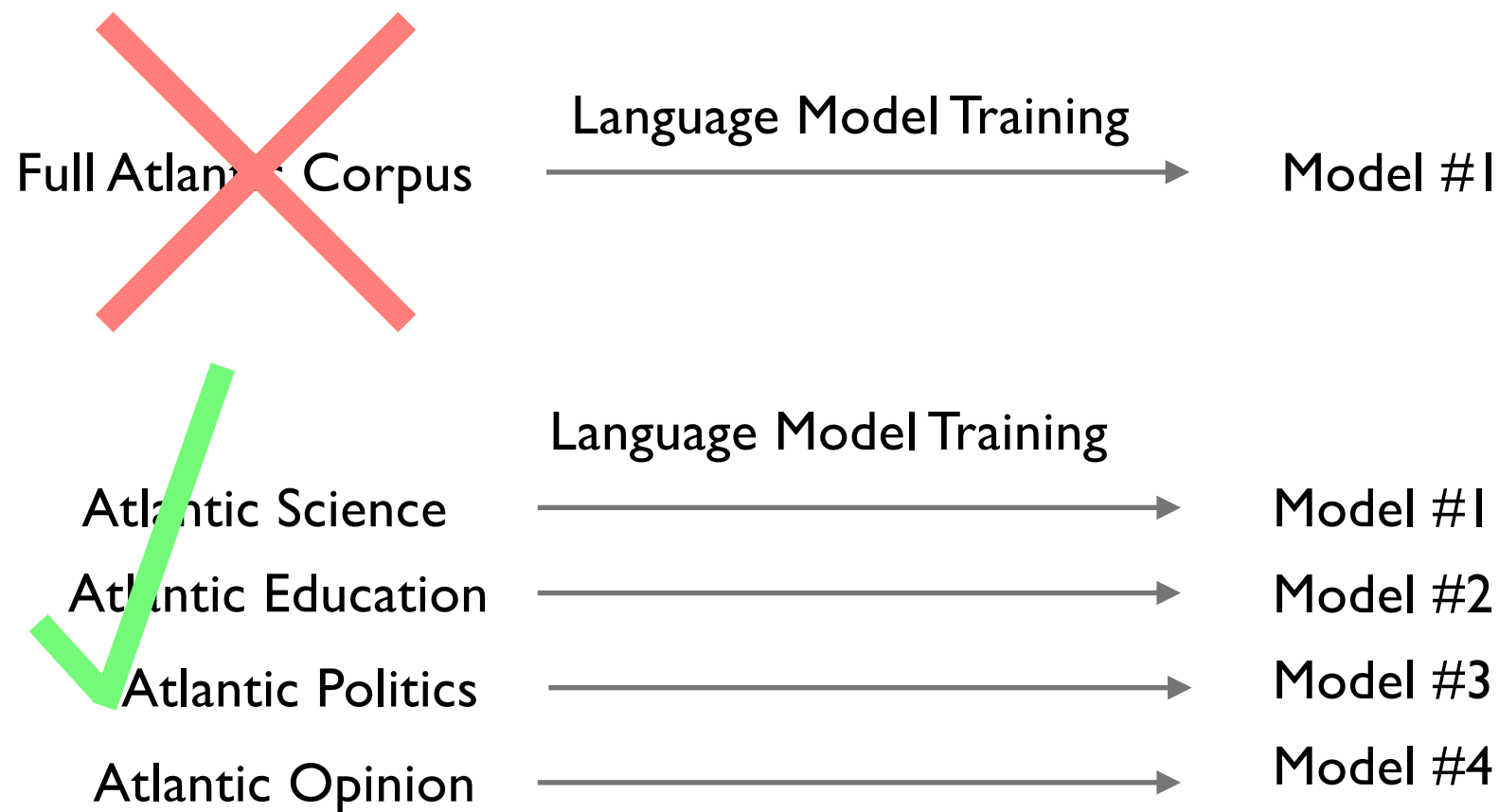
Problems with selecting these corpora

1. There wasn't a sufficiently good use case to motivate these specific corpora
2. The diversity in topics is too much in the direction of the original GPT2 data set of 40 GB of internet text (i.e., no real advantage in re-training)

Corpora and Aspects of Training

I later decided it would make more sense to select corpora according to topic rather than publication

This way the application's essay completions/ generations are more specific to a user chosen class and hence more useful to the user



*Sacrificed ideological diversity for output specificity

Improvements: Find a way to train a single model, rather than 2 X 11 different models.

Corpora and Aspects of Training

2000+
Articles

Nine Article Categories
from *The Atlantic*

- Business
- Education
- Entertainment
- Health
- Ideas
- International
- Politics
- Science
- Technology

Introduced **short story** and **Project Gutenberg** corpora and trained models to give the user more options.

160 Short
Stories

62 Authors

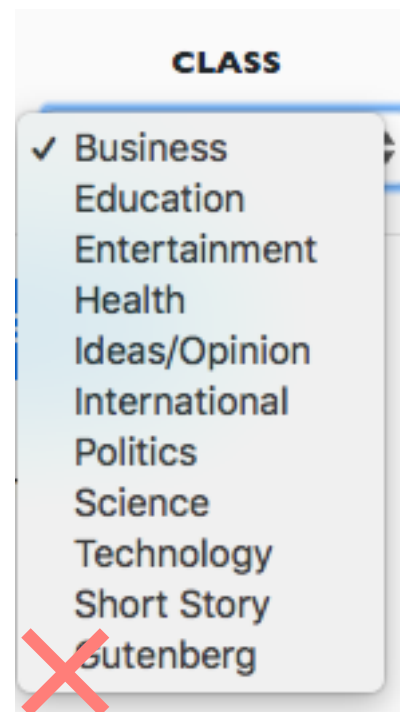
- Anton Chekov
- Henry David Thoreau
- Jack London
- H.P. Lovecraft

20 Books
(Gutenberg)

Including

- Pride and Prejudice
- The Count of Monte Cristo
- Robinson Crusoe
- Moby Dick

Improvements: In a second version of the model, I would **remove the Gutenberg class** since its texts are structurally different from the other texts

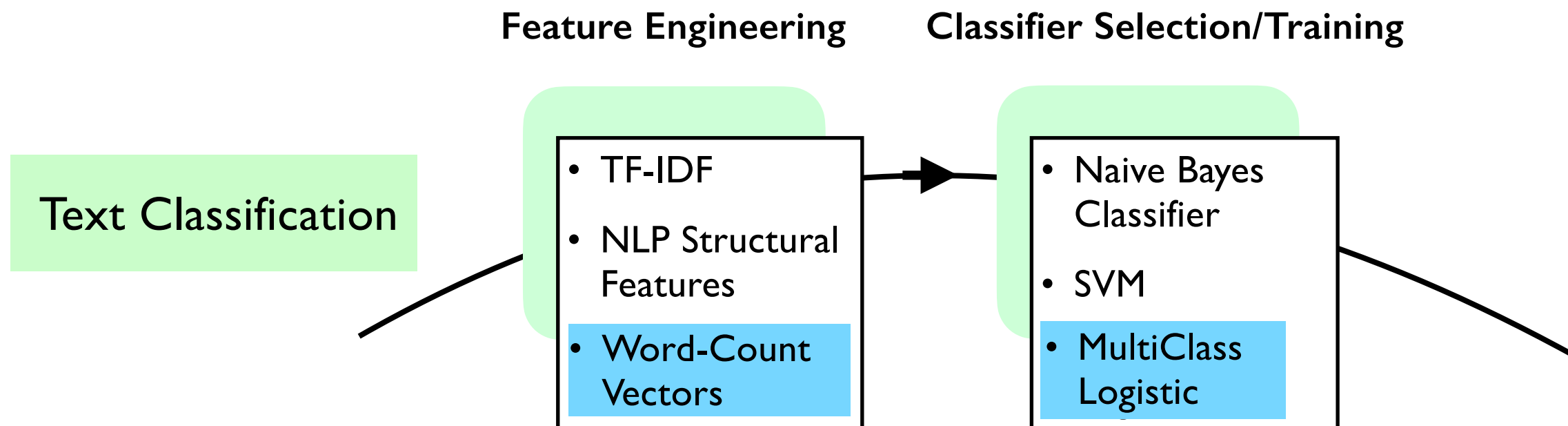


Designing Phantom Pen: Back End

Aspects of the Design

2. Back End

1. Language Model
2. Corpora and Aspects of Training
3. Classification Model
4. Extractive Summary

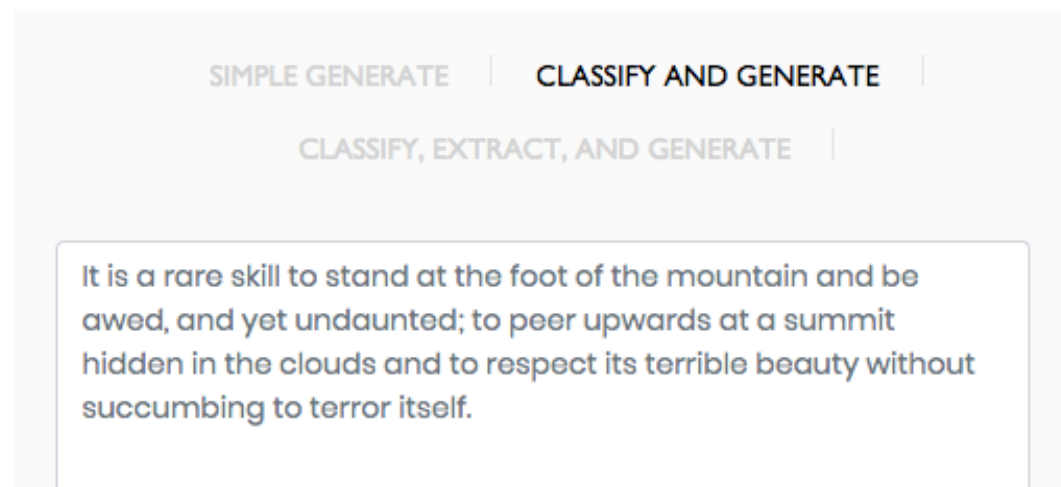


Classifying Text for the User

Beyond simple text generation, I wanted to add additional functions to allow for more flexible user input

Additional Functionality #2:
Determine the class for the user

Program screenshot



Text Classification Problem

We want to determine which of the 10 classes the user's entered text is most similar to

Feature Engineering

Word Count Vectors

Word-Embedding

TF-IDF

Classifier Model

Naive Bayesian Classifier

Support Vector Machine

Logistic Regression

Classifying Text For the User

SVM; Acc = 0.6525

Feature Engineering

Classifier Model

Word Count Vectors

Naive Bayesian Classifier

Word-Embeddings

Support Vector Machine

TF-IDF

Logistic Regression

- Simplest model had the highest accuracy

https://github.com/mowillia/phantom_pen/blob/master/classification_comparisons.ipynb

Logistic Regression; Acc = 0.831

```
accuracy 0.8308823529411765
precision recall f1-score support
business 0.70 0.79 0.74 47
education 0.93 0.82 0.87 62
entertainment 0.82 0.90 0.86 62
health 0.89 0.76 0.82 62
ideas 0.71 0.60 0.65 57
international 0.81 0.86 0.83 29
politics 0.84 0.89 0.87 55
science 0.84 0.95 0.89 56
short-story 0.94 1.00 0.97 47
technology 0.82 0.79 0.80 67

accuracy 0.83
macro avg 0.83 0.84 0.83 544
weighted avg 0.83 0.83 0.83 544
```

```
accuracy 0.6525735294117647
precision recall f1-score support
business 0.72 0.49 0.58 47
education 0.57 0.95 0.72 62
entertainment 0.96 0.73 0.83 62
health 1.00 0.21 0.35 62
ideas 0.30 0.56 0.40 57
international 1.00 0.03 0.07 29
politics 0.61 0.87 0.72 55
science 0.72 0.84 0.78 56
short-story 1.00 0.96 0.98 47
technology 0.78 0.63 0.69 67

accuracy 0.65
macro avg 0.77 0.63 0.61 544
weighted avg 0.75 0.65 0.63 544
```

Naive Bayes Classifier; Acc = 0.80

```
accuracy 0.7996323529411765
precision recall f1-score support
business 0.72 0.72 0.72 47
education 0.87 0.85 0.86 62
entertainment 0.78 0.90 0.84 62
health 0.85 0.66 0.75 62
ideas 0.73 0.42 0.53 57
international 0.83 0.86 0.85 29
politics 0.75 0.91 0.82 55
science 0.77 0.96 0.86 56
short-story 0.90 1.00 0.95 47
technology 0.80 0.76 0.78 67

accuracy 0.80
macro avg 0.80 0.81 0.80 544
weighted avg 0.80 0.80 0.79 544
```

Designing Phantom Pen: Back End

Aspects of the Design

2. Back End

1. Language Model
2. Corpora and Aspects of Training
3. Classification Model
4. Extractive Summary



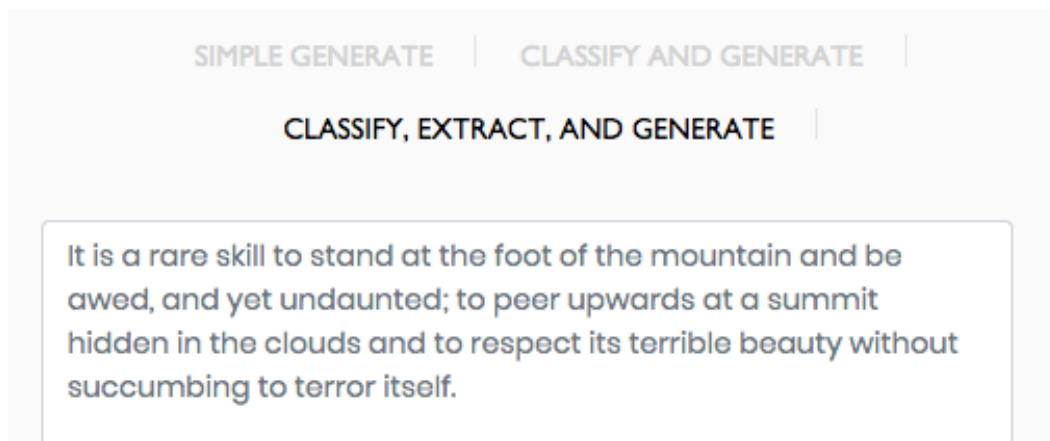
Extracting Important Sentences from Text

The program performed less well on longer texts, so I wanted a way to condense longer texts into shorter text

Additional Functionality #3:

Extract important sentences, classify entire text, and generate new text from extraction

Program screenshot



https://github.com/mowillia/phantom_pen/tree/master

https://github.com/mowillia/phantom_pen/blob/master/text_generation_function.ipynb

Text Extraction Problem

We want to condense a long essay into a shorter one while keeping as much as the important information as possible.

Extractive Summarization

Pick out the sentences that are most important

I considered three approaches

- Word Frequency
- Word Embeddings and Cosine Similarity
- Word-count vectors and Cosine Similarity

*Metric of performance based on how well it picks out sentences I considered important

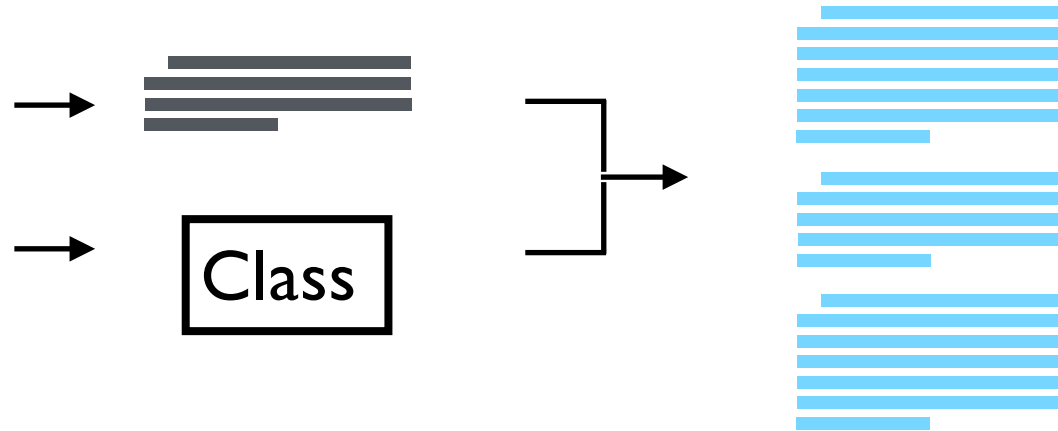
— More rigorous approach: Hand labeling and then checking accuracy

Functions of application

MAIN FUNCTION

SIMPLE GENERATE

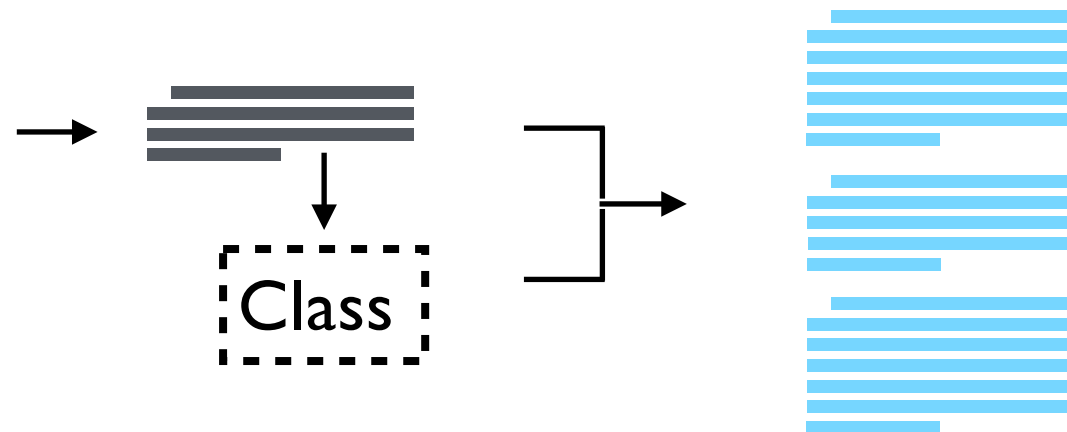
Generates an essay given a **prompt** and a chosen **class**



ADDITIONAL FUNCTION #1

CLASSIFY AND GENERATE

Generates an essay given a **prompt** and the class determined from the prompt

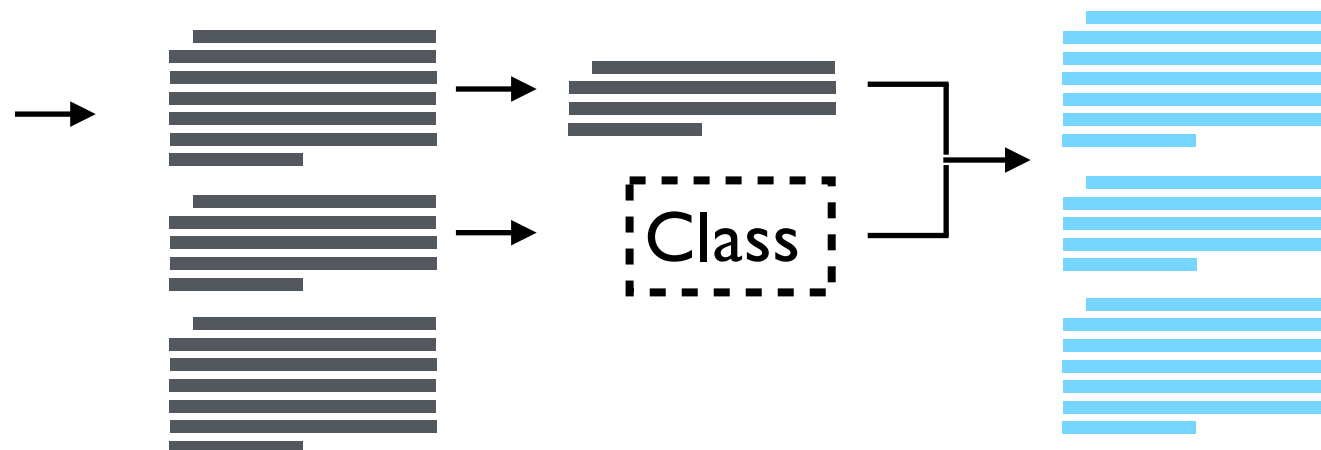


ADDITIONAL FUNCTION #2

CLASSIFY, EXTRACT, AND GENERATE

Generates an essay given a **previous essay**

- Classifies and extracts the important sentences and generates essay from both





END