Guide to Cluster Naming and Cluster Name Evaluation

Table of Contents

Definitions

- **Cluster**: a set of documents which a machine learning algorithm has grouped together due to commonality in the numerical representation of their text's semantic content.
- **Cluster Name**: a short phrase that describes the topic of a cluster.
- **Rubric**: a scoring tool to evaluate something (in our case, a cluster name) in a consistent, structured way.
- **Topic**: the underlying concept which a cluster of documents represents. Sometimes confusingly used as a synonym for "cluster" – I'm trying to remove instances of that throughout this project, but you may still encounter some.

Acronyms

- **LLM**: large language model
- **NSF**: National Science Foundation
- **PMC**: PubMed Central

Project Overview

Text clustering is a way to organize unstructured text data by identifying groups within a set of documents. We use text clustering for a variety of problems, such as qualitative coding of open-text survey responses. Once clusters are generated, they must be named and described to be useful. This often involves manual review by subject matter experts, which is costly and slow. The goal of this project is to assess the feasibility of using large language models (LLMs) like ChatGPT to automate the cluster naming process.

The objectives of the project are:

1. Source two relevant, public text datasets,
2. Cluster the datasets using a workflow typical of RTI text clustering projects,
3. **Manually generate names for a sample of clusters,**
4. Use LLMs to automatically generate names for the same sample of clusters,
5. Develop an evaluation rubric to assess cluster name quality, and
6. **Assess human-generated and model-generated names against the rubric.**

The annotation tasks described in this guide – cluster naming and cluster name evaluation – correspond to objectives 3 and 6. *The evaluation process is where this project adds the most value.* Lots of people are exploring the use of LLMs for all kinds of tasks, but few are rigorously evaluating their performance.

Datasets

We sourced two publicly available datasets to use as benchmarks for this project.

- **NSF Award Abstracts** – these are the abstracts from proposals funded by the National Science Foundation (NSF) in 2022. They outline proposed research across a wide variety of fields. This dataset contains about 7,500 texts, which resulted in 123 clusters.
- **PMC Clinical Notes** – these are patient notes extracted from articles published on PubMed Central (PMC). They describe conditions, treatments, and outcomes across medicine. This dataset contains about 167,000 texts, which resulted in 274 clusters.

Annotation Tasks

This section describes the annotation tasks at a high level. Detailed guidance on completing these tasks is below in "Annotation Guidance".

Cluster Naming

The first annotation task is cluster naming—the task that this project ultimately hopes to streamline. In current text clustering projects, we generate cluster names manually. In this project, we will generate cluster names manually, as well as using LLMs, and then compare the human-generated and model-generated cluster names.

A ***cluster name*** is a short phrase that describes the topic of a cluster. The format can vary across domains, but in general, a cluster name should have the following traits:

- Short (anywhere from a single word up to a short phrase),
- Easy to read and understand,
- Sums up the important traits of documents in the cluster,
- Does not include traits that are only true for a minority of documents in the cluster, and
- The right level of detail for the cluster.

For the cluster naming annotation task, annotators will review documents from each cluster and assign a name to the cluster. Examples and detailed instructions are in the Annotation Guidance section below.

Cluster Name Evaluation

The second annotation task is cluster name evaluation—rating the quality of human-generated and model-generated cluster names. After the cluster naming phase, we will have several candidate names for each cluster, one generated by a human, and several others generated by LLMs. To evaluate the cluster names, annotators will score each cluster name on a ***cluster name evaluation rubric***. The rubric assesses various aspects of the cluster name's quality. Annotators will also pick the overall best name for the cluster and assess the quality of the cluster itself.

The cluster names will be blinded, so annotators do not know whether they were generated by a human or an LLM. Annotators will only review names for clusters which were named by another annotator (i.e., they will never evaluate their own cluster names). Examples and detailed instructions are in the Annotation Guidance section below.

Resources

Naming spreadsheets

- Each annotator has a separate spreadsheet to name the clusters assigned to them.
- Data dictionary:
  - Cluster: the cluster number, which you can use to look up the cluster in the application
  - Count: the number of documents in the cluster
  - Keywords: terms which distinguish this cluster from others, as identified by the clustering algorithm
  - Cluster name: enter your cluster name here
  - Notes: enter any notes related to the naming process or logic here

Evaluation spreadsheets

- Each annotator has a separate spreadsheet to evaluate cluster names.
- Data dictionary:
  - cluster: the cluster number, which you can use to look up the cluster in the application
  - name_id: unique identifier for the cluster name
  - name: the cluster name
  - **See "Evaluation Rubric" below for details on the rest of the columns**

Annotation Guidance

This section describes the annotation tasks in detail, with instructions and examples.

Cluster Naming

Process Overview

The suggested workflow for cluster naming is as follows.

1. Open the naming spreadsheet with your name in the file name.
2. Open the NSF and PMC Cluster Explorer Application.
3. For each row:
   a. Read the cluster keywords to get a first impression of the cluster's content.
   b. Note the number of documents in the cluster.
   c. Look up the cluster in the NSF and PMC Cluster Explorer Application.

d. Read the default sample of 10 documents.

   i. By default, the first three documents shown are "representative documents", which the clustering algorithm thinks are most representative of the cluster's topic. Pay special attention to these, but don't get tunnel vision – they may not represent the full range of the cluster's content, especially in larger clusters.

e. If, after reading the first 10 documents, you are still unsure of the cluster's topic, continue to read more documents until you either feel sure of the cluster's topic or are getting diminishing returns.

   i. The larger the cluster, the more documents you should read to make sure you are getting a representative sample.

f. Once you are as confident as you think you can be, enter a cluster name in the spreadsheet.

g. Feel free to add any notes about your naming process or logic in the Notes column (optional).

Examples

Example 1 (NSF)

- Cluster traits
  - Dataset: NSF Proposal Abstracts
  - Cluster ID: 5
  - Number of documents in cluster: 125
  - Cluster keywords: mathematics, teachers, teaching, student learning, courses, reasoning, stem education, program supports, research development, engineering students
  - Sample documents:
    - *Representative Document 1:* "Applying and Refining a Model for Dynamic, Discussion-Based Professional Development for Middle School Teachers about Fractions, Ratios and Proportions: This project explores the effectiveness of two different versions of professional development (PD) designed to enhance middle school mathematics teachers' understanding of fractions and proportions, and their teaching of these mathematical concepts to students.."
    - *Representative Document 2:* "Developing and Testing a Learning Progression for Middle School Physical Science incorporating Disciplinary Core Ideas, Science and Engineering Practices, and Crosscutting Concepts: This project will develop and test a learning progression for middle school physical science that incorporates the three dimensions identified in Next Generation of Science Standards…"
    - *Representative Document 3:* "Reframing Students' Graph Literacy with a Focus on Students' Thinking: Graph literacy, the ability to comprehend, interpret, and use graphical representations, is critical

for students to learn mathematics, to succeed in STEM coursework and careers, and to engage in informed participation in society…”

    - *Other Document 1:* “Comparing student success, team dynamics, and cost in three different active learning formats in undergraduate physics education: This project aims to serve the national interest by promoting student success in an introductory college physics course. This project plans to implement and compare three different active learning formats in physics teaching…:”
    - *Other Document 2:* “Assessing Student Satisfaction and Engagement in Teams (ASSET): An Empirical Review and Scale Development: Prior research suggests that group differences, such as gender, race/ethnicity, and academic discipline influence the way that students perceive course interactions such as teamwork. These differences may lead to inequities in team experiences, student engagement, and learning for different groups of students. It is helpful to the aims of NSF’s Broadening Participation in Engineering program, to the field of engineering, and to society as a whole to better understand these differences…”
    - *Additional documents not shown to save space.*

- Discussion
  - The keywords and sample documents show a clear pattern. These abstracts propose research related to the teaching of STEM fields. However, this is a fairly large cluster, containing 125 documents. The annotator could try reading a few more documents to ensure that there are no other aspects of the cluster that the sample documents have missed. However, the sample documents are close enough that generating a name based on this sample would also be a reasonable decision.
  - **Suggested name: STEM Education**

Example 2 (PMC)

- Cluster traits
  - Dataset: PMC Clinical Notes
  - Cluster ID: 10
  - Number of documents in cluster: 1,151
  - Cluster keywords: bladder, ureter, renal, ureteral, stone, left kidney, cystoscopy, urothelial, nephrectomy, renal pelvis
  - Sample documents:
    - *Representative Document 1:* “A 65-year-old male presented to our hospital with a 30 years history of right flank pain… Abdominal CT demonstrated significant dilatation of the right renal pelvis and the right upper and mid-ureter with multiple calculi… Then a right-side percutaneous nephrolithotomy (PCNL) was performed…”
    - *Representative Document 2:* “A 60-year-old male patient presented with complaints of pain in abdomen for 15 days… Calculus

        (20 mm) seen in the left renal pelvis caused moderate left hydronephrosis… Hence, the decision for laparoscopic minimally invasive PCNL (Mini-PERC) was taken. Laparoscopy was performed…”

        - *Representative Document 3:* “A 57-year-old male presented with right side abdominal pain for two months. He had the history of right side pelvicaliceal calculus of 26 x 18 mm2 removed percutaneously by nephrolithotomy 10 years before the visit. He also had the history of recurrent renal calculi thereafter and managed accordingly…”
        - *Other Document 1:* “A 45-year-old man was admitted for the evaluation of recurrent abdominal pain. He had a 10-year history of hemodialysis for chronic renal failure due to ADPKD… A right radical nephrectomy was performed with the presumptive clinical diagnosis of RCC… 16 more solid tumor masses were scattered throughout the kidney”
        - *Other Document 2:* “An 8-month-old girl without specific past medical history was admitted to the hospital due to nausea and diarrhea. On physical examination, a firm mass was detected in the left upper quadrant of the abdomen. Abdominal computed tomography (CT) revealed a left renal mass without lymph node enlargement. There were no suspicious metastatic lesions in chest CT. Laboratory data showed leukocytosis. Radical nephrectomy with lymph nodes dissection was performed…”
        - *Additional documents not shown to save space.*
- Discussion
    - Looking at the representative documents alone would suggest that this cluster could be named “kidney stones” However, the other documents discuss other forms of kidney surgery. Also, this is a very large cluster, with 1,151 documents (the PMC clusters in general are much larger than the NSF clusters). The annotator should read several more sampled documents to assess a few questions:
        - What proportion of documents discuss kidney stones vs. other renal surgeries?
        - “Bladder” is the top keyword – do other documents discuss bladder-specific things or is that just due to its relation to the kidneys?
        - Do all documents discuss surgery, or are other renal topics discussed?
    - **Suggested name (assuming additional documents are similar to the first sample): Renal surgery.**

Edge Cases

- **Large Cluster**
    - Large clusters tend to have more variance in the documents than smaller ones.

    - The larger the cluster, the more documents you should read to make sure you are getting a representative sample.
    - See Example 2.
- **Poor Quality Cluster**
    - Some clusters are better than others!
    - Poor quality clusters might include documents that aren't obviously related to one another.
    - Clusters might include documents from more than one topic, e.g., particle physics and astronomy, or pancreatic cancer and colon cancer.
        - In these cases, consider names which reflect the multiple topics.
        - "Particle Physics and Astronomy" and "Pancreatic and Colon Cancer" are both reasonable cluster names for these cases.
    - Alternatively, a cluster might have some documents with a common topic, but many outlier documents which are all over the place in terms of topic.
        - In these cases, focus the name on the documents which have a common topic.
        - For example, if half the documents discuss particle physics and the other half discuss a wide variety of unrelated topics, name the cluster "Particle Physics".
    - Just do your best! It's hard to name poor quality clusters. The cluster name evaluation step will include an evaluation of each cluster's quality, so we will be able to correlate name quality with cluster quality.
- **Outlier Documents**
    - This is closely related to poor quality clusters. The more outlier documents, the lower the cluster quality.
    - If you notice outlier documents, read more documents to get a sense of how common the outliers are, and whether they form a common theme.
    - See Example 2 above, where kidney-related topics other than kidney stones might initially appear to be outliers, but in fact demonstrate that the cluster topic is broader than just kidney stones.
- **Unfamiliar Topic/Jargon**
    - You may encounter documents or entire clusters that use terminology or discuss a topic that you are not familiar with.
    - Google is your friend here. Search for unfamiliar terms until you get at least a basic understanding of the topic.
    - Again, just do your best. Even if you are not confident that you fully understand a cluster, any attempted name is better than "I don't know."
    - You may be able to name a cluster without fully understanding its content. For example, if a cluster of patient notes all discuss "pneumothorax", you could probably name that cluster without knowing what a pneumothorax is.

Cluster Name Evaluation

Process Overview

The suggested workflow for cluster name evaluation is as follows.

1. Open the evaluation spreadsheet with your name in the file name.
2. Open the NSF and PMC Cluster Explorer Application.
3. For each cluster:
    a. Filter the evaluation spreadsheet so you're only viewing the names for a single cluster.
    b. Look up the cluster in the NSF and PMC Cluster Explorer Application.
    c. Follow the same process as in the naming phase to get a sense of the cluster's content. I.e.:
        i. Note the keywords and cluster size.
        ii. Read the default sample of 10 documents.
        iii. Continue reading documents until you're as confident as you can be in your understanding of the cluster's content.
    d. Rate the cluster's overall quality in the **cluster_quality** column. Consider:
        i. Cohesiveness
        ii. Frequency of outliers
        iii. Whether the cluster could be split into multiple topics
    e. Rate each cluster name across the **name_** columns.
        i. Refer to the rubric for definitions of each column.
    f. Pick the name you think is best and enter a **1** in the **best_name** column for that row. Enter a **0** in the **best_name** column for all other rows, or just leave them blank.
    g. Feel free to add any notes about your evaluation process or logic in the **notes** column (optional).

Examples

Example 1 (NSF)

- Cluster traits for this example are the same as Example 1 in the Naming section.
- Names
    - STEM education
    - Science education
    - Mathematics education and science education
    - Middle school STEM education
    - Education
- Discussion
    - Cluster quality: this seems to be a very cohesive cluster with a single, specific topic and few outliers. Assuming there is not a higher rate of outliers or another topic emerging upon reading further documents, cluster quality is **Very Good**.
    - Name evaluation:

        - Some names may be very similar. **STEM education**, **Science education**, and **Mathematics education and science education** are all adequate, but **STEM education** is slightly better, because it is both more concise and more comprehensive. It should rate higher on completeness and fluency.
        - **Middle school STEM education** is too specific; not all documents in this cluster relate to middle school. It should rate lower on completeness and overall quality.
        - **Education** is too broad. It should rate lower on relevance and overall quality.

Example 2 (PMC)

- Cluster traits for this example are the same as Example 2 in the Naming section.
- Names
    - Kidney stones
    - Kidney diseases
    - Renal surgery
    - Kidney stones or kidney cancers or radial nephrectomy
    - Kidney and bladder topics
- Discussion
    - Cluster quality:
        - Let's assume that upon reading further documents, about half the documents discuss kidney stones, and the other half discuss other kidney-related diseases. Almost all the documents involve surgery of some kind.
        - This cluster should probably be split into two clusters: kidney stones and other kidney surgery.
        - However, there are few outliers beyond those two topics.
        - Cluster quality should be **Neutral**: in one way it is poor (should be two clusters), but in another way it is good (cohesive kidney surgery topic).
    - Name evaluation:
        - **Kidney stones** is too specific because many documents discuss other kidney surgeries.
        - **Kidney diseases** and **Kidney and bladder topics** are too broad because they don't capture that all these documents discuss surgery.
        - **Kidney stones or kidney cancers or radial nephrectomy** is accurate but verbose. This sort of compound name can be the best in some situations, but if a more concise name (like **renal surgery**) can capture the content of the cluster equally well, the more concise name is preferred, and should rater higher on fluency and overall quality.

Edge Cases

- **Identical names**
  - Sometimes multiple naming methods may have generated the same name.
  - Ensure that identical names have the same fluency, consistency, relevance, completeness, and overall name quality ratings.
  - If the best name was generated by more than one method, enter 1 in **best_name** for each instance. This is the only circumstance in which you should enter 1 for more than one name per cluster!
- **Unfamiliar Topic/Jargon**
  - You may encounter cluster names, documents, or entire clusters that use terminology or discuss a topic that you are not familiar with.
  - Google is your friend here. Search for unfamiliar terms until you get at least a basic understanding of the topic.

Evaluation Rubric

The evaluation rubric provides more detail on the domains assessed in the cluster name evaluation process. It also serves as a data dictionary for the cluster name evaluation spreadsheets.

Our evaluation domains are drawn from the document understanding and summarization literature. Because cluster naming is slightly different from summarization, we have tweaked the domains slightly. Sources include Dang 2005; Kryscinski et al 2020; and Fabbri et al 2021.

Evaluation Domains

- Cluster Quality
  - Is the cluster coherent enough to give it a high-quality name?
  - Negative formulation: is the cluster too broad, disjointed, etc., to give a high-quality name?
- Name Fluency
  - Is the quality of the writing good?
  - Negative formulation: are there formatting problems, grammatical errors, or anything that makes the name difficult to read or understand?
- Name Consistency
  - Is the name factually aligned with the source documents?

    - Negative formulation: does the name imply anything that is not supported by the source documents?

- Name Relevance

    - Does the name reflect ONLY important content from the source documents?
    - Negative formulation: does the name include tangential aspects of the source documents, or aspects that only apply to a small portion of the source documents?

- Name Completeness

    - Does the name reflect ALL important content from the source documents?
    - Negative formulation: is the name missing key aspects of the source documents?

- Overall Name Quality

    - Is the name reasonable to use in place of a human-generated name?
    - Does the name give a good overall sense of the cluster's contents?

- Best Name

    - Considering all the above domains, which name is the best overall?

Scales

Cluster quality and name quality are evaluated on a five-point Likert scale: **Very good, Good, Neutral, Poor, Very Poor**.

Because each name gets its own row, there are five rows per cluster. The cluster quality rating should be the same for all five rows.

The best name is evaluated as a binary 0/1. Pick one and only one name as the best for each cluster. Enter a 1 in that row. Enter 0 in the other four rows or leave them blank. Choosing the best cluster name is a forced choice. Even if it's difficult to pick between several similar names, you must choose one as the best.

The only exception is if multiple names are identical, in which case you may enter 1 for multiple rows.