VIP-ITS Chatbot

About ITS Chatbot

- Help TAs to handle high volumes of questions during the course and especially before deadlines and exams

- Provide a more personalized experience

Target Problem

1. Enhance the accuracy of the existing predictive model by incorporating additional features that can be used to calculate a new relevance metric
2. Increase the accuracy of chatbot responses using a new transformer-based model that generates answers for student questions based on the textbook, Piazza posts, and other available input
3. Add speech-to-text functionality to increase accessibility and explore further applications of audio input.

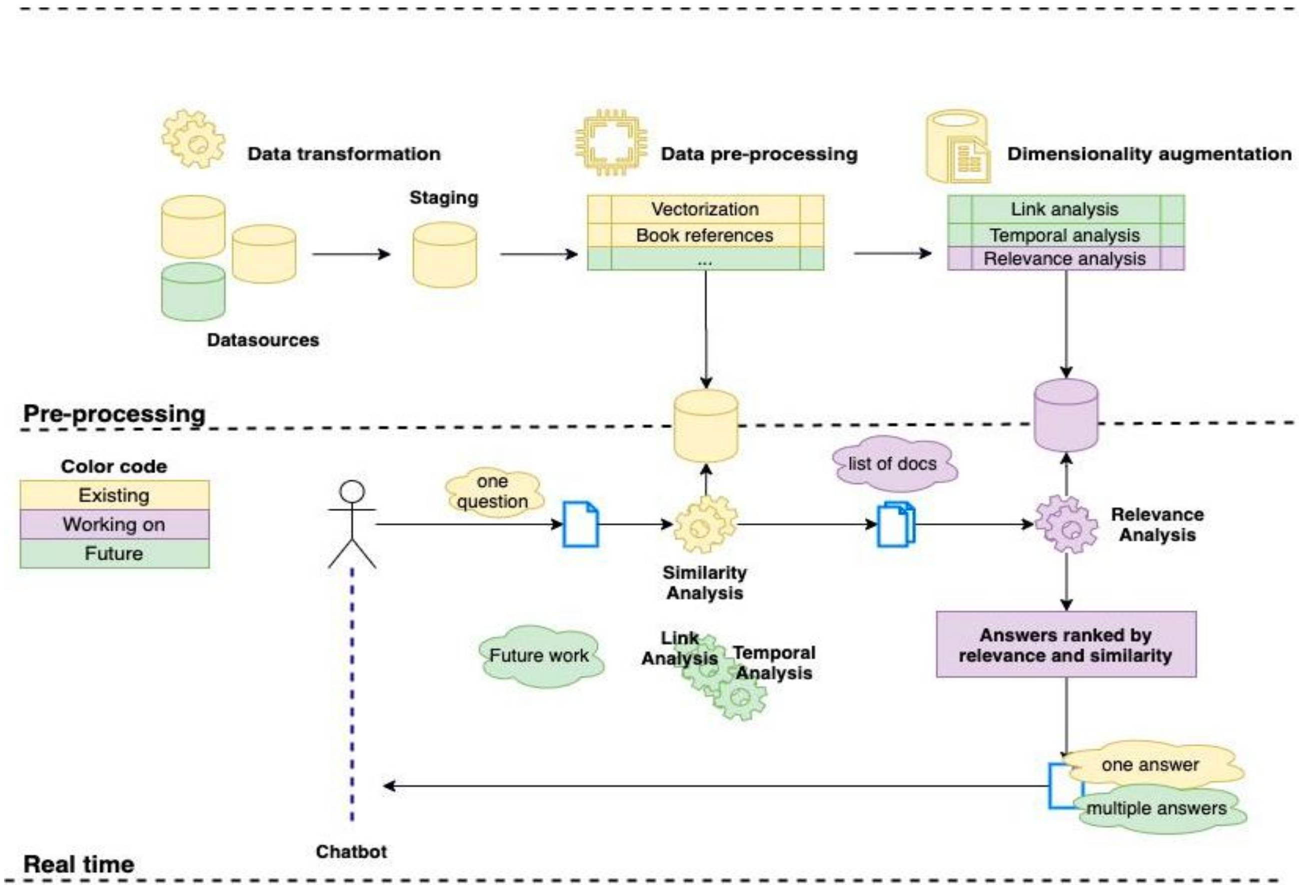
Predictive Model

Enhance the accuracy of the existing predictive model by incorporating additional features that can be used to calculate a new relevance metric.

Problem

- Chatbot prior: giving responses based exclusively on the similarity of the words
  - Similarity algorithm does not account for other factors/fields that would be useful to determine if a given response is relevant or not
- Goal: improve chatbot response accuracy by adding a relevance factor, so that multiple similar answers can be discriminated based on these other features
- How?
  - Analyze Piazza dataset for fields that were relevant
  - Use those fields in a relevance algorithm to determine more relevant responses for the chatbot

Layout of the Semester



Data transformation

Datasources

Staging

Data pre-processing

Vectorization

Book references

...

Dimensionality augmentation

Link analysis

Temporal analysis

Relevance analysis

Pre-processing

Color code

- Existing
- Working on
- Future

one question

Similarity Analysis

list of docs

Relevance Analysis

Future work

Link Analysis

Temporal Analysis

Answers ranked by relevance and similarity

one answer

multiple answers

Chatbot

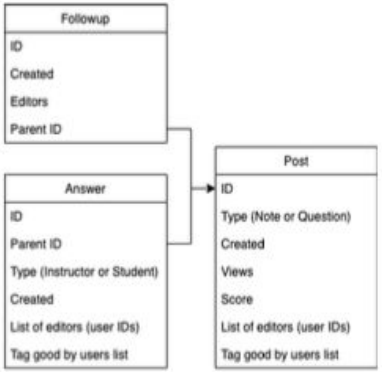
Real time

Steps We Took

- Phase I - Explore the Data Set
    - Gathered information on useful fields to use as metrics for our algorithm
- Phase 2 - Relevance Definition
    - Used fields from Phase I to formulate an algorithm based on field importance
- Phase 3 - Relevance Analysis
    - Added the Relevance metric to the chabot using Piazza data which can be used to rank responses based in Similarity and Relevance

Data transformation

Piazza data model

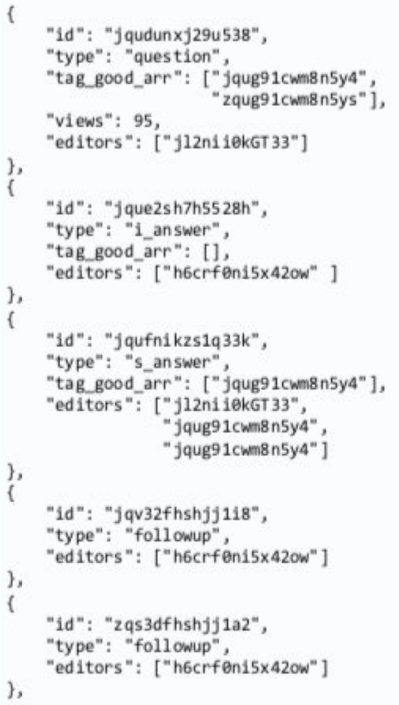


Followup
- ID
- Created
- Editors
- Parent ID

Answer
- ID
- Parent ID
- Type (Instructor or Student)
- Created
- List of editors (user IDs)
- Tag good by users list

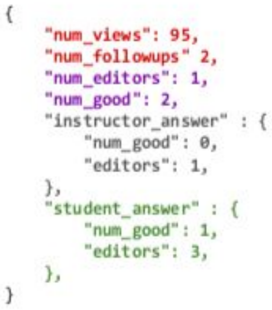
Post
- ID
- Type (Note or Question)
- Created
- Views
- Score
- List of editors (user IDs)
- Tag good by users list

Piazza Raw Data



```
{
    "id": "jqudunxj29u538",
    "type": "question",
    "tag_good_arr": ["jqug91cwm8n5y4",
                     "zqug91cwm8n5ys"],
    "views": 95,
    "editors": ["jl2nii0kGT33"]
},
{
    "id": "jque2sh7h5528h",
    "type": "i_answer",
    "tag_good_arr": [],
    "editors": ["h6crf0ni5x42ow" ]
},
{
    "id": "jqufnikzs1q33k",
    "type": "s_answer",
    "tag_good_arr": ["jqug91cwm8n5y4"],
    "editors": ["jl2nii0kGT33",
                "jqug91cwm8n5y4",
                "jqug91cwm8n5y4"]
},
{
    "id": "jqv32fhshjj1i8",
    "type": "followup",
    "editors": ["h6crf0ni5x42ow"]
},
{
    "id": "zqs3dfhshjj1a2",
    "type": "followup",
    "editors": ["h6crf0ni5x42ow"]
},
```

Algorithm output



```
{
    "num_views": 95,
    "num_followups" 2,
    "num_editors": 1,
    "num_good": 2,
    "instructor_answer" : {
        "num_good": 0,
        "editors": 1,
    },
    "student_answer" : {
        "num_good": 1,
        "editors": 3,
    },
}
```

Problems We Ran Into

- Wanted to use previous semester’s similarity algorithm within our relevance definition
- After testing found that this algorithm was not very accurate - it currently operates as a “bag of words”. It does not take semantics into account which makes it hard to find “similar” questions

What We Tried

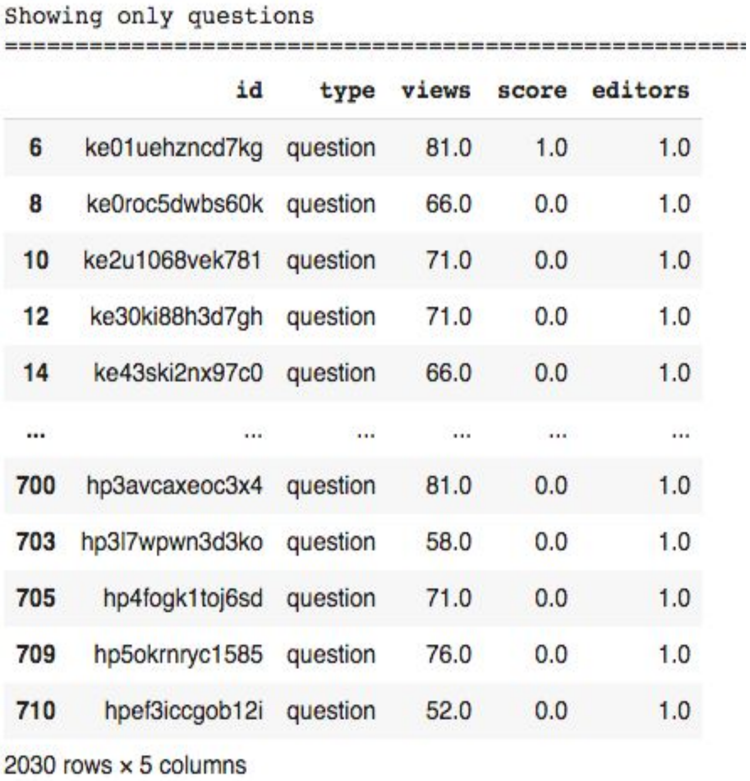
- Sorted data set using the “subject” field instead of the “content” field
- Yielded much smaller distances since answers are not based on semantics, responses are very limited

Documentation

Google Collab Notebooks:

- Data Visualization
  - https://colab.research.google.com/drive/13AEOT_aE6Am4Z95rnJtdQt70CTHF4Vsw?usp=sharing
- Creating a Testing Set
  - https://colab.research.google.com/drive/1ZoxbpYomV3VZsNOh7cMDxErjYKGnAqjZ?usp=sharing
- Relevance Metric
  - https://colab.research.google.com/drive/1nI-HTkZ9Ud5ISRGrNpvWEvo5y56aBoTt?usp=sharing

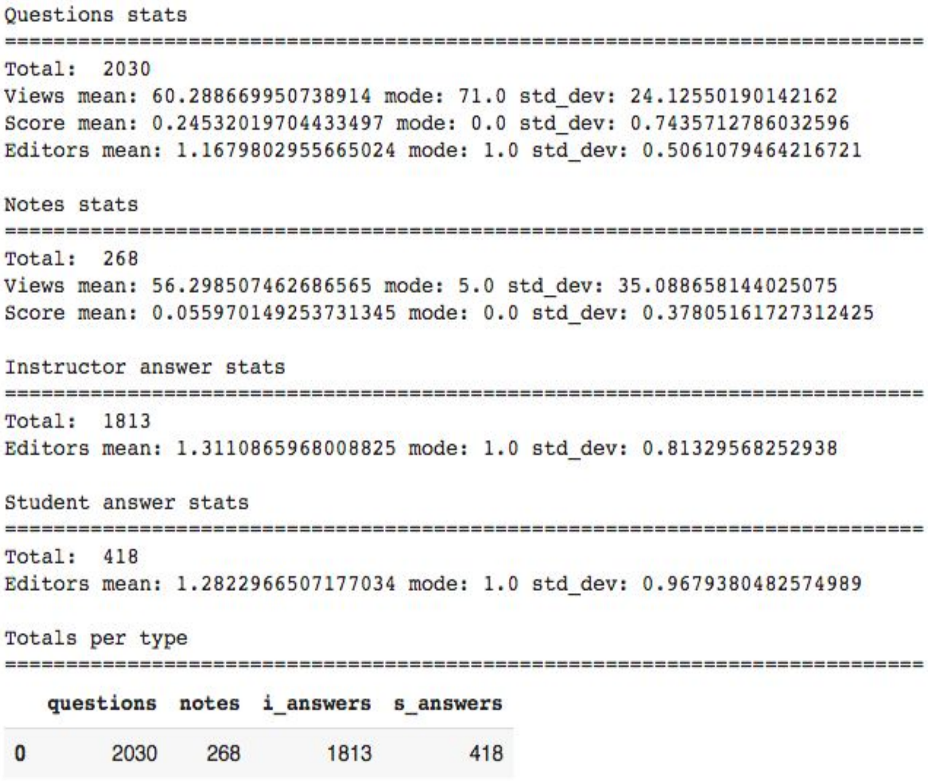
Data Visualization



```
Showing only questions
==================================================================
```

	id	type	views	score	editors
6	ke01uehzncd7kg	question	81.0	1.0	1.0
8	ke0roc5dwbs60k	question	66.0	0.0	1.0
10	ke2u1068vek781	question	71.0	0.0	1.0
12	ke30ki88h3d7gh	question	71.0	0.0	1.0
14	ke43ski2nx97c0	question	66.0	0.0	1.0
...	...	...	...	...	...
700	hp3avcaxeoc3x4	question	81.0	0.0	1.0
703	hp3l7wpwn3d3ko	question	58.0	0.0	1.0
705	hp4fogk1toj6sd	question	71.0	0.0	1.0
709	hp5okrnryc1585	question	76.0	0.0	1.0
710	hpef3iccgob12i	question	52.0	0.0	1.0

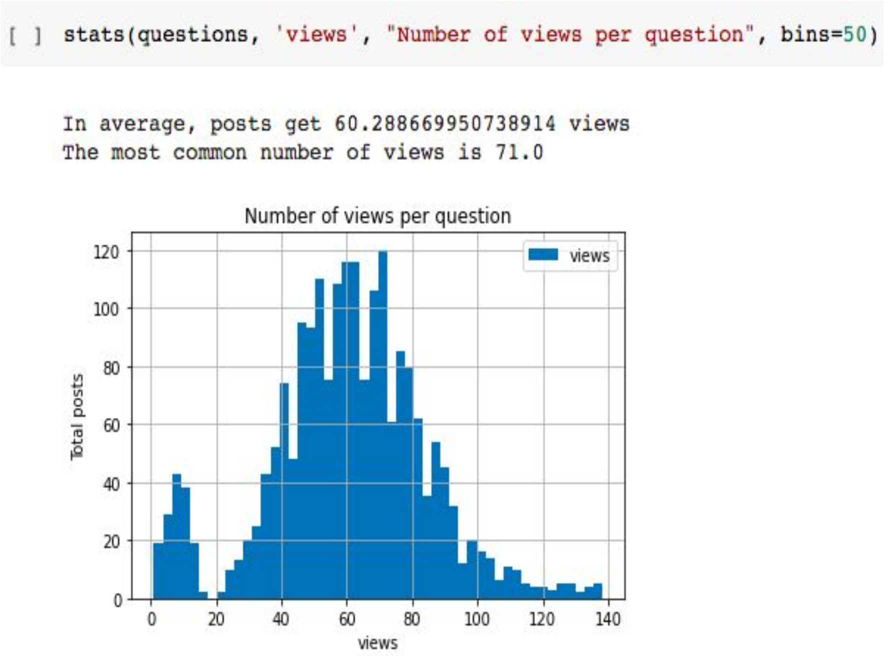
2030 rows × 5 columns



```
Questions stats
============================================================================
Total:  2030
Views mean: 60.288669950738914 mode: 71.0 std_dev: 24.12550190142162
Score mean: 0.24532019704433497 mode: 0.0 std_dev: 0.7435712786032596
Editors mean: 1.1679802955665024 mode: 1.0 std_dev: 0.5061079464216721

Notes stats
============================================================================
Total:  268
Views mean: 56.298507462686565 mode: 5.0 std_dev: 35.088658144025075
Score mean: 0.055970149253731345 mode: 0.0 std_dev: 0.37805161727312425

Instructor answer stats
============================================================================
Total:  1813
Editors mean: 1.3110865968008825 mode: 1.0 std_dev: 0.81329568252938

Student answer stats
============================================================================
Total:  418
Editors mean: 1.2822966507177034 mode: 1.0 std_dev: 0.9679380482574989

Totals per type
============================================================================
```

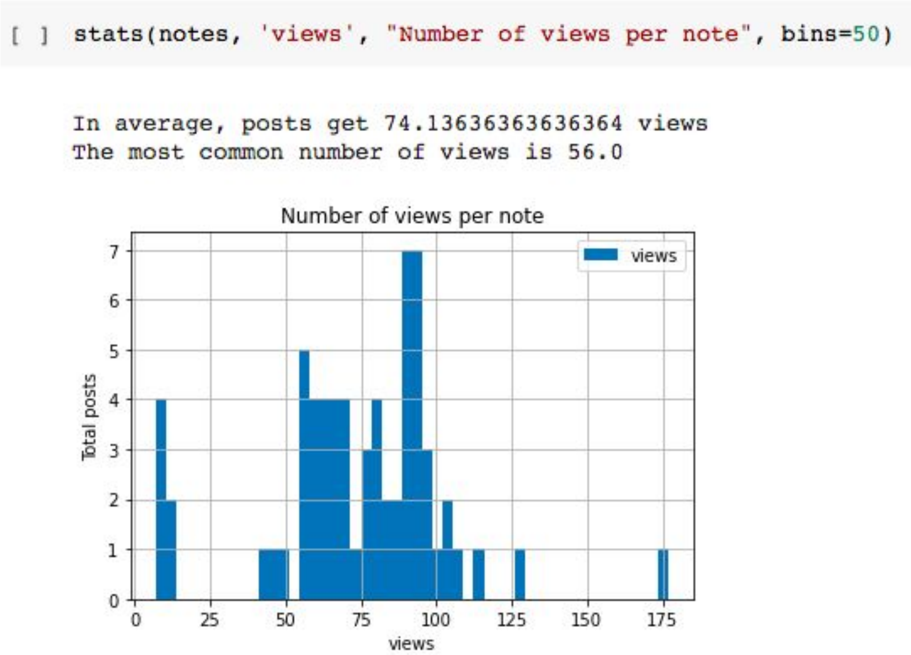
	questions	notes	i_answers	s_answers
0	2030	268	1813	418

Data Visualization



```
[ ] stats(questions, 'views', "Number of views per question", bins=50)
```

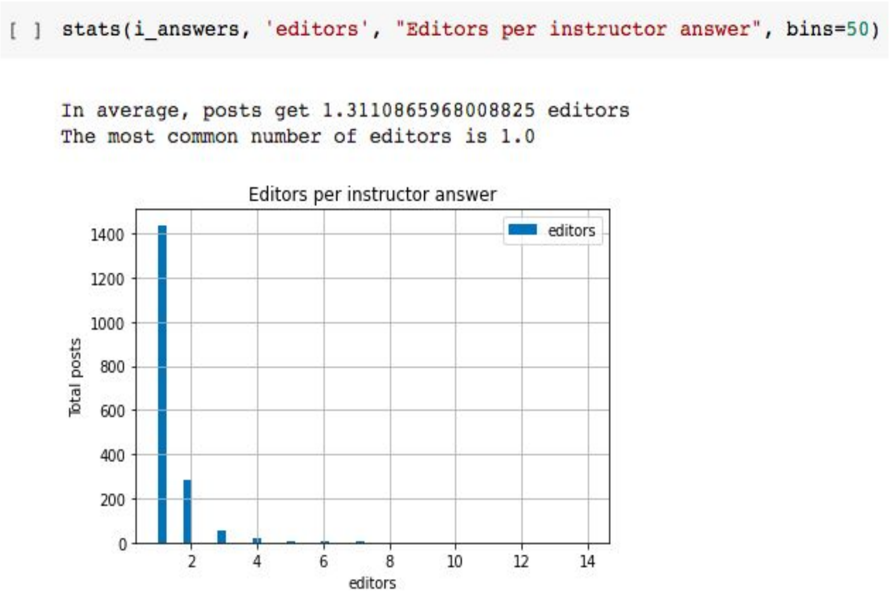
```
In average, posts get 60.288669950738914 views
The most common number of views is 71.0
```



```
[ ] stats(notes, 'views', "Number of views per note", bins=50)
```

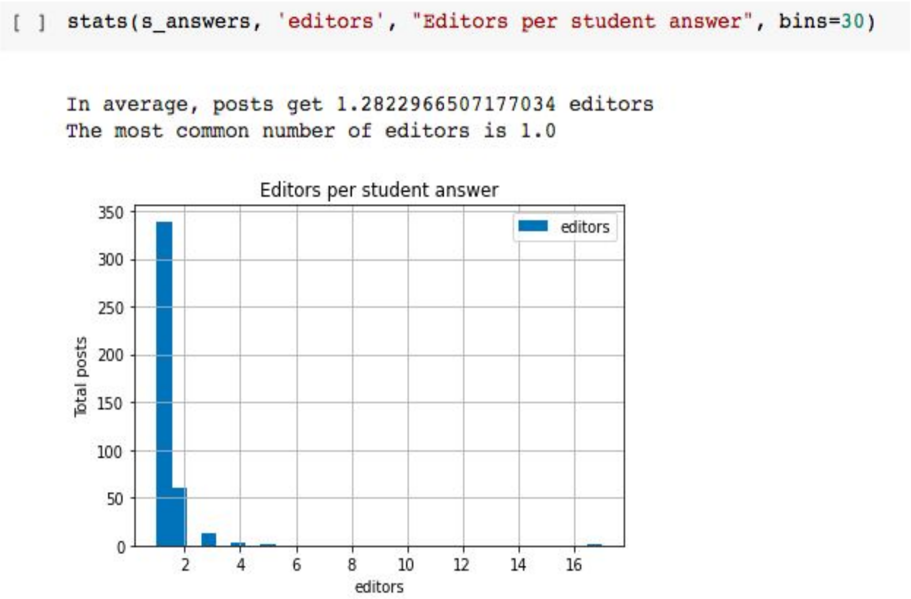
```
In average, posts get 74.13636363636364 views
The most common number of views is 56.0
```

Data Visualization



```
[ ] stats(i_answers, 'editors', "Editors per instructor answer", bins=50)
```

```
In average, posts get 1.3110865968008825 editors
The most common number of editors is 1.0
```

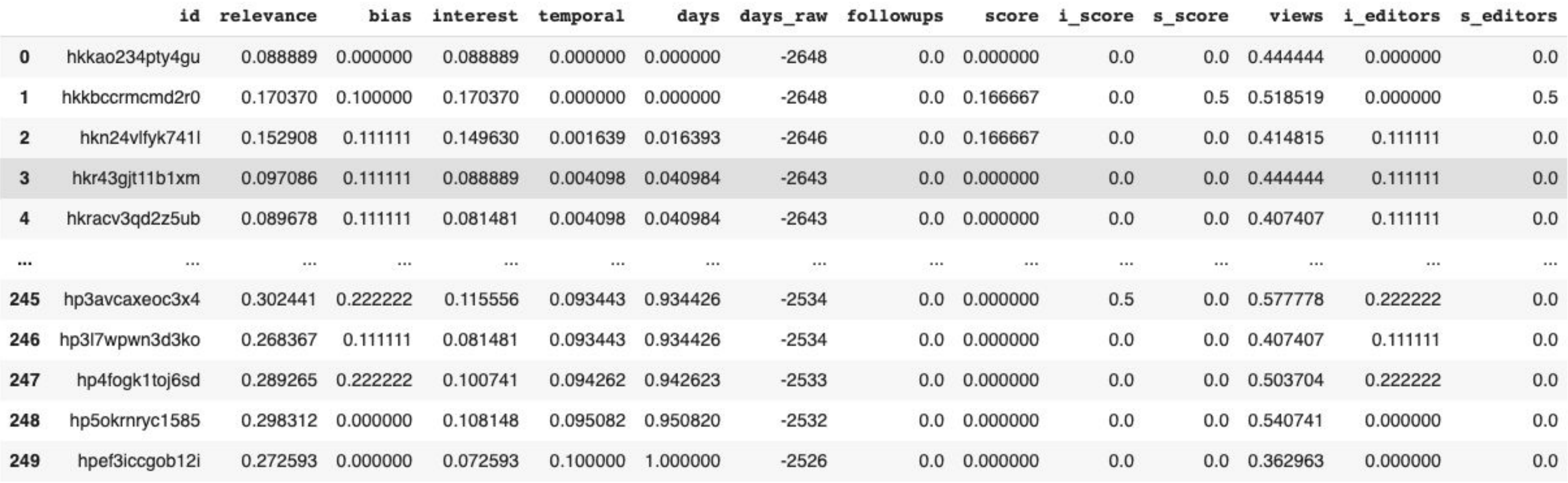


```
[ ] stats(s_answers, 'editors', "Editors per student answer", bins=30)
```

```
In average, posts get 1.2822966507177034 editors
The most common number of editors is 1.0
```

Results

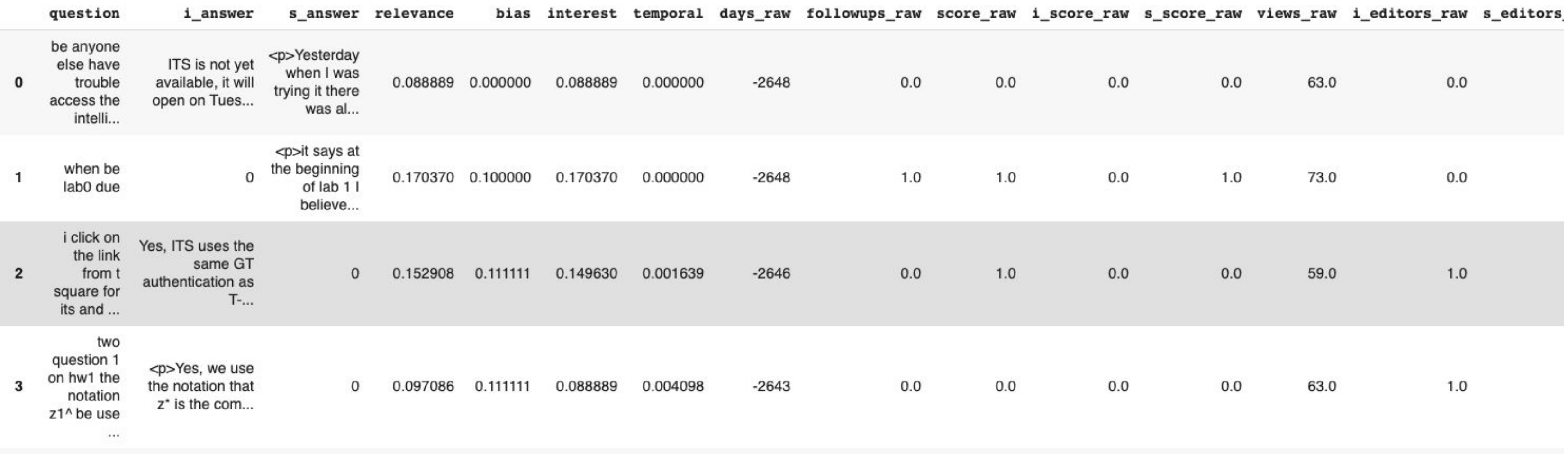
- New dimensions to the chatbot dataset which are used to calculate a Relevance metric



	id	relevance	bias	interest	temporal	days	days_raw	followups	score	i_score	s_score	views	i_editors	s_editors
0	hkkao234pty4gu	0.088889	0.000000	0.088889	0.000000	0.000000	-2648	0.0	0.000000	0.0	0.0	0.444444	0.000000	0.0
1	hkkbccrmcmd2r0	0.170370	0.100000	0.170370	0.000000	0.000000	-2648	0.0	0.166667	0.0	0.5	0.518519	0.000000	0.5
2	hkn24vlfyk741l	0.152908	0.111111	0.149630	0.001639	0.016393	-2646	0.0	0.166667	0.0	0.0	0.414815	0.111111	0.0
3	hkr43gjt11b1xm	0.097086	0.111111	0.088889	0.004098	0.040984	-2643	0.0	0.000000	0.0	0.0	0.444444	0.111111	0.0
4	hkracv3qd2z5ub	0.089678	0.111111	0.081481	0.004098	0.040984	-2643	0.0	0.000000	0.0	0.0	0.407407	0.111111	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
245	hp3avcaxeoc3x4	0.302441	0.222222	0.115556	0.093443	0.934426	-2534	0.0	0.000000	0.5	0.0	0.577778	0.222222	0.0
246	hp3l7wpwn3d3ko	0.268367	0.111111	0.081481	0.093443	0.934426	-2534	0.0	0.000000	0.0	0.0	0.407407	0.111111	0.0
247	hp4fogk1toj6sd	0.289265	0.222222	0.100741	0.094262	0.942623	-2533	0.0	0.000000	0.0	0.0	0.503704	0.222222	0.0
248	hp5okrnryc1585	0.298312	0.000000	0.108148	0.095082	0.950820	-2532	0.0	0.000000	0.0	0.0	0.540741	0.000000	0.0
249	hpef3iccgob12i	0.272593	0.000000	0.072593	0.100000	1.000000	-2526	0.0	0.000000	0.0	0.0	0.362963	0.000000	0.0

Results

- A single dataframe that contains all information required for Similarity and Relevance analysis along with the actual Piazza text. Makes it easier to work with the chatbot



	question	i_answer	s_answer	relevance	bias	interest	temporal	days_raw	followups_raw	score_raw	i_score_raw	s_score_raw	views_raw	i_editors_raw	s_editors
0	be anyone else have trouble access the intelli...	ITS is not yet available, it will open on Tues...	<p>Yesterday when I was trying it there was al...	0.088889	0.000000	0.088889	0.000000	-2648	0.0	0.0	0.0	0.0	63.0	0.0	
1	when be lab0 due	0	<p>it says at the beginning of lab 1 I believe...	0.170370	0.100000	0.170370	0.000000	-2648	1.0	1.0	0.0	1.0	73.0	0.0	
2	i click on the link from t square for its and ...	Yes, ITS uses the same GT authentication as T-...	0	0.152908	0.111111	0.149630	0.001639	-2646	0.0	1.0	0.0	0.0	59.0	1.0	
3	two question 1 on hw1 the notation z1^ be use ...	<p>Yes, we use the notation that z* is the com...	0	0.097086	0.111111	0.088889	0.004098	-2643	0.0	0.0	0.0	0.0	63.0	1.0	

Results

- Re-implemented Similarity Engine using **Vectorized Matrix Operations (with Pandas)** that significantly improves performance

Previous implementation:

Up and running in 2+ minutes

Memory usage tops 4.5GB while loading the data

New implementation:

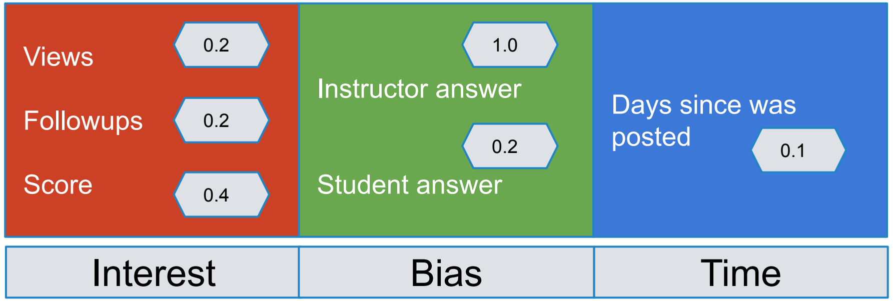
Up and running in 20+ seconds

Memory usage tops 2.5GB while loading the data

Relevance function

The following dimensions have been added and categorized

Each is assigned a Weight



The metrics are normalized using the min/max method

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

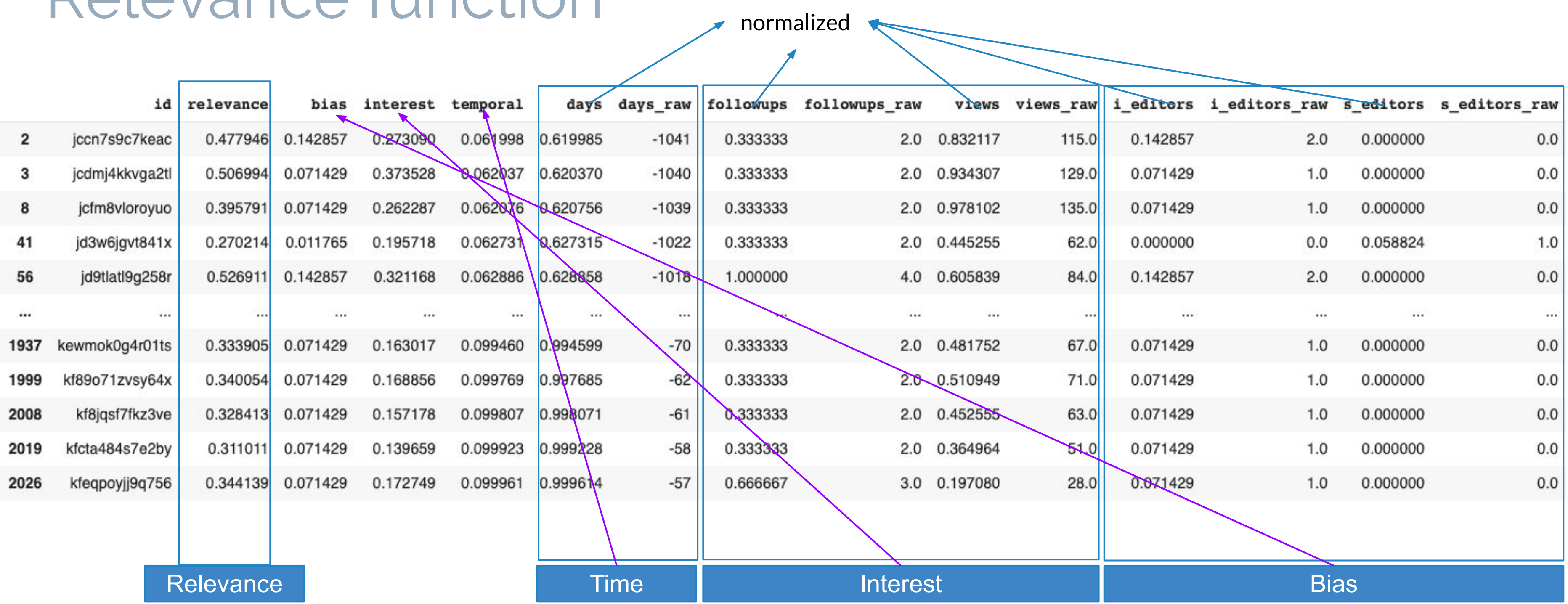
E.g. if number of views max value is 35 (x_{max}), min value is 1 (X_{min}) and a given post had 12 views (x), then this particular post's number of views is normalized to

(12 - 1) / (35 - 1) = 0.32

We apply the weight to this value

```
interest= views     * weights[views] +
          followups * weights[followups] +
          score     * weights[score] +

bias     = i_answers * weights[instructors] +
           s_answers * weights[students] +

time     = days * weights[time]

relevance = interest + bias + time
```

Relevance function



	id	relevance	bias	interest	temporal	days	days_raw	followups	followups_raw	views	views_raw	i_editors	i_editors_raw	s_editors	s_editors_raw
2	jccn7s9c7keac	0.477946	0.142857	0.273090	0.061998	0.619985	-1041	0.333333	2.0	0.832117	115.0	0.142857	2.0	0.000000	0.0
3	jcdmj4kkvga2tl	0.506994	0.071429	0.373528	0.062037	0.620370	-1040	0.333333	2.0	0.934307	129.0	0.071429	1.0	0.000000	0.0
8	jcfm8vloroyuo	0.395791	0.071429	0.262287	0.062076	0.620756	-1039	0.333333	2.0	0.978102	135.0	0.071429	1.0	0.000000	0.0
41	jd3w6jgvt841x	0.270214	0.011765	0.195718	0.062731	0.627315	-1022	0.333333	2.0	0.445255	62.0	0.000000	0.0	0.058824	1.0
56	jd9tlatl9g258r	0.526911	0.142857	0.321168	0.062886	0.628858	-1018	1.000000	4.0	0.605839	84.0	0.142857	2.0	0.000000	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1937	kewmok0g4r01ts	0.333905	0.071429	0.163017	0.099460	0.994599	-70	0.333333	2.0	0.481752	67.0	0.071429	1.0	0.000000	0.0
1999	kf89o71zvsy64x	0.340054	0.071429	0.168856	0.099769	0.997685	-62	0.333333	2.0	0.510949	71.0	0.071429	1.0	0.000000	0.0
2008	kf8jqsf7fkz3ve	0.328413	0.071429	0.157178	0.099807	0.998071	-61	0.333333	2.0	0.452555	63.0	0.071429	1.0	0.000000	0.0
2019	kfcta484s7e2by	0.311011	0.071429	0.139659	0.099923	0.999228	-58	0.333333	2.0	0.364964	51.0	0.071429	1.0	0.000000	0.0
2026	kfeqpoyjj9q756	0.344139	0.071429	0.172749	0.099961	0.999614	-57	0.666667	3.0	0.197080	28.0	0.071429	1.0	0.000000	0.0

normalized

Relevance

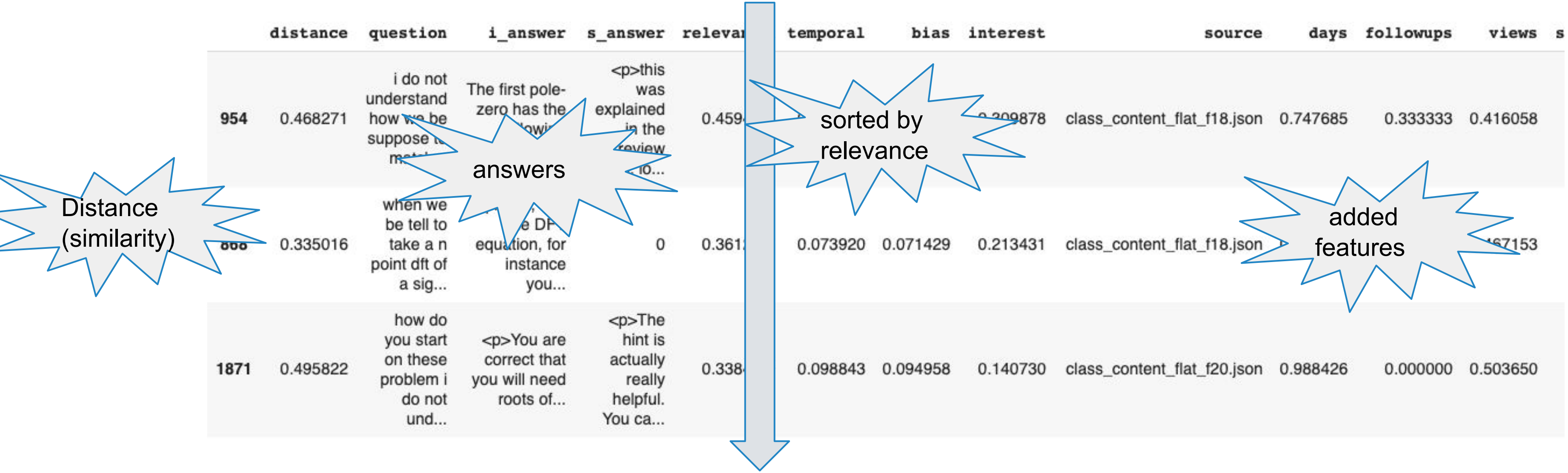
Time

Interest

Bias

Relevance library (Python)

```
from SearchByRelevanceAndSimilarity import SearchByRelevanceAndSimilarity
search_questions = SearchByRelevanceAndSimilarity(data_location "./piazza_data")
questions = search_questions.get_similar_questions(question)
columns = ['distance', 'question', 'relevance', 'temporal', 'bias', 'interest', ...# Filter by columns
questions[columns] # Use the columns
(e.g. for question = "how do we write the expression of the frequency")
```



	distance	question	i_answer	s_answer	relevar	temporal	bias	interest	source	days	followups	views	s
954	0.468271	i do not understand how we be suppose t[illegible] m[illegible]	The first pole-zero has the [illegible]owi[illegible]	<p>this was explained in the review [illegible]o...	0.459	[illegible]	[illegible]	0.309878	class_content_flat_f18.json	0.747685	0.333333	0.416058	
868	0.335016	when we be tell to take a n point dft of a sig...	[illegible] e DF equation, for instance you...	0	0.361	0.073920	0.071429	0.213431	class_content_flat_f18.json	[illegible]	[illegible]	[illegible]67153	
1871	0.495822	how do you start on these problem i do not und...	<p>You are correct that you will need roots of...	<p>The hint is actually really helpful. You ca...	0.338	0.098843	0.094958	0.140730	class_content_flat_f20.json	0.988426	0.000000	0.503650	

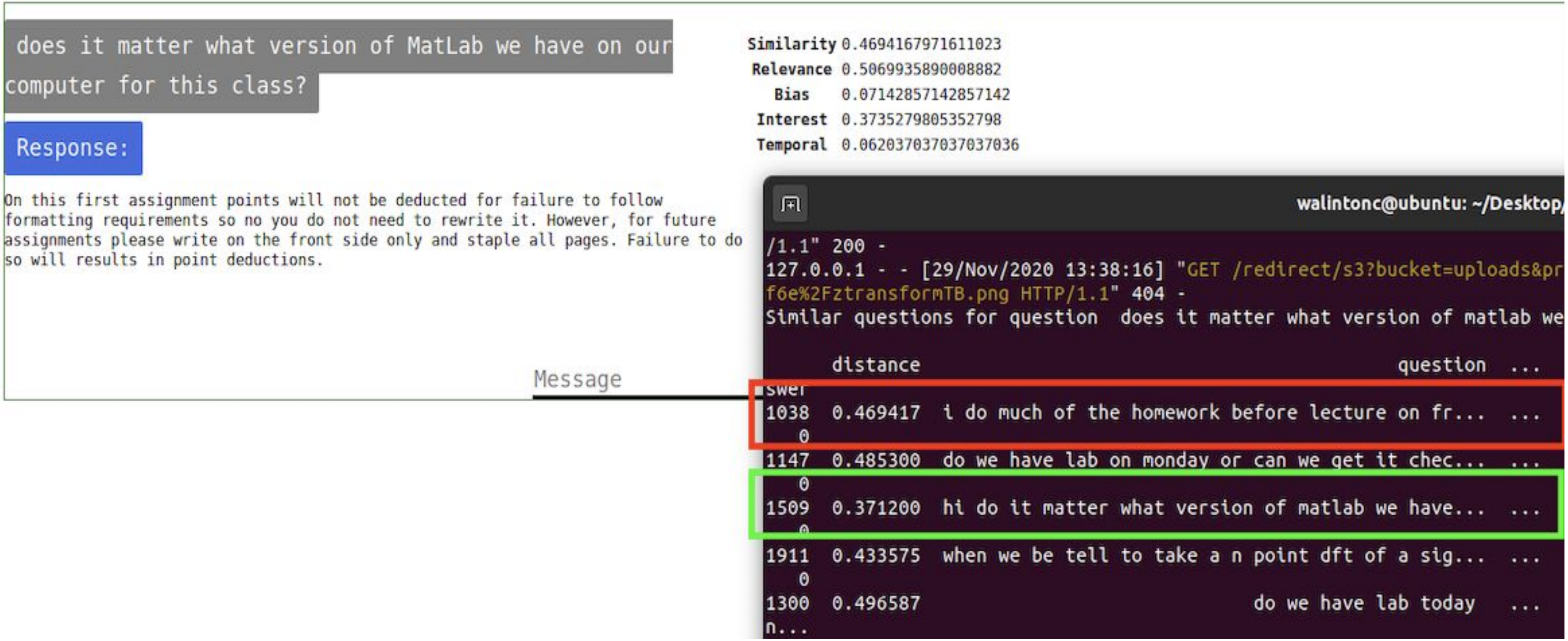
Distance (similarity)

answers

sorted by relevance

added features

Chatbot example



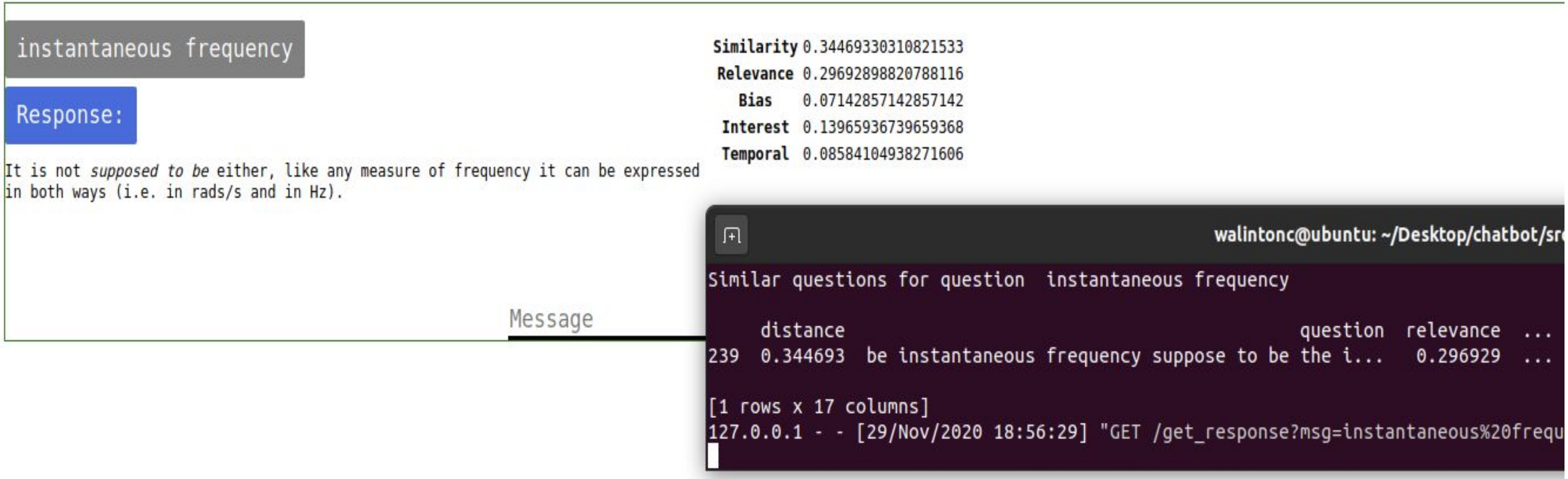
does it matter what version of MatLab we have on our computer for this class?

On this first assignment points will not be deducted for failure to follow formatting requirements so no you do not need to rewrite it. However, for future assignments please write on the front side only and staple all pages. Failure to do so will results in point deductions.

```
Similarity 0.4694167971611023
 Relevance 0.5069935890008882
      Bias 0.07142857142857142
  Interest 0.3735279805352798
  Temporal 0.062037037037037036
```

```
/1.1" 200 -
127.0.0.1 - - [29/Nov/2020 13:38:16] "GET /redirect/s3?bucket=uploads&pr
f6e%2FztransformTB.png HTTP/1.1" 404 -
Similar questions for question  does it matter what version of matlab we

     distance                                             question  ...
swer
1038  0.469417  i do much of the homework before lecture on fr...  ...
   0
1147  0.485300  do we have lab on monday or can we get it chec...  ...
   0
1509  0.371200  hi do it matter what version of matlab we have...  ...
   0
1911  0.433575  when we be tell to take a n point dft of a sig...  ...
   0
1300  0.496587                                do we have lab today  ...
n...
```

Chatbot example

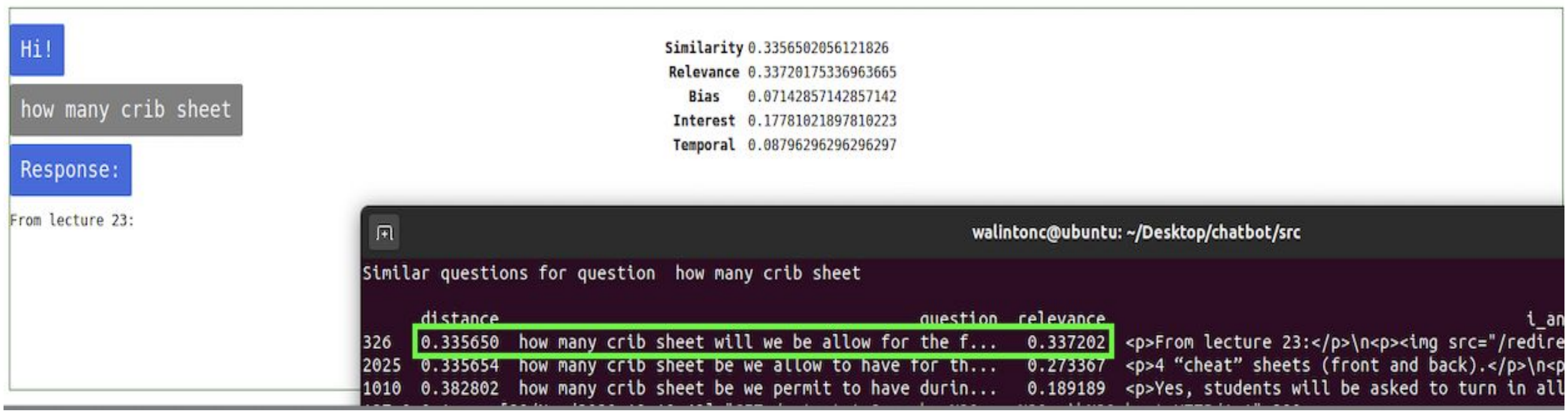


It is not *supposed to be* either, like any measure of frequency it can be expressed in both ways (i.e. in rads/s and in Hz).

Similarity	0.34469330310821533
Relevance	0.29692898820788116
Bias	0.07142857142857142
Interest	0.13965936739659368
Temporal	0.08584104938271606

```
Similar questions for question  instantaneous frequency

     distance                                          question  relevance  ...
239  0.344693  be instantaneous frequency suppose to be the i...   0.296929  ...

[1 rows x 17 columns]
127.0.0.1 - - [29/Nov/2020 18:56:29] "GET /get_response?msg=instantaneous%20frequ
```

Chatbot example



From lecture 23:

```
Similarity 0.3356502056121826
Relevance  0.33720175336963665
Bias       0.07142857142857142
Interest   0.17781021897810223
Temporal   0.08796296296296297
```

walintonc@ubuntu: ~/Desktop/chatbot/src

```
Similar questions for question  how many crib sheet

      distance                                      question  relevance                                                            i_an
326   0.335650  how many crib sheet will we be allow for the f...   0.337202  <p>From lecture 23:</p>\n<p><img src="/redire
2025  0.335654  how many crib sheet be we allow to have for th...   0.273367  <p>4 “cheat” sheets (front and back).</p>\n<p
1010  0.382802  how many crib sheet be we permit to have durin...   0.189189  <p>Yes, students will be asked to turn in all
```

Next?

- **Improve Similarity Engine**
- Normalize data on a per-semester and per-user count
- Need to complete the development of a training-set to find weights for each dimensions using ML mechanisms
- Link answers from Piazza to textbook which can offer better/more detailed responses
- Link analysis (new dimension)
- Incorporate feedback we’ve received

Generative Model

Increase the accuracy of chatbot responses using a new transformer-based model that generates answers using Piazza data and DSP First textbook paragraphs as contexts.

What We Had Before This Semester

- A Word2Vec model that:
  - Converts the training data (Piazza and textbook) into vectors
  - Given user query, convert the query to query vector ***Q***
  - Find the training data vector ***V*** that has the highest cosine similarity with ***Q***
  - Use ***V*** to reference back to the original entry in the training data
  - If the entry is a Piazza question, return the corresponding answer based on the Piazza thread. Otherwise, the entry is a paragraph(s) from the textbook; return it directly.

Why Transformer Model?

- Word2Vec only encodes occurences of words, but not semantics
- A transformer model captures the sequential relationships of words in a text and learns to focus on relevant words
- We hope a transformer model can understand more symbols and formulae and respond to questions more concisely and precisely

Word2Vec-Transformer Model

- It takes time for a Transformer model to retrieve top n relevant contexts
- Instead, use Word2Vec to retrieve most relevant candidate contexts, from which the Transformer extract candidate answers

Results

- We realized two directions to improve the performance
  - Improve on context retrieval
    - Finetune data preprocessing
    - Use a more robust model than Word2Vec (e.g. another Transformer)
  - Improve on answer extraction given contexts
    - Filter Transformer-generated answers

Dual-Transformer Model

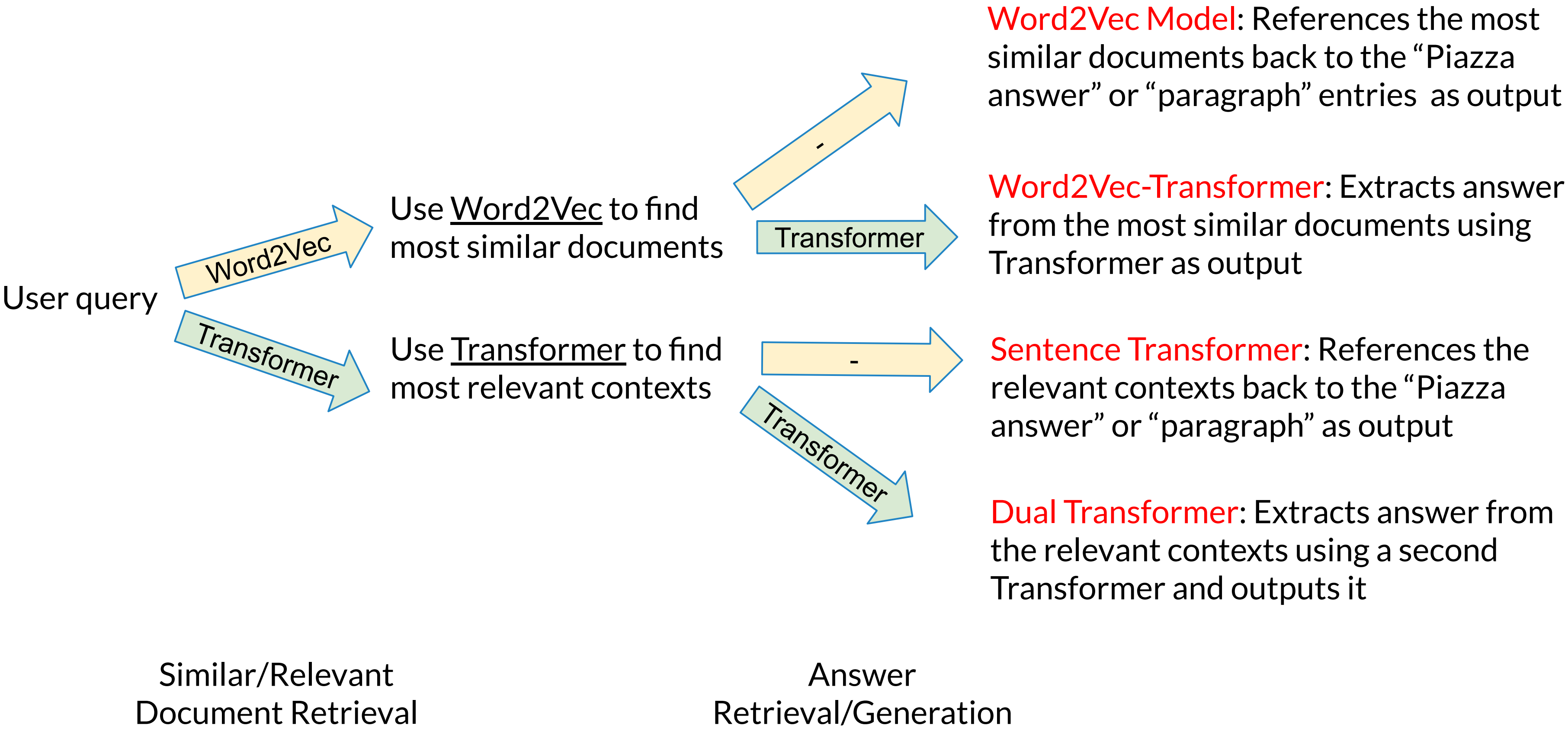
- We experimented a Dual-Transformer model which:
  - Retrieves most relevant candidate contexts with Transformer A.
  - Extracts answers from candidate contexts with another Transformer B.

- **IDEA:**
  What if we only use Transformer A and build a model like the old Word2Vec?

Sentence Transformer

- We experimented a Sentence Transformer model which:
  - Retrieves most relevant entry from the training data using a Transformer.
  - If the entry is a Piazza question, return the corresponding answer based on the Piazza thread. Otherwise, the entry is a paragraph(s) from the textbook; return it directly.

To sum up…



User query

Use Word2Vec to find most similar documents

Use Transformer to find most relevant contexts

Word2Vec Model: References the most similar documents back to the “Piazza answer” or “paragraph” entries as output

Word2Vec-Transformer: Extracts answer from the most similar documents using Transformer as output

Sentence Transformer: References the relevant contexts back to the “Piazza answer” or “paragraph” as output

Dual Transformer: Extracts answer from the relevant contexts using a second Transformer and outputs it

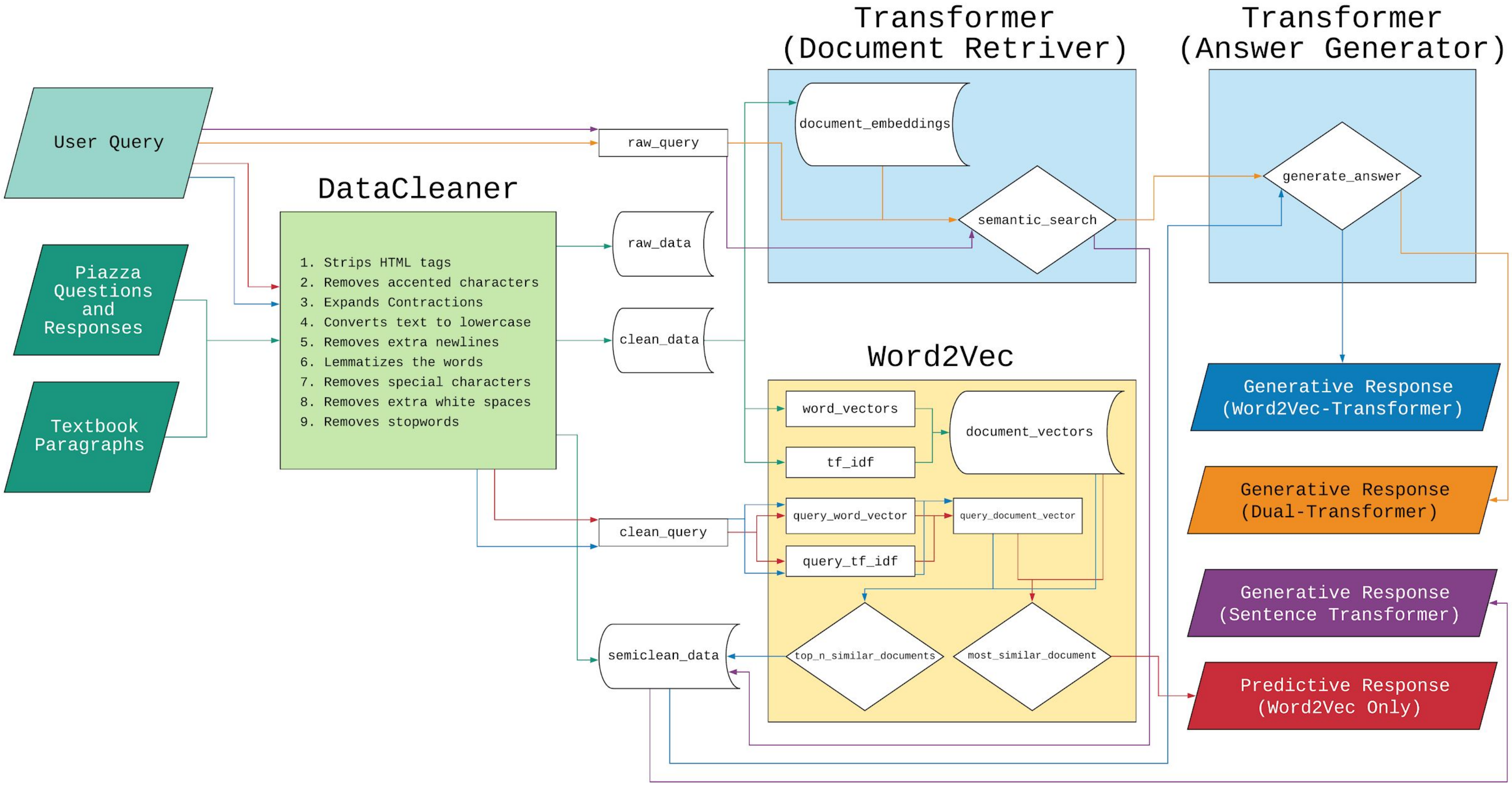
Similar/Relevant Document Retrieval

Answer Retrieval/Generation

What's the difference?

- The old model, namely Word2Vec, consists of two parts:
  - Find the most similar entry from training data
  - If the entry is a Piazza question, return the corresponding answer based on the Piazza thread. Otherwise, the entry is a paragraph(s) from the textbook; return it directly.
- A Transformer-based Model, also works in two parts:
  - Retrieves most relevant entries (we call them "candidate contexts")
  - Extracts answers from candidate contexts and return the best one

What We Have Now



User Query

Piazza Questions and Responses

Textbook Paragraphs

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

raw_query

raw_data

clean_data

clean_query

semiclean_data

Transformer (Document Retriver)

document_embeddings

semantic_search

Transformer (Answer Generator)

generate_answer

Word2Vec

word_vectors

tf_idf

document_vectors

query_word_vector

query_tf_idf

query_document_vector

top_n_similar_documents

most_similar_document

Generative Response (Word2Vec-Transformer)

Generative Response (Dual-Transformer)

Generative Response (Sentence Transformer)

Predictive Response (Word2Vec Only)

Things We've Tried

- ▷ Developed a Word2Vec-Transformer Model
- ▷ Developed a Dual-Transformer Model
- ▷ Developed a Sentence Transformer Model
- ▷ Evaluated Transformer answer confidence

Relevant Readings

- Word2Vec Explained: http://jalammar.github.io/illustrated-word2vec/
- Transformer Explained: http://jalammar.github.io/illustrated-transformer/
- Reading on BERT (a Transformer QA model): https://towardsdatascience.com/bert-nlp-how-to-build-a-question-answering-bot-98b1d1594d7b
- Another reading on BERT: https://medium.com/saarthi-ai/build-a-smart-question-answering-system-with-fine-tuned-bert-b586e4cfa5f5

Recall our plan to improve Word2Vec-Transformer

- We realized two directions to improve the performance
  - Improve on context retrieval
    - Finetune data preprocessing
    - Use a more robust model than Word2Vec (e.g. another Transformer)

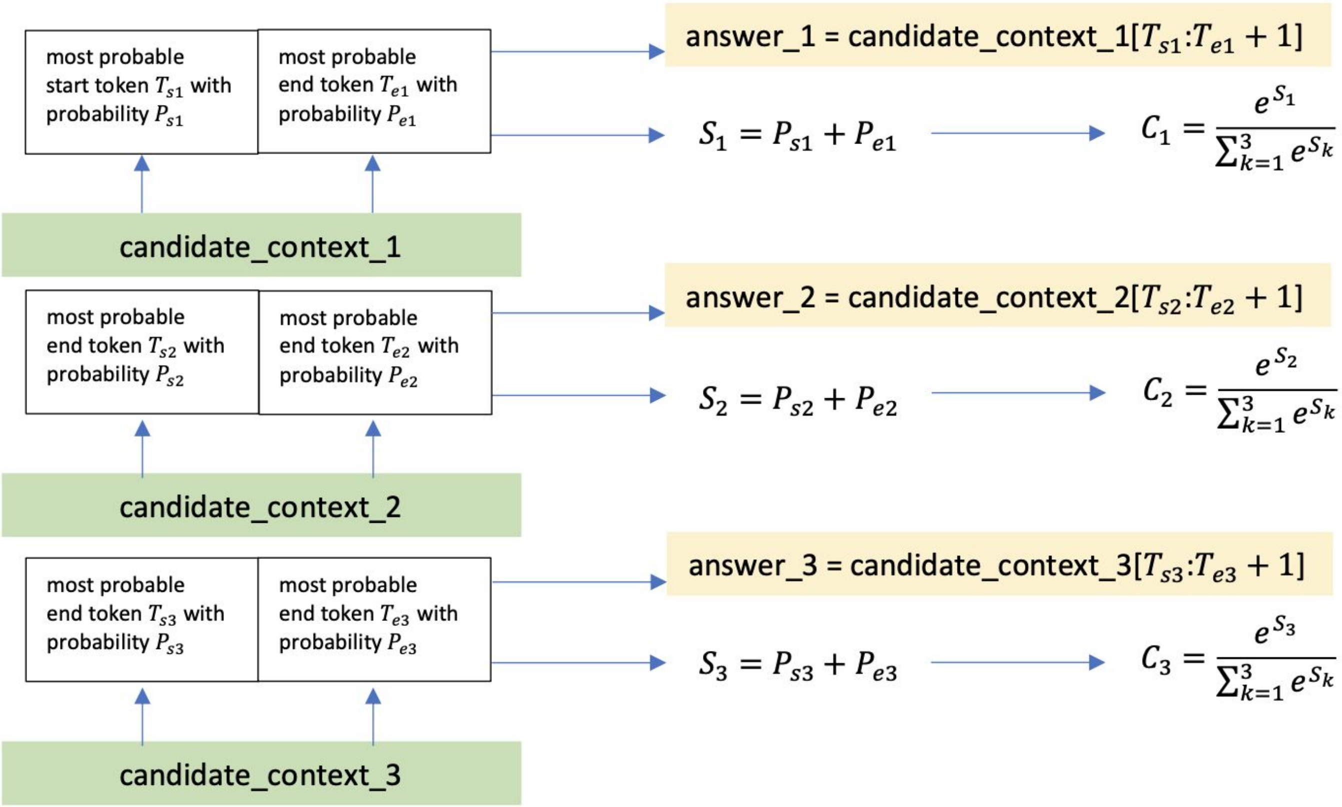
      → Sentence Transformer and Dual Transformer
  - Improve on answer extraction given contexts
    - Filter Transformer-generated answers

Filter Transformer-generated answer with confidence scores

- Generate confidence scores for the *n* answers generated from *n* candidate contexts

- Hoped to filter the answers based on confidence scores

Implementation



most probable start token T_{s1} with probability P_{s1}

most probable end token T_{e1} with probability P_{e1}

candidate_context_1

answer_1 = candidate_context_1[T_{s1}:$T_{e1} + 1$]

$$S_1 = P_{s1} + P_{e1} \longrightarrow C_1 = \frac{e^{S_1}}{\sum_{k=1}^{3} e^{S_k}}$$

most probable end token T_{s2} with probability P_{s2}

most probable end token T_{e2} with probability P_{e2}

candidate_context_2

answer_2 = candidate_context_2[T_{s2}:$T_{e2} + 1$]

$$S_2 = P_{s2} + P_{e2} \longrightarrow C_2 = \frac{e^{S_2}}{\sum_{k=1}^{3} e^{S_k}}$$

most probable end token T_{s3} with probability P_{s3}

most probable end token T_{e3} with probability P_{e3}

candidate_context_3

answer_3 = candidate_context_3[T_{s3}:$T_{e3} + 1$]

$$S_3 = P_{s3} + P_{e3} \longrightarrow C_3 = \frac{e^{S_3}}{\sum_{k=1}^{3} e^{S_k}}$$

Examples: “What are FIR filters?”

Answer: removes certain frequencies

Confidence: [1.]

Answer: to remove rapid fluctuations in signals

Confidence: [4.65888615e-15]

Answer: finite impulse response

Confidence: [5.38018616e-32]

Answer: each output sample is the sum of a finite number of weighted samples of the input sequence

Confidence: [2.74878501e-43]

Examples: “Are calculators allowed for the exams?”

Answer: it says on the package that it is acceptable for sat / act / ap tests

Confidence: [1.]

Answer: no graphing is allowed

Confidence: [1.56288219e-18]

Answer: calculators are allowed

Confidence: [2.05388455e-85]

Answer: calculators are allowed

Confidence: [2.05388455e-85]

Answer: simple computations (with or without a calculator) do not require any justification

Confidence: [3.76182078e-87]

Results

- There are 1-2 (usually 1) answer with extremely high confidence compared to the rest.
- The model is not confused between multiple very likely answers
- Candidate contexts determines the relevance of the answer, and most candidate contexts are not as “good” as the “best” one

Examples: “What are FIR filters?”

RED: candidate contexts

Word2Vec

What function should we use when trying to apply an IIR Filter to an input signal in Matlab, since firfilt is just for FIR filters?

Word2Vec-Transformer

1. If there are any poles not at the origin or infinity, you have a IIR, so you can automatically rule out the last two because they are FIR filters. When the zero is at the origin: If there is a single pole along the positive x axis, your impulse response is b(a^n)u[n] where a < 1, so you get a decaying response. If there is a single pole along the negative x axis, the impulse response is b(a^n)u[n], where -1 < a < 0, so you get something that looks like m: the magnitude decays but the sign alternates. When

Sentence Transformer

… FIR filters can be used to remove rapid fluctuations in signals… In Chapter~※, we will further develop our understanding of FIR systems.

(DSP First paragraph)

Dual-Transformer

1. … the second approach is an FIR filter that also removes certain frequencies…
2. … FIR filters can be used to remove rapid fluctuations in signals…
3. … FIR filters have a finite impulse response…
4. for which each output sample is the sum of a finite number of weighted samples of the input sequence. We will define the basic input output structure of the FIR filter…
5. the general class of feedback systems… since output samples are computed in terms of previously computed…

Examples: “What are FIR filters?”

BLUE: generated answers

Word2Vec

We learn about it in Lab 11. From the pdf: 3.3 IIR Filter Implementation In MATLAB the function that does IIR filtering is called filter. It requires the numerator (num) and denominator (den) coefficients,yy = filter(num, den, xx)...

Word2Vec-Transformer

1. if there are any poles not at the origin or infinity , you have a iir , so you can automatically rule out the last two because they are fir filters

Sentence Transformer

... FIR filters can be used to remove rapid fluctuations in signals... In Chapter~※, we will further develop our understanding of FIR systems.

(DSP First paragraph)

Dual-Transformer

1. removes certain frequencies
2. to remove rapid fluctuations in signals
3. finite impulse response
4. each output sample is the sum of a finite number of weighted samples of the input sequence
5. feedback systems

Examples: “What are finite-impulse-response filters?”

Word2Vec

<p>If you have an IIR Filter, say y[n] = y[n-1]&#43;x[n-5], how would we find the impulse response? For (b0(a1)^n)*u[n] to work, we have to have b0x[argument of y[n&#43;1]], right? so for the above would the impulse response be zero?</p>

Word2Vec-Transformer

1. The difference between an IIR and FIR lowpass filter can be best understood from the pole-zero plot, since the frequency response plots may look identical. FIR is represented by a finite number of coefficients, hence the peaks would look "cosine-like", whereas…
2. Note the difference between FIR and IIR filters, and think about how you could construct and simplify an overall system function $$H(z)$$: $$x[n]...

Sentence Transformer

… the impulse response \(h[n]\) of the FIR filter is simply the sequence of difference equation coefficients. Since \(h[n] = 0\) for \(n<0\) and for \(n>M\), the length of the impulse response sequence \(h[n]\) is finite. This is why the system is called a finite impulse response, (FIR) system…

(DSP First paragraph)

Dual-Transformer

1. … the impulse response \(h[n]\) of the FIR filter is simply the sequence of difference equation coefficients… This is why the system is called a finite impulse response, (FIR) system…
2. … For an FIR filter, the pole/zero plot will have all of its poles at the origin.
3. … FIR filters have a finite impulse response, such as something that can be written as a finite series of b(k)={… } values .

Examples: “What are finite-impulse-response filters?”

Word2Vec

Lecture 23 has an example worked out of an IIR filter's impulse response. Specifically, look at how the example on slide 31 uses a time delay property for the relevant terms.

Word2Vec-Transformer

1. fir is represented by a finite number of coefficients
2. fir

Sentence Transformer

… the impulse response \(h[n]\) of the FIR filter is simply the sequence of difference equation coefficients. Since \(h[n] = 0\) for \(n<0\) and for \(n>M\), the length of the impulse response sequence \(h[n]\) is finite. This is why the system is called a finite impulse response, (FIR) system…

(DSP First paragraph)

Dual-Transformer

1. the sequence of difference equation coefficients
2. fir filter , the pole / zero plot will have all of its poles at the origin
3. something that can be written as a finite series of b (k) = { . . . } values

Examples: “Explain continuous-to-discrete conversion.”

Word2Vec

… A-to-D converters differ from ideal C-to-D converters because of real-world problems such as amplitude quantization to 12 or 16 bits, jitter in the sampling times, and other factors that are difficult to analyze….

(DSP paragraph)

Word2Vec-Transformer

1. … Clearly this isn't the response of a system to an input; applying the DTFT to the input does something else. In fact, it calculates the spectrum of $$x[n]$$ over a continuum of frequencies $$\hat\omega$$.

Sentence Transformer

… How does the D-to-C converter work? In this section, we explain how the D-to-C converter does interpolation, and then describe a practical system that is nearly the same as the ideal D-to-C converter…

(DSP First paragraph)

Dual-Transformer

(None)

Examples: “Explain continuous-to-discrete conversion.”

Word2Vec

… A-to-D converters differ from ideal C-to-D converters because of real-world problems such as amplitude quantization to 12 or 16 bits, jitter in the sampling times, and other factors that are difficult to analyze….

(DSP paragraph)

Word2Vec-Transformer

1. it calculates the spectrum of $ $ x [n] $ $ over a continuum of frequencies

Sentence Transformer

… How does the D-to-C converter work? In this section, we explain how the D-to-C converter does interpolation, and then describe a practical system that is nearly the same as the ideal D-to-C converter…

(DSP First paragraph)

Dual-Transformer

(None)

Examples: “Explain C-to-D conversion.”

Word2Vec

<p>Do we explain our thought process for each individual problem? (for example 1.1a, 1.1b, 1.1c...) or can we explain it for the problem as a whole (just 1.1)?</p>

Word2Vec-Transformer

(None)

Sentence Transformer

An A-to-D does two things to a continuous-time signal $$x(t)$$: It samples, say $$x[n] = x(n / f_s)$$It rounds each sample to one of $$2^b$$ values, where $$b$$ is the number of bits of precision. The C-to-D does only the first step, without any rounding (quantization). You can think of the C-to-D as an A-to-D with infinite precision ($$b = \infty$$).

(DSP First paragraph)

Dual-Transformer

1. An A-to-D does two things to a continuous-time signal $$x(t)$$: It samples, say $$x[n] = x(n / f_s)$$It rounds each sample to one of $$2^b$$ values, where $$b$$ is the number of bits of precision. The C-to-D does only the first step, without any rounding (quantization). You can think of the C-to-D as an A-to-D with infinite precision ($$b = \infty$$).

Examples: “Explain C-to-D conversion.”

Word2Vec

Since 1.1 asks you to do the same task for each subsection you can just have one explanation for the whole problem.

Word2Vec-Transformer

(None)

Sentence Transformer

An A-to-D does two things to a continuous-time signal $$x(t)$$: It samples, say $$x[n] = x(n / f_s)$$It rounds each sample to one of $$2^b$$ values, where $$b$$ is the number of bits of precision. The C-to-D does only the first step, without any rounding (quantization). You can think of the C-to-D as an A-to-D with infinite precision ($$b = \infty$$).

Dual-Transformer

1. an a - to - d does two things to a continuous - time signal $ $ x (t) $ $: it samples , say $ $ x [n] = x (n / f _ s) $ $ it rounds each sample to one of $ $ 2 ^ b $ $ values , where $ $ b $ $ is the number of bits of precision . the c - to - d does only the first step , without any rounding (quantization) . you can think of the c - to - d as an a - to - d with infinite precision

(DSP First paragraph)

Examples: “What is phase difference?”

Word2Vec

In 3.3.1, what value should I have to put for phase1 and phase2?

Word2Vec-Transformer

(None)

Sentence Transformer

Just for clarity can someone explain the difference between phase, frequency and period

Dual-Transformer

1. Period is the amount of time in one cycle of the sinusoid, and can measured as the distance between the peaks of the sinusoid. Frequency is the number of cycles in a second, and is the inverse of the period. Phase is the distance that the sinusoid is shifted from zero.

Examples: “What is phase difference?”

Word2Vec

It doesn't matter. When you take the derivative to find the instantaneous frequency, the phase is a constant, so it goes away.

Word2Vec-Transformer

(None)

Sentence Transformer

Period is the amount of time in one cycle of the sinusoid, and can measured as the distance between the peaks of the sinusoid. Frequency is the number of cycles in a second, and is the inverse of the period. Phase is the distance that the sinusoid is shifted from zero.

Dual-Transformer

1. the distance that the sinusoid is shifted from zero

Examples: “Are calculators allowed for the exams?”

Word2Vec

<p>What kind of calculator should we have for the exam? I have a scientific calculator, do I need to have a graphing calculator for the exam?</p>

Word2Vec-Transformer

1. ... in AP tests, some problems allow or require you to use the graphing capability of the calculator while other problems specifically prohibit you from using graphing. In our tests, no graphing is allowed.
2. Not the focus of this Exam, however both of these concepts should be understood.
3. Yes, calculators that cannot connect to the internet are allowed.... A simple "scientific" calculator should be enough to calculate the trigonometric functions required for the course.

Sentence Transformer

It is fine to use your calculator.

Dual-Transformer

1. ... I bought a graphing calculator... It says on the package that it is acceptable for SAT/ACT/AP tests...
2. ... In our tests, no graphing is allowed.
3. Quiz 1 open note, open book. Calculators are allowed. MATLAB is allowed.
4. Quiz 1 open note, open book. Calculators are allowed. MATLAB is allowed.
5. Simple computations (with or without a calculator) do not require any justification. Try to provide guidance to your grading TA...

Examples: “Are calculators allowed for the exams?”

Word2Vec

Any calc that has polar, cartesian form calculations and common trig functions should be sufficient.

Word2Vec-Transformer

1. no graphing is allowed
2. not the focus of this exam
3. yes , calculators that cannot connect to the internet are allowed

Sentence Transformer

It is fine to use your calculator.

Dual-Transformer

1. it says on the package that it is acceptable for sat / act / ap tests
2. no graphing is allowed
3. calculators are allowed
4. calculators are allowed
5. simple computations (with or without a calculator) do not require any justification

Results

- All models perform similarly on logistical questions.
- Sentence Transformer or Dual-Transformer spends more time to output responses than a Word2Vec or Word2Vec-Transformer. However, they perform better than the other two.
- Word2Vec-Transformer seems to be largely dependent on the relevance of the candidate contexts retrieved by the Word2Vec part of it, which does not perform well

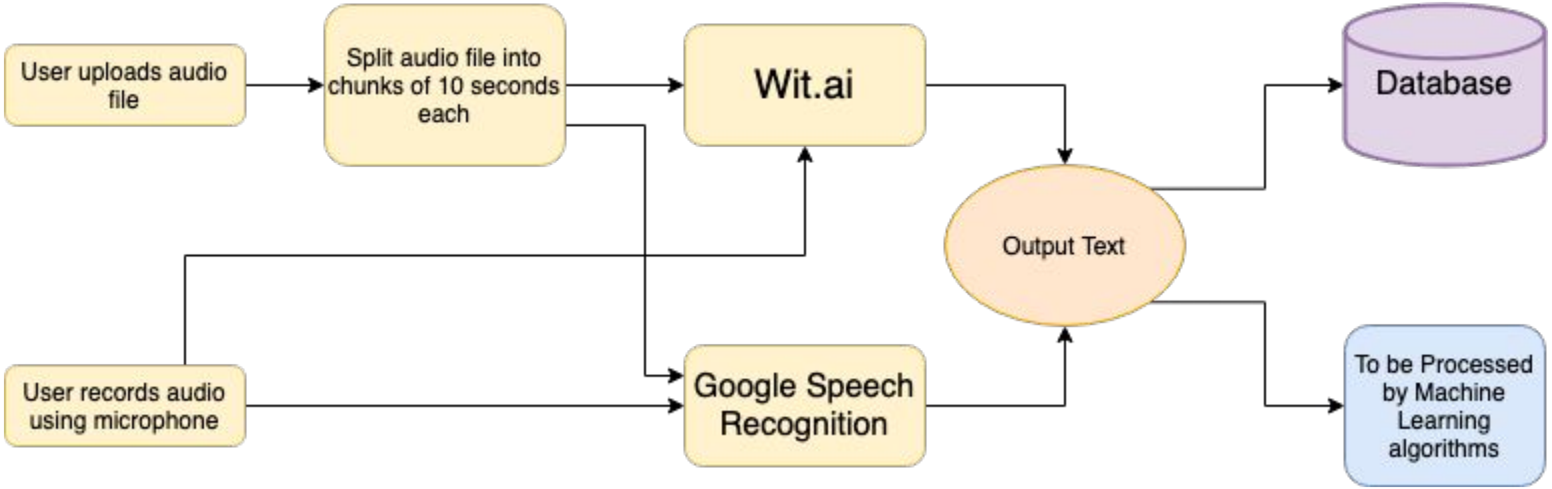
Plans for future semesters

- Explore more annotated data in this domain
- Improve on the run-time of Sentence-Transformer and Dual-Transformer
- Train the pre-trained model further with our data
- Compare current models with the work done by Predictive Team

Front-end

Add speech-to-text functionality to increase accessibility and explore further applications of audio input.

What we have now


User uploads audio file
Split audio file into chunks of 10 seconds each
Wit.ai
User records audio using microphone
Google Speech Recognition
Output Text
Database
To be Processed by Machine Learning algorithms

Things we have tried

- Explored two APIs to transcribe audio
  - Wit.ai
  - Google Speech Recognition
- Split the audio files into chunks of constant time each to better handle long audio files
  - Pydub
- Developed a service that transcribes audio files and users recordings from microphones
- Stored the transcripts and timestamps to the SQLite database

Audio Processing using Pydub

- Split the audio file into chunks of 10 seconds each
  - Might interrupt sentences in between and the API might not be able to recognize incomplete words
- Split the audio file based on silence in between words
  - Process the audio file sentence by sentence
  - Will not cause any interruptions
- Split the audio file into small chunks of a constant interval
  - Slicing is done with overlap so that the next chunk will begin from a constant time backward
  - If any word gets interrupted, it can be covered by this overlap

Microphone Input - Motivation

- User Experience (UX)
- Improves accessibility and ease-of-use
- Accessibility guidelines set by American Disability Act (ADA)

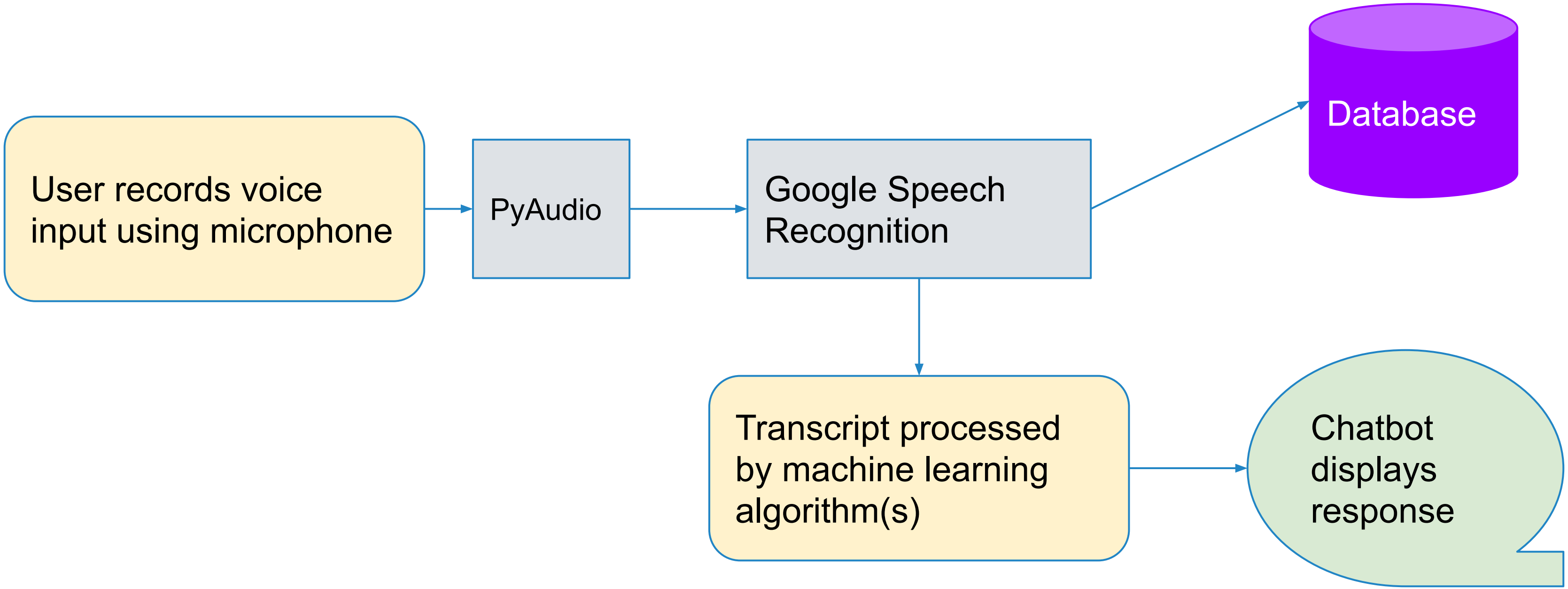
Microphone Input - What we have tried

- PyAudio library for audio input
- 3 APIs for Speech-to-Text were tested
- Microsoft Azure Paid free trial
- Wit.ai - open source library
- **Uses Google Speech Recognition (free version)**
- Sending transcripts to SQLite database

Microphone Input - Key features

- Fast real-time voice input via PyAudio library
- Fast transcription using Google Speech Recognition
- Audio length is dynamic; prints transcription once user is done speaking
- Transcripts stored in database

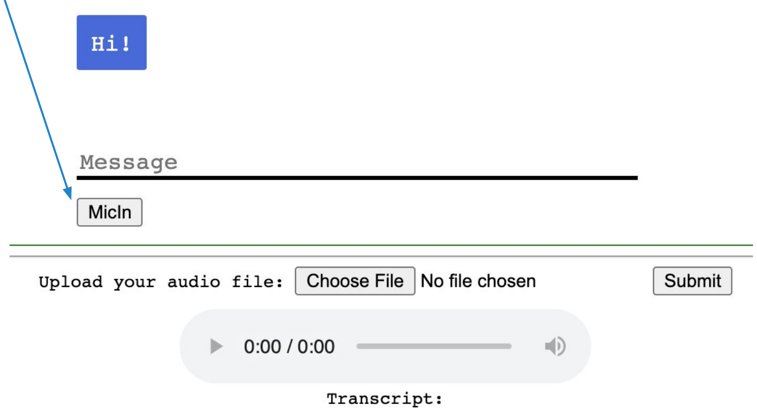
Microphone Input - Block Flow Diagram



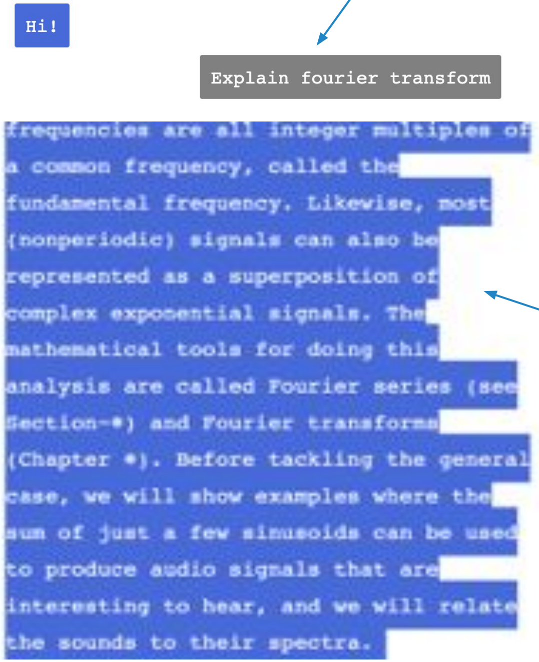
User records voice input using microphone
PyAudio
Google Speech Recognition
Database
Transcript processed by machine learning algorithm(s)
Chatbot displays response

Microphone Input - User Interface Demonstration

Step 1: User clicks Mic Input button


Hi!

Step 2: Chatbot enters input text after processed by speech-to-text


Hi!

Explain fourier transform

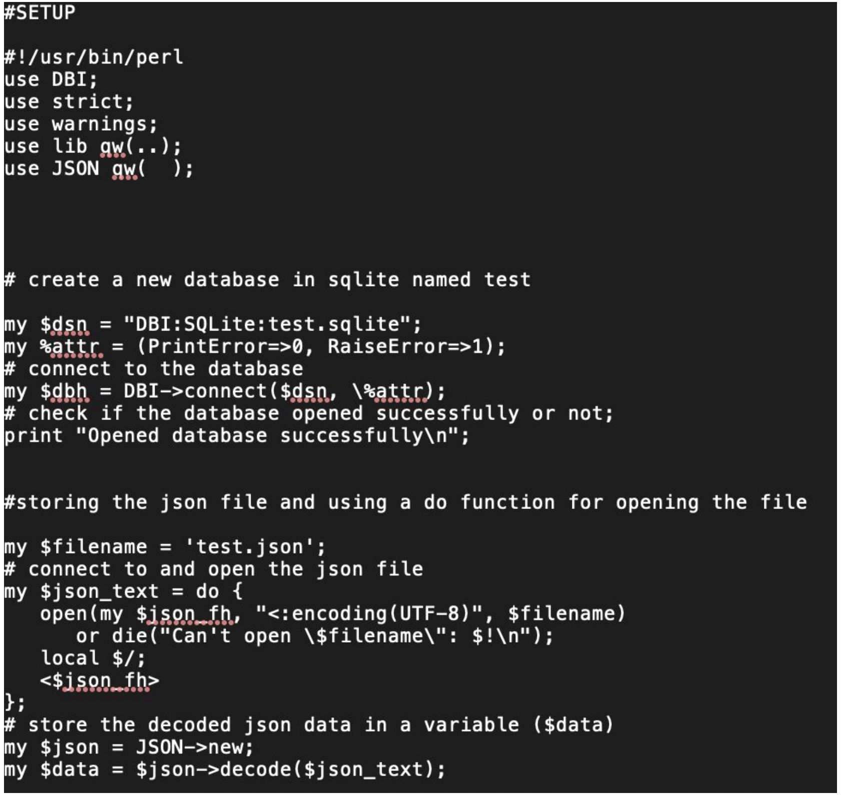
frequencies are all integer multiples of a common frequency, called the fundamental frequency. Likewise, most (nonperiodic) signals can also be represented as a superposition of complex exponential signals. The mathematical tools for doing this analysis are called Fourier series (see Section-*) and Fourier transforms (Chapter *). Before tackling the general case, we will show examples where the sum of just a few sinusoids can be used to produce audio signals that are interesting to hear, and we will relate the sounds to their spectra.

Step 3: Input processed by chatbot and response is given

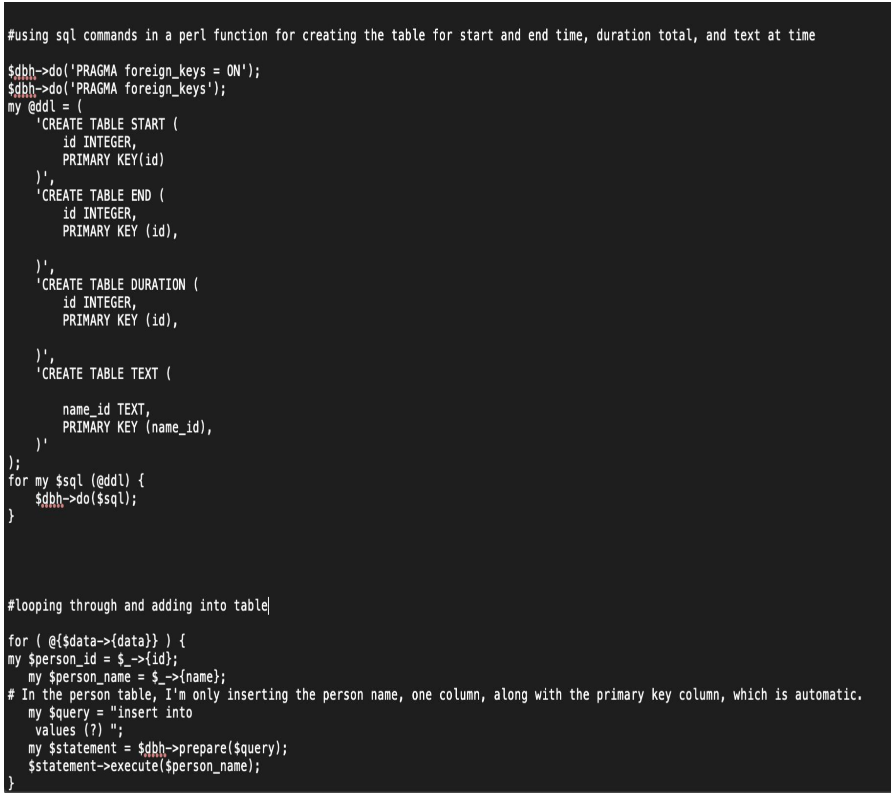
Storing Transcripts in SQLite database

- Transcripts
- Timestamps for when the audio was transcribed
- Archive allows for larger file storage with atomic incremental updating (faster querying)
- Store small sized sound files as BLOB fields
- Using json1 extension for storing JSON files (with transcription) as ordinary text

SQLite3 Code Sample Framework



```
#SETUP

#!/usr/bin/perl
use DBI;
use strict;
use warnings;
use lib qw(..);
use JSON qw(  );



# create a new database in sqlite named test

my $dsn = "DBI:SQLite:test.sqlite";
my %attr = (PrintError=>0, RaiseError=>1);
# connect to the database
my $dbh = DBI->connect($dsn, \%attr);
# check if the database opened successfully or not;
print "Opened database successfully\n";


#storing the json file and using a do function for opening the file

my $filename = 'test.json';
# connect to and open the json file
my $json_text = do {
   open(my $json_fh, "<:encoding(UTF-8)", $filename)
      or die("Can't open \$filename\": $!\n");
   local $/;
   <$json_fh>
};
# store the decoded json data in a variable ($data)
my $json = JSON->new;
my $data = $json->decode($json_text);
```



```
#using sql commands in a perl function for creating the table for start and end time, duration total, and text at time

$dbh->do('PRAGMA foreign_keys = ON');
$dbh->do('PRAGMA foreign_keys');
my @ddl = (
    'CREATE TABLE START (
        id INTEGER,
        PRIMARY KEY(id)
    )',
    'CREATE TABLE END (
        id INTEGER,
        PRIMARY KEY (id),

    )',
    'CREATE TABLE DURATION (
        id INTEGER,
        PRIMARY KEY (id),

    )',
    'CREATE TABLE TEXT (

        name_id TEXT,
        PRIMARY KEY (name_id),
    )'
);
for my $sql (@ddl) {
    $dbh->do($sql);
}




#looping through and adding into table

for ( @{$data->{data}} ) {
my $person_id = $_->{id};
   my $person_name = $_->{name};
# In the person table, I'm only inserting the person name, one column, along with the primary key column, which is automatic.
   my $query = "insert into
    values (?) ";
   my $statement = $dbh->prepare($query);
   $statement->execute($person_name);
}
```

Issues we ran into

- While splitting audio files, sentences are interrupted due to the program using the splitting into chunks of constant size
- Splitting audio files based on silence is difficult as we need to know the dBFS of the audio files in order to set a threshold to consider which parts of the audio files are silent
- Originally without splitting audio files, long audio files caused the program to time out occasionally

Plans for Subsequent Semesters

- Looking further into storing transcription data into SQLite database
  - Streamlining storage process of JSON files
- Adding a numerical feedback system to facilitate long-term improvement for the chatbot