Using Human Interaction with Natural Language Processing Techniques to Reinforce Vocabulary Comprehension and Usage

(Note: some slides removed due to file size restrictions)

Ph.D. Candidate: Tabitha Samuel

Major Advisor: Dr. Michael W. Berry

Committee: Dr. Audris Mockus (EECS), Dr. Amir Sadovnik (EECS),
Dr. Jillian McCarthy (Department of Audiology and Speech Pathology,UTHSC)

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

What is SENCE?

❑ SENCE: SENtence Curation and Evaluation

❑ Natural Language Processing aid to be used in a vocabulary teaching setting for children

❑ Used to reinforce meaning of words already taught

❑ SENCE also assesses students’ capacity to retain vocabulary comprehension from familiar and unfamiliar sources

SENCE Goals

❑ Lower manual intervention for educators in identifying critical words to teach

❑ Assist educators in identifying and grouping sentences for assessment

❑ Aid students in learning and retaining meaning of words

Motivation

Vocabulary Coordinator - VocaCoord

- Novel software that uses speech to text to display vocabulary in real time
- It takes academic vocabulary words spoken by a teacher and displays them in written form providing a static visual representation of the word being taught.
- An image of the word is shown alongside the word itself
- The additional cue helps the student remember the meaning of the word faster as only opposed to the audio cue

Research Objectives

Provide accessible interfaces to an NLP solution

Blend human input with NLP responses

Assess how NLP tools can parse different media formats

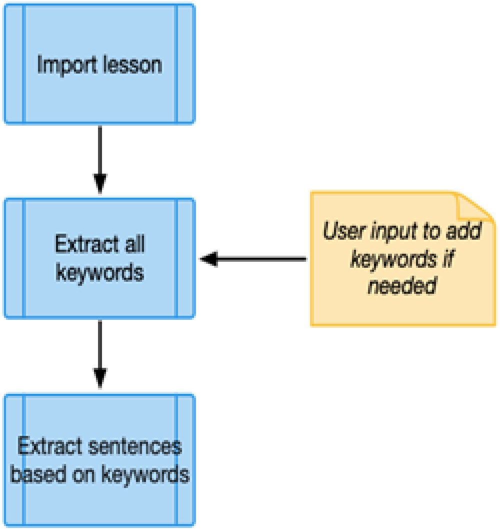
Scaffold complexity to reinforce vocabulary knowledge

Architecture: Corpus parsing and storage

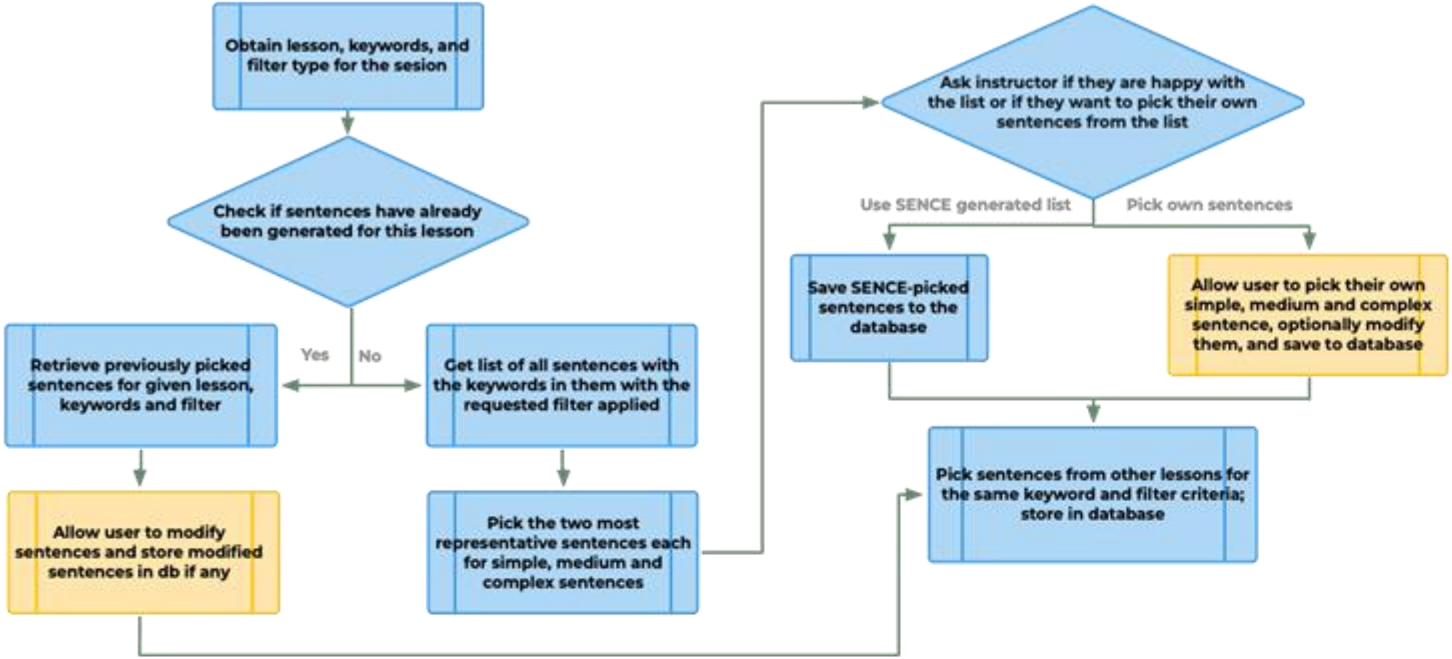
Parse and store passage

Capture keywords from the passage

Capture sentences associated with the keywords being taught


Import lesson
Extract all keywords
User input to add keywords if needed
Extract sentences based on keywords

Architecture: Lesson-specific keyword and sentence selection and storage



Obtain lesson, keywords, and filter type for the sesion

Check if sentences have already been generated for this lesson

Yes

No

Retrieve previously picked sentences for given lesson, keywords and filter

Allow user to modify sentences and store modified sentences in db if any

Get list of all sentences with the keywords in them with the requested filter applied

Pick the two most representative sentences each for simple, medium and complex sentences

Ask instructor if they are happy with the list or if they want to pick their own sentences from the list

Use SENCE generated list

Pick own sentences

Save SENCE-picked sentences to the database

Allow user to pick their own simple, medium and complex sentence, optionally modify them, and save to database

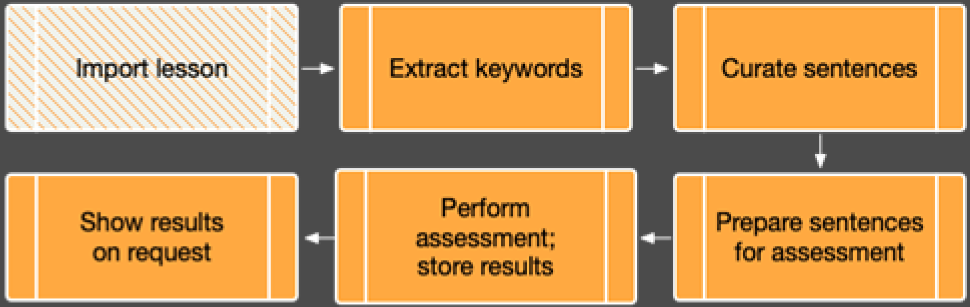
Pick sentences from other lessons for the same keyword and filter criteria; store in database

Architecture: Major Libraries and Software Used

- spaCy - used for most NLP tasks
- Stanza - used specifically for lemmatization
- MongoDB - noSQL database
- pyInputPlus - for user-friendly interactions and custom validations of input

Methodology and Results

Step 1: Import lesson



Methods & Results: Corpus parsing and storage

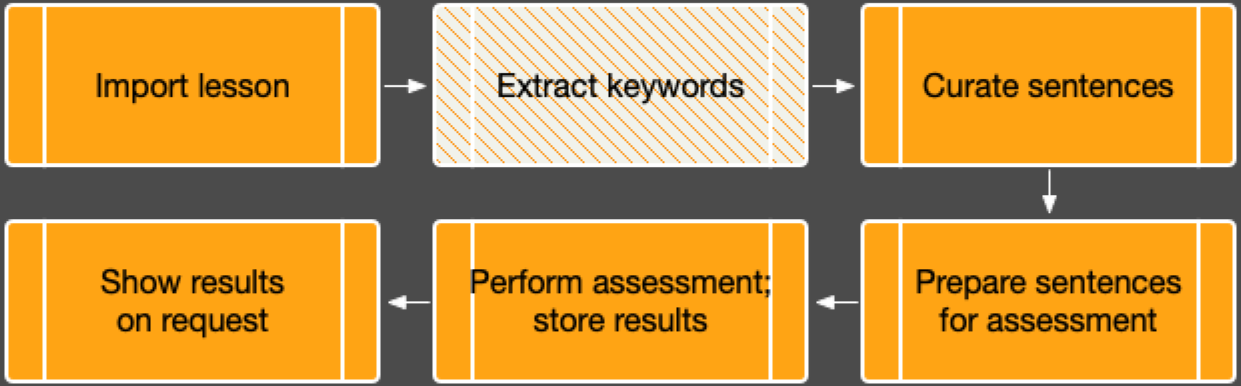
- First step is to extract text, clean and parse it
- Text is heavily formatted
- Automated and manual pre-processing
- Manual:
  - text copied over to a plain text editor to remove formatting
  - Metadata removed manually

Methods & Results: Corpus parsing and storage

Automated pre-processing

- Remove extra whitespace
- Meta information evaluated - word count, sentence count, average sentence length, length of longest sentence
- Passage and metadata is stored in MongoDB in the ‘passage’ collection
- Error checking

Step 2: Extract Keywords



Methods & Results: Keyword extraction

- Keywords are critical words essential to the context of the passage
- Subset of keywords to be taught in a lesson and is the vocabulary being reinforced through a session of SENCE
- On a passage about volcanoes, keywords could be lava, mountain, spew, molten
- Objective in this segment of SENCE:
  - Automate keyword extraction
  - Adding human curation to keyword list generated

Methods & Results: Keyword extraction

- Text is first converted to lower-case to ensure that SENCE treats words such as “hello”, “Hello”, and “HELLO” the same
- Text is then lemmatized
- Remove stopwords
- Part of speech tagging
  - SENCE only accepts nouns, adjectives, verbs, adverbs, and proper nouns as keywords in this version
- spaCy is used to then generate the top 10 keywords

Methods & Results: Evaluation of different keyword extraction methods

spaCy, Yake, Rake, Bert, Textacy, SgRank, and TextRank were evaluated for keyword extraction

NLP Package	Underlying approach
Rake	Unsupervised learning focusing on statistical properties
Yake	Unsupervised learning focusing on statistical properties
Textacy	Graph based method
TextRank	Graph based method based on Google's Page Rank algorithm
SgRank	Combines statistical and graph-based methods
BERT	Uses transformer-based models
spaCy	uses both statistical models and transformer-based models

Methods & Results: Evaluation of different keyword extraction methods

Statistical methods: Solely rely on statistical properties of a text, eg word frequencies

Graph-based methods: build a graph where nodes represent words or phrases and edges represent relationships

Transformer based methods: Uses deep learning to understand context and semantics. Results in more accurate set of keyword that represent the richer semantic meaning of the text

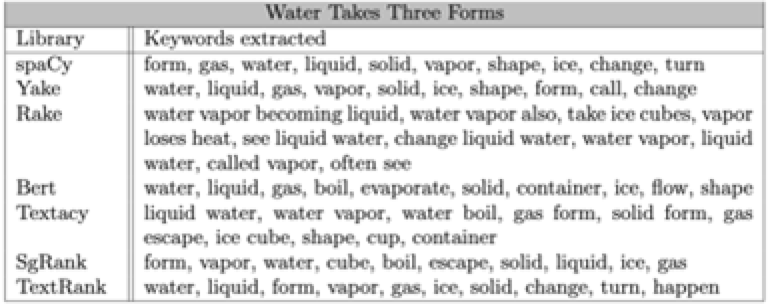
Methods & Results: Evaluation of different keyword extraction methods

❑ Experiment to capture which package performs the best over a variety of subjects and passages.

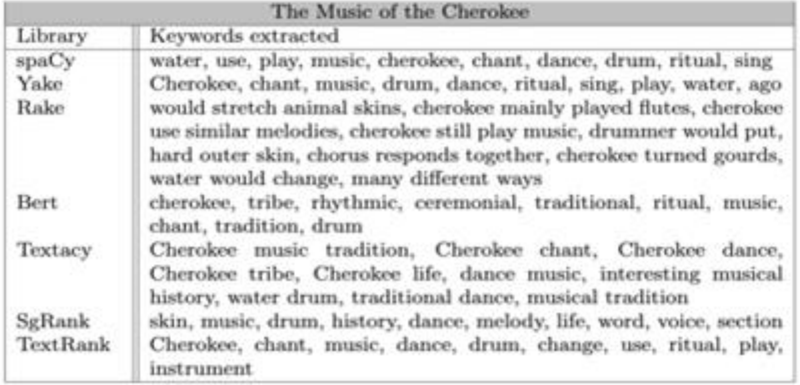
❑ Nine passages chosen - 3 each from science-, social studies- and arts-based passages.

❑ A subject matter expert in language and speech helped assess the quality of results and determine which packages most closely matched human choices

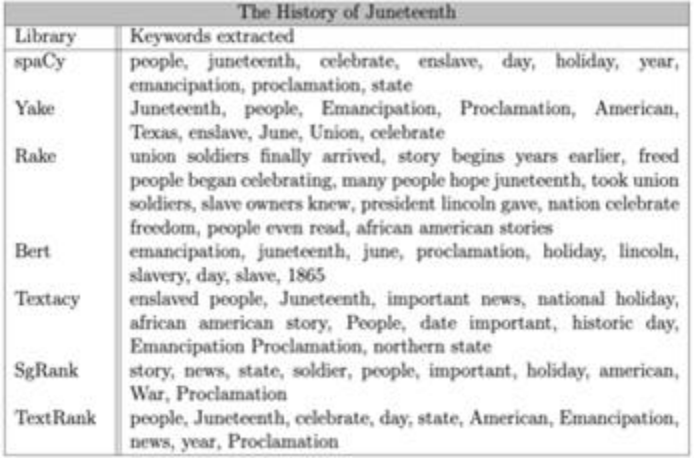
Methods & Results: Evaluation of different keyword extraction methods



Water Takes Three Forms	
Library	Keywords extracted
spaCy	form, gas, water, liquid, solid, vapor, shape, ice, change, turn
Yake	water, liquid, gas, vapor, solid, ice, shape, form, call, change
Rake	water vapor becoming liquid, water vapor also, take ice cubes, vapor loses heat, see liquid water, change liquid water, water vapor, liquid water, called vapor, often see
Bert	water, liquid, gas, boil, evaporate, solid, container, ice, flow, shape
Textacy	liquid water, water vapor, water boil, gas form, solid form, gas escape, ice cube, shape, cup, container
SgRank	form, vapor, water, cube, boil, escape, solid, liquid, ice, gas
TextRank	water, liquid, form, vapor, gas, ice, solid, change, turn, happen



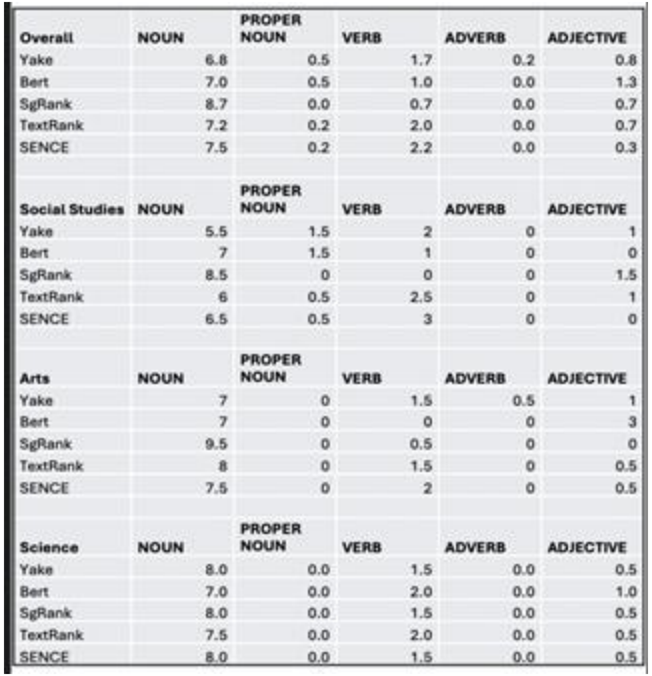
The Music of the Cherokee	
Library	Keywords extracted
spaCy	water, use, play, music, cherokee, chant, dance, drum, ritual, sing
Yake	Cherokee, chant, music, drum, dance, ritual, sing, play, water, ago
Rake	would stretch animal skins, cherokee mainly played flutes, cherokee use similar melodies, cherokee still play music, drummer would put, hard outer skin, chorus responds together, cherokee turned gourds, water would change, many different ways
Bert	cherokee, tribe, rhythmic, ceremonial, traditional, ritual, music, chant, tradition, drum
Textacy	Cherokee music tradition, Cherokee chant, Cherokee dance, Cherokee tribe, Cherokee life, dance music, interesting musical history, water drum, traditional dance, musical tradition
SgRank	skin, music, drum, history, dance, melody, life, word, voice, section
TextRank	Cherokee, chant, music, dance, drum, change, use, ritual, play, instrument



The History of Juneteenth	
Library	Keywords extracted
spaCy	people, juneteenth, celebrate, enslave, day, holiday, year, emancipation, proclamation, state
Yake	Juneteenth, people, Emancipation, Proclamation, American, Texas, enslave, June, Union, celebrate
Rake	union soldiers finally arrived, story begins years earlier, freed people began celebrating, many people hope juneteenth, took union soldiers, slave owners knew, president lincoln gave, nation celebrate freedom, people even read, african american stories
Bert	emancipation, juneteenth, june, proclamation, holiday, lincoln, slavery, day, slave, 1865
Textacy	enslaved people, Juneteenth, important news, national holiday, african american story, People, date important, historic day, Emancipation Proclamation, northern state
SgRank	story, news, state, soldier, people, important, holiday, american, War, Proclamation
TextRank	people, Juneteenth, celebrate, day, state, American, Emancipation, news, year, Proclamation

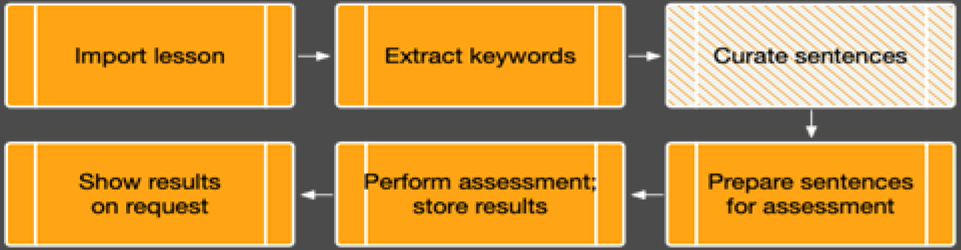
Methods & Results: Evaluation of different keyword extraction methods

- BERT and SgRank prefer nouns, proper nouns and adjectives over verbs and adverbs
  - Transformer based models use contextual information for keyword extraction
  - Nouns and adjectives carry more semantic weight in a sentence
- Yake, TextRank and SENCE perform similarly in science based texts, but SENCE has more variety of parts-of-speech in non-science based texts



Overall	NOUN	PROPER NOUN	VERB	ADVERB	ADJECTIVE
Yake	6.8	0.5	1.7	0.2	0.8
Bert	7.0	0.5	1.0	0.0	1.3
SgRank	8.7	0.0	0.7	0.0	0.7
TextRank	7.2	0.2	2.0	0.0	0.7
SENCE	7.5	0.2	2.2	0.0	0.3
Social Studies	**NOUN**	**PROPER NOUN**	**VERB**	**ADVERB**	**ADJECTIVE**
Yake	5.5	1.5	2	0	1
Bert	7	1.5	1	0	0
SgRank	8.5	0	0	0	1.5
TextRank	6	0.5	2.5	0	1
SENCE	6.5	0.5	3	0	0
Arts	**NOUN**	**PROPER NOUN**	**VERB**	**ADVERB**	**ADJECTIVE**
Yake	7	0	1.5	0.5	1
Bert	7	0	0	0	3
SgRank	9.5	0	0.5	0	0
TextRank	8	0	1.5	0	0.5
SENCE	7.5	0	2	0	0.5
Science	**NOUN**	**PROPER NOUN**	**VERB**	**ADVERB**	**ADJECTIVE**
Yake	8.0	0.0	1.5	0.0	0.5
Bert	7.0	0.0	2.0	0.0	1.0
SgRank	8.0	0.0	1.5	0.0	0.5
TextRank	7.5	0.0	2.0	0.0	0.5
SENCE	8.0	0.0	1.5	0.0	0.5

Step 3: Sentence Curation



Import lesson

Extract keywords

Curate sentences

Prepare sentences for assessment

Perform assessment; store results

Show results on request

Methods & Results: Sentence retrieval based on keywords

- Ask the instructor to choose the passage
- Retrieve keywords
- If keywords are found:
    - Lemmatize the sentence
    - Check each lemmatized sentence against list of keywords
    - Store results

Choice of Lemmatizer:spaCy vs Stanza
Original sentence: ‘I love Dogs and Cats’
spaCy lemmatization: ‘I love Dogs and cat’
Stanza lemmatization: “I love dog and cat’

Methods & Results: Sentence curation

- Each group of keywords for a lesson can produce several sentences.
- There are 35 sentences with at least one of the words ‘volcano’, ‘lava’, ‘mountain’, ‘ash’ from a single passage.
- Need a way to pick only the most relevant sentences.
- Introducing: SENCE Sorts
  - Sentence Length
  - Number of keywords
  - Syntactic dependency of keyword relative to the sentence
  - Number of tier 2 and tier 3 words

Methods & Results: Sorts - Sentence Length & Number of keywords

- Approach:
  - Establish upper and lower bounds for simple, medium and complex sentences
  - Determine either sentence length or number of keywords in each sentence and classify appropriately
  - Pick top two of simple, medium and complex each to be stored for the lesson, keywords chosen and sort type.

Bounds for Sentence Length Sort

simple_upperbound = math.ceil(max_length_of_sentence/3)+2
*medium_upperbound = (simple - 2) * 2*

Bounds for Number of keywords Sort

bucket size = math.floor(num kw/3)
simple upperbound = bucket size
medium upperbound = simple upperbound + bucket size + 1

Methods & Results: Sorts - Determining top ‘x’ results

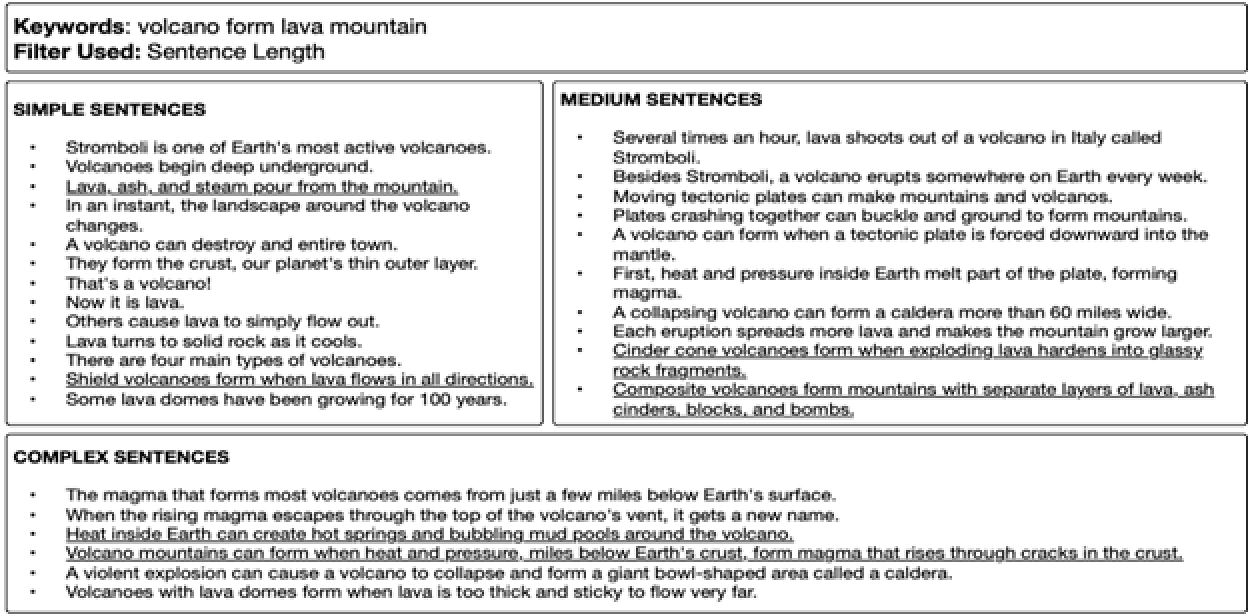
Approach:

- How does one find the most representative sentence(s) from a group of sentences?
    - Applications in text summarization, key phrase extraction
- Approach used in SENCE:
    - Calculate cosine similarity between the keywords string and each picked sentence
    - The top ‘x’ match scores are returned as matches

For keywords: lava, ash, mountain, volcano

Sentence	**Similarity Score**
Several times an hour, lava shoots out of a volcano in Italy called Stromboli.	0.2936643870643254
Lava, ash, and steam pour from the mountain.	0.4417674864719228
Besides Stromboli, a volcano erupts somewhere on Earth every week.	0.24732109352596257
The magma that forms most volcanoes comes from just a few miles below Earth's surface.	0.01811698354476031

Methods & Results: Sorts - Sentence Length Results



Keywords: volcano form lava mountain
Filter Used: Sentence Length

SIMPLE SENTENCES

- Stromboli is one of Earth's most active volcanoes.
- Volcanoes begin deep underground.
- Lava, ash, and steam pour from the mountain.
- In an instant, the landscape around the volcano changes.
- A volcano can destroy and entire town.
- They form the crust, our planet's thin outer layer.
- That's a volcano!
- Now it is lava.
- Others cause lava to simply flow out.
- Lava turns to solid rock as it cools.
- There are four main types of volcanoes.
- Shield volcanoes form when lava flows in all directions.
- Some lava domes have been growing for 100 years.

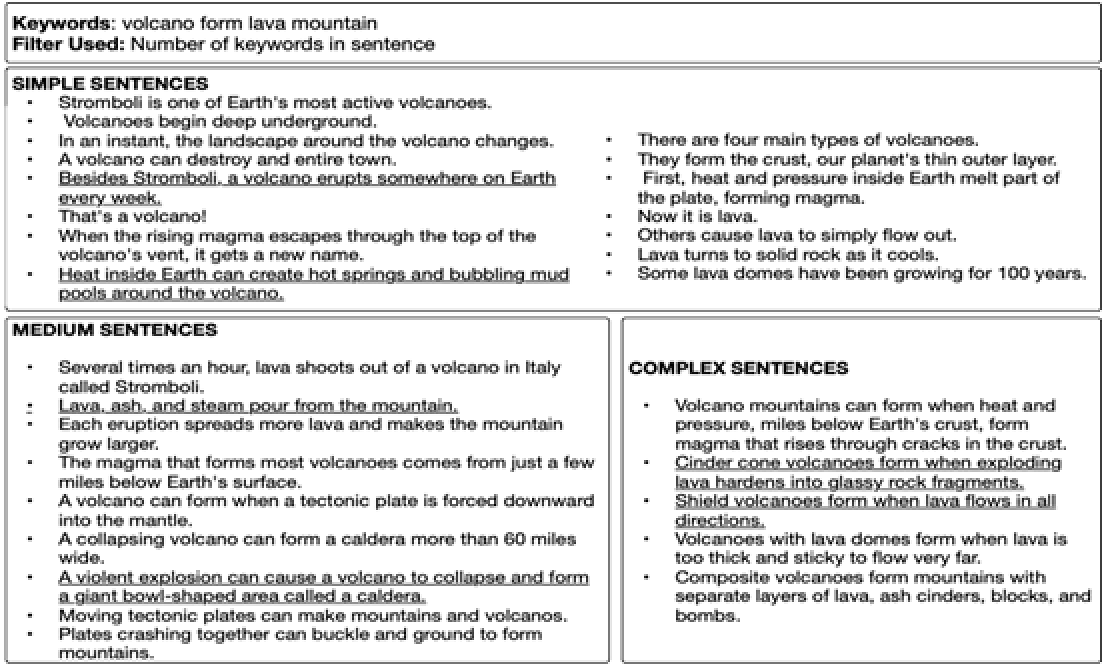
MEDIUM SENTENCES

- Several times an hour, lava shoots out of a volcano in Italy called Stromboli.
- Besides Stromboli, a volcano erupts somewhere on Earth every week.
- Moving tectonic plates can make mountains and volcanos.
- Plates crashing together can buckle and ground to form mountains.
- A volcano can form when a tectonic plate is forced downward into the mantle.
- First, heat and pressure inside Earth melt part of the plate, forming magma.
- A collapsing volcano can form a caldera more than 60 miles wide.
- Each eruption spreads more lava and makes the mountain grow larger.
- Cinder cone volcanoes form when exploding lava hardens into glassy rock fragments.
- Composite volcanoes form mountains with separate layers of lava, ash cinders, blocks, and bombs.

COMPLEX SENTENCES

- The magma that forms most volcanoes comes from just a few miles below Earth's surface.
- When the rising magma escapes through the top of the volcano's vent, it gets a new name.
- Heat inside Earth can create hot springs and bubbling mud pools around the volcano.
- Volcano mountains can form when heat and pressure, miles below Earth's crust, form magma that rises through cracks in the crust.
- A violent explosion can cause a volcano to collapse and form a giant bowl-shaped area called a caldera.
- Volcanoes with lava domes form when lava is too thick and sticky to flow very far.

Methods & Results: Sorts - Number of Keywords in a sentence Results



Keywords: volcano form lava mountain
Filter Used: Number of keywords in sentence

SIMPLE SENTENCES

- Stromboli is one of Earth's most active volcanoes.
- Volcanoes begin deep underground.
- In an instant, the landscape around the volcano changes.
- A volcano can destroy and entire town.
- Besides Stromboli, a volcano erupts somewhere on Earth every week.
- That's a volcano!
- When the rising magma escapes through the top of the volcano's vent, it gets a new name.
- Heat inside Earth can create hot springs and bubbling mud pools around the volcano.
- There are four main types of volcanoes.
- They form the crust, our planet's thin outer layer.
- First, heat and pressure inside Earth melt part of the plate, forming magma.
- Now it is lava.
- Others cause lava to simply flow out.
- Lava turns to solid rock as it cools.
- Some lava domes have been growing for 100 years.

MEDIUM SENTENCES

- Several times an hour, lava shoots out of a volcano in Italy called Stromboli.
- Lava, ash, and steam pour from the mountain.
- Each eruption spreads more lava and makes the mountain grow larger.
- The magma that forms most volcanoes comes from just a few miles below Earth's surface.
- A volcano can form when a tectonic plate is forced downward into the mantle.
- A collapsing volcano can form a caldera more than 60 miles wide.
- A violent explosion can cause a volcano to collapse and form a giant bowl-shaped area called a caldera.
- Moving tectonic plates can make mountains and volcanos.
- Plates crashing together can buckle and ground to form mountains.

COMPLEX SENTENCES

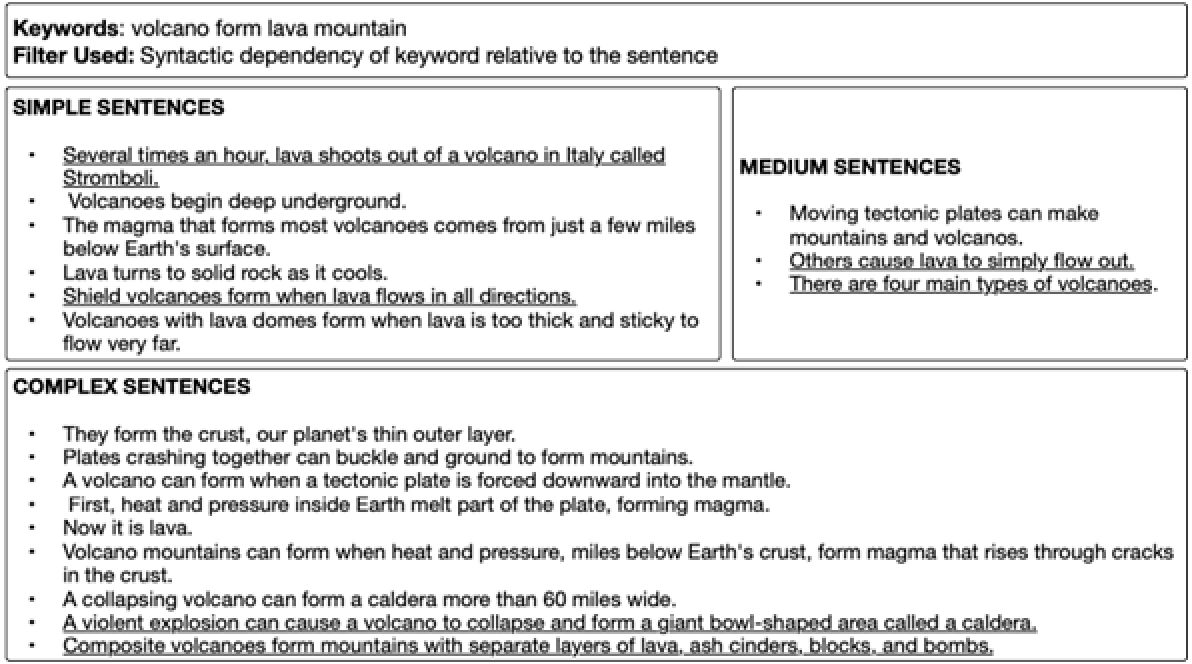
- Volcano mountains can form when heat and pressure, miles below Earth's crust, form magma that rises through cracks in the crust.
- Cinder cone volcanoes form when exploding lava hardens into glassy rock fragments.
- Shield volcanoes form when lava flows in all directions.
- Volcanoes with lava domes form when lava is too thick and sticky to flow very far.
- Composite volcanoes form mountains with separate layers of lava, ash cinders, blocks, and bombs.

Methods & Results: Syntactic dependency of keyword relative to the sentence

Approach:

- ❑ Convert the sentence into a spaCy document.
  - ❑ spaCy document includes metadata such as parts-of-speech and syntactic dependency information
- ❑ Use spaCy method to identify noun chunks in sentence
- ❑ spaCy does not have a in-built method to discover verb chunks - wrote one similar to behavior of the noun chunking method
- ❑ Iterate over lemmatized version of each chunk to identify presence of keywords if any
- ❑ If keyword is present:
  - ❑ If it is subject: classify as simple sentence
  - ❑ If it is the object: classify as medium sentence
  - ❑ Everything else: complex sentence

Methods & Results: Syntactic dependency of keyword relative to the sentence Results



Keywords: volcano form lava mountain
Filter Used: Syntactic dependency of keyword relative to the sentence

SIMPLE SENTENCES

- Several times an hour, lava shoots out of a volcano in Italy called Stromboli.
- Volcanoes begin deep underground.
- The magma that forms most volcanoes comes from just a few miles below Earth's surface.
- Lava turns to solid rock as it cools.
- Shield volcanoes form when lava flows in all directions.
- Volcanoes with lava domes form when lava is too thick and sticky to flow very far.

MEDIUM SENTENCES

- Moving tectonic plates can make mountains and volcanos.
- Others cause lava to simply flow out.
- There are four main types of volcanoes.

COMPLEX SENTENCES

- They form the crust, our planet's thin outer layer.
- Plates crashing together can buckle and ground to form mountains.
- A volcano can form when a tectonic plate is forced downward into the mantle.
- First, heat and pressure inside Earth melt part of the plate, forming magma.
- Now it is lava.
- Volcano mountains can form when heat and pressure, miles below Earth's crust, form magma that rises through cracks in the crust.
- A collapsing volcano can form a caldera more than 60 miles wide.
- A violent explosion can cause a volcano to collapse and form a giant bowl-shaped area called a caldera.
- Composite volcanoes form mountains with separate layers of lava, ash cinders, blocks, and bombs.

Methods & Results: Number of Tier 2 and 3 words in a sentence

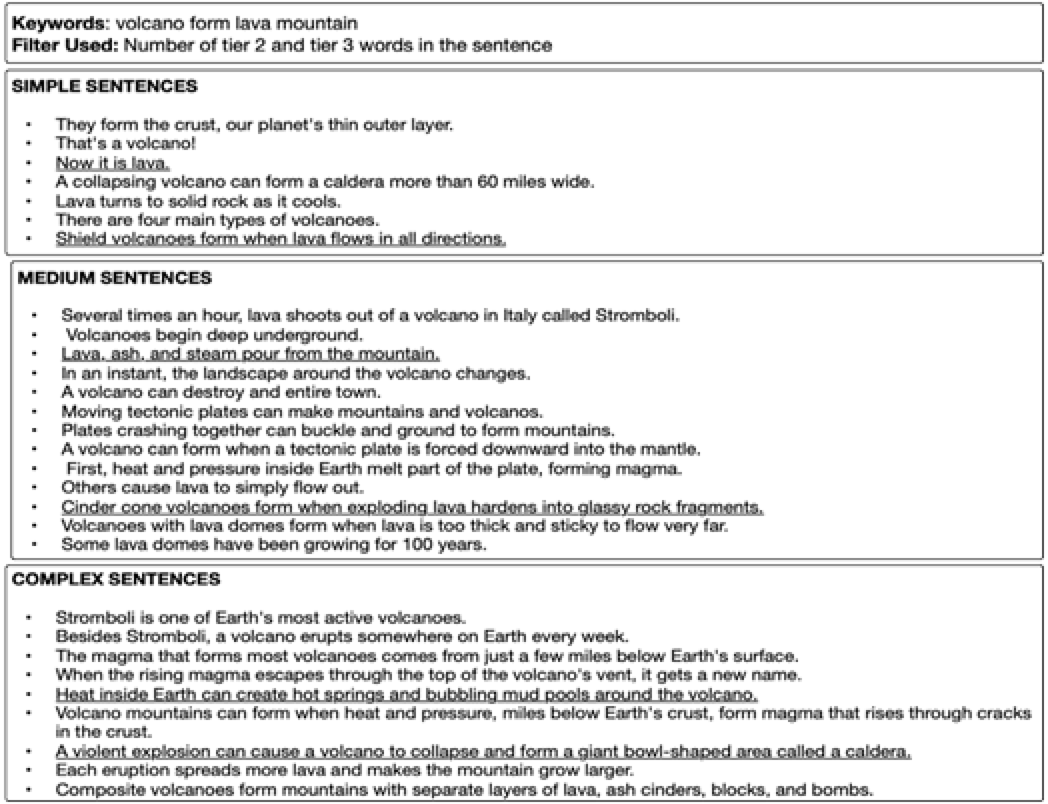
- Most children begin first grade with ~3,000 words of spoken vocabulary
- They will learn ~3,000 words per year through third grade
- How do we know which words to teach?
- Tier words
  - Tier 1 - Basic vocabulary
  - Tier 2 - High frequency / multiple meaning words
  - Tier 3 - Subject related
- SENCE has a collection of ~2500 tier 2 and 3 words
- Instructors can add to this collection

Methods & Results:
Number of Tier 2 and 3 words in a sentence

Approach:

- Sentences are retrieved from MongoDB
- Each sentence is then tokenized
- All stop words are removed
- Each token is lemmatized
- Tokens are matched to tier words
- Instructor is shown list of tier words in the selected sentences and allowed to add more
- Additional sentences based on new tier words are retrieved
- Sentences are classified as simple, medium and complex depending on how many tier words are in the sentence
- Top 2 from each bucket is chosen to be stored

Methods & Results: Number of Tier 2 and 3 words in a sentence - Results



Keywords: volcano form lava mountain
Filter Used: Number of tier 2 and tier 3 words in the sentence

SIMPLE SENTENCES

- They form the crust, our planet's thin outer layer.
- That's a volcano!
- Now it is lava.
- A collapsing volcano can form a caldera more than 60 miles wide.
- Lava turns to solid rock as it cools.
- There are four main types of volcanoes.
- Shield volcanoes form when lava flows in all directions.

MEDIUM SENTENCES

- Several times an hour, lava shoots out of a volcano in Italy called Stromboli.
- Volcanoes begin deep underground.
- Lava, ash, and steam pour from the mountain.
- In an instant, the landscape around the volcano changes.
- A volcano can destroy and entire town.
- Moving tectonic plates can make mountains and volcanos.
- Plates crashing together can buckle and ground to form mountains.
- A volcano can form when a tectonic plate is forced downward into the mantle.
- First, heat and pressure inside Earth melt part of the plate, forming magma.
- Others cause lava to simply flow out.
- Cinder cone volcanoes form when exploding lava hardens into glassy rock fragments.
- Volcanoes with lava domes form when lava is too thick and sticky to flow very far.
- Some lava domes have been growing for 100 years.

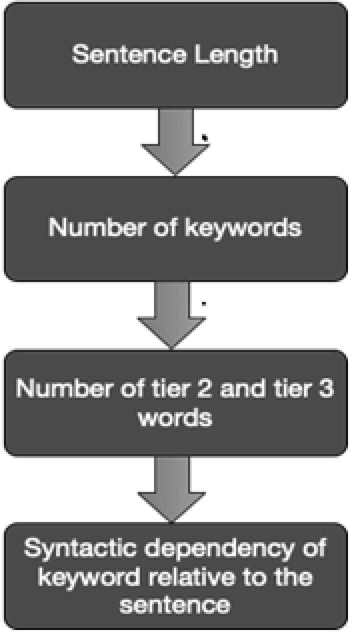
COMPLEX SENTENCES

- Stromboli is one of Earth's most active volcanoes.
- Besides Stromboli, a volcano erupts somewhere on Earth every week.
- The magma that forms most volcanoes comes from just a few miles below Earth's surface.
- When the rising magma escapes through the top of the volcano's vent, it gets a new name.
- Heat inside Earth can create hot springs and bubbling mud pools around the volcano.
- Volcano mountains can form when heat and pressure, miles below Earth's crust, form magma that rises through cracks in the crust.
- A violent explosion can cause a volcano to collapse and form a giant bowl-shaped area called a caldera.
- Each eruption spreads more lava and makes the mountain grow larger.
- Composite volcanoes form mountains with separate layers of lava, ash cinders, blocks, and bombs.

Methods & Results: Sort types

- ❑ The four types are examples of the types of sorts that can be used to rank sentences
- ❑ No means exhaustive

Proposed order of use



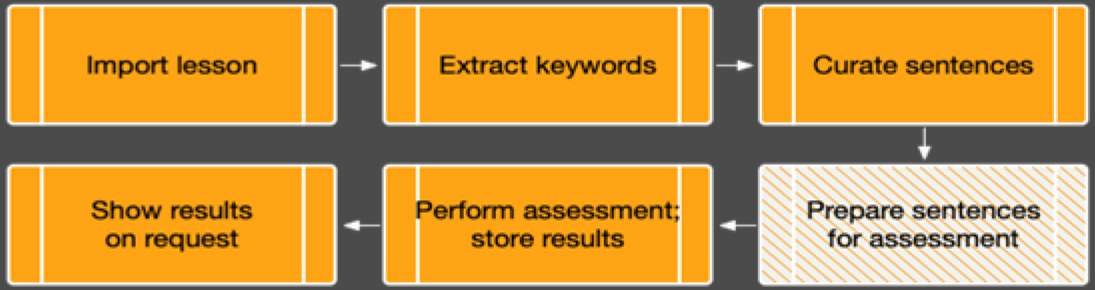
Sentence Length

Number of keywords

Number of tier 2 and tier 3 words

Syntactic dependency of keyword relative to the sentence

Step 4: Prepare Sentences for Assessment



Import lesson

Extract keywords

Curate sentences

Prepare sentences for assessment

Perform assessment; store results

Show results on request

Methods & Results: Prepare sentences for assessment

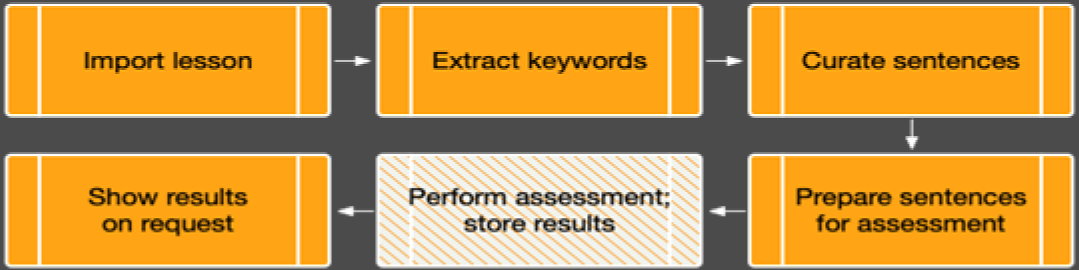
Assessment example

Student is asked to fill in the blanks using the words *‘volcano’* and *‘lava’*

Sentence presented to student is:
Cinder cone ——— form when exploding ——— hardens into glassy rock fragments.

- ❑ Use spaCy’s Matcher tool to find start indices of matched keywords
- ❑ Matcher > Regex as Matcher can find all forms of a word eg volcanoes will return as a match for the match word ‘volcano’.
- ❑ Tokens are then replaced by dashes and stored in the database

Step 5: Perform assessment and show results



Import lesson

Extract keywords

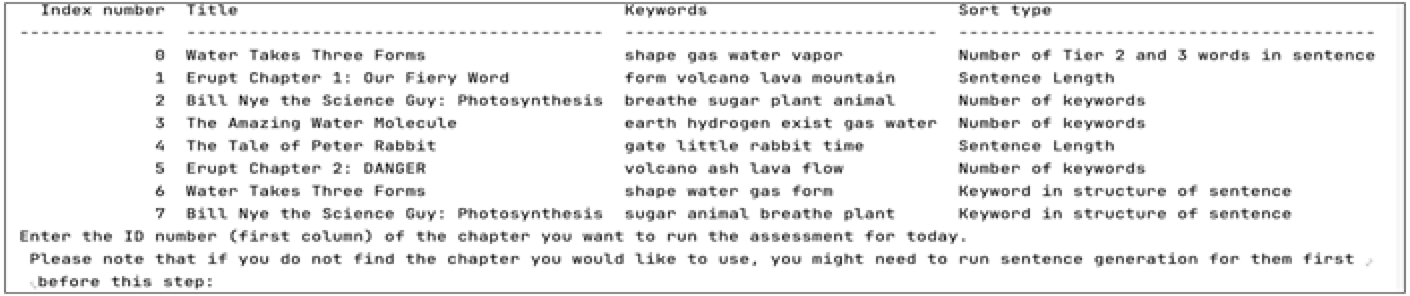
Curate sentences

Prepare sentences for assessment

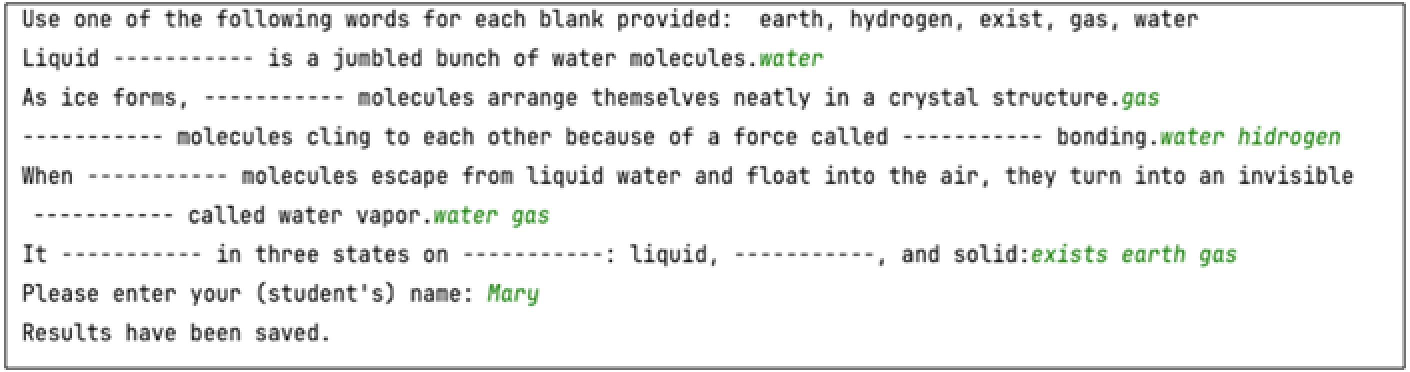
Perform assessment; store results

Show results on request

Methods & Results: Perform assessment



```
  Index number  Title                                     Keywords                          Sort type
--------------  ----------------------------------------  --------------------------------  ------------------------------------------
             0  Water Takes Three Forms                   shape gas water vapor             Number of Tier 2 and 3 words in sentence
             1  Erupt Chapter 1: Our Fiery Word           form volcano lava mountain        Sentence Length
             2  Bill Nye the Science Guy: Photosynthesis  breathe sugar plant animal        Number of keywords
             3  The Amazing Water Molecule                earth hydrogen exist gas water    Number of keywords
             4  The Tale of Peter Rabbit                  gate little rabbit time           Sentence Length
             5  Erupt Chapter 2: DANGER                   volcano ash lava flow             Number of keywords
             6  Water Takes Three Forms                   shape water gas form              Keyword in structure of sentence
             7  Bill Nye the Science Guy: Photosynthesis  sugar animal breathe plant        Keyword in structure of sentence
Enter the ID number (first column) of the chapter you want to run the assessment for today.
 Please note that if you do not find the chapter you would like to use, you might need to run sentence generation for them first
 before this step:
```



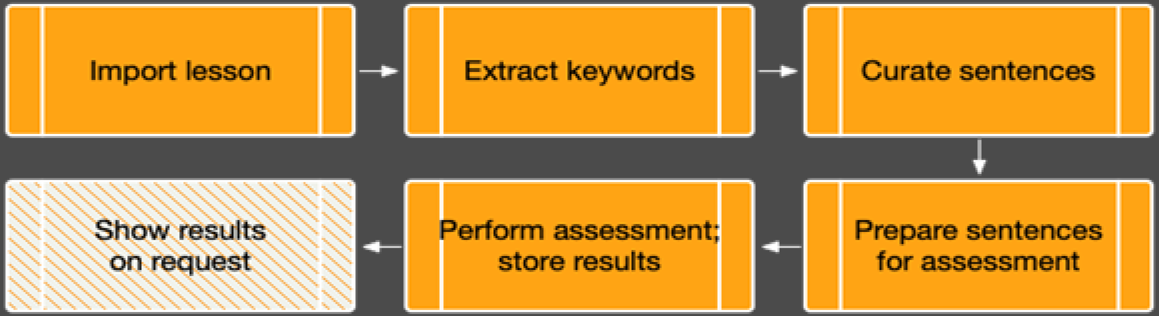
```
Use one of the following words for each blank provided:  earth, hydrogen, exist, gas, water
Liquid ----------- is a jumbled bunch of water molecules.water
As ice forms, ----------- molecules arrange themselves neatly in a crystal structure.gas
----------- molecules cling to each other because of a force called ----------- bonding.water hidrogen
When ----------- molecules escape from liquid water and float into the air, they turn into an invisible
 ----------- called water vapor.water gas
It ----------- in three states on -----------: liquid, -----------, and solid:exists earth gas
Please enter your (student's) name: Mary
Results have been saved.
```

Methods & Results: Perform assessment

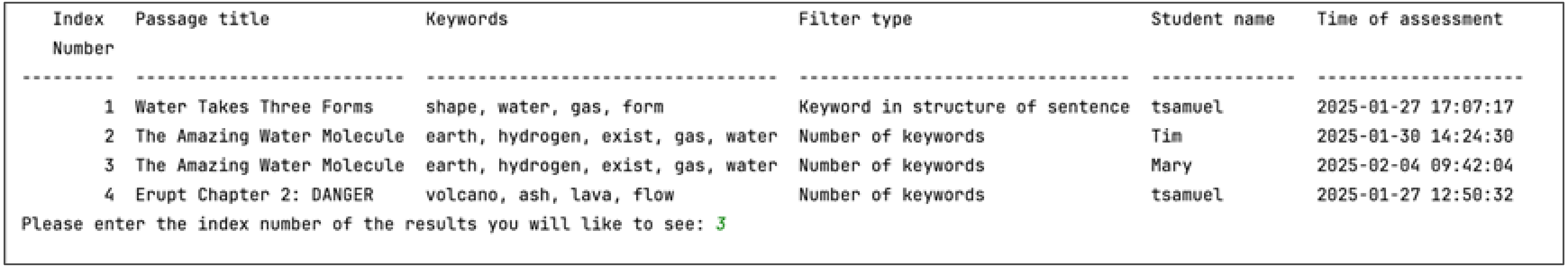
Determining correctness of response:

- First replaces the dashes version of sentence with responses
- Checks if sentence matches the original sentence - if yes, marks it as a correct response
- If there is no direct match:
    - SENCE goes through each response word and uses edit distance to check for typos comparing to expected keyword
    - If edit distance is two or less, it replaces the response word with the original keyword
    - This is done to account for typos
    - Sentence is checked again. If it is a match with the original sentence, the response is marked correct, else wrong.

Step 6: Show results

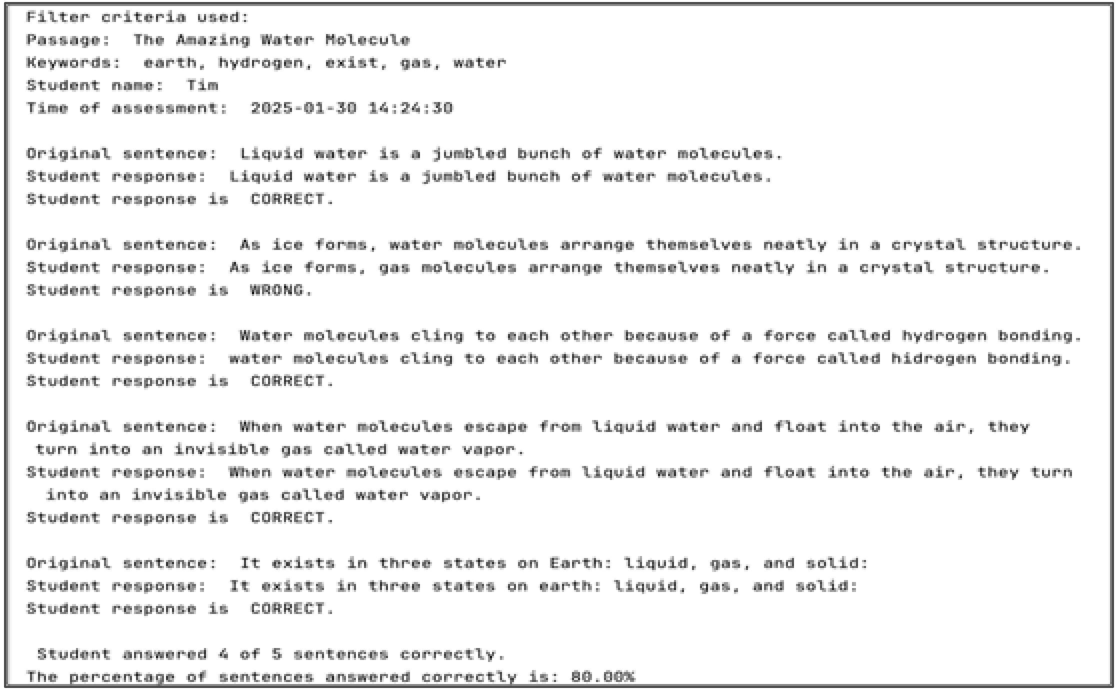


Methods & Results: Show results



```
   Index   Passage title              Keywords                            Filter type                       Student name    Time of assessment
   Number
---------  --------------------------  ----------------------------------  --------------------------------  --------------  --------------------
        1  Water Takes Three Forms     shape, water, gas, form             Keyword in structure of sentence  tsamuel         2025-01-27 17:07:17
        2  The Amazing Water Molecule  earth, hydrogen, exist, gas, water  Number of keywords                Tim             2025-01-30 14:24:30
        3  The Amazing Water Molecule  earth, hydrogen, exist, gas, water  Number of keywords                Mary            2025-02-04 09:42:04
        4  Erupt Chapter 2: DANGER     volcano, ash, lava, flow            Number of keywords                tsamuel         2025-01-27 12:50:32
Please enter the index number of the results you will like to see: 3
```

Methods & Results: Show results

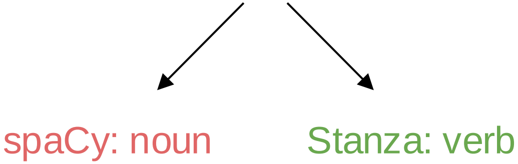


```
Filter criteria used:
Passage:  The Amazing Water Molecule
Keywords:  earth, hydrogen, exist, gas, water
Student name:  Tim
Time of assessment:  2025-01-30 14:24:30

Original sentence:  Liquid water is a jumbled bunch of water molecules.
Student response:  Liquid water is a jumbled bunch of water molecules.
Student response is  CORRECT.

Original sentence:  As ice forms, water molecules arrange themselves neatly in a crystal structure.
Student response:  As ice forms, gas molecules arrange themselves neatly in a crystal structure.
Student response is  WRONG.

Original sentence:  Water molecules cling to each other because of a force called hydrogen bonding.
Student response:  water molecules cling to each other because of a force called hidrogen bonding.
Student response is  CORRECT.

Original sentence:  When water molecules escape from liquid water and float into the air, they
 turn into an invisible gas called water vapor.
Student response:  When water molecules escape from liquid water and float into the air, they turn
  into an invisible gas called water vapor.
Student response is  CORRECT.

Original sentence:  It exists in three states on Earth: liquid, gas, and solid:
Student response:  It exists in three states on earth: liquid, gas, and solid:
Student response is  CORRECT.

 Student answered 4 of 5 sentences correctly.
The percentage of sentences answered correctly is: 80.00%
```

Conclusions and Broader Impacts

- Assessment of off-the-shelf NLP tools:
  - Many libraries available as commercial and open-sourced products
  - Tricky to determine which is best suited for a NLP task
  - Compute and storage costs need to be taken into consideration
  - Even while doing the same function, not all have similar results

Several times an hour, lava shoots out of a volcano in Italy called Stromboli



spaCy: noun

Stanza: verb

Conclusions

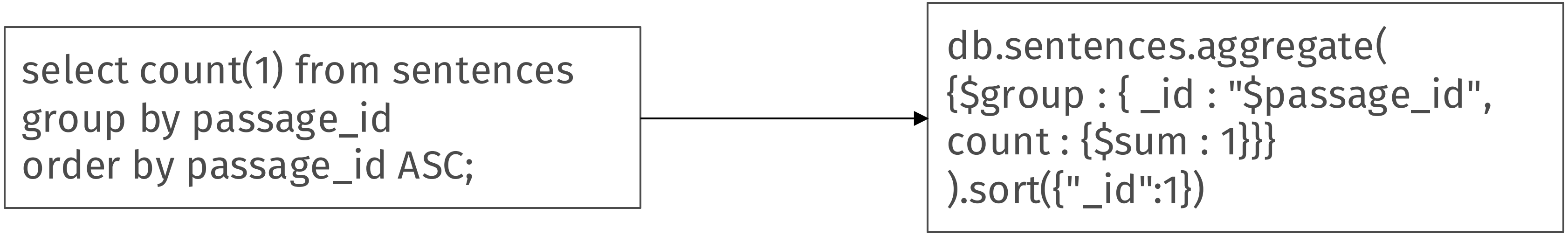
- Assessment of off-the-shelf NLP tools:
    - Different keyterm extraction algorithms produce different results - and have wide range of flexibility in fine-tuning
    - Comprehensive packages like spaCy and NLTK can do much of the common NLP tasks such as word and sentence tokenization, collect basic sentence metrics and finds parts of speech and lemmas.
    - Sometimes libraries treat the same process differently between multiple methods in the same library.
        - spaCy Word tokenizer treats Mr. as one token
        - spaCy Matcher treats Mr. as two tokens

Conclusions

- **Comparing keyterm extraction algorithms**
  - All need pre-processing such as removing stop words and lemmatizing
  - Some generate key phrases (3-4 words long) instead of key terms
  - Transformer based methods performed better than graph-only or statistical-only methods
- Comparing sentence sort types
  - Sentence length → Number of keywords in a sentence → Number of tier 2 and 3 words in a sentence → Syntactic dependency of keyword in sentence

Conclusions

- **NoSQL databases** are a completely different approach to storing data than relational databases
  - Easy to manipulate and retrieve text based information
  - Not meant for heavy in-database analysis



```
select count(1) from sentences
group by passage_id
order by passage_id ASC;
```

SQL

```
db.sentences.aggregate(
{$group : { _id : "$passage_id",
count : {$sum : 1}}}
).sort({"_id":1})
```

MongoDB

Conclusions

- **Volume of text needed for reasonable sentence retrieval:**
    - There needs to be a base minimum number of keywords specified to retrieve sufficient simple, medium and complex sentences
    - There also needs to be a limit on the number of keywords a student is asked to fill out per sentence
- Storybooks are not ideal candidates for sentence retrieval:
    - While storybooks have much larger lengths of text, there were not many sentences retrieved for top ten keywords.
        - Not many highly repeated words in story books compared to lesson-based texts

Broader Impacts

- ❑ SENCE provides a straightforward, user friendly interface to NLP applications for non NLP proficient users
- ❑ Unlike many other AI based applications, does not need an internet connection; the application can be run locally to a system. This has broad applications in rural and remote areas
- ❑ Open to community contributions since it is open-sourced

Future Work

- Large Language Models:
  - LLMs can provide more extensive outside corpora
  - Retrieval Augmented Generation can be used to teach students vocabulary comprehension based on synonyms, and associated vocabulary usage.
- Scaling:
  - Graphics work needed to provide a more captivating interface
  - Web development work needed to capture user and session information for a shared instructor version - where instructors can add, edit keywords and retrieved sentences and results are shared between instructors
- Multi-word keyterms