Analyzing teacher's speech using topics inferred by unsupervised modeling from textbooks

Using textbooks to analyze classroom sessions

03-10-2018

Hello!

I am Catalina Espinoza

I am a research assistant at the Centro de Investigación Avanzada en Educación (**CIAE**).

Universidad de Chile.

You can reach me at catalina.espinoza@ciae.uchile.cl

Smart Speech Project

Centro de Investigación Avanzada en Educación
Universidad de Chile

JYVÄSKYLÄN YLIOPISTO

TAMPEREEN YLIOPISTO

Chilean collaborators:

- *Roberto Araya*
- *Daniela Caballero*
- *Raúl Gormáz*
- *Abelino Jiménez*
- *Catalina Espinoza*

CONICYT
Ministerio de Educación
Gobierno de Chile

Finnish collaborators:

- *Jouni Viiri*
- *Sami Lehesvuori*
- *Toni Pikkarainen*

SUOMEN AKATEMIA

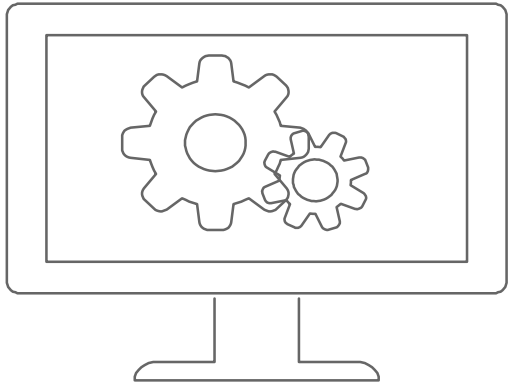
Smart Speech App

Teacher Module

Observer Module

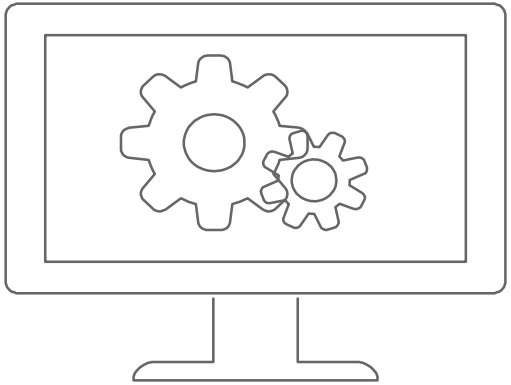
Smart Speech App - Teacher module

The teacher can record the audio of a session, and upload it to the web platform

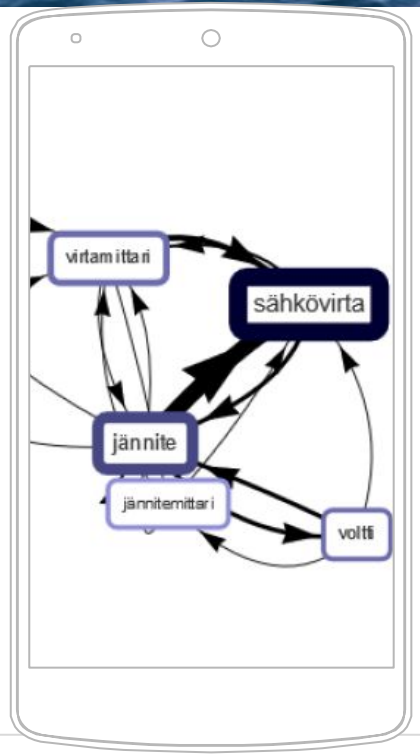


The audio record is transcribed using an Automatic Speech Recognition service

Smart Speech App – Teacher module

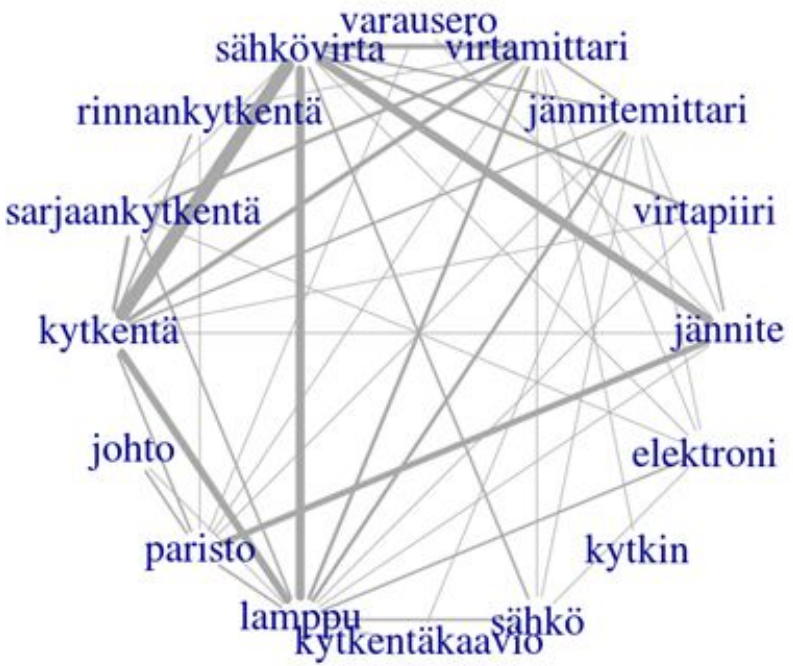


Once the transcription is complete, we can visualize the main features of the lesson



The teacher gets feedback of the lesson

Concept networks visualizations



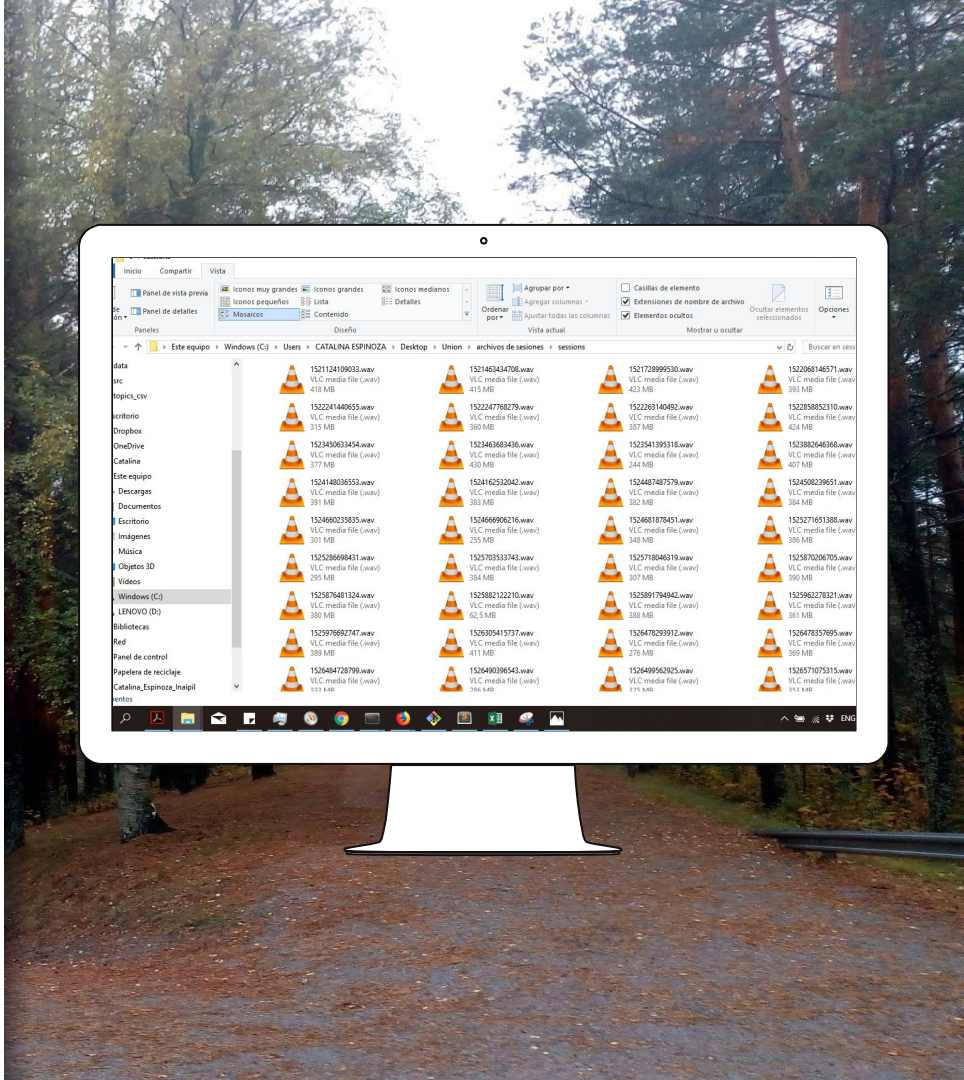
References

Helaakoski, J., & Viiri, J. (2011). A graph-theoretic perspective on the content structure of physics lessons and its relation to student learning gains. *Oppiminen, opetus ja opettajaksi kasvu ainedidaktisen tutkimuksen valossa*, 55.

Caballero, D., Araya, R., Kronholm, H., Viiri, J., Mansikkaniemi, A., Lehesvuori, S., ... & Kurimo, M. (2017, September). ASR in classroom today: automatic visualization of conceptual network in science classrooms. In *European Conference on Technology Enhanced Learning* (pp. 541-544). Springer, Cham.

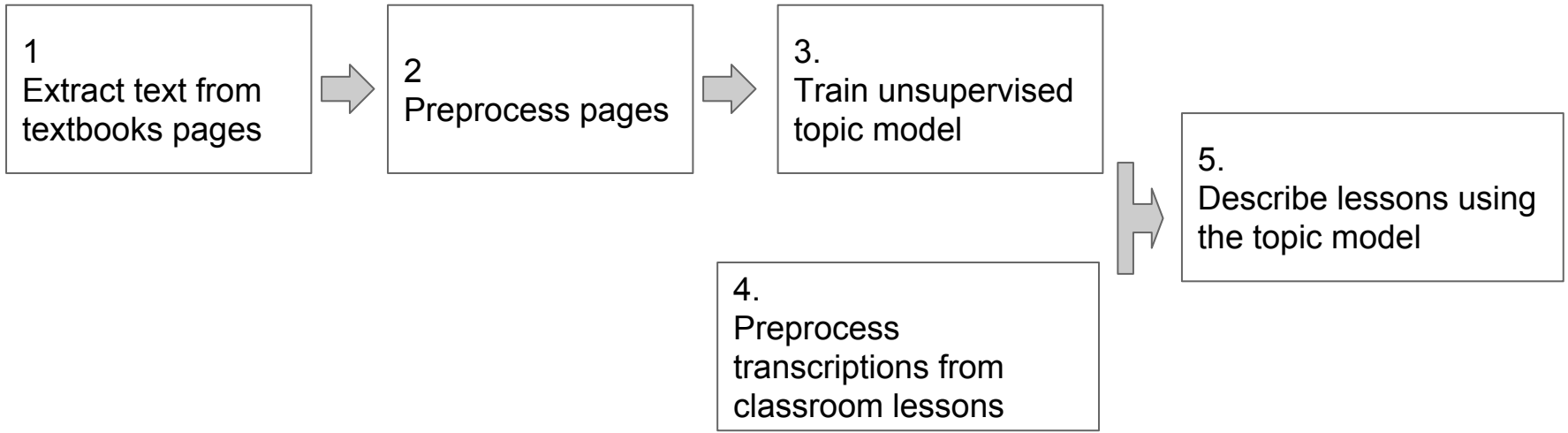
We are gathering transcriptions from classroom sessions

And we are looking for ways to analyze them automatically



From textbooks to lessons

General view of the pipeline



THE PIPELINE

1. Extract text from textbooks pages

- We collected 8 school textbooks (content and exercises) in different formats (pdf, word, images)
- We extracted the text from each page (~ 2000 pages)
- We considered each page as an individual document

2. Preprocess pages

- General cleaning
- Remove stop words
- Detect names
- Replace numbers with a tag NUMBER
- Remove and replace symbols
- Stemming (?)

3. Train unsupervised topic model (LDA)

- Latent Dirichlet Allocation (LDA) model (Blei, Ng and Jordan, 2003)
- LDA infers topics from a corpus of documents
- Topics are groups of words that occur together

Assumptions:

- Topics are probability distributions over a dictionary
- Documents are bags of words
- Documents are described by a mix of topics

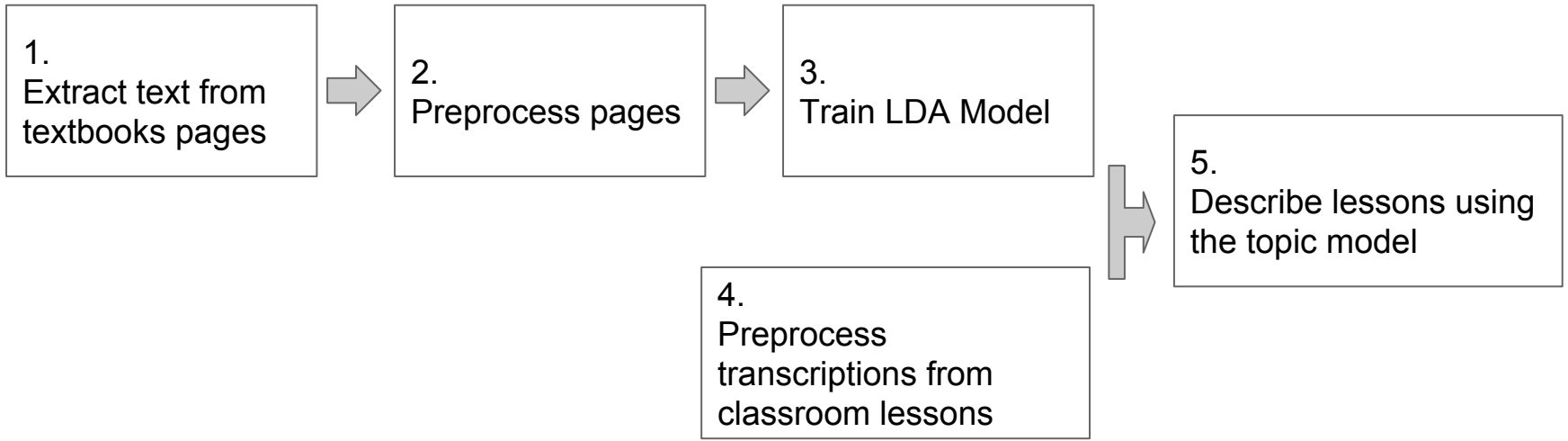
30 topics LDA model trained with finnish textbooks

Example of top-10 words from 5 topics

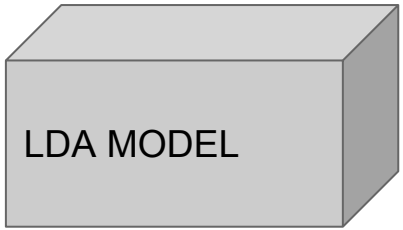
Topic 2	Topic 3	Topic 11	Topic 13	Topic 29
nopeus	jäsen	painopiste	jännite	valo
aika	lukujonon	g	pariston	valon
matka	kirjoita	tasapainossa	virtapiirissä	valoa
kiihtyvyys	suhde	kappale	sarjaan	linssin
kuvaaja	jäsentä	tarvitaan	sähkövirtaa	linssi
suora	tilavuudesta	laatikkoa	kytketään	kuva
koordinaatistossa	riipu	pysyy	virtapiirin	kupera
tasaisesti	Aritmeettinen	asteista	virtapiiri	heijastuu
Ilmanvastus	kokoa	jarrutusmatka	käämin	kuvan
kerroin	tulostaa	kappaleen	napojen	peili

4. Preprocess transcriptions from classrooms sessions

- General cleaning
- Remove stop words
- Detect names
- Replace numbers with a tag NUMBER
- Remove and replace symbols
- Stemming (?)

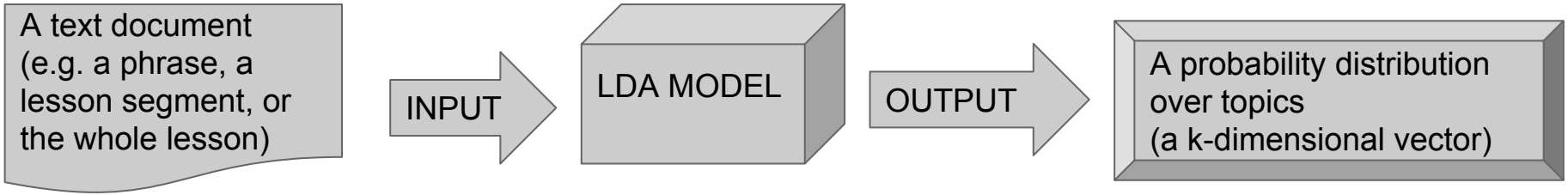


THE PIPELINE



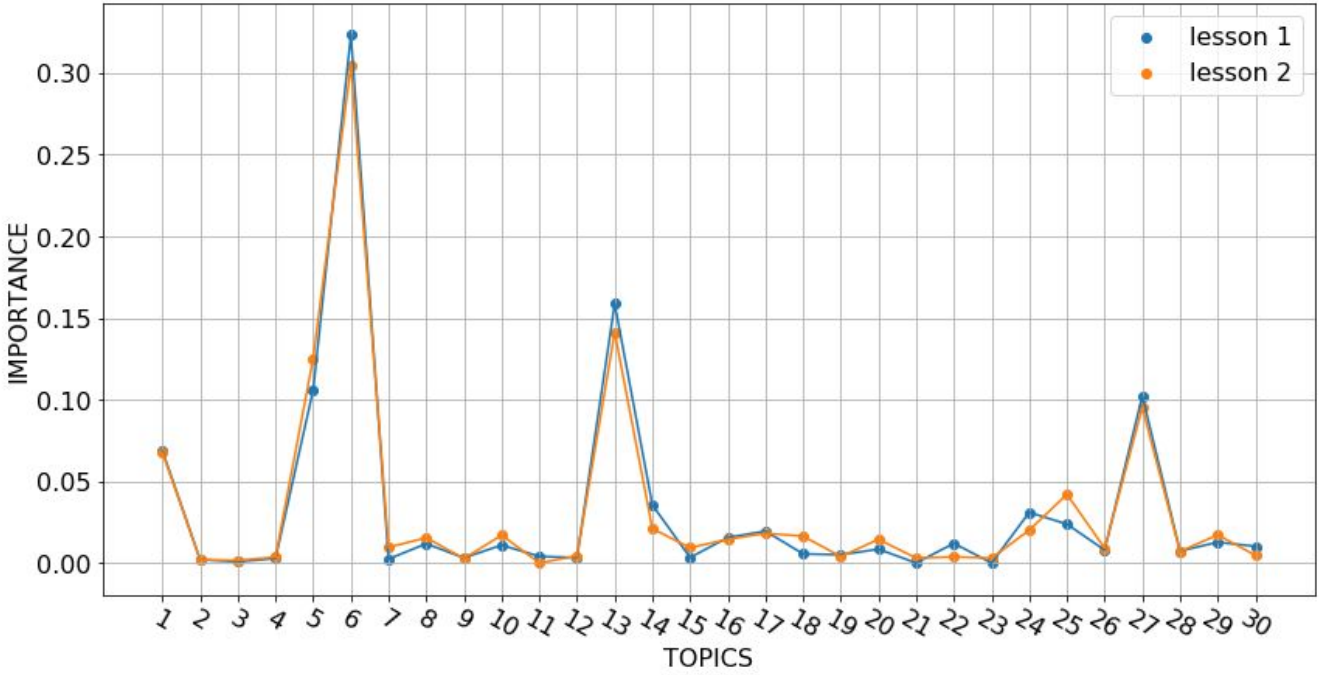
Available methods in **gensim** library (https://radimrehurek.com/gensim/)

- get_document_topics(*bow*, *minimum_probability=None*, *minimum_phi_value=None*, *per_word_topics=False*)
- ...



What can we do with the LDA model?

5. Describe sessions using the topic model



Words that activate topics in lesson 1

For topic 5

lampun
eli
sähkövirta
aika
kuinka
suuri
lamppu

For topic 6

eli
jo
esimerkiksi
paljon
usein
hyvin
joten
vastaavasti

For topic 13

jännite
pariston
kytketty
sarjaan
rinnan
jännitemittari
kaksi
Virtapiirissä
virtamittari
jännitettä
kytketään
virtapiiri
napojen
paristo
kytkentäkaavio
sähkövirtaa
paristot
jännitteen

For Topic 27

tee
voit
tarkoittaa

Words that activate topics in lesson 2

For topic 5

sähkövirta
aika
lampun
eli
suuri
lamppu
kuinka
yksikkö

For topic 6

eli
sähkö
enemmän
jo
paljon
hyvin
elektroneja
vastaavasti
jolloin
fysiikan
lisäksi
sisällä

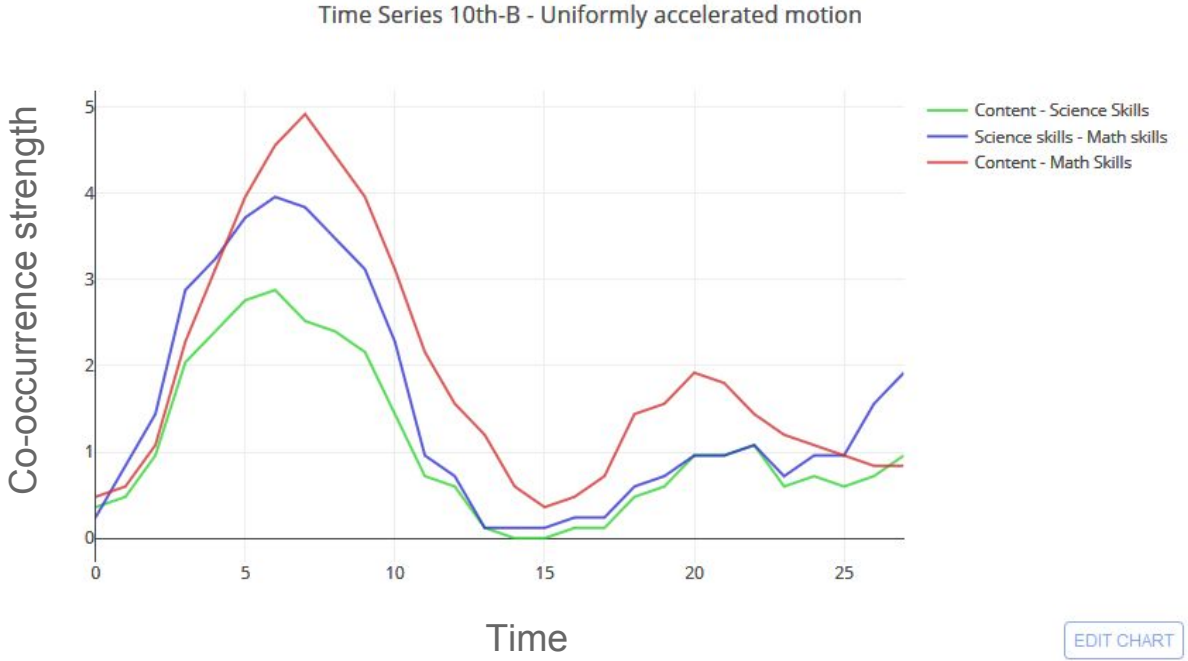
For topic 13

jännite
sarjaan
rinnan
pariston
paristo
sähkövirtaa
kytketään
kytketty
Virtapiirissä
jännitteen
jännitettä
paristot
kaksi
kytkentäkaavio
hehkulamppu

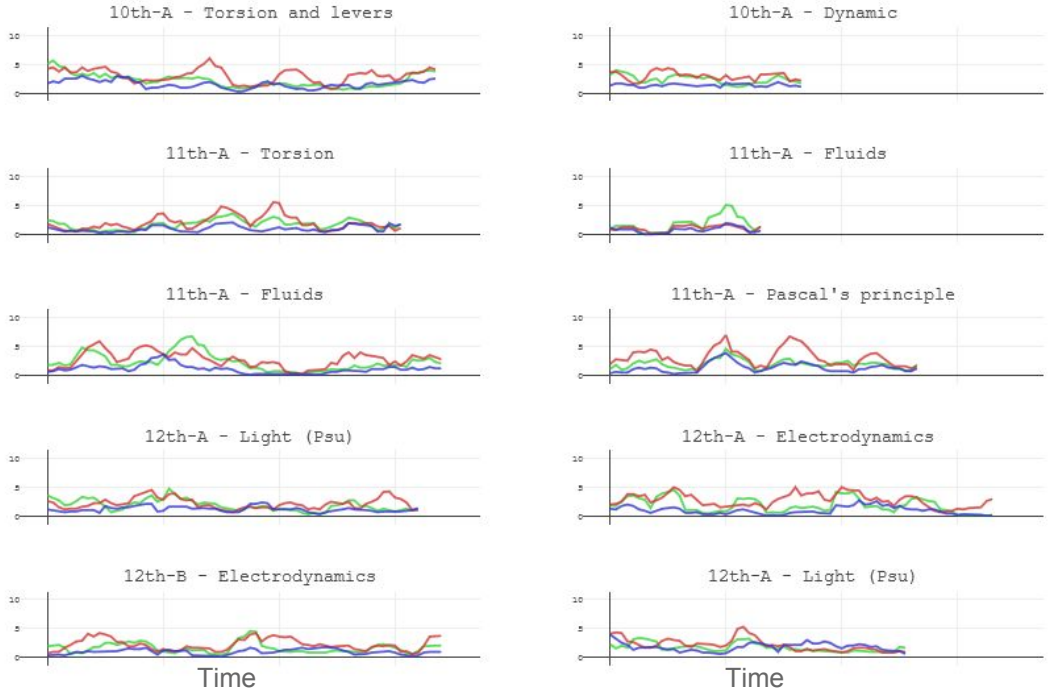
For Topic 27

voit
suuntaan
tarkoittaa

Temporal topic co-occurrence (Chilean lessons)

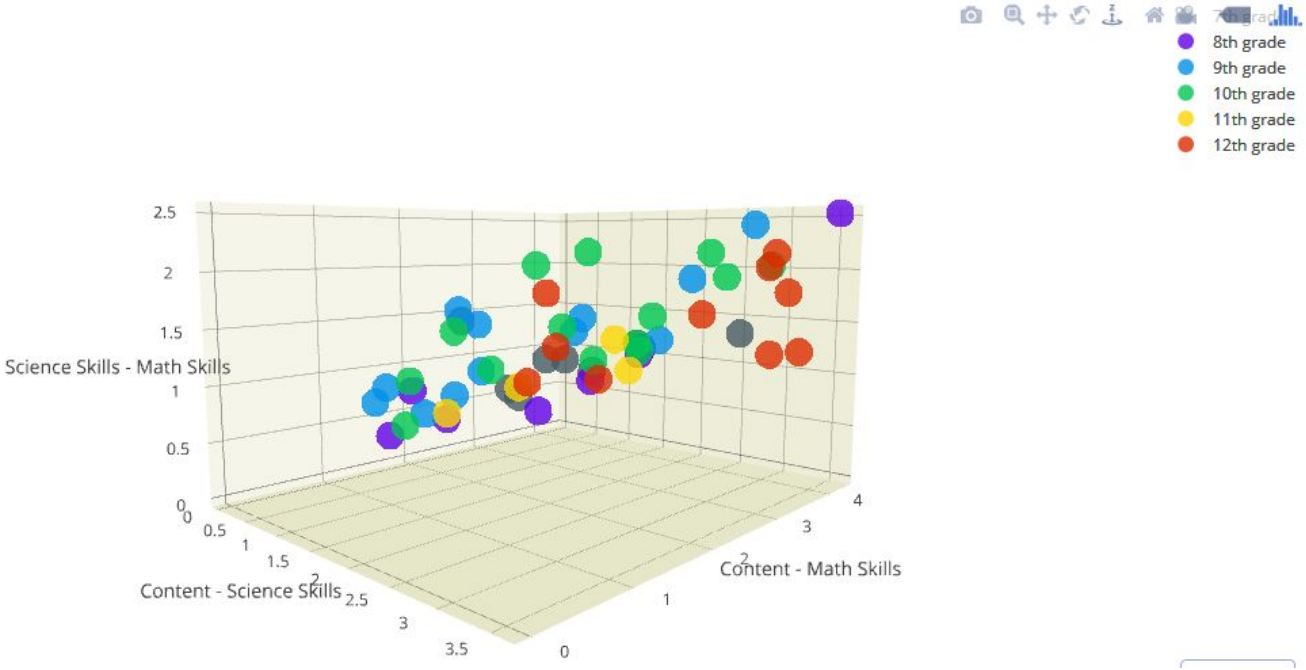


Temporal topic co-occurrence in different sessions



Content - Science Skills
Content - Math Skills
Science skills - Math skills

Defining a space to compare different sessions



In summary:

- We are describing classroom sessions using topics extracted from textbooks
- We want to be able to compare sessions from different content, levels, teachers, and countries.
- We are looking for visualizations that are useful to teachers and researchers

Thanks!

Any questions?

You can reach me at:

- catalina.espinoza@ciae.uchile.cl

Credits

Special thanks to all the people who made and released these awesome resources for free:

- Presentation template by SlidesCarnival
- Photographs by Unsplash and myself

Thanks!

Any questions?

You can reach me at:

- catalina.espinoza@ciae.uchile.cl