

Step by step User Guide

Contents

1	Dashboard	1
2	Data File Interface:	2
2.1	Step 1: Upload data files (left panel)	3
2.2	Step 2: Overview of the uploaded files (right panel)	3
3	EDA Interface:	4
3.1	Step 3: Summary statistics and plots	4
3.2	Step 4: Pre-processing	5
4	Topic Modeling and LDAvis Interfaces	7
4.1	Step 5: Train LDA model with the desired number of topics	8
4.2	Step 6: Model interpretation & Refitting	8
4.2.1	Top words and word cloud	8
4.2.2	Document classification for a given topic	9
4.2.3	LDAvis	10
4.2.4	Sentiment Analysis and Interaction with other Variables	11
4.2.5	Refitting and Seed words	11
4.3	Step 7: Download the results	12

1 Dashboard

Note: some areas in the figures of this user guide have been erased to avoid leaking sensitive information.

The dashboard provides an entire workflow to analyze survey comments, including uploading files, preprocessing the corpus, performing topic modeling, and interpretation of the model.

Firstly, you can look up this *User Guide* by clicking the green button, which will redirect to the *User Guide* page.

Except for the *User Guide*, the dashboard includes 4 main interfaces:

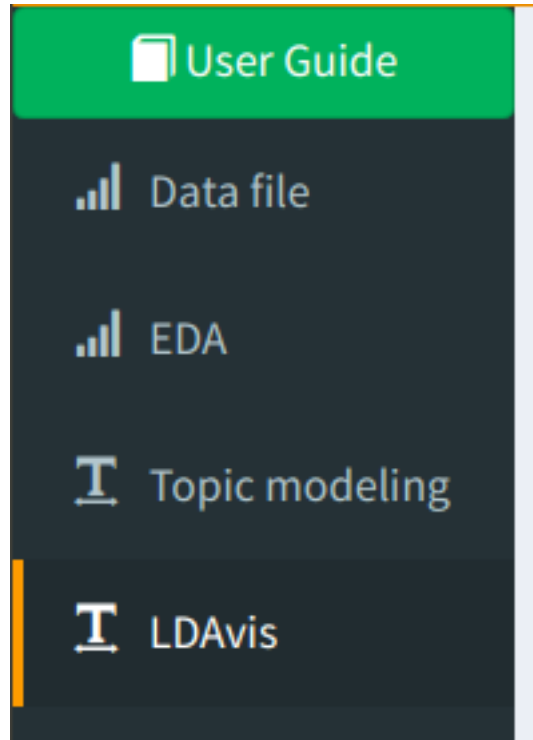


Figure 1: Dashboard Interfaces

- **Datafile:** allows users to upload multiple survey data files, and select fields to include in the final analysis.
- **EDA:** provides some descriptive statistics about the corpus, and also allow user-specified pre-processing operations.
- **Topic modeling:** this is the main interface where users perform topic modeling using LDA, interpret the model, and download results.
- **LDAvis:** Interactive visualization for interpreting topic model.

2 Data File Interface:

The screenshot shows the 'Interactive Topic Modeling' web application. The sidebar on the left contains navigation links: 'User Guide', 'Data file', 'EDA', 'Topic modeling', and 'LDAvis'. The 'Data File' section is active, showing a list of uploaded data files and a 'Select columns' dropdown menu. The main area displays a table of data entries with the following columns: ID, Type, Tokens, Sentences, Year, AGE_RNG_AGR, GENDER, and CIP_CLUSTER_EXPANDED. The table shows 10 entries, each with a unique ID and associated data. At the bottom, there is a pagination bar indicating 'Showing 1 to 10 of 1,235 entries' and navigation links for 'Previous', '1', '2', '3', '4', '5', '124', and 'Next'.

Figure 2: Data File Interface

2.1 Step 1: Upload data files (left panel)

1. Use **Upload Data files** button to upload the files for analysis. It only accepts CSV, and you can upload multiple files at the same time, making sure that files have the same columns.
2. **Select columns**: select the columns for included in the analysis.
3. **ID**: select the unique ID column for each comment
4. **Text**: select the text column.
5. The checkbox **Reshape to sentences** allow to break the comment into sentences.
6. After the above selections, you can **apply** and finalize the settings.

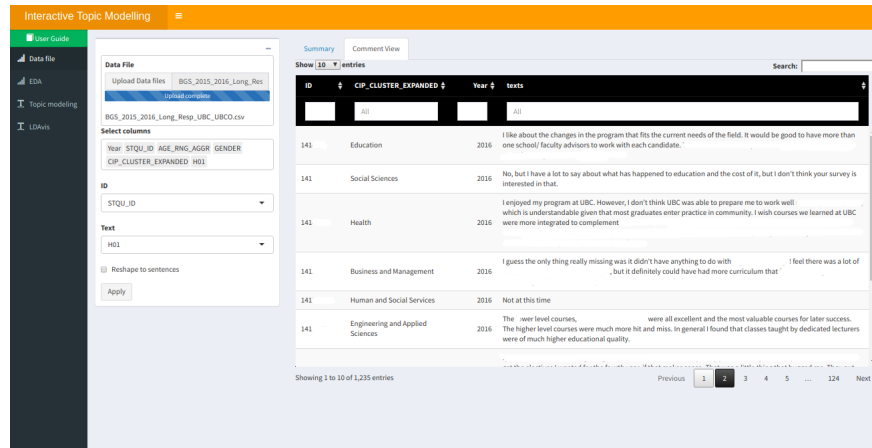


Figure 3: Comment View

2.2 Step 2: Overview of the uploaded files (right panel)

1. After **apply**, the right panel will show data table that gives a preview of the files.
2. The **Summary** tab shows all the selected columns, with *Text* column broken into 3 columns *Types*, *Tokens*, and *Sentences*. - *Tokens*: words, numbers, punctuations - *Types*: unique tokens - *Sentences*: number of sentences
3. To have a better idea of the actual comment, you can switch to the **Comment View** tab, where the detail comments are shown in the *texts* column.

After overviewing the files, you can go to the second interface, **EDA**.

3 EDA Interface:

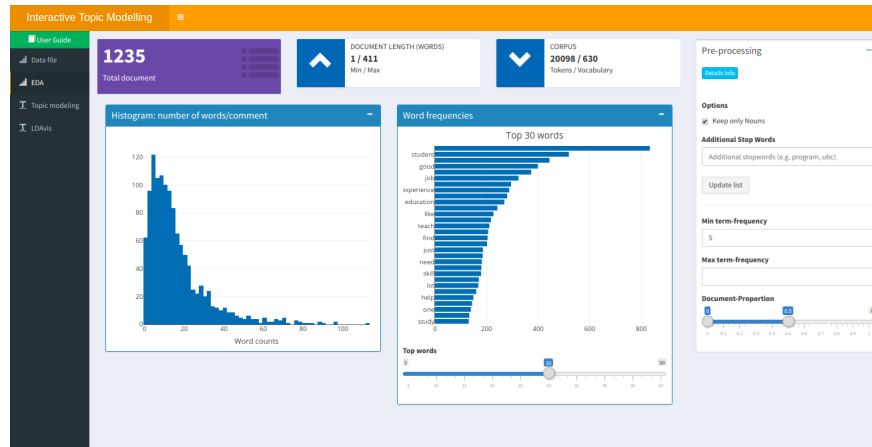


Figure 4: EDA Interface

In this interface, you can:

- Get a brief idea of the corpus by looking at some descriptive statistics and plots.
- Pre-process the corpus with user-specific inputs.

3.1 Step 3: Summary statistics and plots

At the upper part, you can know about:

- **Total documents** of the corpus (the number may change if you *reshape to sentences* at the previous interface)
- Maximum/minimum **Document Length** in terms of words
- Total number of **tokens** and **vocabulary** (i.e. unique number of tokens) of the corpus

At the bottom part, you can know about:

- The distribution of word counts for each document - **histogram**
- The sorted word frequencies of words. The slider below controls the number of top words to show.

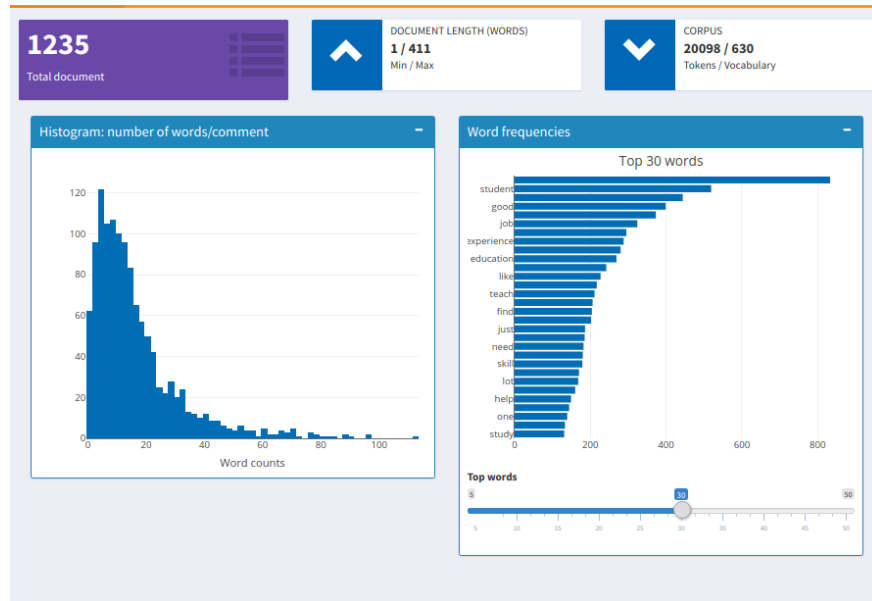


Figure 5: Summary Statistics

3.2 Step 4: Pre-processing

By default, some standard *Natural Language Processing* techniques have been applied to the corpus. It includes **tokenization**, **remove numbers**, **remove punctuations**, **lemmatization**, **remove English stop words**. Also, words **less than 3 characters** are removed.

More options are available:

- **Keep only nouns.** That is to remove verbs, adjectives, adverbs, e.t.c.
- **Provide user-specific stop words.** This gives the flexibility to filter out words that convey little meanings based on domain knowledge. As shown in the figure below, *'program'*, *'student'*, and *'course'* are supplied as additional stop words.
- **Remove words based on maximum/minimum term frequencies and document proportion** (values range from 0 to 1). In this case, words with a frequency lower/higher than 5/200 will be removed. Also, words that appear at more than 0.6 of the documents will be removed.

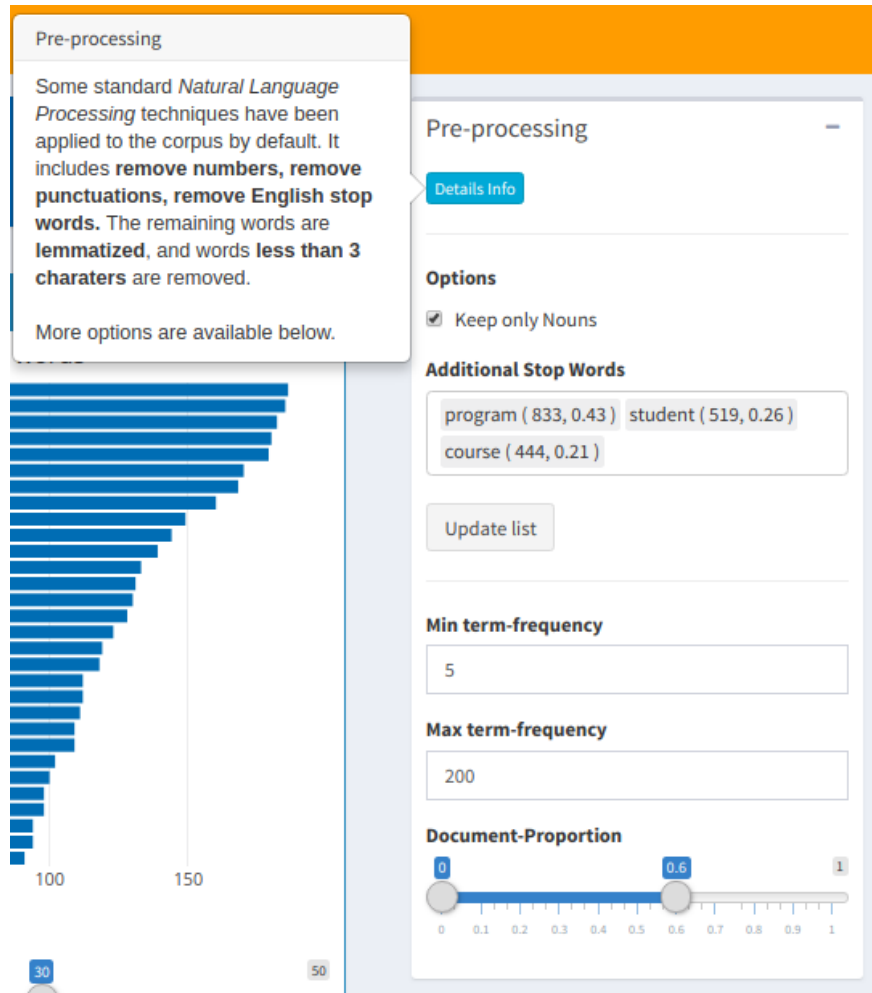


Figure 6: Pre-processing

4 Topic Modeling and LDAvis Interfaces

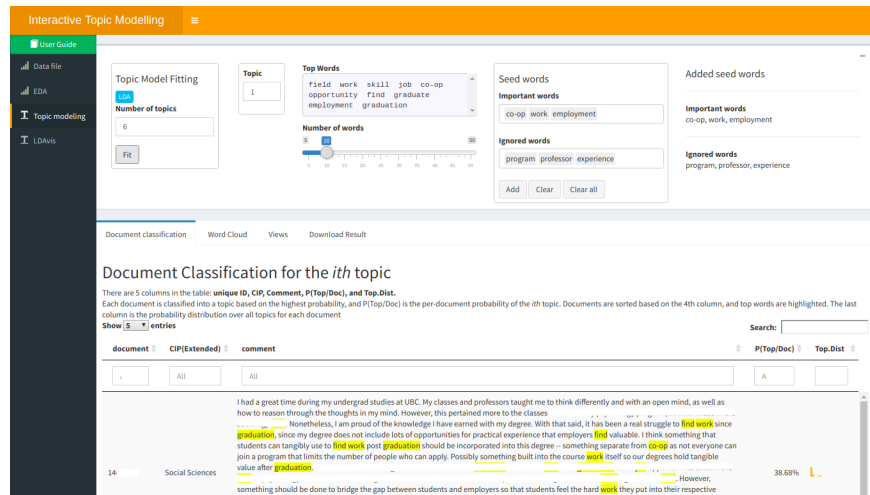


Figure 7: Topic Modeling Interface

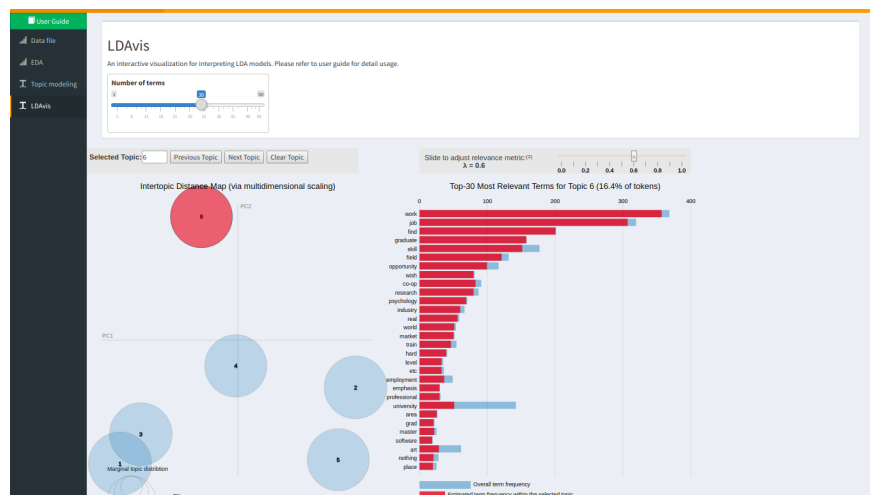


Figure 8: LDAvis Interface

In these two interfaces, you can:

- Fit a topic model with LDA of a fixed number of topics (K)
- Interpreting the model with different tools
- Download the results file

4.1 Step 5: Train LDA model with the desired number of topics

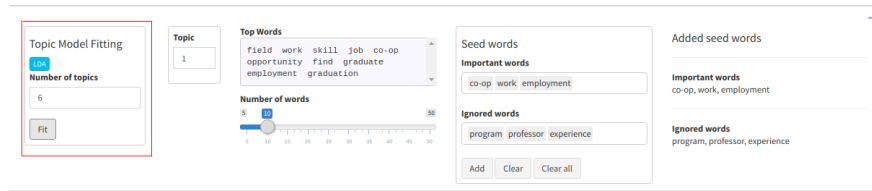


Figure 9: Model Fitting (left part)

1. First, choose the **number of topics** (default $K = 6$).
2. Click **“Fit”** to fit Latent Dirichlet Allocation on the corpus.

Note:

- Try to go with default $K = 6$ first, and then gradually increase K to refine the model. $K = 6 \sim 10$ is most interpretable model for the BGS 2015 - 2016 data.
- Be cautious, running time increase with increasing number of topics. $K \geq 50$ can take really long time to run.

4.2 Step 6: Model interpretation & Refitting

After fitting the model, you need to interpret the model to understand the dataset. We provide tools for model interpretation:

- Inspecting top words and the word clouds generated from these top words
- Document classification for a given topic
- LDAvis

Besides, we also provide functionalities for augmenting human efforts to analyze the data. It includes:

- Sentiment Analysis
- Interaction with other variables: CIP, Age, and Gender

4.2.1 Top words and word cloud

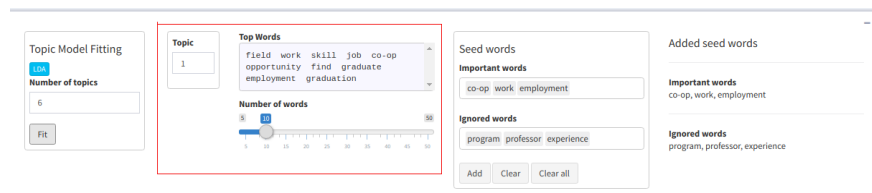


Figure 10: Top words

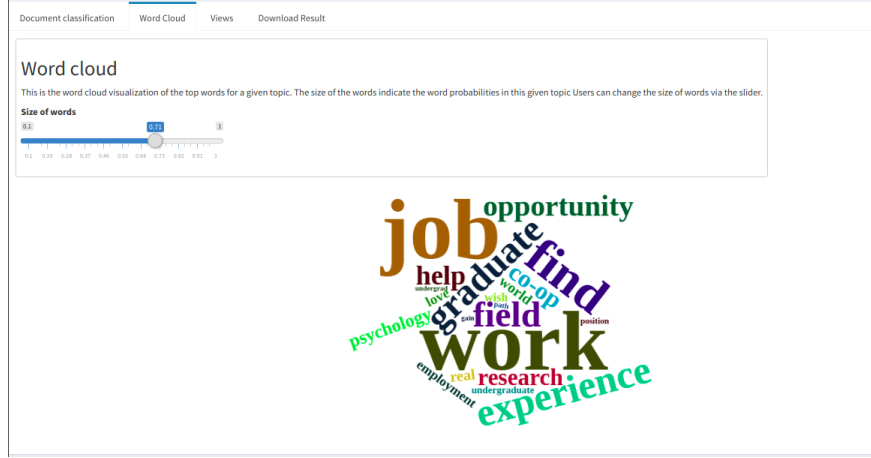


Figure 11: Word cloud

Top words are the most representative words for each topic. You can first inspect the top words to interpret these topics.

At the upper panel, you can use the **Topic** widget to select the i th topic from the fitted model, and the associated top words will be rendered in the **Top Words** box. You can use the slider to choose the **number of top words** to show.

At the lower panel, you can switch to the **word cloud** tab which visualizes the top words for the i th topic as word clouds. We encode the estimated word probabilities for a given topic as the size of each word. For example, "job" and "work" (biggest two words) are most likely generated from this topic. Besides, The **size of words** slider controls the overall size of the word clouds.

4.2.2 Document classification for a given topic

However, top words may not always be intuitive for humans to interpret them properly.

To solve this, we further provide **document classification** for a given topic. It allows users to interpret the topic by browsing through some of the most likely comments in this topic.

There are 5 columns in the table:

- **unique ID**: unique ID for each comment
- **CIP**: CIP code for each ID
- **Comment**: actual comment with key words highlighted (yellow background)
- **P(Top/Doc)**: per-document probability of the given topic
- **Top.Dist**: probability distribution over all topics

Document classification Word Cloud Views Download Result

Document Classification for the *ith* topic

There are 5 columns in the table: **unique ID**, **CIP**, **Comment**, **P(Top/Doc)**, and **Top.Dist**. Each document is classified into a topic based on the highest probability, and P(Top/Doc) is the per-document probability of the *ith* topic. Documents are sorted based on the 4th column, and top words are highlighted. The last column is the probability distribution over all topics for each document

Show 5 entries

Search:

document	CIP(Extended)	comment	P(Top/Doc)	Top.Dist
12	Sciences	Department should have its own co-op program. I did the co-op program and the types of job postings was so broad that it was hard to find a position that was relevant to the field especially when competing against hundreds of other co-op students for the same job . I wish I had gained more experience in the field to make the transition into finding work in the field post-graduation more attainable, especially when many jobs require 2-3 years experience . such a broad field that a co-op program would be a great way for students to explore what kind of job opportunities would fit their interests.	36.33%	...
12	Health	I would not recommend the co-op program to anyone. co-op does not provide students with the necessary skills to work after graduation. Other programs at UBC provide their students with co-op opportunities and job opportunities. Other programs hold multiple career fairs and networking events for students through out the year. They create opportunities for students to get involved in projects that will help them make connections and build their resume. For my masters degree, I switched my major to biology at UBC. Seeing their undergrad students' job opportunities, I do regret not doing my undergraduate in the biology department.	30.47%	...
13	Science	Please encourage professors to talk to students about opportunities after undergrad and how to plan for them. This should be required. There was only one who briefly suggested talking to her for letters of reference. I felt like the only opportunity in biology and after volunteering in even a fascinating lab, I was not interested and felt like there were few options after graduating. This spurred a lot of depressive tendencies. I am sure I am not alone. Also, the biology department should encourage students to look into the counselling department at UBC, which is where I found the most relevant experiential training. It was what I was expecting from my education in biology . We learned how to listen and apply biology to the real world . This is important. With such a reputable university worldwide , why are not putting honour into the biology program? Why are there so few job opportunities that the biology students can apply for?	29.34%	...

Showing 1 to 5 of 293 entries

Previous 1 2 3 4 5 ... 59 Next

Figure 12: Document classification

Each document is classified into a topic based on the highest probability across all topics, and then sorted descendingly based on $P(Top/Doc)$. The highlighted keywords help users to read through and interpret the comments quickly. And, the last column help users understand how confident is the topic assignment.

For example, the highlighted keywords help quickly identify the student is talking about job prospect with focus on co-op. The student may also mention other topics as indicated by the last column.

4.2.3 LDAvis

LDAvis is an interactive visualization tool designed to help interpret the topics in a topic model. Here we just describe the functionalities. For methodology behind LDAvis, please refer to the [original paper](#).

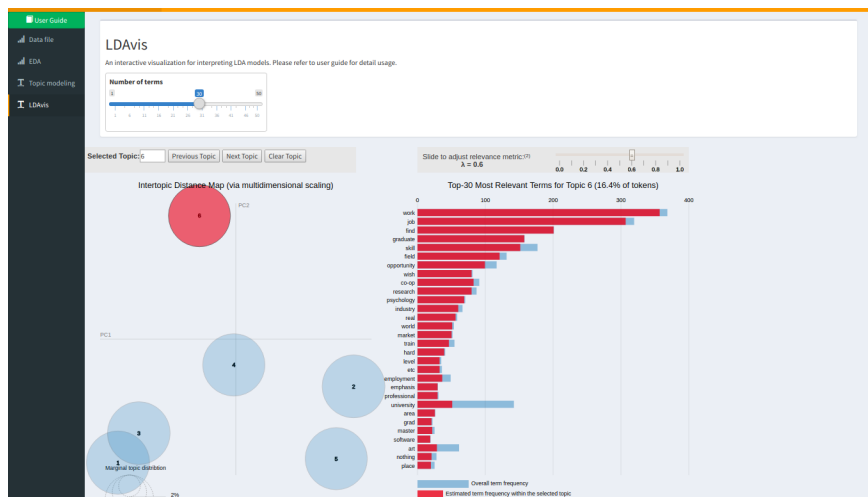


Figure 13: LDAvis Interface

The left panel is a **intertopic distance map** that indicate the distance and overlap between each topic. The size of the bubbles indicates the number of tokens related to a specific topic. In our example, the sizes of the bubbles are most or less the same. Besides, the distance shows the distances and overlaps between topics. In this case, the 2 topics at the left bottom corner are overlapping.

The right panel is the **most relevant terms** for a given topic. Relevant is a combination between estimated term frequency for a given topic, the same as *top words*, and topic-specific terms. By changing the value of λ , terms are reordered based on estimated term frequency for a given topic ($\lambda = 1$) or topic-specific terms ($\lambda = 0$). You can change different values of λ to reorder words to interpret the model better. According to the original paper, $\lambda = 0.6$ is the best to interpret topics.

On top of the visualization, a *slider* controls the number of words to display.

4.2.4 Sentiment Analysis and Interaction with other Variables

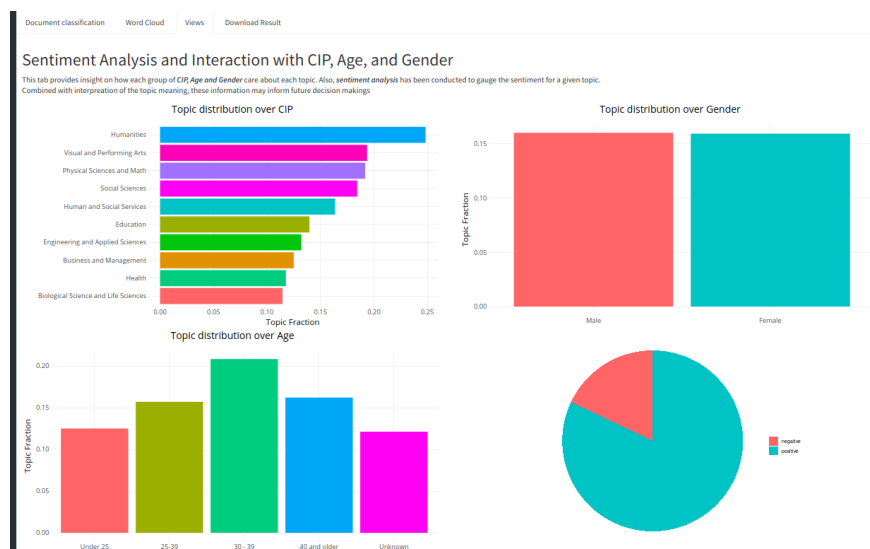


Figure 14: Sentiment Analysis and Interaction with other Variables

We also include **sentiment analysis** and **interaction with other variables** for each topic to help analyze the comments.

Sentiment Analysis: The process of analyzing users' attitude related to the comment or feedback is called sentiment analysis. We use the open-source library [TextBlob](#) to perform the task. It will take a sentence and give a score between -1 and 1, with -1 being totally unhappy and 1 being totally happy. The result is further visualized as a pie chart for each topic. You may relate this to the current topic to analyze the text comment. For the current topic, you can know that the comments are generally positive.

Interaction with other variables: We have also related each topic with CIP, Age, and Gender via visualizations. You can connect these to the current topic to identify the interests or concerns of a particular group. For example, **Humanities** students concern the most about the current topic across CIP groups.

4.2.5 Refitting and Seed words

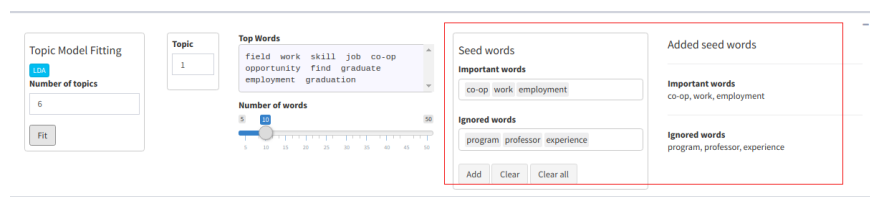


Figure 15: Insert seed words

The model may sometimes cluster words into a topic that users may not agree on. Therefore, you may want to refine the model for better topics discovery. You can achieve this by incorporating seed words for each topic.

- **Important words:** words that you want to **include** in this topic
- **Ignored words:** words that you want to **exclude** in this topic

After choosing a list of important/ignored words, you can click **Add** to add the seed words for this topic. The **Added seed words** part shows the seed words being added to this topic. You can remove the seed words for a topic by clicking **Clear** and for all topics by clicking **Clear all**. For example, **co-op**, **work**, **employment** are added as important words, while **program**, **professor**, **experience** are added as ignored words for topic 1.

You can switch to each topic and add seed words and then refit the model. You can switch to each topic and add seed words, and then refit and interpret the model. This process may repeat several times until the result is satisfying.

4.3 Step 7: Download the results

Download result data frame

Here provides the download of the result data frame from the topic modeling process. It is the original data frame with the selected columns during the data uploading process. Plus 1 column of the **sentiment** of the text, ranged from -1 to 1, with 1 being totally happy and -1 total unhappy. It also includes the per-document-per-topic probabilities, indicated as **topic_***.

[Download Result](#)

Result preview

Show 5 entries

	document	texts	Year	AGE_RNG_AGGR	GENDER	CIP_CLUSTER_EXPANDED	sentiment	topic_1	topic_2	topic_3	topic_4	topic_5	topic_6
1	141	I like about the changes in th...	2016			Education	0.2400	0.1414	0.1869	0.1717	0.1869	0.1414	0.1717
2	141	No, but I have a lot to say ab...	2016			Social Sciences	0.2500	0.1388	0.1462	0.2164	0.1462	0.1462	0.1462
3	141	I enjoyed my program at UBC. H...	2016			Health	0.4543	0.1814	0.1961	0.1520	0.1225	0.1373	0.2108
4	141	I guess the only thing really ...	2016			Business and Management...	0.0625	0.1566	0.1869	0.1414	0.2323	0.1414	0.1414
5	141	Not at this time	2016			Human and Social Services...	0.0000	0.1830	0.1634	0.1634	0.1634	0.1634	0.1634

Figure 16: Download results

When you are satisfied with the model, you can download the results as a **CSV** file.

The downloaded file contains the original uploaded data frame in Step 1, with additional 1 column of the **sentiment** of the text, ranged from -1 (unhappy) to 1 (happy), and the per-document-per-topic probabilities indicated as **topic_***.