(v. 2) Module 2 - Examine representative topics (and clusters)

Start of Block: INFO ABOUT YOU & EXERCISE

# IS THIS A TRIAL?

Before you begin, please check the choice below if this is just a trial or experimental exercise that should be ignored in collecting outputs from this module.

[ ]  Check this field if this is just a trial.

# OPEN RESEARCH POLICY STATEMENT

The modules of the WE1S Interpretation Protocol create records of research that support the project's aim of advancing "open, generalizable, and replicable digital humanities methodology" (see project ["About"](http://we1s.ucsb.edu/about/) statement). This aim is related to current practices of ["open science"](https://en.wikipedia.org/wiki/Open_science) and ["open-notebook science."](https://en.wikipedia.org/wiki/Open-notebook_science) The basic idea is that the materials, methods, steps, and findings of research should be transparent so that other researchers know how conclusions were reached and can test or reproduce the process--an ideal that is especially needed for understanding such machine-learning methods as topic modeling (see [WE1S bibliography on machine-learning interpretability](https://we1s.ucsb.edu/research/we1s-bibliography/bibliography-interpretability-and-explainability/)).

In completing this and other modules of the WE1S Interpretation Protocol you are creating an "open notebook" that in principle can be shared with others as a public record of research. *Please be sure that all descriptions, observations, and notes you write in this module are consistent with being part of public record--that is, appropriate in tone, respectful to others, and not in violation of privacy or copyright restrictions*.

*However please also do not let the ideal of open-notebook science inhibit you from speculative or tentative in-progress commentary*. It is understood that research, open or otherwise, is a *process* of finding knowledge with many stages and levels of confidence. (At the conclusion of this and other modules of the Interpretation Protocol, you will be asked to rate your confidence level in the current exercise.)

*In addition, lab notebooks are allowed to be messy as a part of the research process.* It is not expected that notes written into these modules be copyedited as if for formal publication. (The time for polishing is later when writing up a report synthesizing results from one or more modules of this Interpretation Protocol.)

# EXERCISE INFO

**Please enter the following the following information, which will generate an "Exercise ID".** (Tab from input field to input field to force calculation of the Exercise ID.)

* **Your name or team name**. (If team, then use the format "Team1", "Team2", etc. Please do not use a space): **[Name]**
* **Interpretation project number** (e.g., "4," indicating that this is the fourth interpretation project for you or your team that the exercise with this module is part of): **[Project#]**
* **Which stage of your present interpretation project does this module represent?** (E.g., if this module is the second you have used in the sequence of modules for the same interpretation project, then write the number "2" here. If multiple team members are working in parallel using various modules, then follow the convention of adding name-abbreviation—e.g., 2AL: **[Stage#]**
* **The *present* Interpretation Protocol module you are using is set as follows**:**2**(This is part of your auto-generated exercise ID below, and cannot be altered in the present module document.)
* **Today's date in the format YYYY-MM-DD (e.g., 2020-07-06)** (use the date-picker calendar): **[Date]**

**Your team folder in the project Google team working space.**
This refers to your team working folder for notes, reports, materials, etc. related to interpretation work in the WE1S Google Team drive’s [\_Team Working Space](https://drive.google.com/drive/u/1/folders/1er05JpagMNlfdnMIvcCVgwbweGkGeSua) folder. Go there to find the folder your team created for your work.

**[FolderURL]**

# Exercise ID

The ID for the present exercise (generated from the information you just provided) is as follows. (If you see a mistake, please go "back" in this survey and correct the information that generated the Exercise ID.)

[Name]**\_** [Project#]\_ [Stage#]\_2\_ [Date]

# Exercise Description

To assist in remembering what the exercise using this module is about, please enter below **a very brief description of what you are currently doing** (e.g., "We are now going to compare the keywords 'humanities' and 'sciences' in the 20190621\_2132\_us-humanities-top-newspapers---topics200 model.")

Click or tap here to enter text.

# Topic model you are studying

For the topic model you are studying, please enter below its **"topic model name"** and "**start page"** URL. A **"topic model name"** looks like this: *"20190621\_2132\_us-humanities-top-newspapers---topics200"*. You can locate this name in the [Registry of WE1S Topic Models](https://docs.google.com/spreadsheets/d/1n9nCzHut5aZEYRiSx5QD_yiiaLO9TeeQgwJJPd1ATt0/edit#gid=0). Note that this is the name of the specific granularity of the model you are studying (number of topics). The **"start page"** is the URL of the overall topic model (including all its granularities and their visualizations) listed on this page on the Harbor 10002 server: [Index of / projects/](http://harbor.english.ucsb.edu:10002/projects/). (You can also find this URL in the Registry of WE1S Topic Models.) A start page URL looks like this: http://harbor.english.ucsb.edu:10002/projects/20190621\_2132\_us-humanities-top-newspapers/

**Topic model name**: Click or tap here to enter text.

**"Start page" URL**: Click or tap here to enter text.

# Research question you are addressing

Please enter below information about the research question you are addressing in this exercise. You can find this information in the "Registry of WE1S Research Questions." (If you are taking an overview of a model prior to working on a specific research question, just enter "Taking an overview")

**Research question**:

Click or tap here to enter text.

**Operationalized form of question**:

Click or tap here to enter text.

**Research question ID** (use the format "team2-q1", representing in this example team 2's first research question): Click or tap here to enter text.

# Previous Interpretation Protocol modules you used to address your research question (if any) [optional]

Enter in sequence below any Interpretation Protocol modules you have already completed in previous steps of your current interpretation project. For example, if for steps 1-3 preceding this exercise (where the current exercise is now step 4 in the sequence) you previously used modules 3a, 3b, and 3c in that order, then enter those module numbers in that sequence.

 **Previous module sequence**: Click or tap here to enter text.

End of Block: INFO ABOUT YOU & EXERCISE (v. 2, created 9 June 2019, rev. 7 July 2019)

Start of Block: WHAT THIS MODULE DOES

# What module 2 does

(Module 2, created 19 June 2019; last rev. 20 June 2019.)

Module 2 of the WE1S topic model Interpretation Protocol guides you through examining representative topics and topic-clusters with more precision than similar exercises in Module 1  ("Take an overview of a model").

"Representative" means that you will be observing a range of topics/clusters from the important to those that are minor or outliers. Together with Module 1, Module2 gives you a good initial sense of a model as context for addressing specific research questions.

End of Block: WHAT THIS MODULE DOES

Start of Block: IDENTIFY THREE TIERS OF TOPICS

# Step 1. Identify selected topics at three levels of proportional weight

This part of this module asks you to identify topics at three levels of proportional weight in the model. “Proportional weight” means basically how important the topic is, or how often it turns up in the corpus of documents. (A bit more exactly, proportional weight refers to the percentage of the total number of words ["tokens"] in the corpus that the modelling process has assigned to a particular topic. It is also sometimes thought of as "conditional probability," referring to how likely it is that a topic will be encountered in a model based on a particular corpus of documents.)

**Why look for three levels of topics by weight?**
Typically, topics in models fall into several levels of proportional weight, which you might imagine as a set of terraces. For example, there will be a high terrace consisting of several topics at extremely high proportional weight (whose percentages are also often near each other). There are often several kinds of such big topics (explained below). Then there might be a noticeable drop-off to the next terrace below, where there is a series of topics whose percentages are in the middle range (and also often near each other in level). Then, beneath there will be lower terraces. Often, you will recognize the difference between terraces by the fact that, for example, one set of topics will have percentages ranging from 2.1 to 2.4; then there will be a noticeable shift in range to the next set with percentages, for example, of 1.5-1.8.

**Methods of finding and sorting topics by weight:** Use the general-purpose visualization interfaces in the WE1S Topic Model Observatory as follows (see [*TMO Guide*](https://we1s.ucsb.edu/research/we1s-tools-and-software/topic-model-observatory/tmo-guide/) for more detailed how-to's):

* In Dfr-Browser, go to the “List” tab under “Overview” and sort the topics in descending order of “proportion of corpus.” To sort, click on the label “proportion of corpus” or the small arrows next to that label at the right. Each topic’s percentage figure (its proportion of the corpus) shows at the right next to a blue bar graph roughly visualizing that percentage.


* In pyLDAvis, look at the relative size of the topic circles in the left-hand “Intertopic Distance Map” panel. Click on any individual topic circle to show its proportion of the corpus (e.g., “5.6% of tokens”) at the top of the right-hand panel of pyLDAvis.


* The size of the topic circles in TopicBubbles serves a similar purpose:



**For the purpose of this module, select for examination topics from three levels by proportional weight (high, middle, low) according to the instructions in the following sequence of questions**:

## ENTER HIGH-WEIGHT TOPICS

Enter in **ranked order** (where first is highest) the topic numbers for the weightiest topics in the **“high”** proportional weight level (up to a maximum of **10** topics).

1. ***Highest***-weight Topic #: Click or tap here to enter #.
2. High-weight Topic #: Click or tap here to enter #.
3. High-weight Topic #: Click or tap here to enter #.
4. High-weight Topic #: Click or tap here to enter #.
5. High-weight Topic #: Click or tap here to enter #.
6. High-weight Topic #: Click or tap here to enter #.
7. High-weight Topic #: Click or tap here to enter #.
8. High-weight Topic #: Click or tap here to enter #.
9. High-weight Topic #: Click or tap here to enter #.
10. High-weight Topic #: Click or tap here to enter #.

## ENTER MIDDLE-WEIGHT TOPICS

Identify the level of topics that seems to be one level down, which for the purpose of this exercise we will call the **“middle”-**weight level. Choose up to **5** such "middle"-weight topics that interest you or that seem representative and enter their topic numbers below:

1. Middle-weight Topic #: Click or tap here to enter #.
2. Middle-weight Topic #: Click or tap here to enter #.
3. Middle-weight Topic #: Click or tap here to enter #.
4. Middle-weight Topic #: Click or tap here to enter #.
5. Middle-weight Topic #: Click or tap here to enter #.

## ENTER LOW-WEIGHT TOPICS

Now in the Dfr-Browser topic "list" view (sorted by proportional weight of topics), jump down to the very bottom of the list to identify the **low** topics by proportional weight. Enter the topic numbers for up to 5 low-weight topics at that level that interest you:

1. Low-weight Topic #: Click or tap here to enter #.
2. Low-weight Topic #: Click or tap here to enter #.
3. Low-weight Topic #: Click or tap here to enter #.
4. Low-weight Topic #: Click or tap here to enter #.
5. Low-weight Topic #: Click or tap here to enter #.

End of Block: IDENTIFY THREE TIERS OF TOPICS

Start of Block: EXAMINE HIGH-WEIGHT TOPICS

# Step 2. Examine high-weight topics

Topics with very high proportional weight in topic models are often hard to interpret and require special assessment procedures. If possible, separate the 10 high-weight topics you identified above into three classes, which can be called **“super-topics,”** **“meta-topics,”** and **“understandable top topics.”**

A **super-topic** is typically an extremely high-weight topic characterized by words that are hard to understand as a cohesive group and often seem to represent a general, existential, daily, ordinary, or broadly common stratum of social, spatial, and temporal experience (e.g., *"day back night man year told home asked room days morning high make hard made"*). A super-topic is hard to understand even after sampling some of the documents associated with the topic.

By contrast, a **meta-topic** is typically characterized by topic words that--while also hard to understand as a cohesive group (e.g., containing the same kinds of existential social, space, and time words)--may represent categories or registers of discourse:  i.e., the discourses of different social classes, media types, genres, or even sections of newspapers [like “Entertainment” or “Sports”]). For instance, if words like "lol" or "sucks" are top words in such topics, then there is a good chance that the topics are *discourse types* rather than *themes* per se related to students, popular culture, Reddit, etc. Looking at documents associated with the topic can give a better sense of these broad kinds or classes of language.

An **understandable top topic** is distinguished from both super- and meta-topics by being a high-weight topic whose topic words (and associated documents) seem to cohere around understandable themes or issues that you can fairly easily label in a short phrase (e.g., *“higher education programs”* or *“business and careers”*).

## PARSE HIGH-WEIGHT TOPICS

For each of the 10 **high-weight** topics you previously identified please examine the topic as follows:

* Look at the 10 most frequent words in the topic. Look quickly at the top 10 articles associated with the topic (and read a bit more closely the top 5 articles).
* Then try to separate out **“super-topics”** from **“meta-topics”** from **“understandable top topics,”** and record their topic numbers below. (Not all fields for each kind of topic need to be filled.)

***Reminder*:** [**High-weight topics you listed in ranked order by weight**
*(jump to previous location in this document)*.](#_ENTER_HIGH-WEIGHT_TOPICS)

### Super-topics (identify by topic number): Click or tap here to enter text.

### Meta-topics (identify by topic number): Click or tap here to enter text.

### Understandable-topics (identify by topic number): Click or tap here to enter text.

## OBSERVE HIGH-WEIGHT TOPICS (minus the non-useful “super-topics)

Now that you have sorted the high-weight topics into the categories of "super-topics," "meta-topics," and "understandable top topics," you will make some observations about the latter two categories.
*(For this purpose, we will now put aside and ignore the "super-topics" unless you see anything interesting there, in which case add a comment in the Running Note below.)*

### OBSERVE META-TOPICS

**Observation on meta-topics**: write a brief observation of the “**meta-topics**," (either generalizing about all the ones you found or focusing on a single example). ***Reminder*:** [**Meta-topics you identified** *(jump to previous location in this document)*.](#_Meta-topics_(identify_by)

*An “observation” should have two parts: a more-or-less objective description in which you have high confidence; and a thesis, hypothesis, interpretation, or suggestion for which you have less confidence before you study the model more closely but that might suggest avenues of research.* [RunningNote1]

### Label meta-topics

Try to label the meta-topics you have discovered.
(*A label should be a word or short phrase representing only the objective part of your observation, or what the topic is “about”. It should not be formed as a sentence, which tends to posit a thesis.)*

|  |  |
| --- | --- |
| **Topic # of meta-topic**(add rows as needed) | **Label** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

### OBSERVE UNDERSTANDABLE TOPICS

**Observation on understandable top topics**: write a brief observation of the “**understandable top topics**," (either generalizing about all the ones you found or focusing on a single example). ***Reminder*:** [**Understandable topics you identified** *(jump to previous location in this document)*.](#_Understandable-topics_(identify_by)

*An “observation” should have two parts: a more-or-less objective description in which you have high confidence; and a thesis, hypothesis, interpretation, or suggestion for which you have less confidence before you study the model more closely but that might suggest avenues of research.* [RunningNote2]

### Label understandable topics

Try to label the understandable top topics you have discovered.
*(A label should be a word or short phrase representing only the objective part of your observation, or what the topic is “about”. It should not be formed as a sentence, which tends to posit a thesis.)*

|  |  |
| --- | --- |
| **Topic # of understandable-topic**(add rows as needed) | **Label** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

End of Block: EXAMINE HIGH-WEIGHT TOPICS

Start of Block: EXAMINE MIDDLE-WEIGHT TOPICS

# Step 3. Examine middle-weight topics

Next, let's look at the middle-weight topics you identified.

## OBSERVE MIDDLE-WEIGHT TOPICS

**Observation on middle-weight topics**: write a brief observation of the middle-weight topics you identified (either generalizing about all the ones you found or focusing on a single example). ***Reminder*:** [**Middle-weight topics you identified** *(jump to previous location in this document)*.](#_ENTER_MIDDLE-WEIGHT_TOPICS)

*An “observation” should have two parts: a more-or-less objective description in which you have high confidence; and a thesis, hypothesis, interpretation, or suggestion for which you have less confidence before you study the model more closely but that might suggest avenues of research.*

[RunningNote3

### Label middle-weight topics

Try to label the middle-weight topics you have discovered.
(*A label should be a word or short phrase representing only the objective part of your observation, or what the topic is “about”. It should not be formed as a sentence, which tends to posit a thesis.)*

|  |  |
| --- | --- |
| **Topic # of middle-weight topic**(add rows as needed) | **Label** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

End of Block: EXAMINE MIDDLE-WEIGHT TOPICS

Start of Block: EXAMINE LOW-WEIGHT TOPICS

# Step 4. Examine low-weight topics

Next, let's look at the low-weight topics you identified.

## OBSERVE LOW-WEIGHT TOPICS

**Observation on low-weight topics**: write a brief observation of the middle-weight topics you identified (either generalizing about all the ones you found or focusing on a single example). ***Reminder*:** [**Low-weight topics you identified** *(jump to previous location in this document)*.](#_ENTER_LOW-WEIGHT_TOPICS)

*An “observation” should have two parts: a more-or-less objective description in which you have high confidence; and a thesis, hypothesis, interpretation, or suggestion for which you have less confidence before you study the model more closely but that might suggest avenues of research.*

[RunningNote4]

### Label low-weight topics

Try to label the middle-weight topics you have discovered.
(*A label should be a word or short phrase representing only the objective part of your observation, or what the topic is “about”. It should not be formed as a sentence, which tends to posit a thesis.)*

|  |  |
| --- | --- |
| **Topic # of low-weight topic**(add rows as needed) | **Label** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

End of Block: EXAMINE LOW-WEIGHT TOPICS

Start of Block: EXAMINE CLUSTERS

# Step 5. Examine clusters of topics in the model

Next, let's see if we can recognize any meaningul "clusters" of topics (topics that are statistically "near" each other in the words and documents associated with them and "far" or separated from other clusters).

Using the specialized "clustering analysis" tools named Clusters7D and DendrogramViewer in the WE1S Topic Model Observatory (see [*TMO Guide*](https://we1s.ucsb.edu/research/we1s-tools-and-software/topic-model-observatory/tmo-guide/) for how-to's and best practices), see if you can identify clusters of topics in the model. A checklist of things you might want to look for:

* Topic-clusters in Clusters7D that bulk large in the model (respresented as larger circles in the left panel)
* Topic-clusters in DendrogramViewer with many "leaves" in a "clade"
* Topic-clusters that seem clearly separated from others

 ***Important methodological note***: The primary value of topic models lies in the topics they surface. You should always have less confidence in the existence and meaningfulness of topic "clusters" shown in visualization interfaces representing in a 2D-graph space the statistical "nearness" of topics. (Clusters7D and DendrogramViewer are more specialized and trustworthy for this task; the "scaled" views in Dfr-browser or TopicBubbles and "intertopic distance map" view in pyLDAvis are less so.) The reason is that by definition "topics" in a topic model are already a *reduction* of the complexity of the underlying corpus of documents (they reduce the myriad relations of thought and language in those documents to a much smaller number of "topics"). "Clusters" of topics are then a *second-level reduction* or *reduction-of-a-reduction* (they reduce the topics into a much smaller set). With each step in the reduction series, we gain meaningfulness ("these documents are 'about' *this*") at the cost of confidence ("does that cluster really exist?").

If you find a topic-cluster that is really interesting or important for your research question, there are a number of things you can do to increase your confidence in the cluster--e.g., look at a low-granularity version of your topic model (say, 50 topics instead of 250) and see if a topic-cluster in the higher-granularity model resolves into a single topic in the lower-granularity model.

(WE1S Interpretation Protocol modules 3.b and 4.b include enhanced, more precise exercises in topic-cluster analysis that try to optimize the interpretive usefulness and confidence-level of "clusters.")

**Once you have looked over the possible clusters in the model, answer the following sequence of questions.**

## OBSERVE CLUSTERS

**Observation on topic clusters:** Write a brief observation about major or interesting clusters of topics you see.

*An “observation” should have two parts: a more-or-less objective description in which you have high confidence; and a thesis, hypothesis, interpretation, or suggestion for which you have less confidence before you study the model more closely but that might suggest avenues of research.*

[Running Note 5]

## LABEL CLUSTERS

**Label topic clusters:** Label up to three of the interesting or major clusters you observed above. ( (A label should be a word or short phrase representing only the objective part of your observation, or what the topic is “about”. It should not be formed as a sentence, which tends to posit a thesis.)

|  |  |
| --- | --- |
| **Cluster** | **Label** |
| 1 |  |
| 2 |  |
| 3 |  |

# Step 5. Summary observations to complete this module

Now that you have conducted the exercises in this Module 2 for exploring representative topics and clusters, write an observation that synthesizes your understanding of this topic model so far.
An “observation” should have two parts: a more-or-less objective description in which you have high confidence; and a thesis, hypothesis, interpretation, or suggestion for which you have less confidence before you study the model more closely but that might bear on your research question.

**To assist you in making a final, summary observation, here are your notes so far from earlier sections of this module. (These are editable; editing will also change your previous entries in this document.)**

[RunningNote1]

[RunningNote2]

[RunningNote3

[RunningNote4]

[Running Note 5]

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

**Enter below your final, summary observations:**

[FinalNote]

End of Block: SUMMARY OBSERVATIONS

Start of Block: RESEARCH TAKEAWAYS (v. 2, created 23 June 2019; rev. 27 June 2019)

|  |
| --- |
|  |

## TAKEAWAYS: Before you finish: Research Takeaways?

After studying your topic model in this exercise, are there any research takeaways you want to record, store, describe, or log (depending on the nature of the takeaway) to provide key evidence, materials, and data that can be used in writing up a report on your research question or for future humanities advocacy? For example, a research takeaway can be:

* Evidence in the form of an important topic(s) or relation between topic(s) and words, documents that you want to describe so that you can find it again;
* A screenshot of something you found while analyzing your topic model with one of the Topic Model Observatory visualization interfaces;
* Citations for or excerpts from key documents associated with a topic (e.g., three newspaper articles you want to use in a report);
* A text copy of words in a topic ranked by proportional weight;
* Etc.

If you have found such research takeaways, record, store, or describe them as appropriate in your team folder on the WE1S Google team drive. You may also want quickly to log or make a note about your takeaways in the field below in order to have a record of them as associated with this survey. That way, when you read over your survey results, you can remember what you took away and where you stored or annotated it. P.S. Some of the Topic Model Observatory interfaces may also allow you to use the [Hypothes.is](https://web.hypothes.is/) web annotation tool to highlight or annotated them in your browser for private or shared use.

## Companion questions?

"Companion questions" may be defined as questions that help provide context for the research question you are currently addressing or that are follow-on questions you might want to pursue later. WE1S asks you periodically to consider if companion questions have occurred to you because it's a good way to prevent the kind of "tunnel vision" that research can sometimes create.

This is best explained by way of example. Suppose that your research question concerns the degree and nature of discussion about the "humanities crisis." A tunnel-vision view of the problem would be to consider only discussion of the "humanities crisis." A contextually wider view would consider companion questions such as the following: what is the degree and nature of discussion about *other* kinds of crisis by comparison? For example, in what other contexts (political, economic, spiritual, etc.) does the word "crisis" tend to appear? Do mentions of "crisis" in all these contexts tend to spike up together, or be concentrated in certain kinds of media or sources? Etc.

Having advanced on your research question using the present module, have any "companion questions" occurred to you? I**f so, please set them down here so that they can be remembered for possible future use:**

End of Block: RESEARCH TAKEAWAYS (v. 2, created 23 June 2019; rev. 27 June 2019)

Start of Block: SURVEY END (v.2, created 20 June 2019, rev. 28 June 2019)

## CONFIDENCE? Final consideration -- Confidence assessment

It is possible after conducting an interpretation exercise to conclude that you have either high or low confidence in the results (with confidence descending all the way down to a null result).

Please rank your level of confidence in the results of this interpretation exercise on the following scale:

1 ----------------------------------------------------- to ------------------------------------------------- 10

 (1 = null result) (10 = full confidence)

Choose a confidence value.

## Suggestions about this module?

Do you have any suggestions about this module of the WE1S Interpretation Protocol (problems, improvements)?

## **Important**: Save this completed module in your team working folder.

Please save this completed module of the Interpretation Protocol in your team’s working folder:

**[FolderURL]**

This module is part of the documentary evidence trail for your future research reports and Key Finding cards.

END **This is the end of the current module of the WE1S Interpretation Protocol.**

End of Block: SURVEY END (v.2, created 20 June 2019, rev. 28 June 2019)