What can be learned from Natural Language Processing of MOOCs?

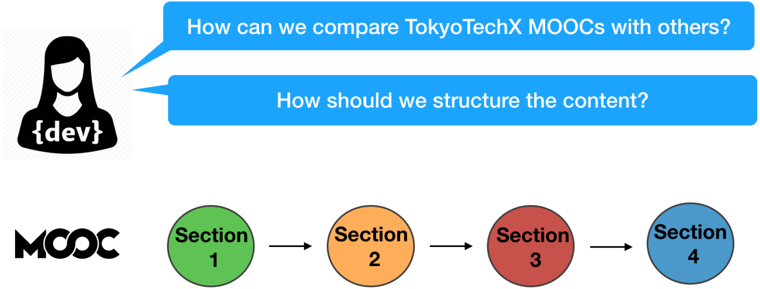
Zarina, Nopphon, Eric, Naoaki & Jeffrey

Online Education Development Office
Center for Innovative Teaching and Learning
Tokyo Institute of Technology

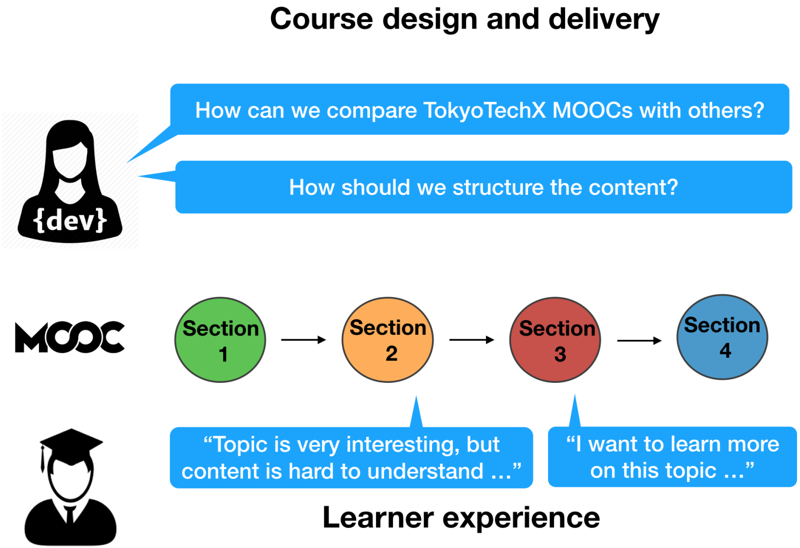
Tokyo Tech

Where did our work start from?

Course design and delivery



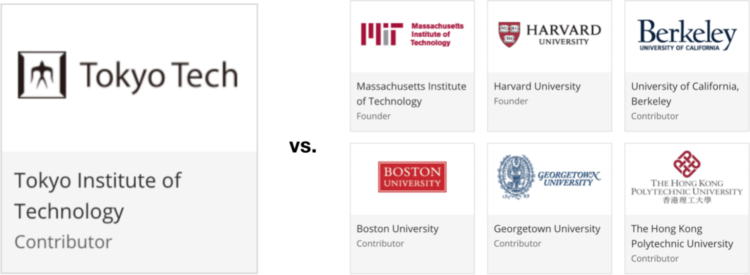
Where did our work start from?



Goal of analysis

To identify course design, delivery and content related elements that might improve MOOC quality and learner experience:

- Define metrics for comparing MOOCs' content



Outline

1. Current state of MOOC analysis using NLP
2. Tokyo Tech edX MOOC Crawler
3. Statistical analysis of crawled courses
4. NLP analysis using document embeddings

Research in MOOCs

Learner activity

- **Dropout prediction**
- **Adaptive real-time support**

Course content

- **Content classification**
- **Content matching**

NLP research in MOOCs

Natural Language Processing (NLP) is a branch of computer science and artificial intelligence that allows computers understand and interpret human language.

Learner activity

- **Dropout prediction**
- **Adaptive real-time support**

Course content

- **Content classification**
- **Content matching**

NLP analysis

Why NLP techniques are useful?

Word embedding

One hot vector representation

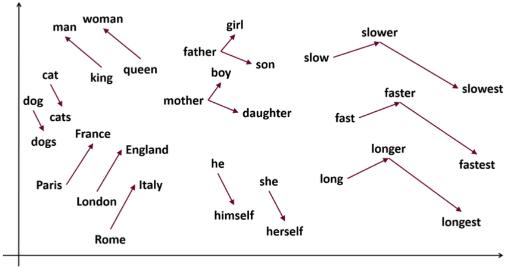
dog 0 1 0 0 0 0 0 0 0 0 0

cat 0 0 1 0 0 0 0 0 0 0 0

Word embedding

dog 0.3 0.3 0.1 0.5

cat 0.3 0.3 0.1 0.5



Why NLP techniques are useful?

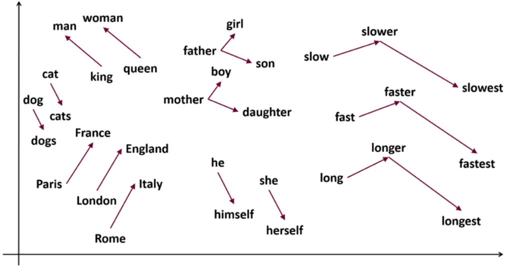
Word embedding

One hot vector representation

dog	0 1 0 0 0 0 0 0 0 0 0
cat	0 0 1 0 0 0 0 0 0 0 0

Word embedding

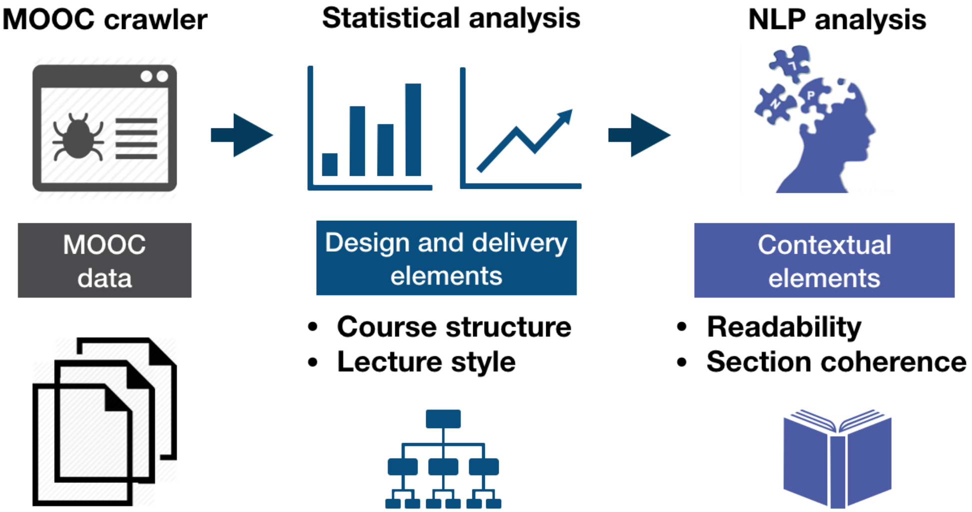
dog	0.3 0.3 0.1 0.5
cat	0.3 0.3 0.1 0.5



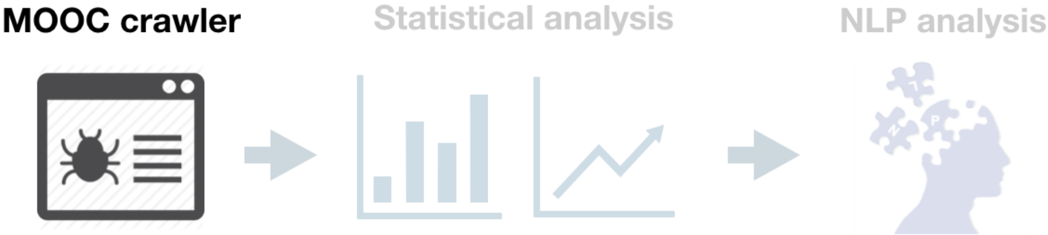
Advantages:

- Model captures semantic similarity
- Model is fast to train
- Human effort for training is minimal (unsupervised learning)

Analysis overview

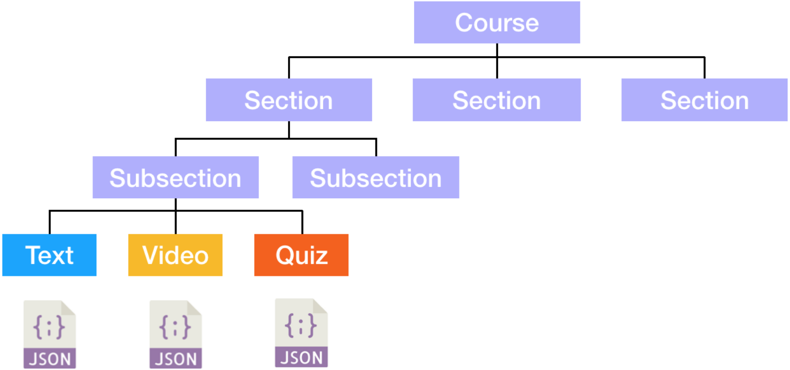


Outline



Tokyo Tech edX MOOC Crawler

- Python-based tool developed for mining text data of edX MOOCs on a user's dashboard



Tokyo Tech

Output examples of edX-crawler

Meta data for text component


```
text_block_01: {
        content: "Welcome to the Autophagy MOOC!
        section: "01-Introduction",
        subsection: "000-Welcome__Course_Navigation",
        unit_idx: "seq_contents_0.txt",
        word_count:  213
}
```


Meta data for video component

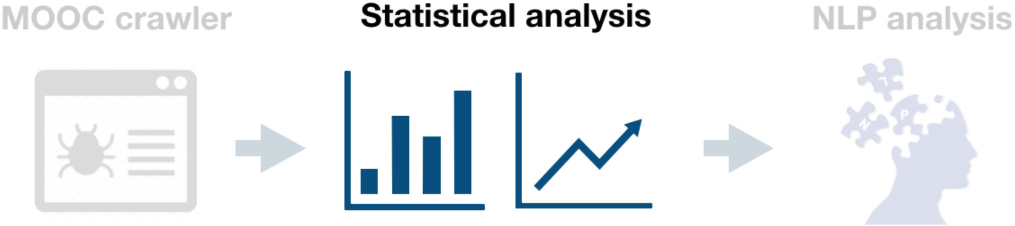

```
video_block_01: {
        section: "02-Week_1._Introduction_to_the_solid_Earth",
        subsection: "Introduction",
        transcript_en: "The name of this course is "Introduction to ...",
        unit_idx: "seq_contents_0",
        video_duration: 249,
        youtube_url: https://youtu.be/35g4lVKXx8I
}
```

Tokyo Tech edX MOOC Crawler

Check out our edX crawler tool available on gitHub:
`https://github.com/TokyoTechX/web-crawler`

We are looking forward for your feedback!

Outline



edX MOOCs vs TokyoTechX MOOCs

308 edX MOOCs

Language: English
Availability: Archived
Subject filters:

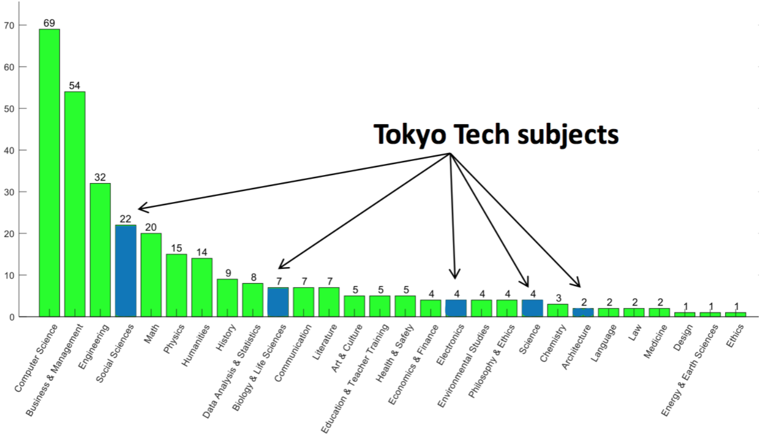
- Business & Management
- Computer Science
- Humanities
- Engineering
- Math
- Physics
- Social Sciences

TokyoTechX MOOCs

- 2 courses in English
  - Autophagy
  - Deep Earth Science
- 2 courses in Japanese & English
  - Intro to Electrical Engineering
  - Modern Japanese Architecture
- 1 course in Japanese
  - Science and Engineering Ethics

Distribution of subjects

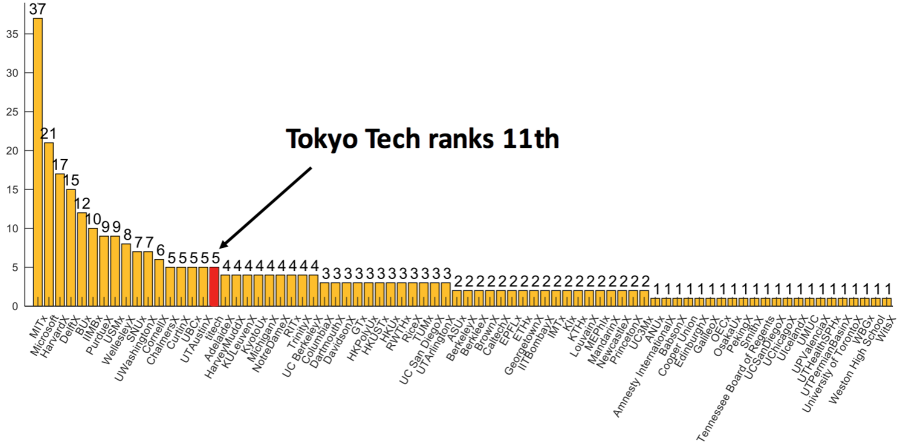
- 28 subjects in total
- Top 5 subjects (63%):
  - Computer science, Business and Management, Engineering, Social science, Math



Tokyo Tech

Distribution of institutions

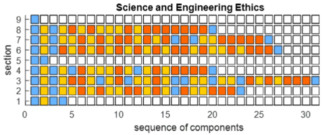
- 78 institutions in total
- Top 5 institutes (33%):
    - MIT, Microsoft, Harvard, Delft, IIMB



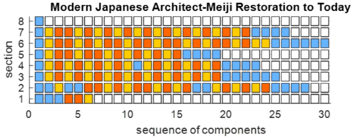
How is MOOC content structured?

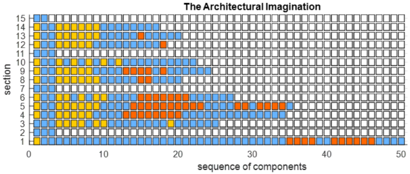
- We focused on 3 types of components
- Each course has different learning sequence
- How much content is in each component?

TokyoTechX



TokyoTechX



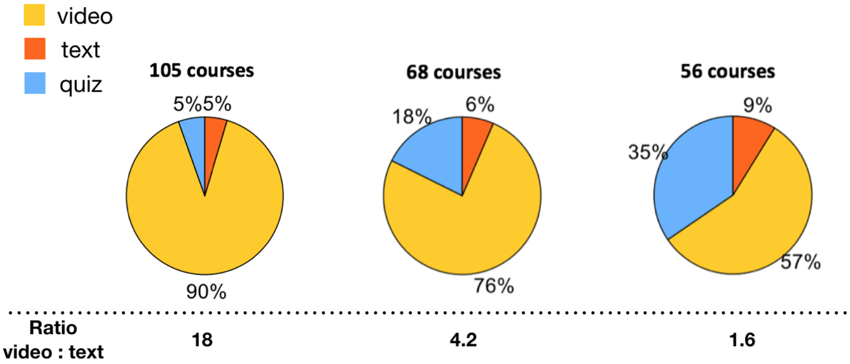


HarvardX

Tokyo Tech

Course content clustering - word count based

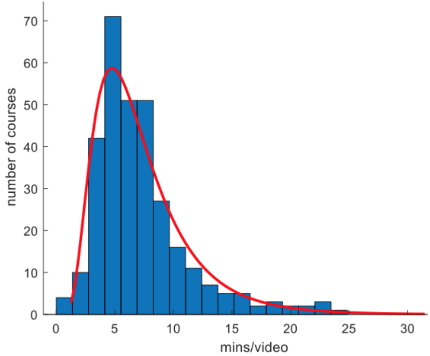
- About 75% of all courses falls into 3 clusters, which were computed using k-means clustering [2]
- 2 TokyoTechX courses in 1st cluster (Autophagy, Japanese Architecture)



Tokyo Tech

Video lecture duration

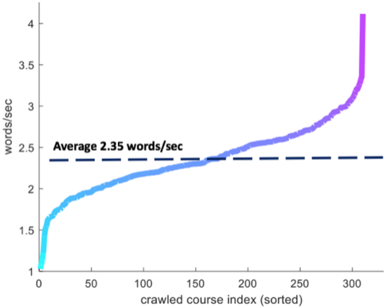
On average, video duration range is 3.3 - 9.1 minutes



TokyoTechX course	Average video duration (mins/video)
Deep Earth Science	4
Autophagy	4.5
Science and Engineering Ethics	6.6
Intro to Electrical Engineering	8.6
Modern Japanese Architect	13

Tokyo Tech

Speaking rate of video lecturers

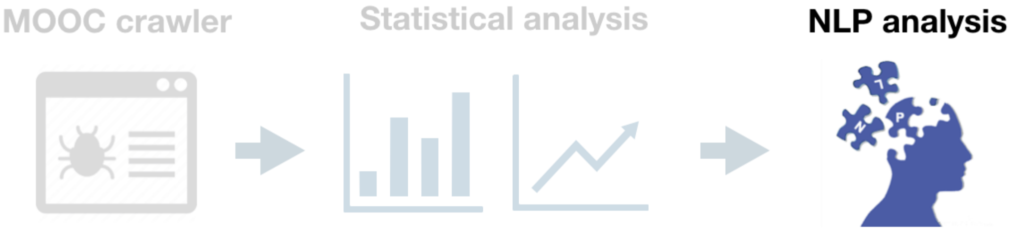


Course	Speaking rate (words/sec)
Deep Earth Science	1.8
Autophagy	1.68
Modern Japanese Architect	1.88

Figure: Speaking rate

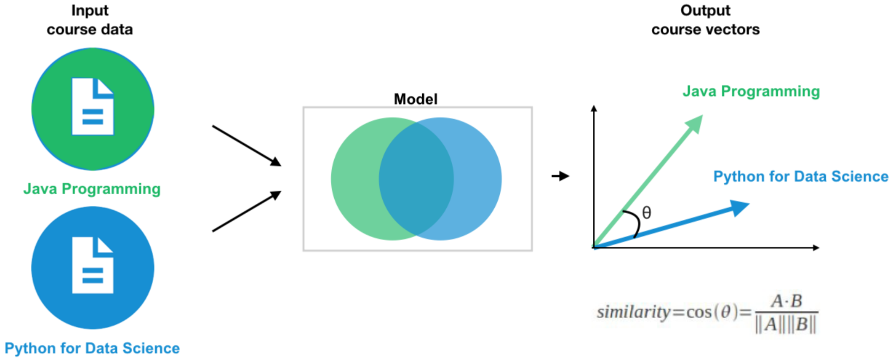
- **Fastest speaking** lecturer in Introduction to Public Speaking (4.1 words/sec)
- **Slowest speaking** lecturer in More Fun with Prime Numbers (1.03 words/sec)

Outline



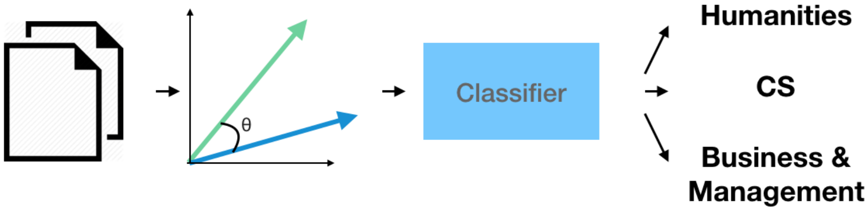
Vector representation of documents

- We need a measure to compare courses with each other
- Doc2vec [3] allows to represent text document as vectors and maps similar documents closer in a vector space



Course classification using document embeddings

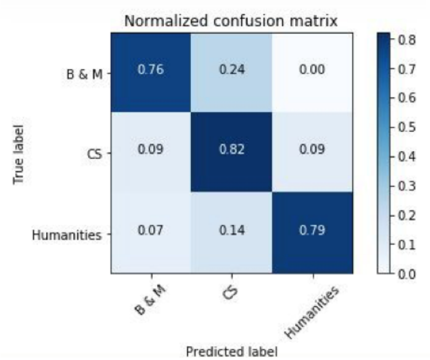
- How accurately can we classify courses into categories/subjects?



Tokyo Tech

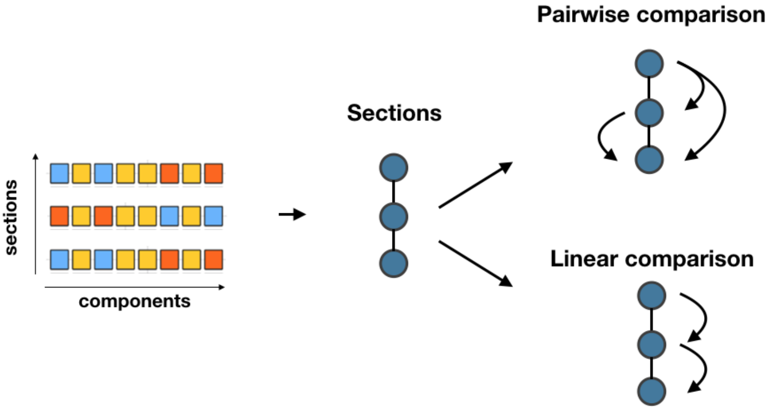
Classification results

- Linear classifier with SGD training
- Accuracy is 80%



- Can we capture similarity between course sections?
- How can we apply it to extend MOOCs readability analysis?

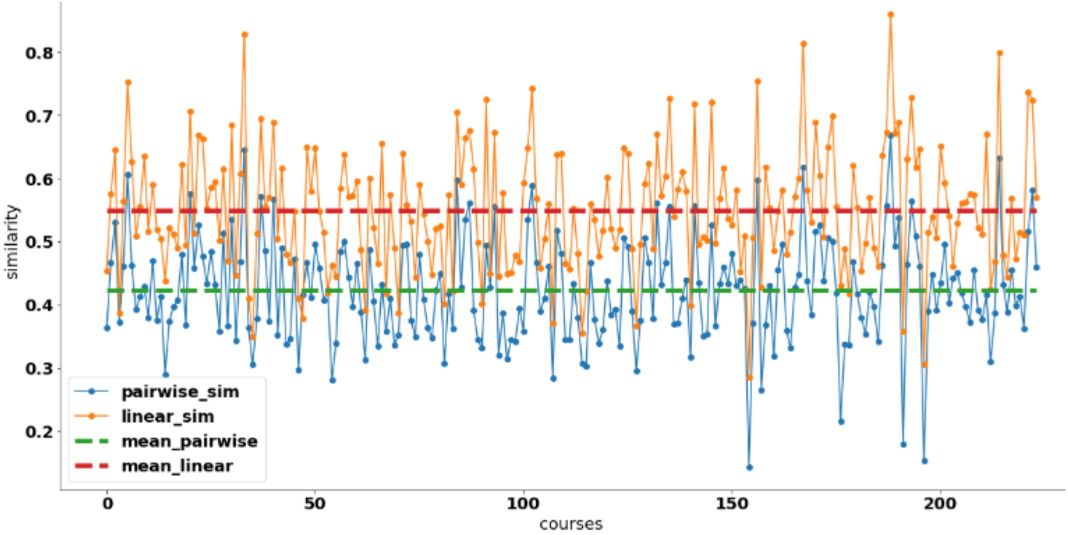
Section comparisons using embeddings



$$score = \frac{\sum_{i \neq j}^{N} s(s_i, s_j)}{\binom{N}{2}}$$

$$score = \frac{\sum_{i=1, j=i+1}^{N-1} s(s_i, s_j)}{N-1}$$

Pairwise cosine similarity vs Linear cosine similarity

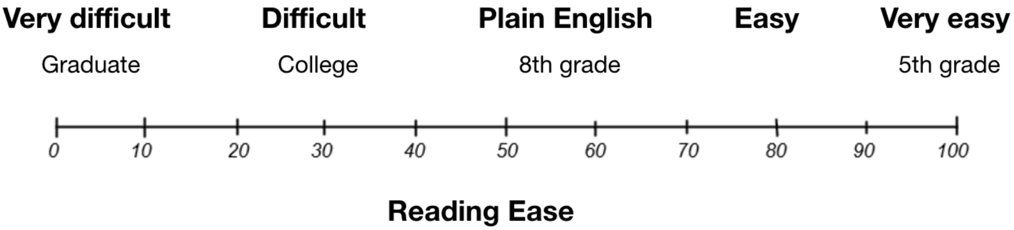


Readability and content flow of the course

Measure content flow and readability using two parameters:

- Flesch-Kincaid reading ease [4]

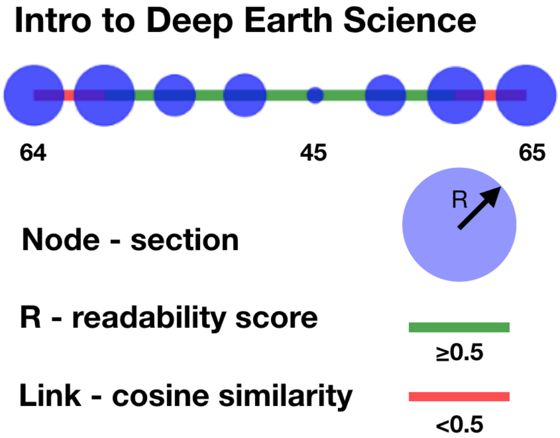
$$\textbf{score} = 206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$



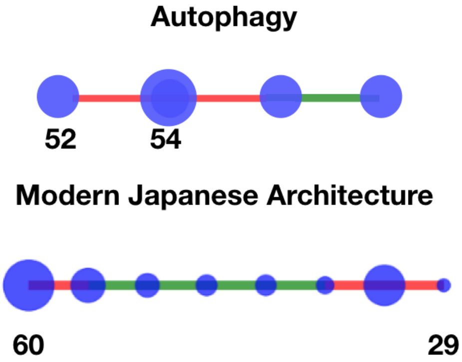
Readability and content flow of the course

Measure content flow and readability using two parameters:

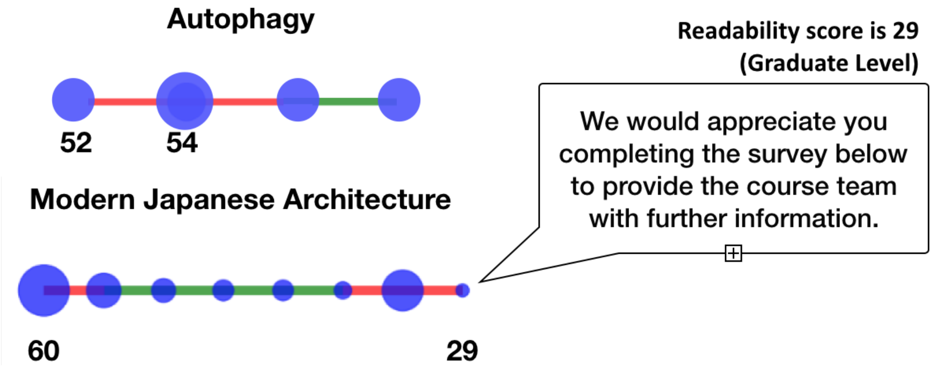
- Flesch-Kincaid reading ease
- Cosine similarity between sections



Readability and content flow of the course

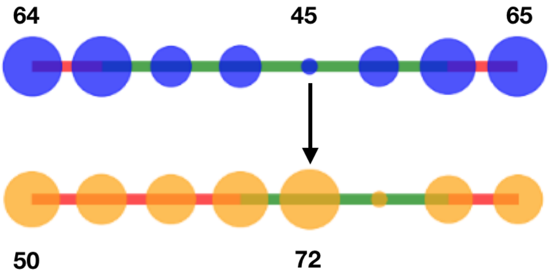


Readability and content flow of the course



Application

- **Provide a feedback on the content at the development stage**
  Identifying low readability score sections
- **Efficient learning**
  Finding similar section with higher readability score using document vectors



Tokyo Tech

Implications

- Serial MOOCs creation process: Develop, Run & Analysis
- Analysis can be done during Develop stage

Conclusion

The purpose of analysis was to identify features for comparing Tokyo Tech MOOCs with other MOOCs.

We learned:

- Most of the edX MOOCs are video-based
- Readability analysis can be useful for developing cohesive and learner-friendly content
- Combination of the MOOC features can be applied to predict course popularity

Future work

- Continue work on MOOC evaluation and data analysis
  Present at JSET conference in Japan in September 2018
- Welcome collaborations on MOOC content analysis
- See Github for our tools:
  - https://github.com/TokyoTechX

References

- Z. A. Pardos, S. Tang, D. Davis, and C. V. Le, "Enabling real-time adaptivity in moocs with a personalized next-step recommendation framework," in *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, pp. 23–32, ACM, 2017.
- D.-J. Kim, Y.-W. Park, and D.-J. PARK, "A novel validity index for determination of the optimal number of clusters," *IEICE Transactions on Information and Systems*, vol. 84, no. 2, pp. 281–285, 2001.
- Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International Conference on Machine Learning*, pp. 1188–1196, 2014.
- J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel," tech. rep., Naval Technical Training Command Millington TN Research Branch, 1975.