TutorBot Backend

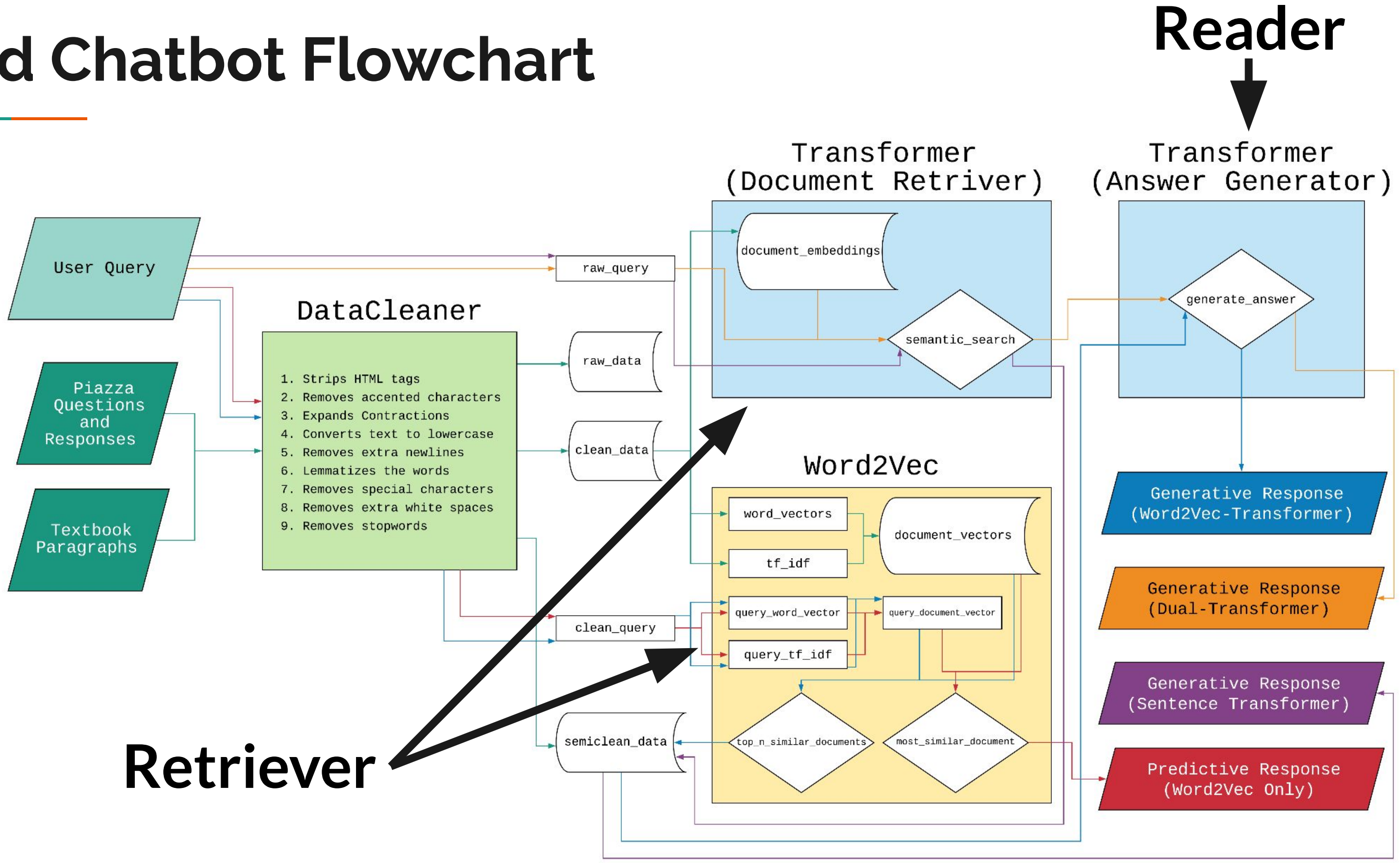
Spring 2022

Aahan Kerawala, Pat Tran, Christina Chen

Introduction

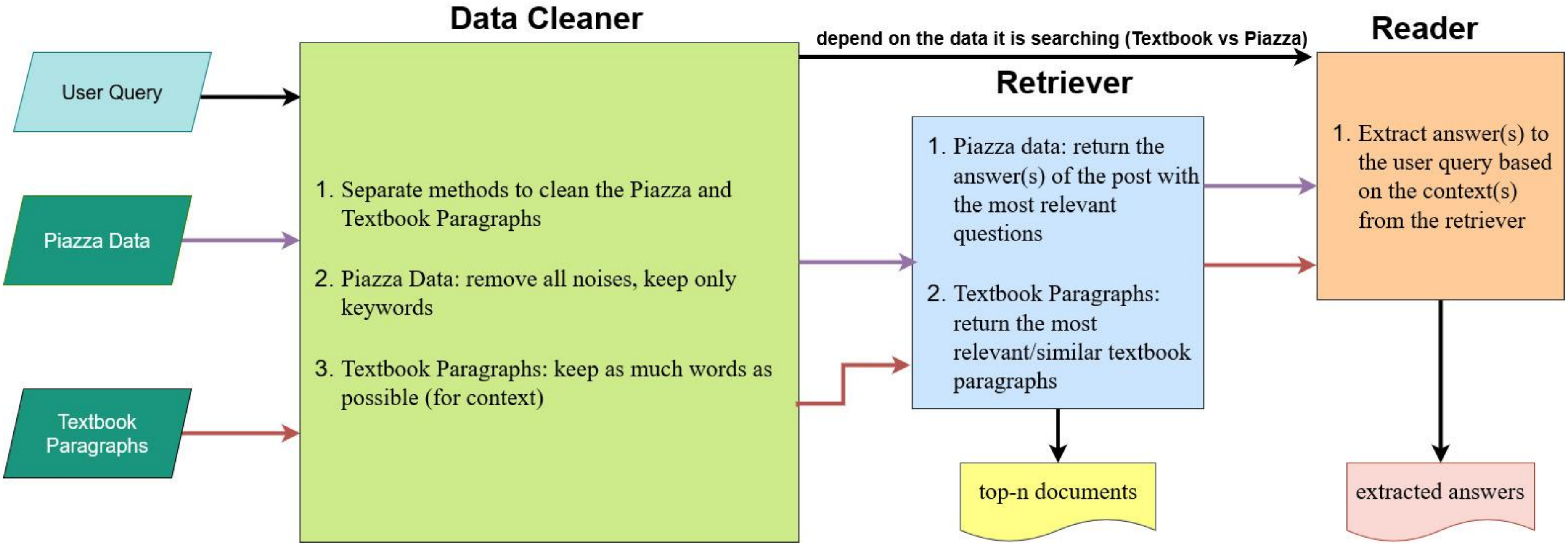
- In previous semesters, we created **TutorBot**, a combination between TutorJS and Chatbot team.
  - The TutorJS's goal was to help electrical engineering students solve signal processing algorithms in Javascript
  - The Chatbot generate answers from the course discussion and textbook from and input question.
- This presentation focuses on the Chatbot:
  - Run on Google Colab
  - Needs GPU support to extract answers quickly

Old Chatbot Flowchart



Reader
Transformer (Document Retriver)
Transformer (Answer Generator)
User Query
raw_query
document_embeddings
semantic_search
generate_answer
DataCleaner
1. Strips HTML tags
2. Removes accented characters
3. Expands Contractions
4. Converts text to lowercase
5. Removes extra newlines
6. Lemmatizes the words
7. Removes special characters
8. Removes extra white spaces
9. Removes stopwords
Piazza Questions and Responses
Textbook Paragraphs
raw_data
clean_data
Word2Vec
word_vectors
tf_idf
document_vectors
query_word_vector
query_tf_idf
query_document_vector
clean_query
semiclean_data
top_n_similar_documents
most_similar_document
Generative Response (Word2Vec-Transformer)
Generative Response (Dual-Transformer)
Generative Response (Sentence Transformer)
Predictive Response (Word2Vec Only)
Retriever

Current (Simplified) Chatbot Flowchart



Data Cleaner

1. Separate methods to clean the Piazza and Textbook Paragraphs
2. Piazza Data: remove all noises, keep only keywords
3. Textbook Paragraphs: keep as much words as possible (for context)

Retriever

1. Piazza data: return the answer(s) of the post with the most relevant questions
2. Textbook Paragraphs: return the most relevant/similar textbook paragraphs

Reader

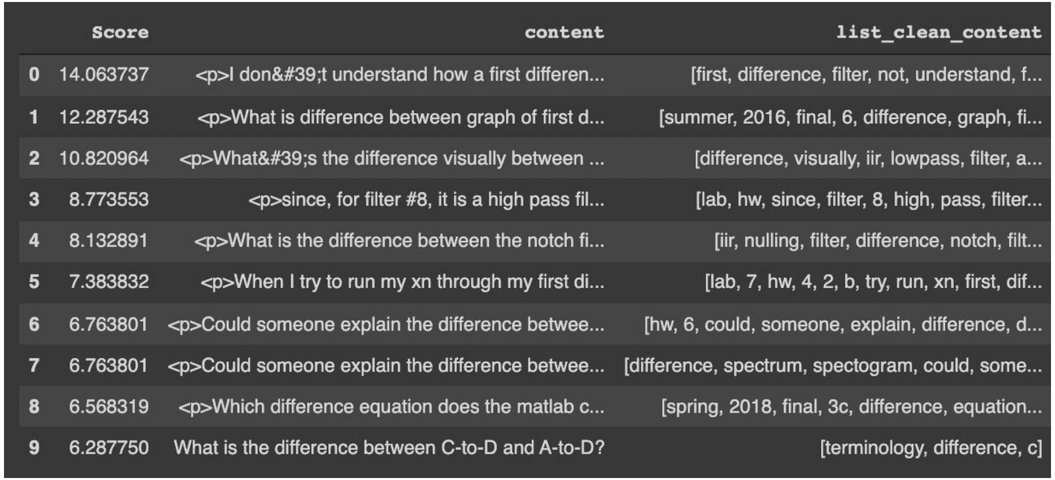
1. Extract answer(s) to the user query based on the context(s) from the retriever

Chatbot: Retriever (Search Engine)

- Model: BM25 or Word2Vec
- Behave similarly to a search engine
- Find top-n relevant document
- Works on the concept of TF/IDF

Question: What is a first-difference filter?

=> ['first', 'difference', 'filter']



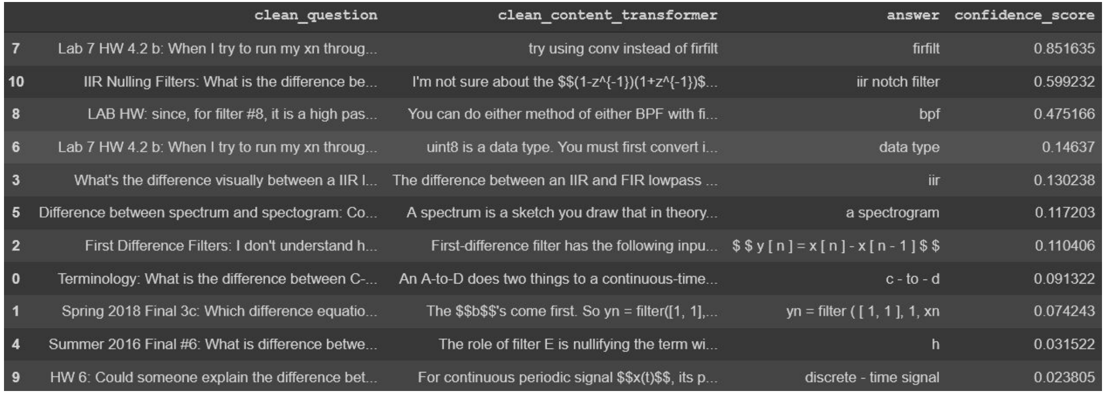
	Score	content	list_clean_content
0	14.063737	<p>I don&#39;t understand how a first differen...	[first, difference, filter, not, understand, f...
1	12.287543	<p>What is difference between graph of first d...	[summer, 2016, final, 6, difference, graph, fi...
2	10.820964	<p>What&#39;s the difference visually between ...	[difference, visually, iir, lowpass, filter, a...
3	8.773553	<p>since, for filter #8, it is a high pass fil...	[lab, hw, since, filter, 8, high, pass, filter...
4	8.132891	<p>What is the difference between the notch fi...	[iir, nulling, filter, difference, notch, filt...
5	7.383832	<p>When I try to run my xn through my first di...	[lab, 7, hw, 4, 2, b, try, run, xn, first, dif...
6	6.763801	<p>Could someone explain the difference betwee...	[hw, 6, could, someone, explain, difference, d...
7	6.763801	<p>Could someone explain the difference betwee...	[difference, spectrum, spectogram, could, some...
8	6.568319	<p>Which difference equation does the matlab c...	[spring, 2018, final, 3c, difference, equation...
9	6.287750	What is the difference between C-to-D and A-to-D?	[terminology, difference, c]

Note: only useful columns are kept to showcase

Chatbot: Reader (Transformers)

- Extract the answer from the top-n relevant documents
  - clean_content_transformer column
- Current pre-trained model: ahotrod/electra_large_discriminator_squad2_512
- Problem 1: slow without a GPU
- Problem 2: the model was trained on SQuAD 2.0 so it does not perform well on DSPFirst textbook

Question: What is a first-difference filter?

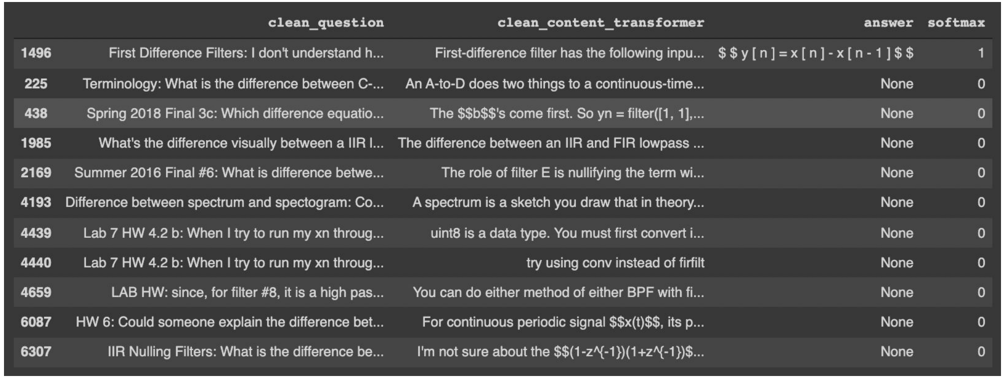


	clean_question	clean_content_transformer	answer	confidence_score
7	Lab 7 HW 4.2 b: When I try to run my xn throug...	try using conv instead of firfilt	firfilt	0.851635
10	IIR Nulling Filters: What is the difference be...	I'm not sure about the $$(1-z^{-1})(1+z^{-1})$...	iir notch filter	0.599232
8	LAB HW: since, for filter #8, it is a high pas...	You can do either method of either BPF with fi...	bpf	0.475166
6	Lab 7 HW 4.2 b: When I try to run my xn throug...	uint8 is a data type. You must first convert i...	data type	0.14637
3	What's the difference visually between a IIR l...	The difference between an IIR and FIR lowpass ...	iir	0.130238
5	Difference between spectrum and spectogram: Co...	A spectrum is a sketch you draw that in theory...	a spectrogram	0.117203
2	First Difference Filters: I don't understand h...	First-difference filter has the following inpu...	$ $ y [n] = x [n] - x [n - 1] $ $	0.110406
0	Terminology: What is the difference between C-...	An A-to-D does two things to a continuous-time...	c - to - d	0.091322
1	Spring 2018 Final 3c: Which difference equatio...	The $$b$$'s come first. So yn = filter([1, 1],...	yn = filter ([1, 1], 1, xn	0.074243
4	Summer 2016 Final #6: What is difference betwe...	The role of filter E is nullifying the term wi...	h	0.031522
9	HW 6: Could someone explain the difference bet...	For continuous periodic signal $$x(t)$$, its p...	discrete - time signal	0.023805

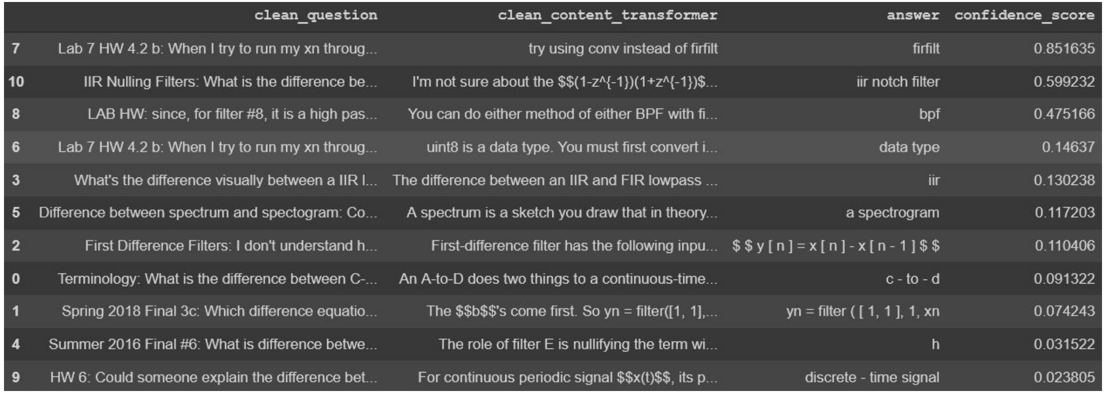
Note: only useful columns are kept to showcase

Old vs. New Transformer Algorithm

Question: What is a first-difference filter?



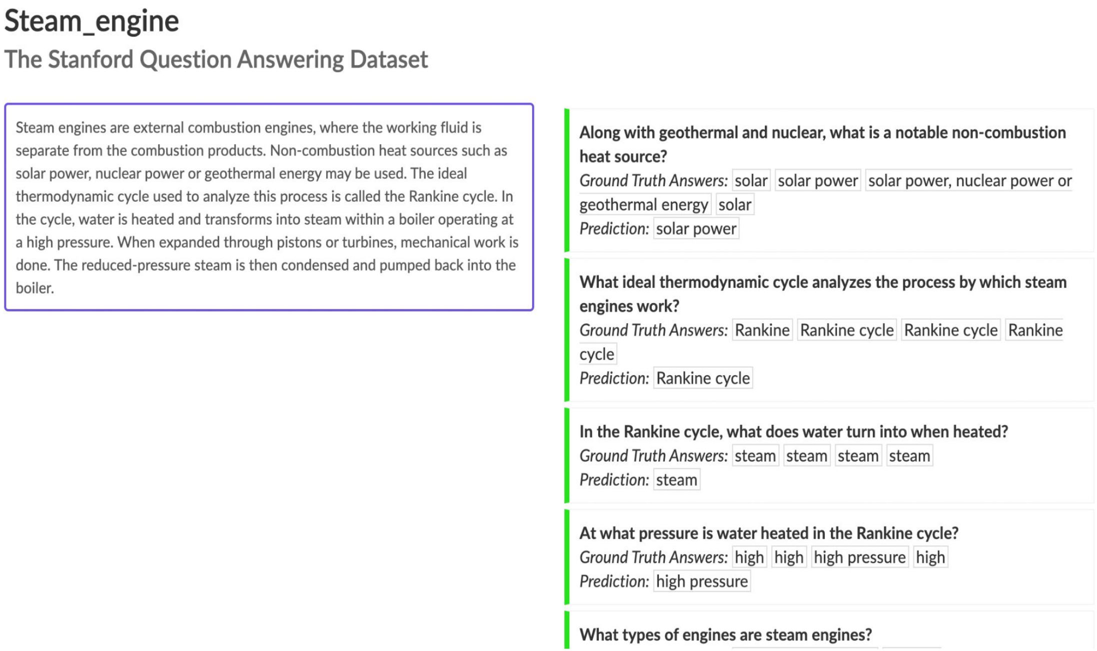
	clean_question	clean_content_transformer	answer	softmax
1496	First Difference Filters: I don't understand h...	First-difference filter has the following inpu...	$ $ y [n] = x [n] - x [n - 1] $ $	1
225	Terminology: What is the difference between C-...	An A-to-D does two things to a continuous-time...	None	0
438	Spring 2018 Final 3c: Which difference equatio...	The $$b$$'s come first. So yn = filter([1, 1],...	None	0
1985	What's the difference visually between a IIR l...	The difference between an IIR and FIR lowpass ...	None	0
2169	Summer 2016 Final #6: What is difference betwe...	The role of filter E is nullifying the term wi...	None	0
4193	Difference between spectrum and spectogram: Co...	A spectrum is a sketch you draw that in theory...	None	0
4439	Lab 7 HW 4.2 b: When I try to run my xn throug...	uint8 is a data type. You must first convert i...	None	0
4440	Lab 7 HW 4.2 b: When I try to run my xn throug...	try using conv instead of firfilt	None	0
4659	LAB HW: since, for filter #8, it is a high pas...	You can do either method of either BPF with fi...	None	0
6087	HW 6: Could someone explain the difference bet...	For continuous periodic signal $$x(t)$$, its p...	None	0
6307	IIR Nulling Filters: What is the difference be...	I'm not sure about the $$(1-z^{-1})(1+z^{-1})$...	None	0



	clean_question	clean_content_transformer	answer	confidence_score
7	Lab 7 HW 4.2 b: When I try to run my xn throug...	try using conv instead of firfilt	firfilt	0.851635
10	IIR Nulling Filters: What is the difference be...	I'm not sure about the $$(1-z^{-1})(1+z^{-1})$...	iir notch filter	0.599232
8	LAB HW: since, for filter #8, it is a high pas...	You can do either method of either BPF with fi...	bpf	0.475166
6	Lab 7 HW 4.2 b: When I try to run my xn throug...	uint8 is a data type. You must first convert i...	data type	0.14637
3	What's the difference visually between a IIR l...	The difference between an IIR and FIR lowpass ...	iir	0.130238
5	Difference between spectrum and spectogram: Co...	A spectrum is a sketch you draw that in theory...	a spectrogram	0.117203
2	First Difference Filters: I don't understand h...	First-difference filter has the following inpu...	$ $ y [n] = x [n] - x [n - 1] $ $	0.110406
0	Terminology: What is the difference between C-...	An A-to-D does two things to a continuous-time...	c - to - d	0.091322
1	Spring 2018 Final 3c: Which difference equatio...	The $$b$$'s come first. So yn = filter([1, 1],...	yn = filter ([1, 1], 1, xn	0.074243
4	Summer 2016 Final #6: What is difference betwe...	The role of filter E is nullifying the term wi...	h	0.031522
9	HW 6: Could someone explain the difference bet...	For continuous periodic signal $$x(t)$$, its p...	discrete - time signal	0.023805

SQuAD 2.0 Dataset

- Biggest problem: not enough data
- Solution: generating questions and answers from DSPFirst textbook
- Similar to SQuAD (The Stanford Question Answering Dataset) format



Steam_engine

The Stanford Question Answering Dataset

Steam engines are external combustion engines, where the working fluid is separate from the combustion products. Non-combustion heat sources such as solar power, nuclear power or geothermal energy may be used. The ideal thermodynamic cycle used to analyze this process is called the Rankine cycle. In the cycle, water is heated and transforms into steam within a boiler operating at a high pressure. When expanded through pistons or turbines, mechanical work is done. The reduced-pressure steam is then condensed and pumped back into the boiler.

Along with geothermal and nuclear, what is a notable non-combustion heat source?
Ground Truth Answers: solar | solar power | solar power, nuclear power or geothermal energy | solar
Prediction: solar power

What ideal thermodynamic cycle analyzes the process by which steam engines work?
Ground Truth Answers: Rankine | Rankine cycle | Rankine cycle | Rankine cycle
Prediction: Rankine cycle

In the Rankine cycle, what does water turn into when heated?
Ground Truth Answers: steam | steam | steam | steam
Prediction: steam

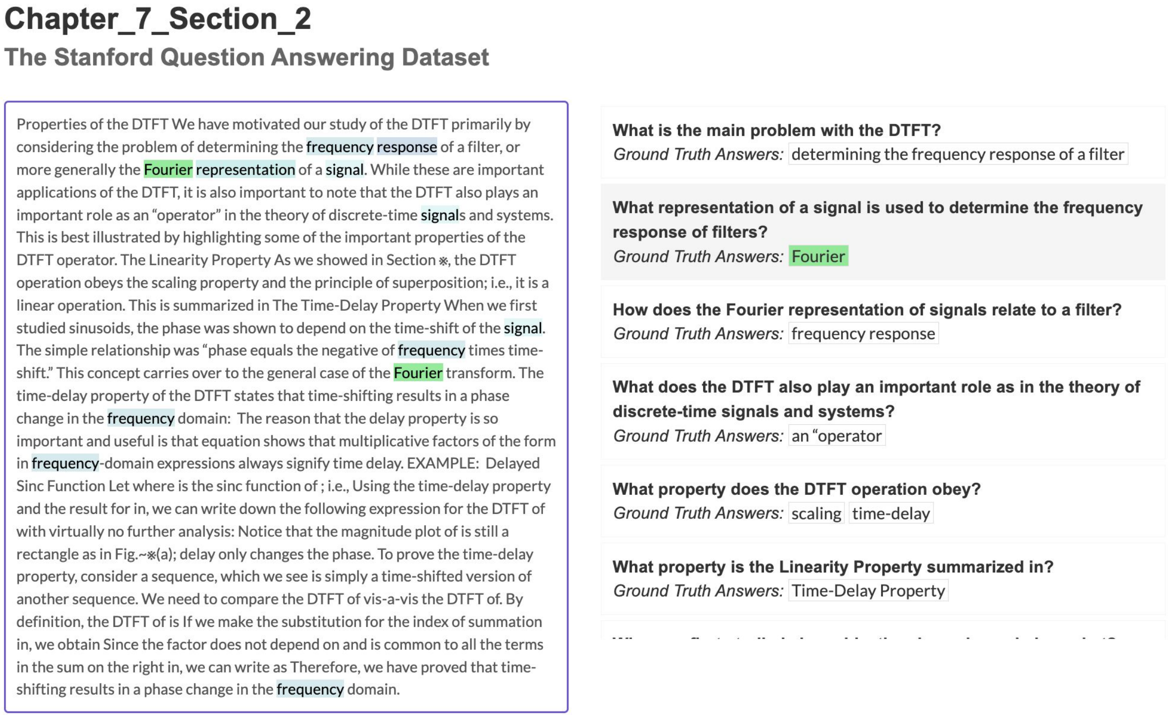
At what pressure is water heated in the Rankine cycle?
Ground Truth Answers: high | high | high pressure | high
Prediction: high pressure

What types of engines are steam engines?

SQuAD 2.0 Dataset: Topic - Steam Engine

DSPFirst Dataset

- Purpose: observe how the machine understand and interpret the cleaned DSPFirst data, and fine-tune the pre-trained model.
- Generates questions using a Transformer pre-trained model (T5) specifically trained for this task
- The answer(s) were also generated from the same T5 model.
- The generated answer(s) can be incorrect



Chapter_7_Section_2

The Stanford Question Answering Dataset

Properties of the DTFT We have motivated our study of the DTFT primarily by considering the problem of determining the frequency response of a filter, or more generally the Fourier representation of a signal. While these are important applications of the DTFT, it is also important to note that the DTFT also plays an important role as an “operator” in the theory of discrete-time signals and systems. This is best illustrated by highlighting some of the important properties of the DTFT operator. The Linearity Property As we showed in Section ⋇, the DTFT operation obeys the scaling property and the principle of superposition; i.e., it is a linear operation. This is summarized in The Time-Delay Property When we first studied sinusoids, the phase was shown to depend on the time-shift of the signal. The simple relationship was “phase equals the negative of frequency times time-shift.” This concept carries over to the general case of the Fourier transform. The time-delay property of the DTFT states that time-shifting results in a phase change in the frequency domain: The reason that the delay property is so important and useful is that equation shows that multiplicative factors of the form in frequency-domain expressions always signify time delay. EXAMPLE: Delayed Sinc Function Let where is the sinc function of ; i.e., Using the time-delay property and the result for in, we can write down the following expression for the DTFT of with virtually no further analysis: Notice that the magnitude plot of is still a rectangle as in Fig.~⋇(a); delay only changes the phase. To prove the time-delay property, consider a sequence, which we see is simply a time-shifted version of another sequence. We need to compare the DTFT of vis-a-vis the DTFT of. By definition, the DTFT of is If we make the substitution for the index of summation in, we obtain Since the factor does not depend on and is common to all the terms in the sum on the right in, we can write as Therefore, we have proved that time-shifting results in a phase change in the frequency domain.

What is the main problem with the DTFT?
Ground Truth Answers: determining the frequency response of a filter

What representation of a signal is used to determine the frequency response of filters?
Ground Truth Answers: Fourier

How does the Fourier representation of signals relate to a filter?
Ground Truth Answers: frequency response

What does the DTFT also play an important role as in the theory of discrete-time signals and systems?
Ground Truth Answers: an “operator

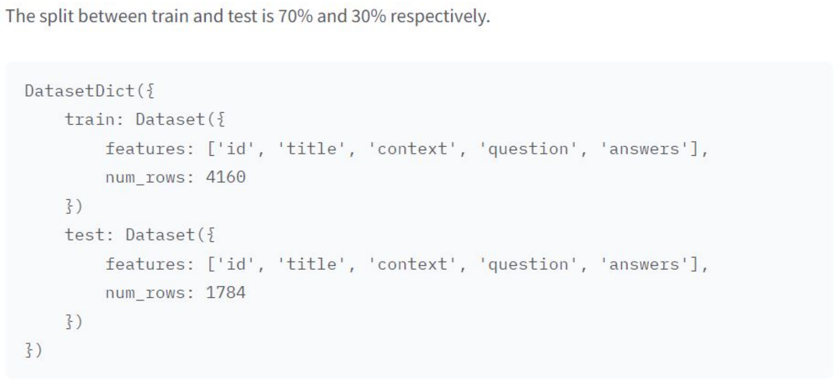
What property does the DTFT operation obey?
Ground Truth Answers: scaling time-delay

What property is the Linearity Property summarized in?
Ground Truth Answers: Time-Delay Property

DSPFirst Dataset: Chapter 7 Section 2

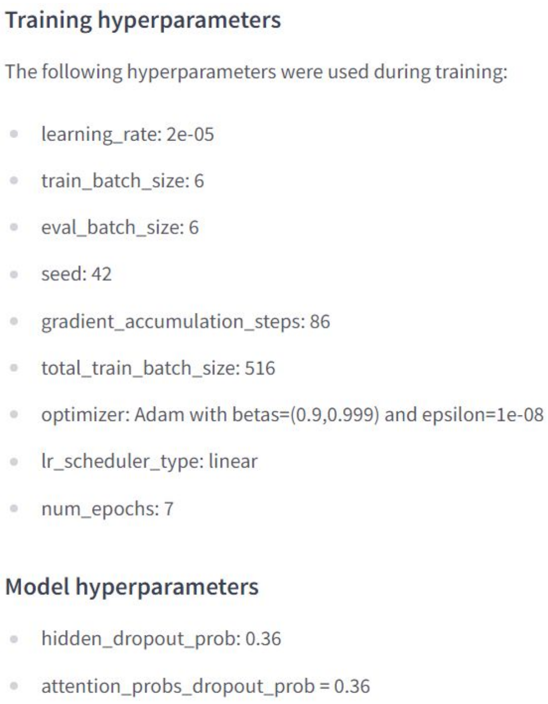
Fine-tuning

- Current pre-trained model: **electra_large_discriminator_squad2_512**
- Utilize the generated Question and Answer Dataset to fine-tune the pre-trained QA model
- per_device_batch_size of 6 results in 14.82 GB VRAM
- Utilizes *gradient_accumulation_steps* to get total batch size to 514
  - Total batch size should be at least 256



The split between train and test is 70% and 30% respectively.

```
DatasetDict({
    train: Dataset({
        features: ['id', 'title', 'context', 'question', 'answers'],
        num_rows: 4160
    })
    test: Dataset({
        features: ['id', 'title', 'context', 'question', 'answers'],
        num_rows: 1784
    })
})
```



Training hyperparameters

The following hyperparameters were used during training:

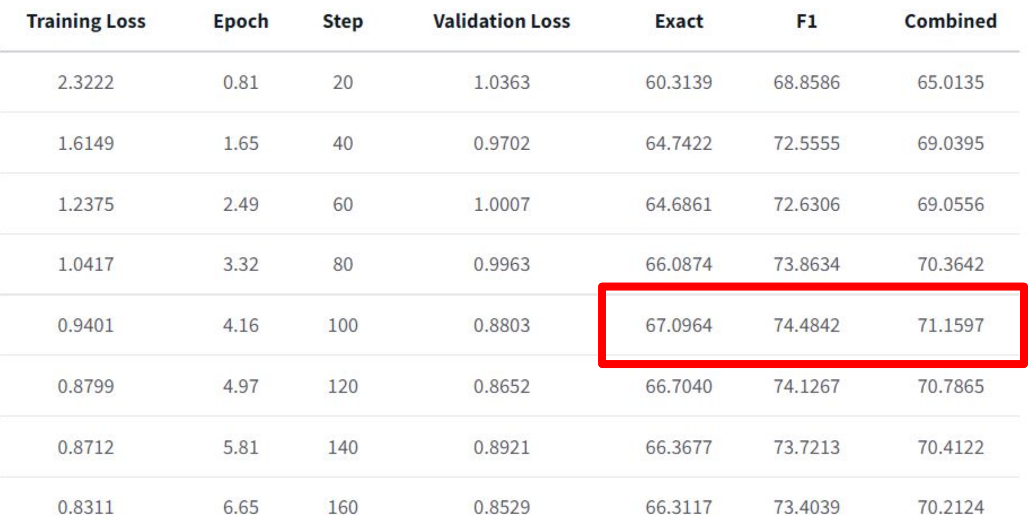
- learning_rate: 2e-05
- train_batch_size: 6
- eval_batch_size: 6
- seed: 42
- gradient_accumulation_steps: 86
- total_train_batch_size: 516
- optimizer: Adam with betas=(0.9,0.999) and epsilon=1e-08
- lr_scheduler_type: linear
- num_epochs: 7

Model hyperparameters

- hidden_dropout_prob: 0.36
- attention_probs_dropout_prob = 0.36

Fine-tuning

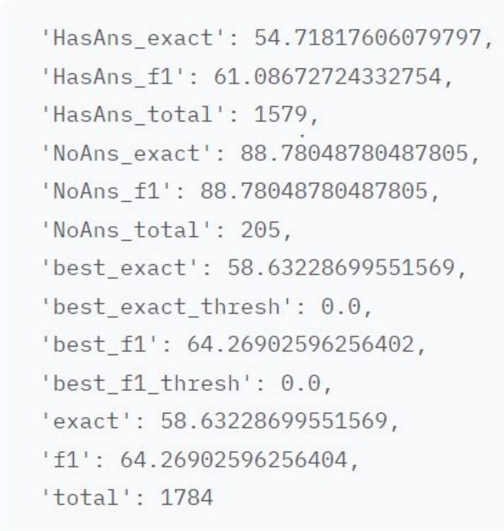
- https://huggingface.co/ptran74/DSPFirst-Finetuning-5
- 'combined' metric: 55% F1 + 45% EM
- Load the state with best 'combined' score at the end
- Note: the F1 and EM metrics are calculated from the generated dataset



Training Loss	Epoch	Step	Validation Loss	Exact	F1	Combined
2.3222	0.81	20	1.0363	60.3139	68.8586	65.0135
1.6149	1.65	40	0.9702	64.7422	72.5555	69.0395
1.2375	2.49	60	1.0007	64.6861	72.6306	69.0556
1.0417	3.32	80	0.9963	66.0874	73.8634	70.3642
0.9401	4.16	100	0.8803	67.0964	74.4842	71.1597
0.8799	4.97	120	0.8652	66.7040	74.1267	70.7865
0.8712	5.81	140	0.8921	66.3677	73.7213	70.4122
0.8311	6.65	160	0.8529	66.3117	73.4039	70.2124

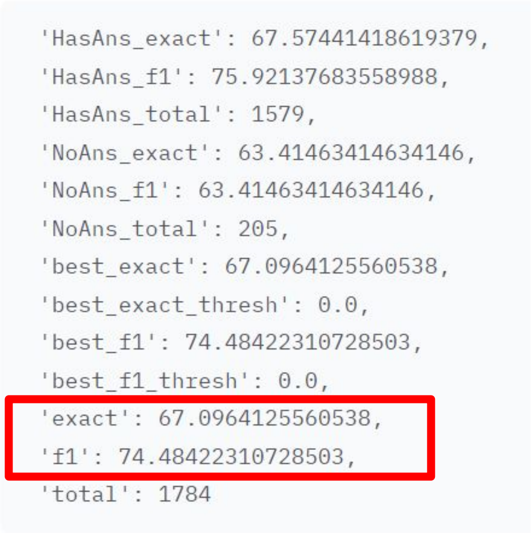
Fine-tuning

Before fine-tuning:



```
'HasAns_exact': 54.71817606079797,
'HasAns_f1': 61.08672724332754,
'HasAns_total': 1579,
'NoAns_exact': 88.78048780487805,
'NoAns_f1': 88.78048780487805,
'NoAns_total': 205,
'best_exact': 58.63228699551569,
'best_exact_thresh': 0.0,
'best_f1': 64.26902596256402,
'best_f1_thresh': 0.0,
'exact': 58.63228699551569,
'f1': 64.26902596256404,
'total': 1784
```

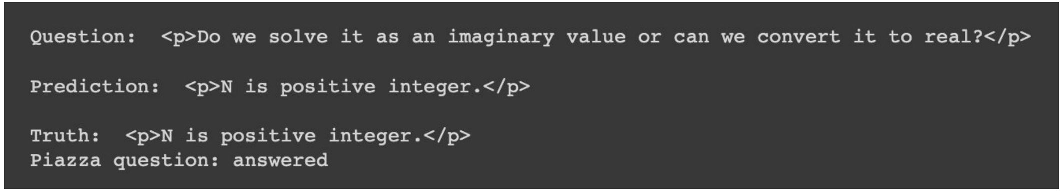
After fine-tuning:



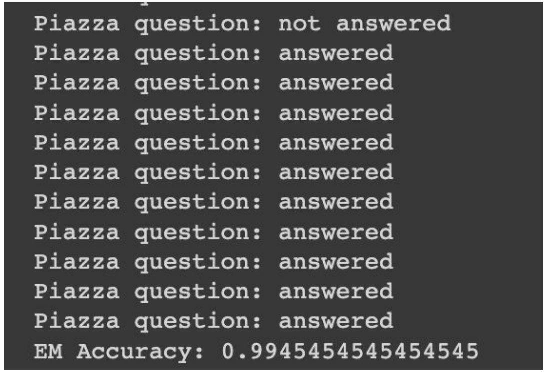
```
'HasAns_exact': 67.57441418619379,
'HasAns_f1': 75.92137683558988,
'HasAns_total': 1579,
'NoAns_exact': 63.41463414634146,
'NoAns_f1': 63.41463414634146,
'NoAns_total': 205,
'best_exact': 67.0964125560538,
'best_exact_thresh': 0.0,
'best_f1': 74.48422310728503,
'best_f1_thresh': 0.0,
'exact': 67.0964125560538,
'f1': 74.48422310728503,
'total': 1784
```

Creating Metric System

- Based on a dataset created we have a set of questions and the desired answer
- These same questions are inputted into our algorithm to see if responses are the same
- The EM Score returned is between 0 and 1 giving us our EM score



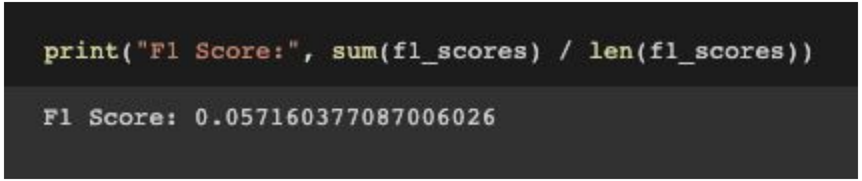
```
Question:  <p>Do we solve it as an imaginary value or can we convert it to real?</p>

Prediction:  <p>N is positive integer.</p>

Truth:  <p>N is positive integer.</p>
Piazza question: answered
```



```
Piazza question: not answered
Piazza question: answered
Piazza question: answered
Piazza question: answered
Piazza question: answered
Piazza question: answered
Piazza question: answered
Piazza question: answered
Piazza question: answered
Piazza question: answered
Piazza question: answered
EM Accuracy: 0.9945454545454545
```

F1 Metric

- The way the F1 score is calculated is by taking the similar words within a statement in regards to the length of the statement
- For example: “My name is Paul” and “Paul” would give a score of 0.4
- A longer statement that is 40 words but matches the first short statement could give us a score of 0.021

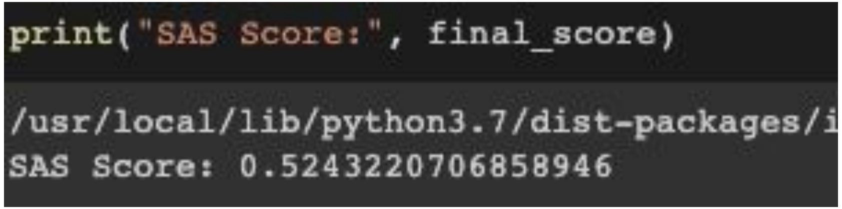


```
print("F1 Score:", sum(f1_scores) / len(f1_scores))
```

```
F1 Score: 0.05716037708700602б
```

Semantic Answer Similarity Metric

- Semantic Answer Similarity takes the important words and compares it between the two statements
- Is able to identify different words with the same meaning as the same word
- Typically ignores unimportant words such as “the”, “and”, etc.

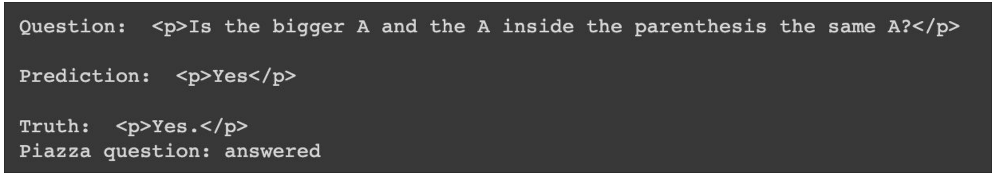


```
print("SAS Score:", final_score)
```

```
/usr/local/lib/python3.7/dist-packages/i
SAS Score: 0.5243220706858946
```

Drawbacks

- ExactMatch has a flaw in which different tokens will confuse the tests and mark a question as not answered
- F1 is hard to judge whether the score is good or not because it heavily depends on the length of the statements
- Semantic Similarity fixes these problems and is the best metric to use out of the three



```
Question:  <p>Is the bigger A and the A inside the parenthesis the same A?</p>

Prediction:  <p>Yes</p>

Truth:  <p>Yes.</p>
Piazza question: answered
```

Next Steps

- Fine-tune Question Generation Model
  - Needs to fine-tune on Natural Questions dataset to generate longer answers
  - Requires TPU v3 (16GB High Bandwidth Memory)
  - Google Colab only provides TPU v2 (8GB HBM)
- Perform data augmentation on the generated questions and answers dataset
  - Increase the dataset size
  - The more data we have, the better performance we can achieve
- Manually review the dataset
  - Some of the generated questions may be wrong
  - Some questions could have been marked as impossible to answer when there is an answer
- Review and adjust the DataCleaner class
  - Current visualization of the cleaned textbook data: https://github.gatech.edu/pages/VIP-ITS/textbook_SQuAD_explore/explore/textbookv1.0/textbook/

Documentation

Google Colab Notebooks:

- Fine-tune Question Generation Model
    - https://colab.research.google.com/drive/1L-5_mzqFXT-Qww51rDzZggm9lgMRA5rY?usp=sharing
- Question Generation from Haystack
    - https://colab.research.google.com/drive/1KGK6bo3fMsrzqiXY_L1NBiUsz1gT1EIJ?usp=sharing
- Fine-tune pre-trained QA model
    - https://colab.research.google.com/drive/1dJXNstk2NSenwzdtl9xA8AqjP4LL-Ks_?usp=sharing

Demo