Note taking is an important practice during lectures. With the rapid development of electronic devices such as smart phones, tablets and laptops, digital note-taking has become an option for many students. In the classroom setting, note takers might fail to record clearly organized notes due to limitations of time and devices. Therefore, users often need to manually organize long paragraphs of notes by dividing them into appropriate sections based on content, or structuring them into lists and tables. The goal of this research is to develop a system that automatically structures unorganized notes. Using topic modeling and automatic summarization, we detect and analyze the “structure” of the notes. We then build a layout based on the structure and format notes into a more readable format.

- Raw text = unstructured texts that need to be organized.
- Users can type or record lectures to obtain raw texts.
- For this research, we use online course subtitles from Coursera as raw texts.¹

1. In the raw text, we categorize each word into different topics. A word belongs to topic -1 if it does not belong to any topic that is detected by our topic model.
   - By calculating each topics’ frequency of appearance, we determine the main topic over a fixed number of texts.
   - When the main topic shifts, we generate breaks that separate the text into topic segments.
   - For example:

   ![Graph showing topic categorization and frequency](image)

   Break text into topic segments

2. Sample Evaluation:
   - Approximately 121 words away from real breaks.
   - Video Segment
   - Generated Breaks
   - Good Breaks
   - Real Breaks
   - Hand-annotated
   - Precision = “No garbage?” = 0.704545
   - Recall = “Got everything?” = 0.574074

3. We summarize each topic segments into short sentences by using sentence extraction.⁴
   - Sentence extraction = identify the most salient sentences of a text.
   - Users can decide the length of the sentences to be extracted for each topic segments.

4. We format the notes into HTML files.
   - For example:

   ![Sample Evaluation](image)

   Main sentences are separated from topic segments for easy review.

Future Study
- Continue testing and evaluating the system.
- Improve precision and recall by adjusting system’s settings.
- Keep developing the user interface.

² Mallet, a Java-based statistical natural language processing package to generate topic models for raw text inputs.³
³ For example, when we run Mallet on an hour-long lecture subtitle, we detect five topics, each shown as a word cloud:
   - 0 processing android device
   - ve run ll running javascript java ios thing simple devices
   - screen ip address mode work
   - 1 width ve speed make symmetry line screen
   - ll idea brush point green map brushes
   - application divided distance mapping simple
   - 2 sound time ll things desktop week start app
   - bit yeah people lot show sounds music good ve kind make
   - 3 mouse color ve draw position drawing line colors red numbers program rectangle screen
   - green blue code basic point lines
   - 4 sound audio maxim play ve sketch file ll code folder data environment create beat device files store speed

⁴ Sentence extraction = identify the most salient sentences of a text.