Topic Modeling Manual

May - August 2022
Intern: Byron Chung (Computer Science - Western University)
Research Assistant: Yaguang Li (Computer Science - Western University)
Volunteer: Andy Liang (McMaster University)
Volunteer: Annie Cho (University of Toronto)
Principal Investigator/Supervisor: Dr. Mi Song Kim (Faculty of Education - Western University)

This study is funded by the PI's research project named "**a topic modeling approach TowaRds Automated aNalySis of research Methods In Teacher daTa Literacies" ('TRANSMITTAL')**.

Drawing upon our previous topic modeling result (PI: Mi Song Kim), this time we have developed an app to conduct the following THREE Steps.

There are 3 different data sources (see Table 1) for using our App: journal articles (PDF files), news articles, and Learning Management System (LMS) (e.g., Sakai at OWL) data. Journal articles are detailed full-text articles which can be found in numerous databases. News articles can be found using the news API function in the topic modeling app when given search terms. Data from LMS are CSV files that contain the abstract of a journal paper which can be found in several databases.

Table 1: Overview of topic modeling using different data sources

	Cse 1	Case 2	Case 3
S1 (Data collection)	Journal articles from database (see Examples 1 & 2 below)	News articles	LMS
S2 (Data Preprocessing and Data Analyzation)	Any duplicates and outliers found in the journal articles will be removed and the number of topics shown can be selected by the user.	The duplicates and outliers from similar news articles will be removed and the number of topics shown can be selected by the user.	Duplicates and outliers from the CSV files will be removed and the number of topics shown can be selected by the user.
S3 (Visualization)	Visualizations are shown in this step.	Visualizations are shown in this step.	Visualizations are shown in this step.

STEP 1: Data Collection

1) Upload a PDF file/CSV file or use the news API to find any relevant keywords.
    a) To choose PDF or CSV files the user will have to download the file from an external database and upload the file through the application
    b) To use the news API, the user needs to input a keyword through our application and use the articles received.
2) Select the database you want to search the data from
3) Type in the keywords you wish to search in the 'Search Data' tab
4) Select a year for 'Year from' and 'Year to' for the articles
5) Select the type of data
6) Select the language
7) If there are other information necessary type in 'Other'

STEP 2: Data Preprocessing and Data Analyzation

1) Select 'Skip Duplicates' if you wish to remove articles that are either identical or related to the same study.
2) Select 'Remove Outliers' if you wish to remove articles that are irrelevant or very dissimilar to the keywords.
3) Select the number of topics you wish to generate with the keywords for the application.

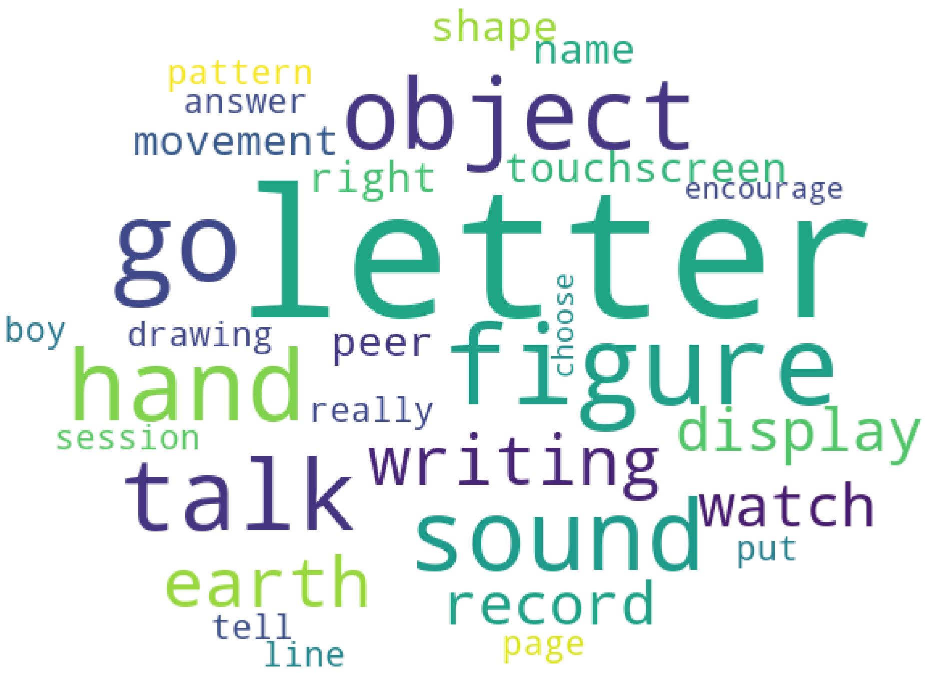
STEP 3: Data Visualization

1) Visualization results will be shown here. You can select different types of visualizations from LDA model, topic overview, Adjacency matrix, word topic frequency, topic network, and word cloud.
2) You are able to download the image results

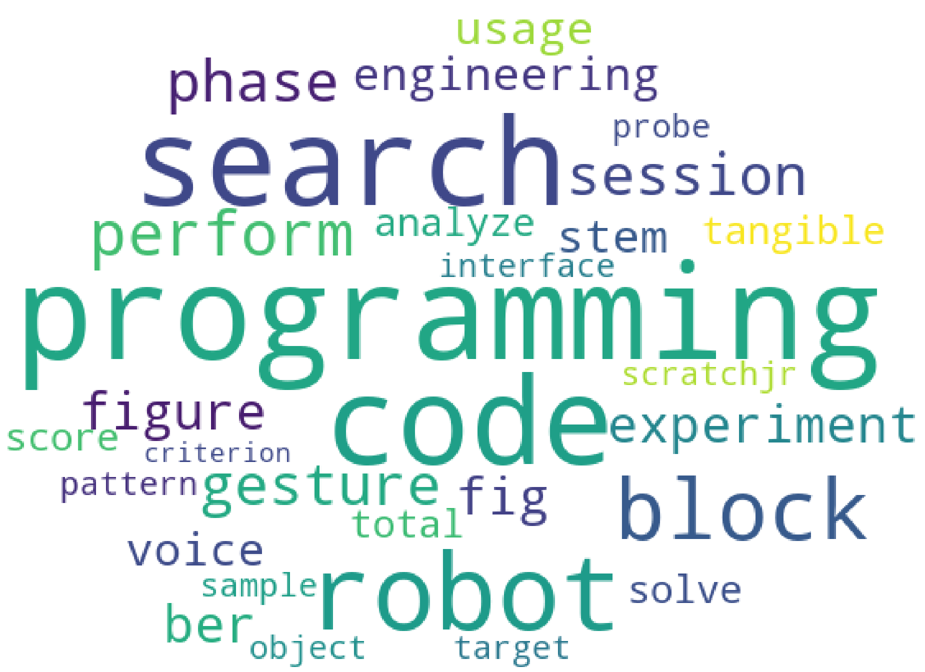
Examples: To display the use of topic modeling we can run the app on several examples.

Example 1

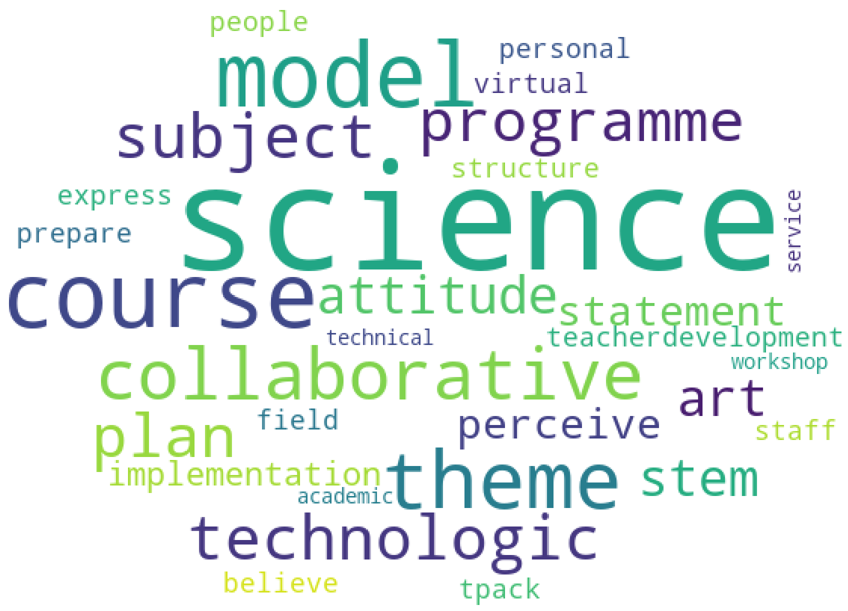
For the first example, the search terms used for the topic modeling app are touch screens, early literacy, early learning, and technology integration. Results from the full-text files retrieved from the Western Libraries choosing only peer-reviewed, English articles, and articles published after 2008 can be seen below resulting in 9 different wordclouds and 6 different key topics from 200 different articles:



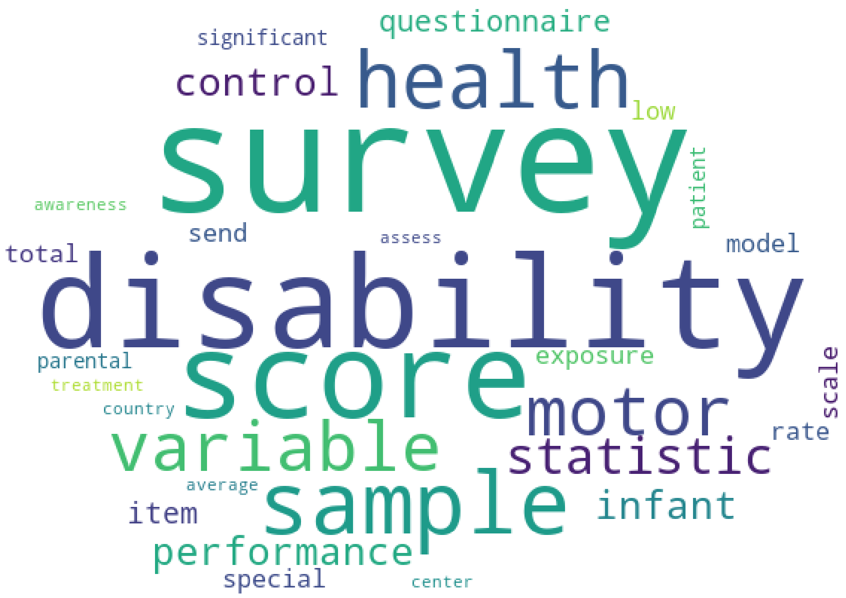
WordCloud 1



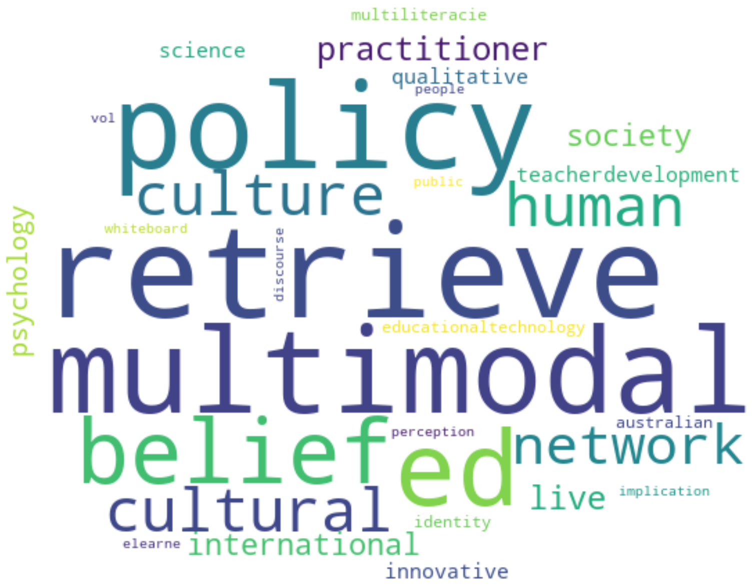
WordCloud 2



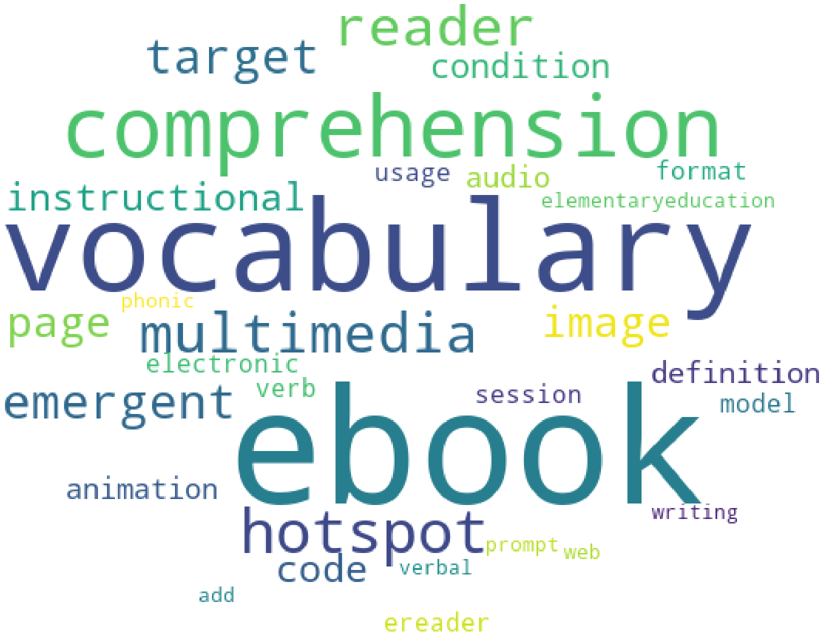
WordCloud 3



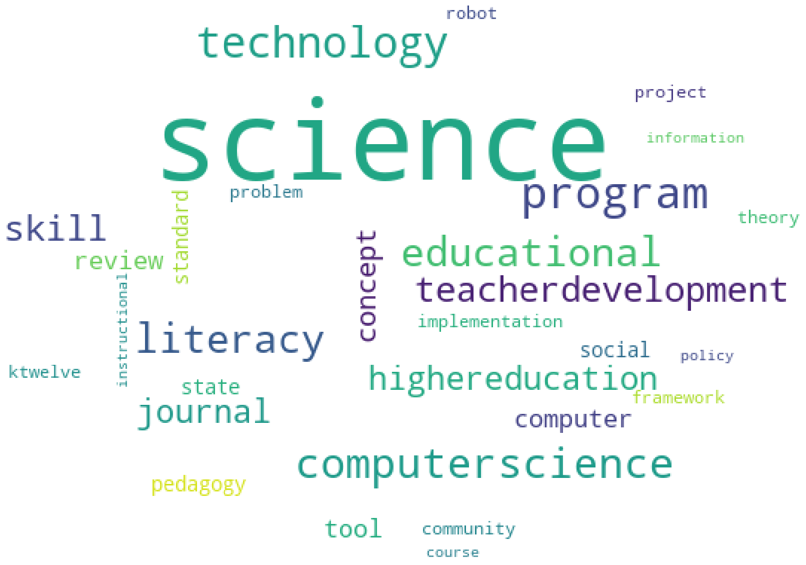
WordCloud 4



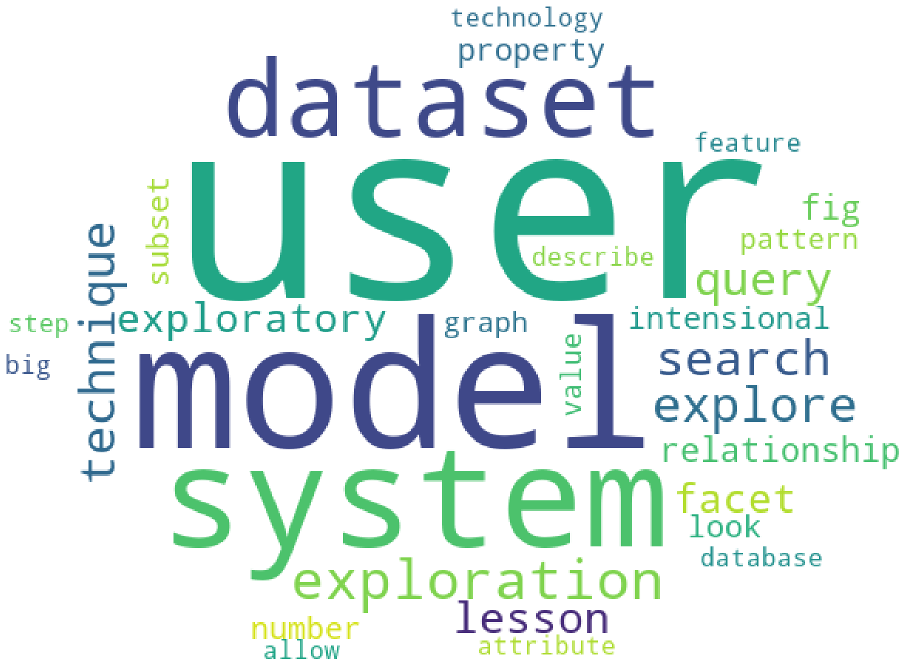
WordCloud 5



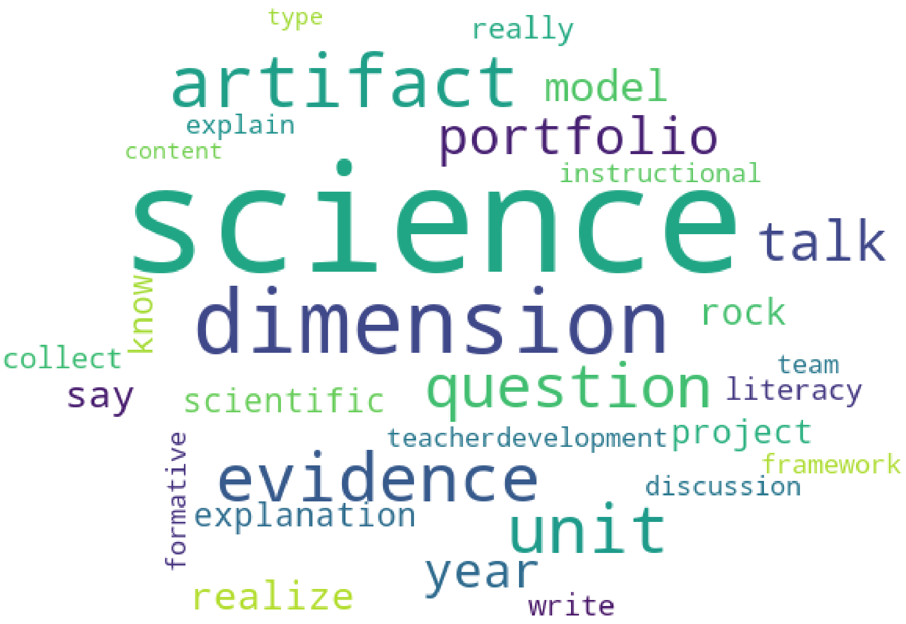
WordCloud 6



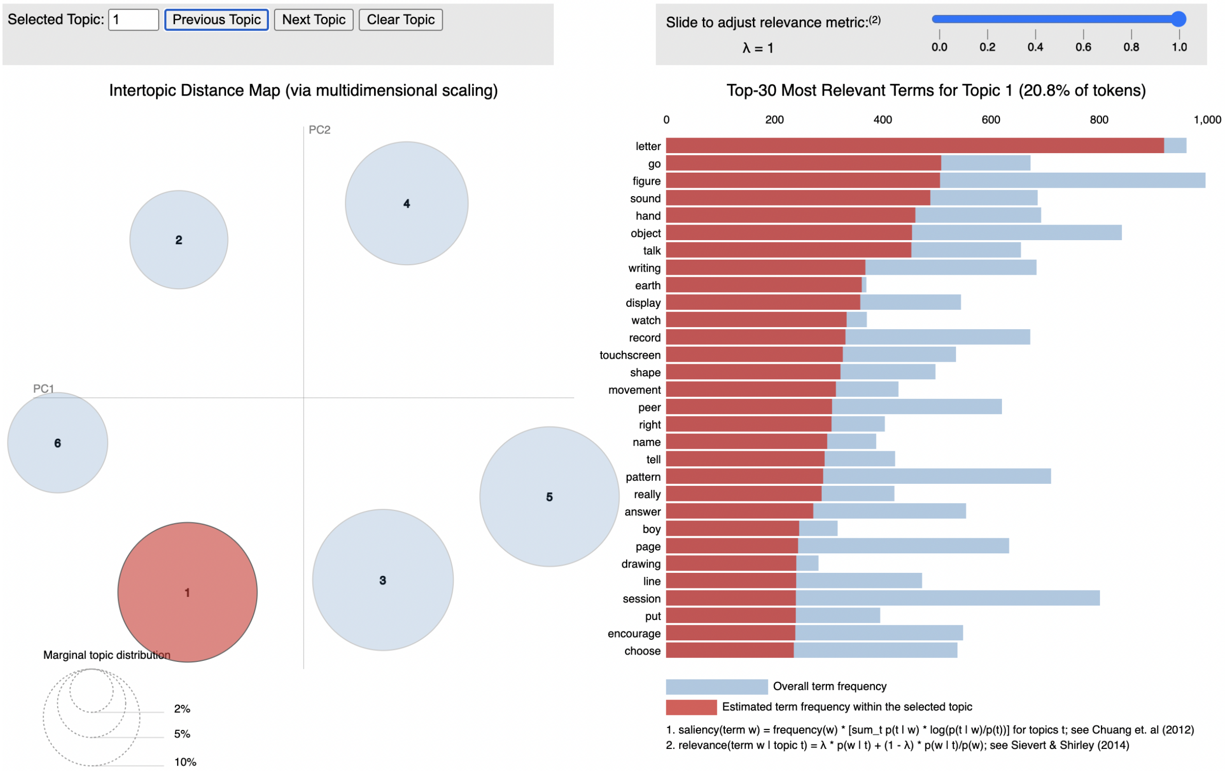
WordCloud 7



WordCloud 8



WordCloud 9



Selected Topic: 1

Slide to adjust relevance metric:(2)

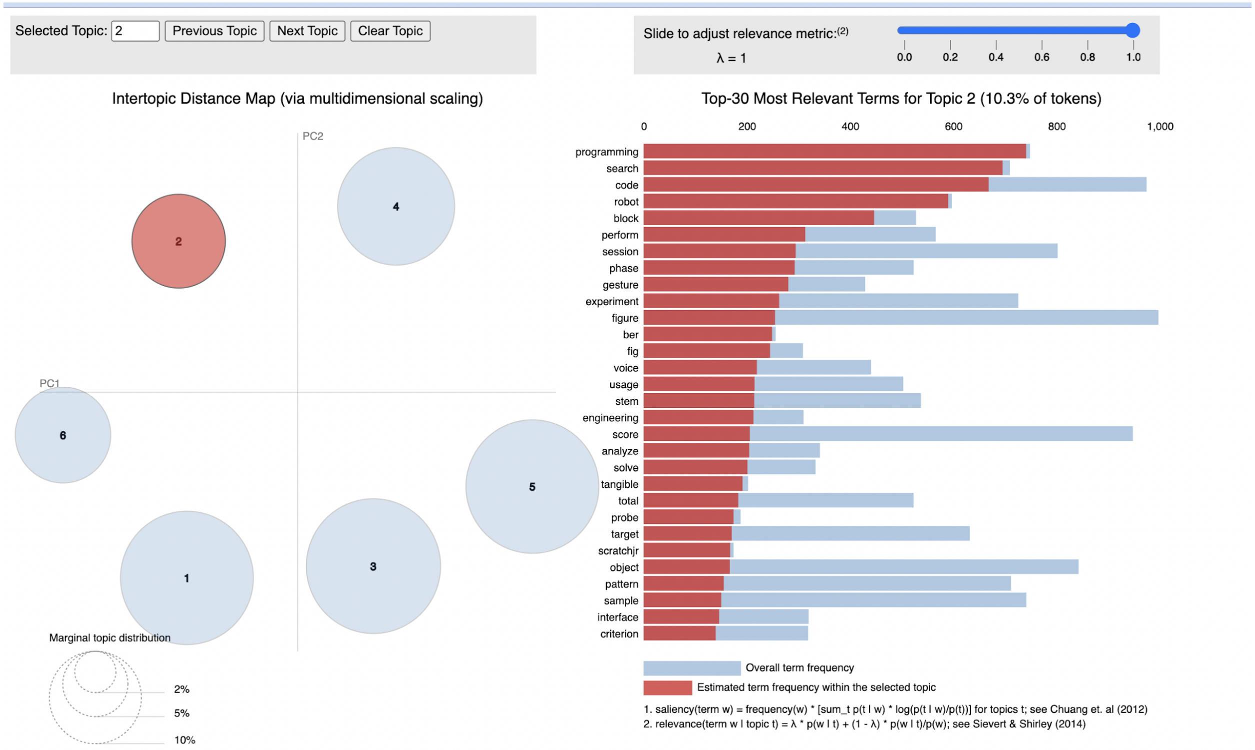
λ = 1

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 1 (20.8% of tokens)

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Topic 1

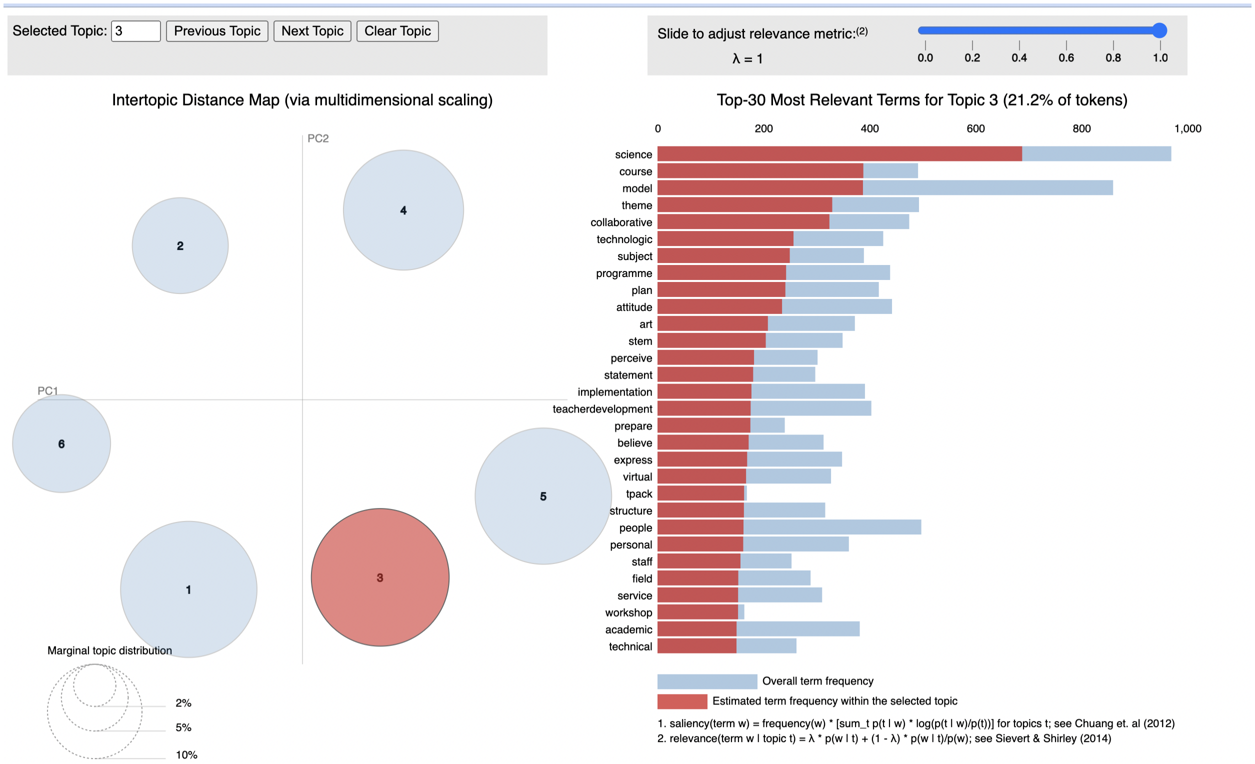


Selected Topic: 2

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t I w) * log(p(t I w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w I topic t) = λ * p(w I t) + (1 - λ) * p(w I t)/p(w); see Sievert & Shirley (2014)

Topic 2

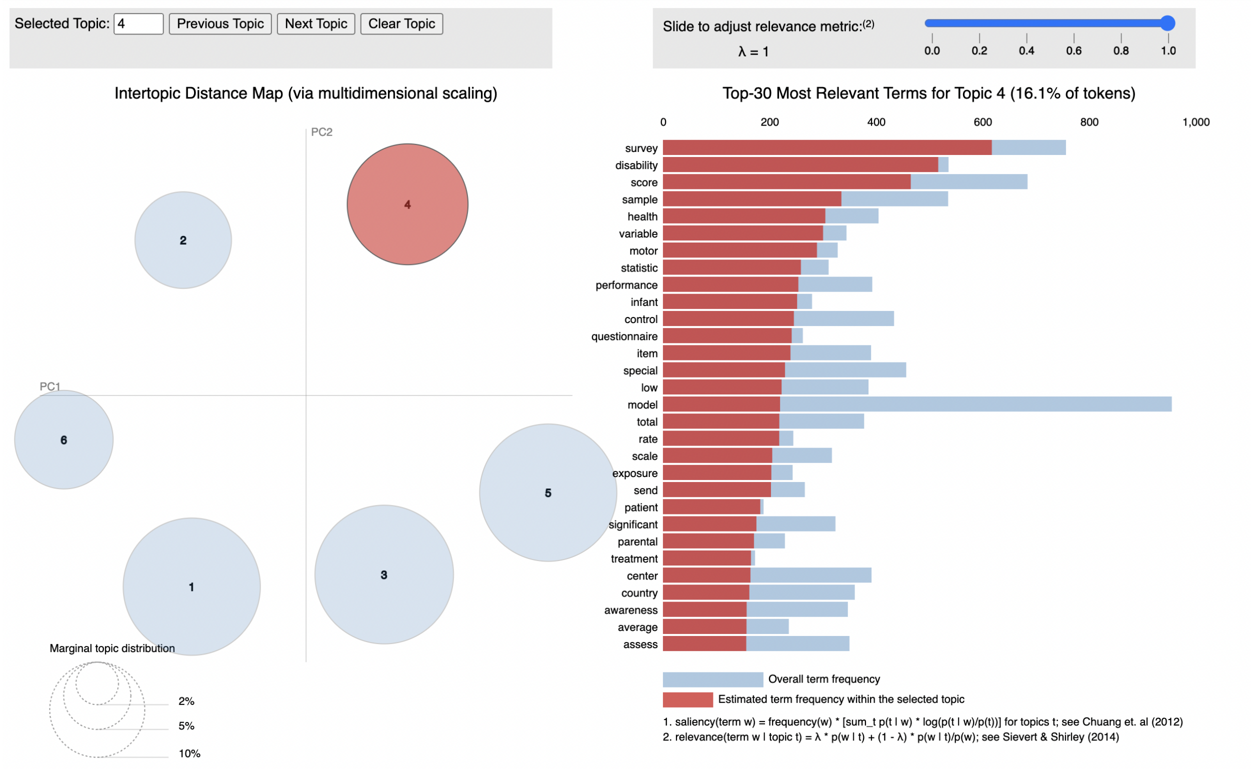


Selected Topic: 3

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t I w) * log(p(t I w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w I topic t) = λ * p(w I t) + (1 - λ) * p(w I t)/p(w); see Sievert & Shirley (2014)

Topic 3



Selected Topic: 4 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:(2)

λ = 1

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 4 (16.1% of tokens)

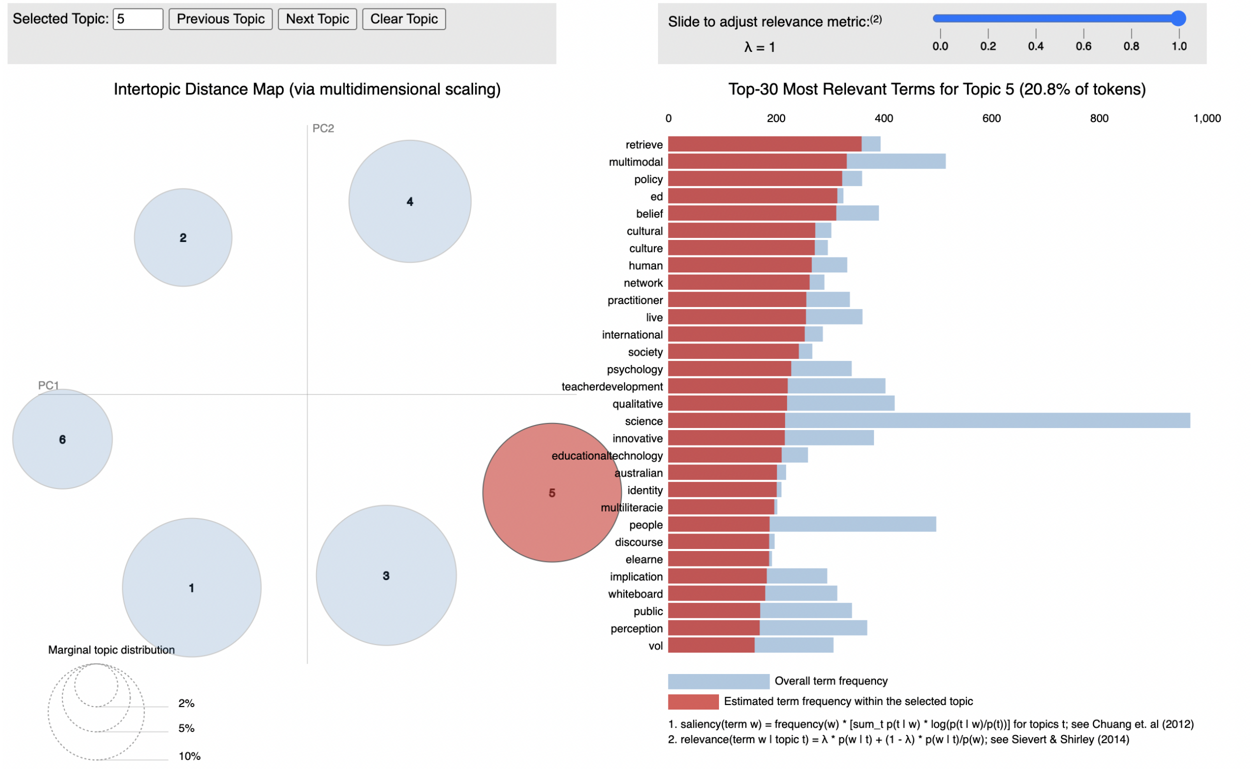
🟦 Overall term frequency

🟥 Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Topic 4



Selected Topic: 5 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:(2)

λ = 1

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 5 (20.8% of tokens)

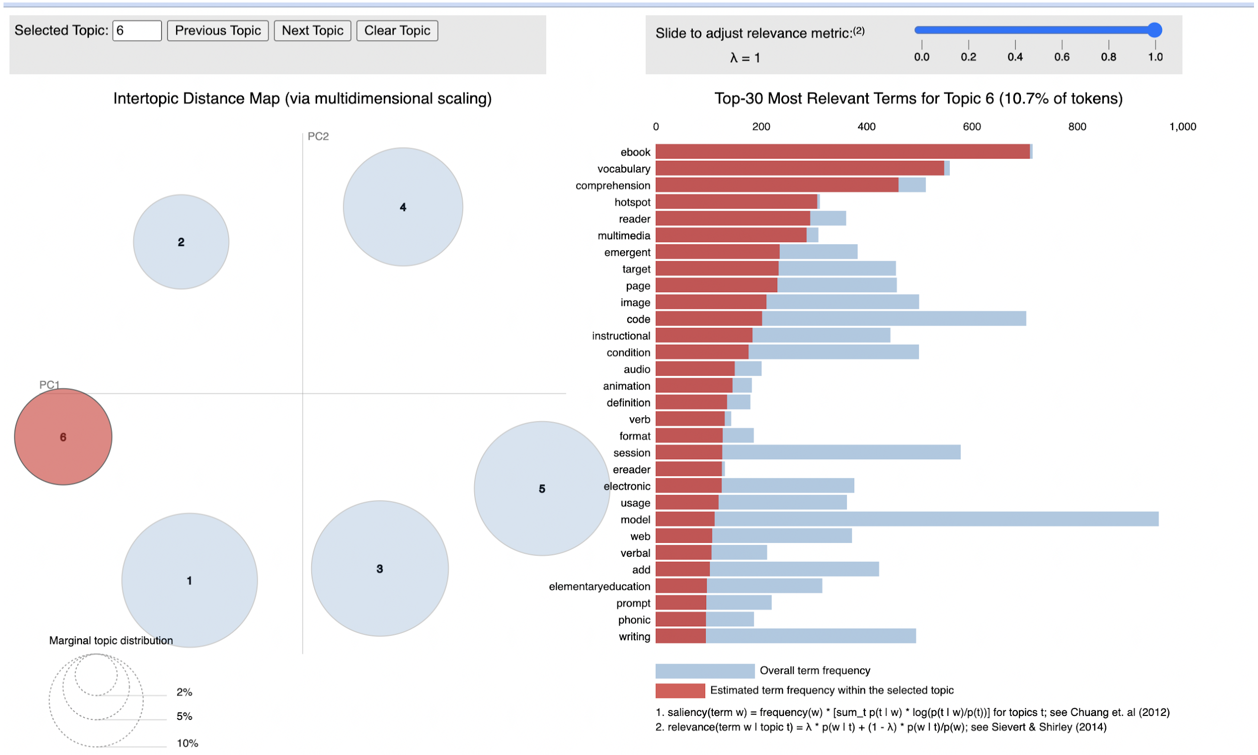
🟦 Overall term frequency

🟥 Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Topic 5



Selected Topic: 6

Slide to adjust relevance metric:(2)

λ = 1

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 6 (10.7% of tokens)

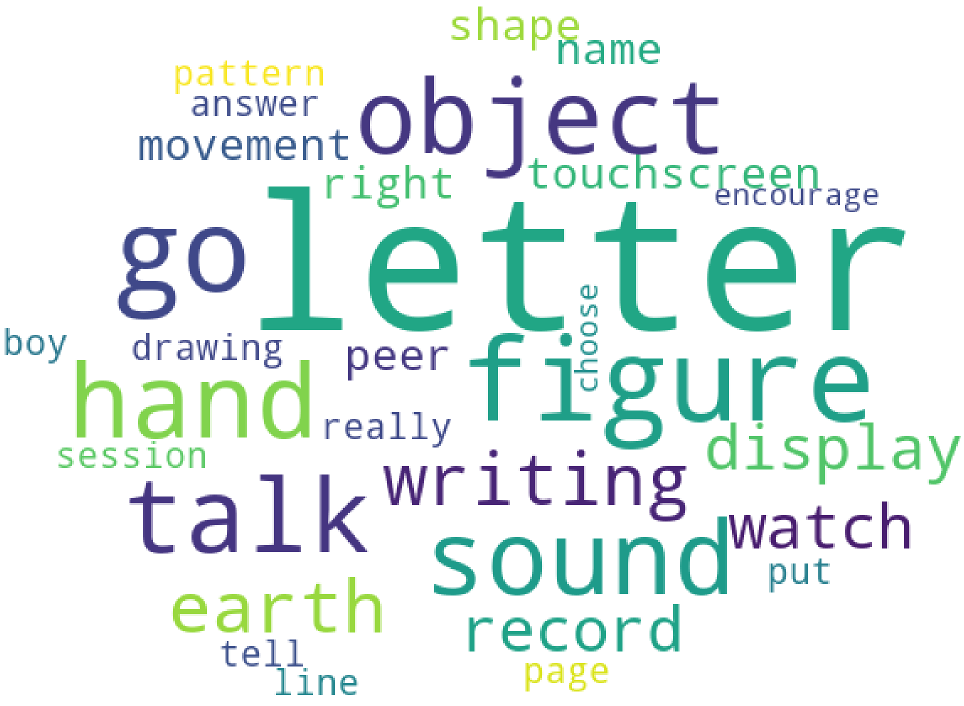
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

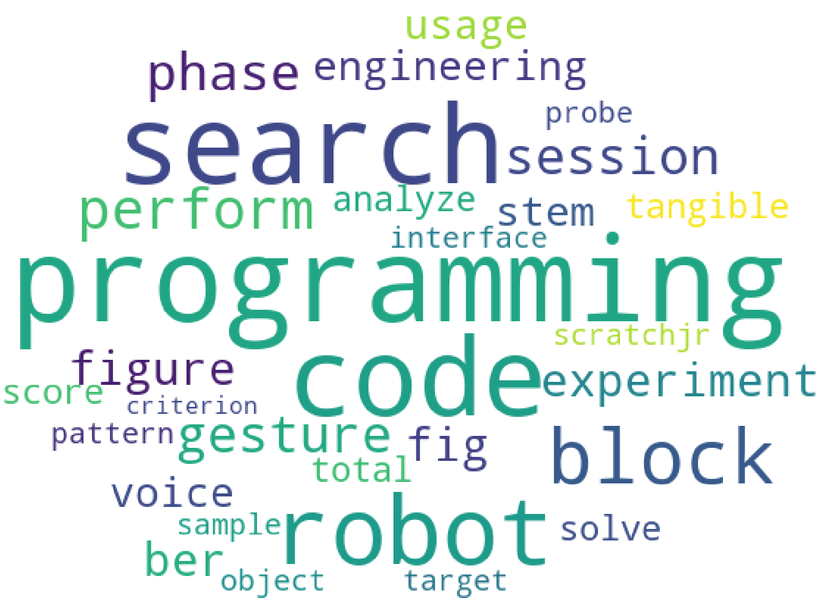
Topic 6

Example 2

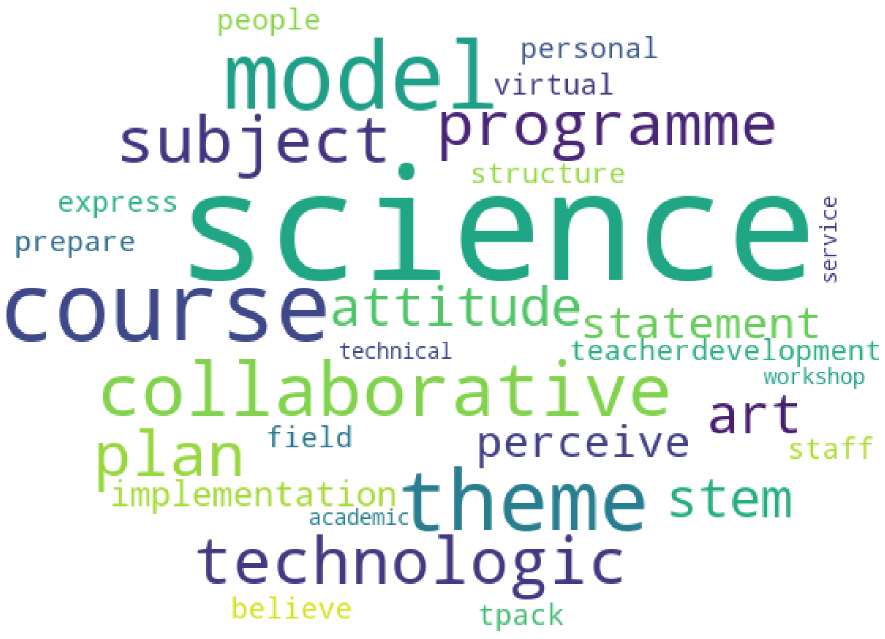
The second example, the search terms used for the topic modeling app are pedagogical documentation, learning analytics, data literacy, and multimodal. Results from the full-text files retrieved from the Western Libraries choosing only peer-reviewed, English articles, and articles published after 2008 can be seen below resulting in 9 different wordclouds and 6 different key topics from 217 different articles:



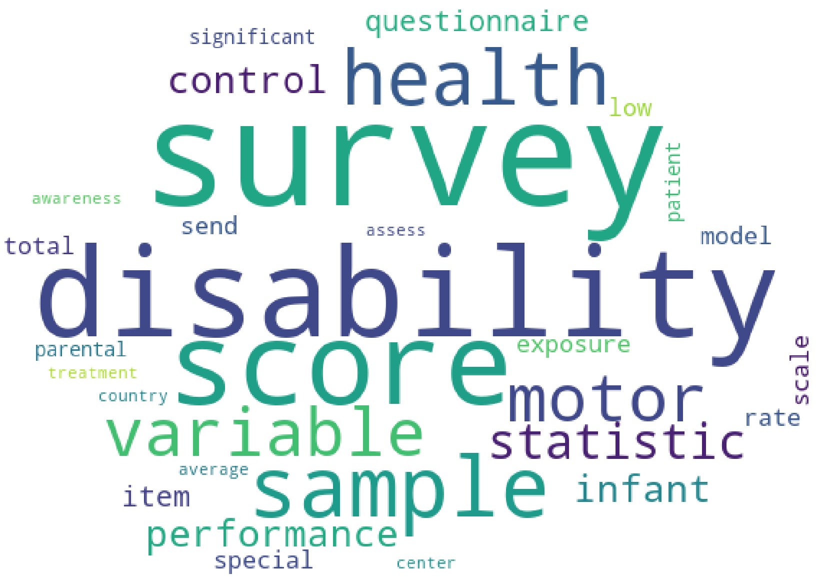
WordCloud 1



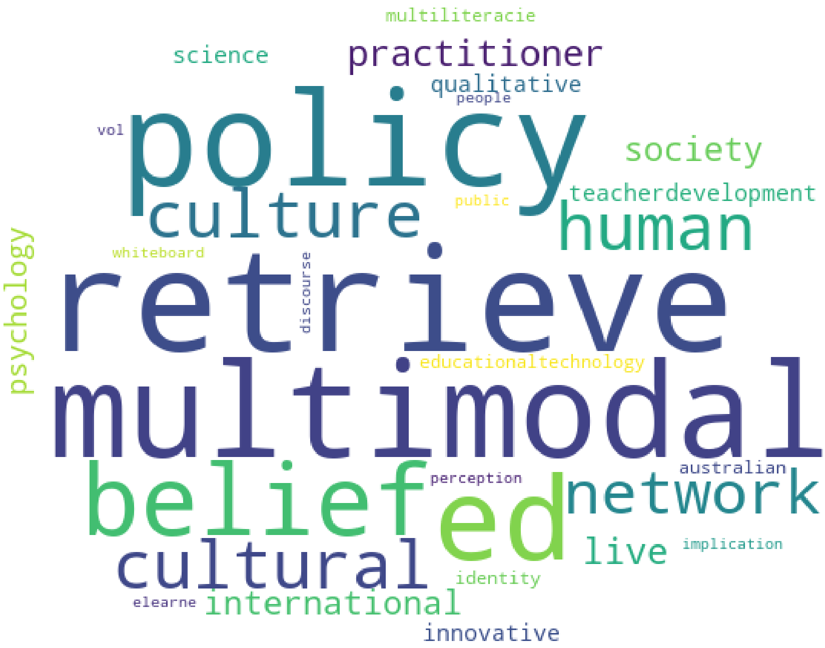
WordCloud 2



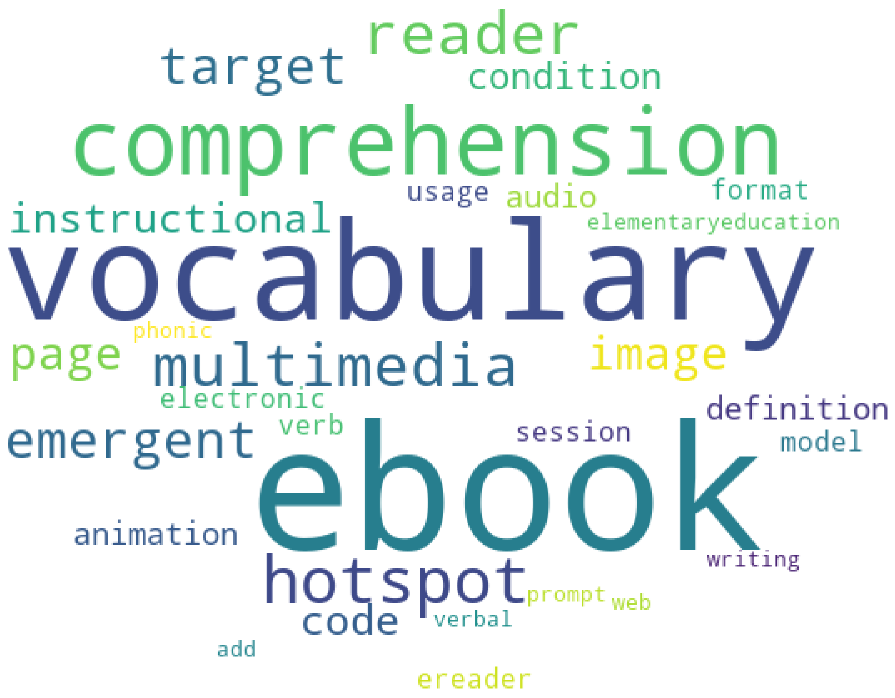
WordCloud 3



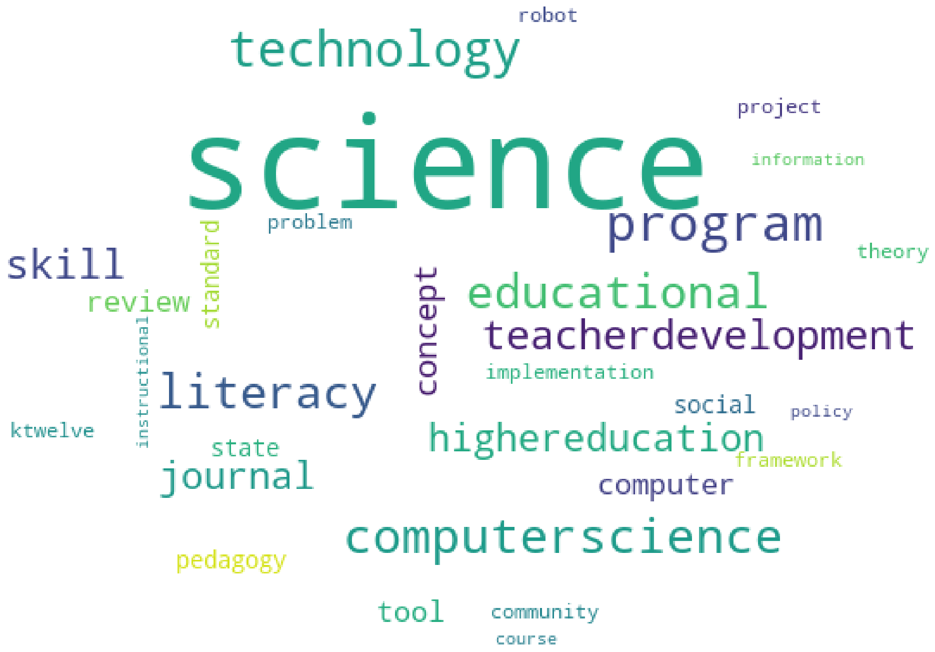
WordCloud 4



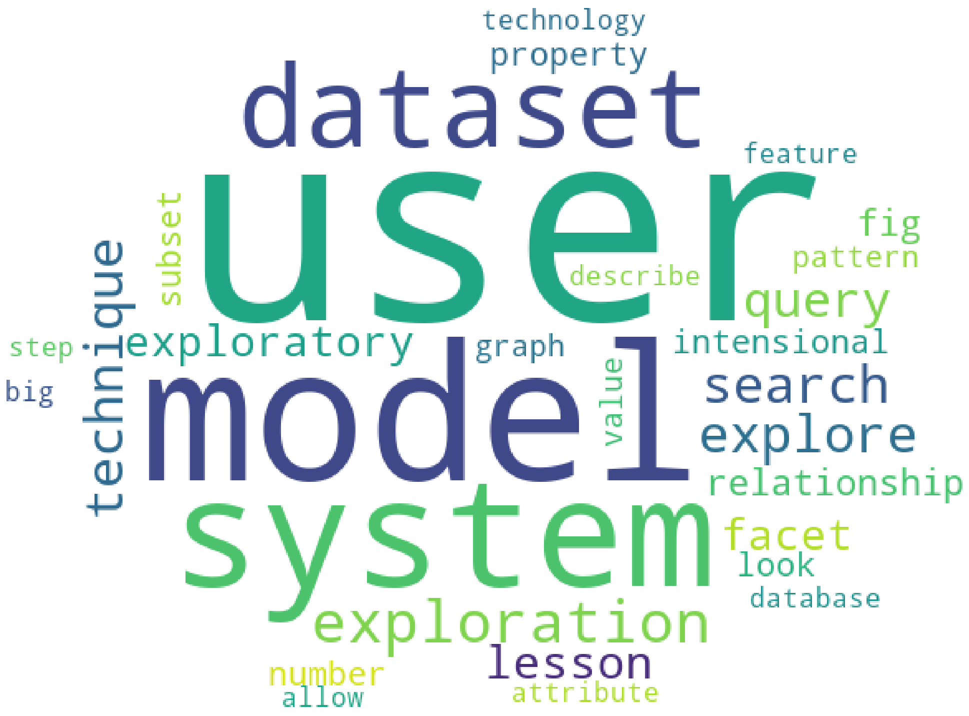
WordCloud 5



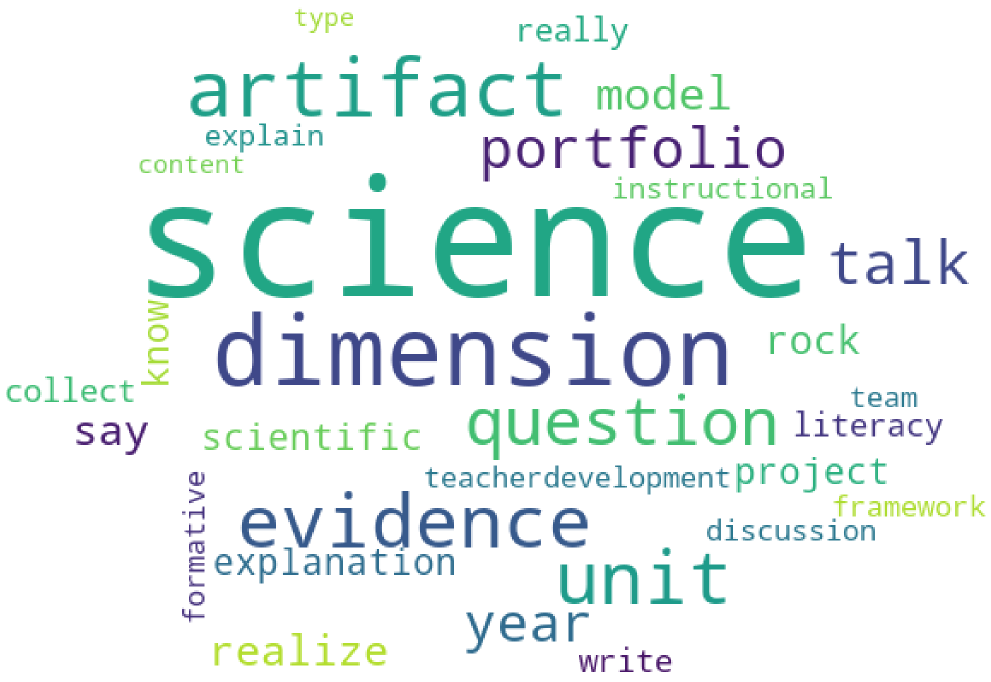
WordCloud 6



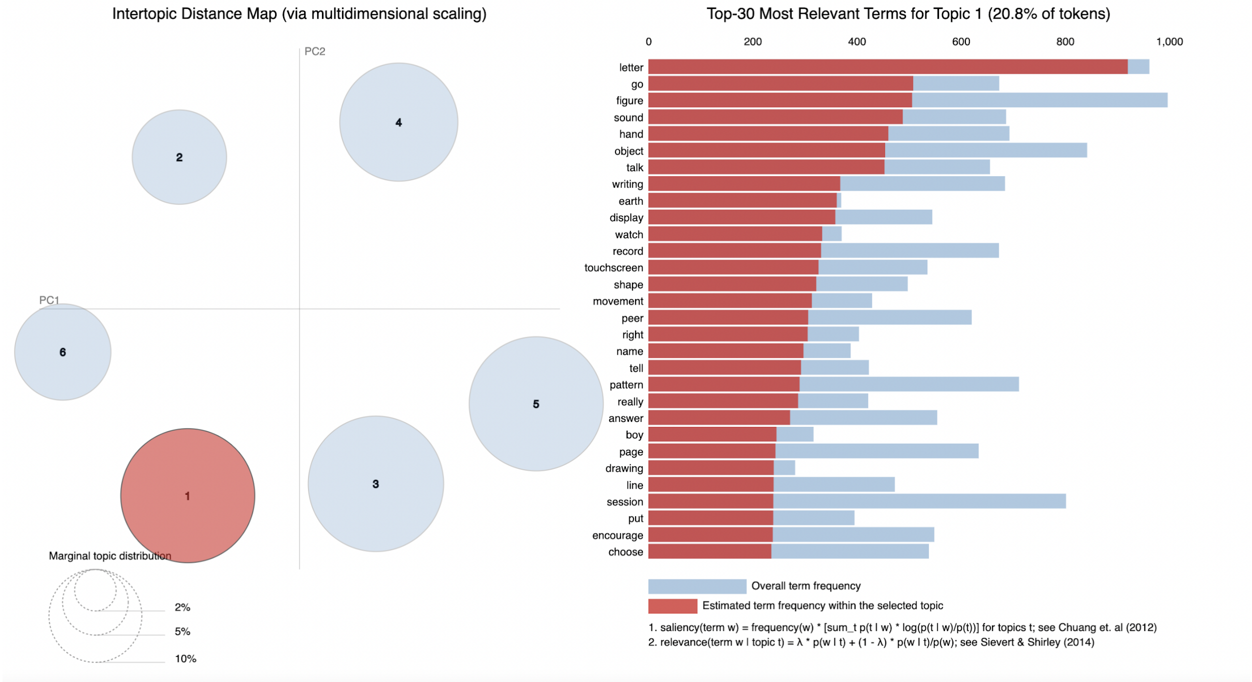
WordCloud 7



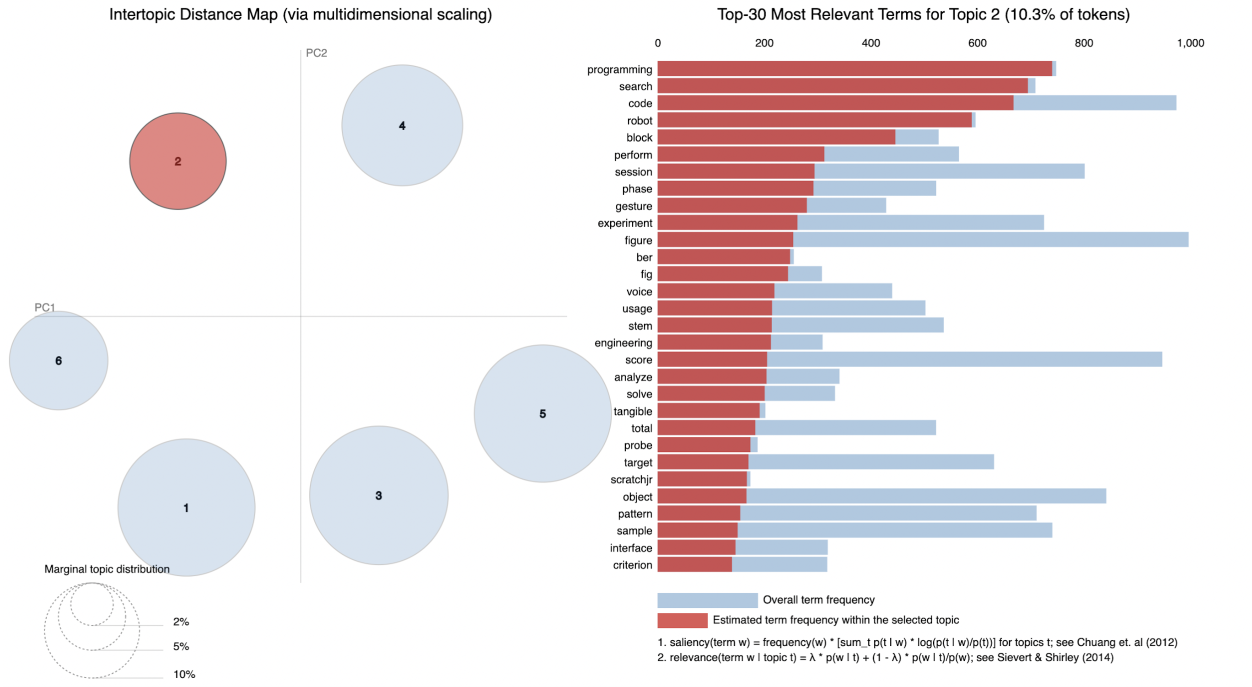
WordCloud 8



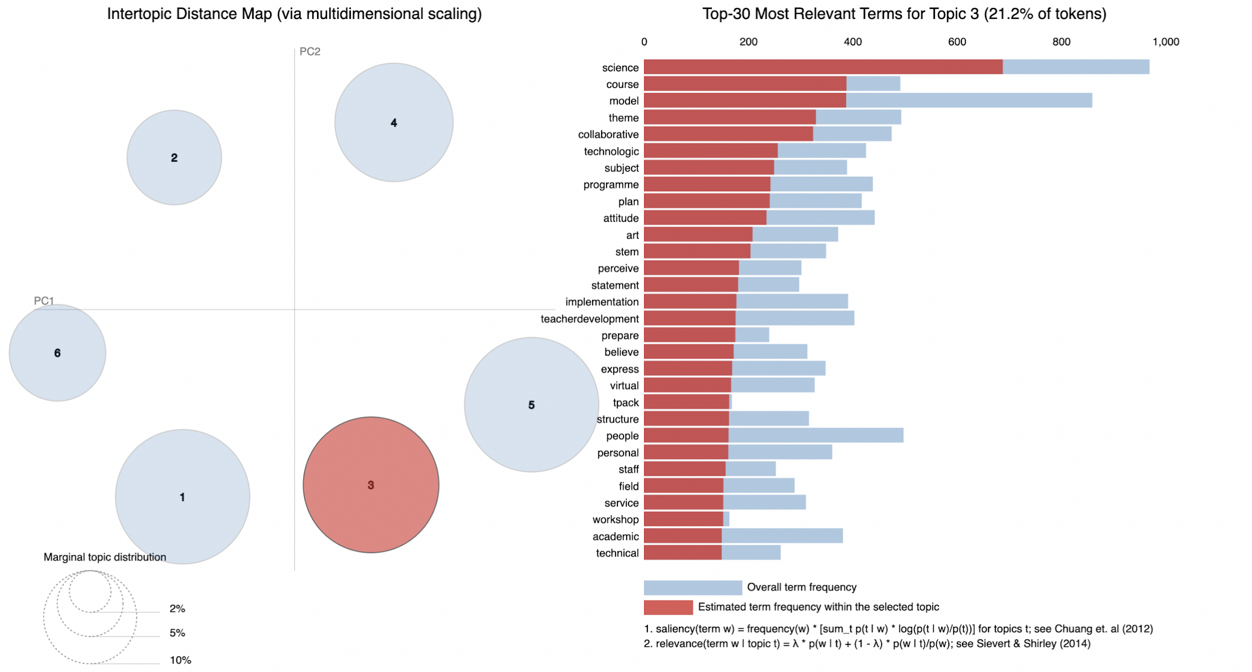
WordCloud 9



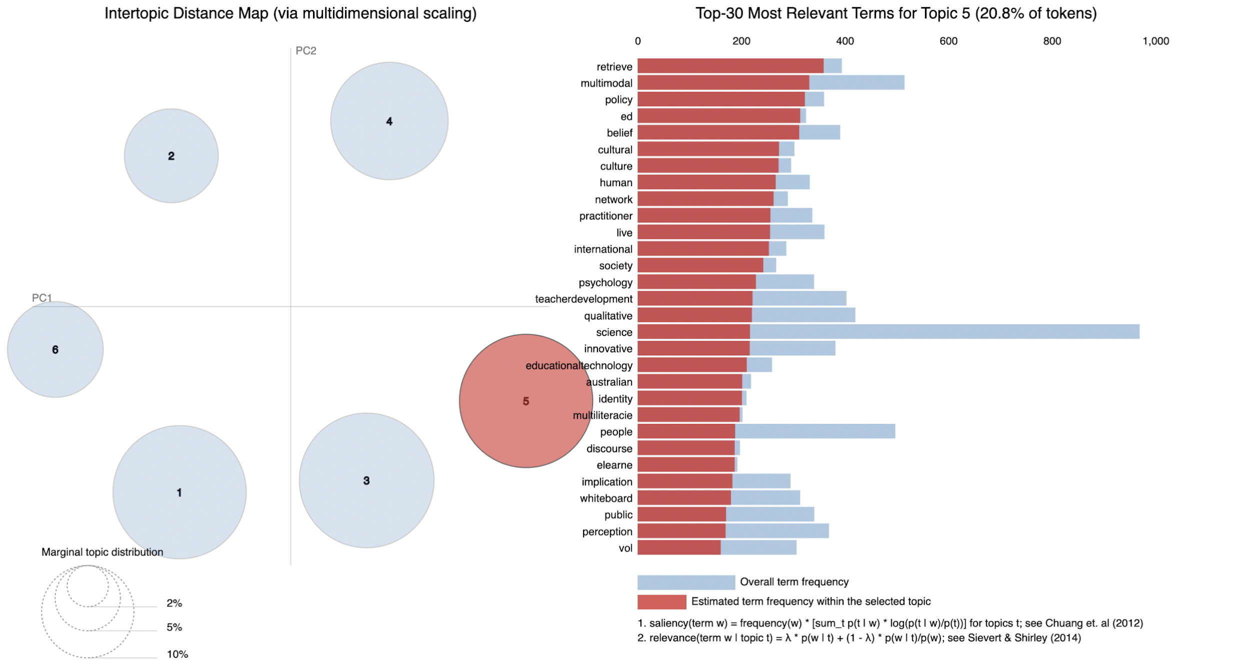
Topic 1



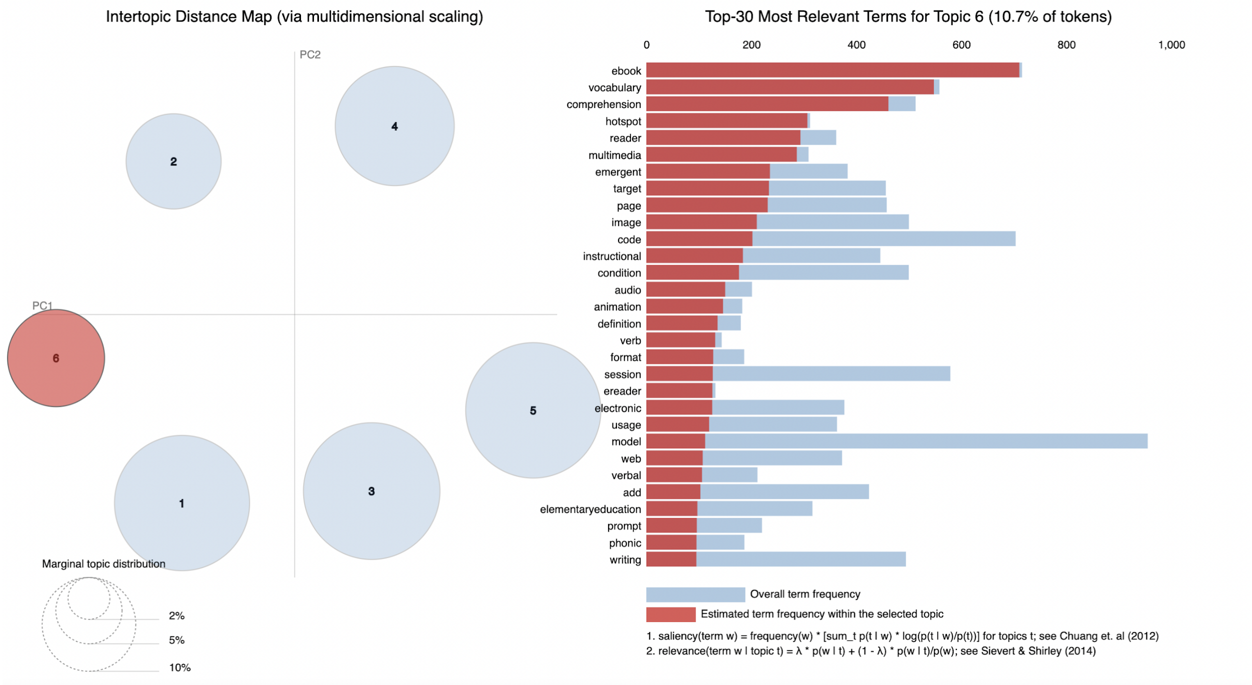
Topic 2



Topic 3



Topic 5



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Topic 6