Tag Evaluation Questionnaire

Dear Evaluator,

Thank you for participating in this tag evaluation study. Please follow the instructions carefully for each section.
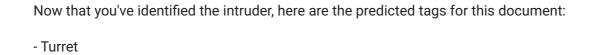
1. What is your age range?

   Mark only one oval.

   - ⬭ 18-24
   - ⬭ 25-34
   - ⬭ 35-44
   - ⬭ 45-54
   - ⬭ 55-64
   - ⬭ 65 or older

2. What is your gender?

   Mark only one oval.

   - ⬭ Male
   - ⬭ Female
   - ⬭ Non-binary
   - ⬭ Prefer not to say
   - ⬭ Other: _____

3. What is the highest level of education you have completed?

   Mark only one oval.

   ◯ High school or equivalent

   ◯ Some college

   ◯ Bachelor's degree

   ◯ Master's degree

   ◯ Doctoral degree

4. What is the highest level of English proficiency you have achieved?

   Mark only one oval.

   ◯ A1 (Beginner)

   ◯ A2 (Elementary)

   ◯ B1 (Intermediate)

   ◯ B2 (Upper Intermediate)

   ◯ C1 (Advanced)

   ◯ C2 (Proficient)

   ◯ Native speaker

5. Please read the following document called **League of Legends Solo Queue Ranked** *
Games carefully:

"League of Legends is a multiplayer online battle arena where two teams, blue and red, face off. There are three lanes, a jungle, and five roles. The goal is to take down the enemy Nexus to win the game.

This dataset contains statistics of ranked games from a Platinum Elo. Each game is unique and can be identified by its gameId, which can be used to fetch more attributes from the Riot API. Each game has features from different time frames, from ten minutes to the end of the game. For example, game one at ten minutes, game one at twelve minutes, game one at fourteen minutes, and so on. In total, there are over two hundred forty thousand game frames. There are fifty-five features collected for the Blue team, including kills, deaths, gold, experience, and level. It is up to the user to do some feature engineering to get more insights."

Now, examine the list of tags below. One of these tags is an 'intruder' that doesn't belong with the other tags. Please identify this intruder tag:

Mark only one oval.

○ Turret

○ Lost

○ Destroyed

○ Passenger

○ Impressions

○ Clicking

○ Cost

○ Advertisement

Now that you've identified the intruder, here are the predicted tags for this document:

- Turret

- Lost

- Destroyed

- Impressions

- Clicking

- Cost

- Advertisement

And here is the document again:

"League of Legends is a multiplayer online battle arena where two teams, blue and red, face off. There are three lanes, a jungle, and five roles. The goal is to take down the enemy Nexus to win the game.

This dataset contains statistics of ranked games from a Platinum Elo. Each game is unique and can be identified by its gameId, which can be used to fetch more attributes from the Riot API. Each game has features from different time frames, from ten minutes to the end of the game. For example, game one at ten minutes, game one at twelve minutes, game one at fourteen minutes, and so on. In total, there are over two hundred forty thousand game frames. There are fifty-five features collected for the Blue team, including kills, deaths, gold, experience, and level. It is up to the user to do some feature engineering to get more insights."

6. How well does each tag represent the main themes of the document? *

Mark only one oval per row.

	1 - Not at all	2 - Somewhat well	3 - Moderately well	4 - Very well	5 - Extremely well
Turret	◯	◯	◯	◯	◯
Lost	◯	◯	◯	◯	◯
Destroyed	◯	◯	◯	◯	◯
Impressions	◯	◯	◯	◯	◯
Clicking	◯	◯	◯	◯	◯
Cost	◯	◯	◯	◯	◯
Advertisement	◯	◯	◯	◯	◯

7. How specific or general is each tag? *

Mark only one oval per row.

	1 - Very specific to this document	2 - Somewhat specific	3 - Balanced	4 - Somewhat general	5 - Very general, could apply to many documents	Not Applicable/Tag is irrelevant
Turret	○	○	○	○	○	○
Lost	○	○	○	○	○	○
Destroyed	○	○	○	○	○	○
Impressions	○	○	○	○	○	○
Clicking	○	○	○	○	○	○
Cost	○	○	○	○	○	○
Advertisement	○	○	○	○	○	○

8. How well do all the tags together cover the main themes of the document? *

1 - Not well at all, misses all major themes
2 - Somewhat well, captures some major themes
3 - Moderately well, captures about half of the major themes
4 - Very well, captures most major themes
5 - Extremely well, captures all major themes

Mark only one oval.

1	2	3	4	5
○	○	○	○	○

9. If you selected any rating other than **5: Extremely well**, please suggest additional tags that would improve the coverage of the document's themes:

_____

10. Please read the following document called **FOREX USD/JPY Minute High** carefully: *

"This dataset contains historical price data of the FOREX USD/JPY from Dukascopy. Each instance, or row, represents one candlestick of one minute. The dataset spans from January first to December thirteenth and does not include weekends, as the FOREX market is not traded on weekends. The timezone of the feature Timestamp is Europe/Amsterdam."

Now, examine the list of tags below. One of these tags is an 'intruder' that doesn't belong with the other tags. Please identify this intruder tag:

Mark only one oval.

○ Thyrotropin

○ Minute

○ USD

○ Jump

○ Releasing

○ High

○ Bid

○ Ask

Now that you've identified the intruder, here are the predicted tags for this document:

- Thyrotropin

- Minute

- USD

- Releasing

- High

- Bid

- Ask

And here is the document again:

"This dataset contains historical price data of the FOREX USD/JPY from Dukascopy. Each instance, or row, represents one candlestick of one minute. The dataset spans from January first to December thirteenth and does not include weekends, as the FOREX market is not traded on weekends. The timezone of the feature Timestamp is Europe/Amsterdam."

11. How well does each tag represent the main themes of the document? *

Mark only one oval per row.

	1 - Not at all	2 - Somewhat well	3 - Moderately well	4 - Very well	5 - Extremely well
Thyrotropin	◯	◯	◯	◯	◯
Minute	◯	◯	◯	◯	◯
USD	◯	◯	◯	◯	◯
Releasing	◯	◯	◯	◯	◯
High	◯	◯	◯	◯	◯
Bid	◯	◯	◯	◯	◯
Ask	◯	◯	◯	◯	◯

12. How specific or general is each tag? *

Mark only one oval per row.

	1 - Very specific to this document	2 - Somewhat specific	3 - Balanced	4 - Somewhat general	5 - Very general, could apply to many documents	Not Applicable/Tag is irrelevant
Thyrotropin	◯	◯	◯	◯	◯	◯
Minute	◯	◯	◯	◯	◯	◯
USD	◯	◯	◯	◯	◯	◯
Releasing	◯	◯	◯	◯	◯	◯
High	◯	◯	◯	◯	◯	◯
Bid	◯	◯	◯	◯	◯	◯
Ask	◯	◯	◯	◯	◯	◯

13. How well do all the tags together cover the main themes of the document? *

1 - Not well at all, misses all major themes
2 - Somewhat well, captures some major themes
3 - Moderately well, captures about half of the major themes
4 - Very well, captures most major themes
5 - Extremely well, captures all major themes

Mark only one oval.

	1	2	3	4	5
	◯	◯	◯	◯	◯

14. If you selected any rating other than **5: Extremely well**, please suggest additional tags that would improve the coverage of the document's themes:

_____

15. Please read the following document called **Lung Cancer** carefully: *

"Lung Cancer, a microarray dataset, consists of 203 instances and 5 classes. This dataset is a classification of human lung carcinomas by mRNA expression profiling, which reveals distinct adenocarcinoma subclasses. The dataset is a comprehensive collection of gene expression profiles, providing valuable insights into the molecular mechanisms underlying lung cancer.

The dataset features a wide range of gene expression profiles, including those related to immune response, cell signaling, and metabolism. These profiles are essential for understanding the complex biology of lung cancer and for identifying potential therapeutic targets."

Now, examine the list of tags below. One of these tags is an 'intruder' that doesn't belong with the other tags. Please identify this intruder tag:

Mark only one oval.

( ) Cancer

( ) Microarray

( ) Prostate

( ) Ovarian

( ) Piece

( ) Gliomas

( ) Lung

( ) Genes

Now that you've identified the intruder, here are the predicted tags for this document:

- Cancer

- Microarray

- Prostate

- Ovarian

- Gliomas

- Lung

- Genes

And here is the document again:

"Lung Cancer Dataset, a microarray dataset, consists of 203 instances and 5 classes. This dataset is a classification of human lung carcinomas by mRNA expression profiling, which reveals distinct adenocarcinoma subclasses. The dataset is a comprehensive collection of gene expression profiles, providing valuable insights into the molecular mechanisms underlying lung cancer.
The dataset features a wide range of gene expression profiles, including those related to immune response, cell signaling, and metabolism. These profiles are essential for understanding the complex biology of lung cancer and for identifying potential therapeutic targets.

The dataset is a valuable resource for researchers and clinicians, providing a comprehensive understanding of lung cancer biology and facilitating the development of personalized treatment strategies."

16. How well does each tag represent the main themes of the document? *

Mark only one oval per row.

	1 - Not at all	2 - Somewhat well	3 - Moderately well	4 - Very well	5 - Extremely well
Cancer	◯	◯	◯	◯	◯
Microarray	◯	◯	◯	◯	◯
Prostate	◯	◯	◯	◯	◯
Ovarian	◯	◯	◯	◯	◯
Gliomas	◯	◯	◯	◯	◯
Lung	◯	◯	◯	◯	◯
Genes	◯	◯	◯	◯	◯

17. How specific or general is each tag? *

Mark only one oval per row.

	1 - Very specific to this document	2 - Somewhat specific	3 - Balanced	4 - Somewhat general	5 - Very general, could apply to many documents	Not Applicable/Tag is irrelevant
Cancer	◯	◯	◯	◯	◯	◯
Microarray	◯	◯	◯	◯	◯	◯
Prostate	◯	◯	◯	◯	◯	◯
Ovarian	◯	◯	◯	◯	◯	◯
Gliomas	◯	◯	◯	◯	◯	◯
Lung	◯	◯	◯	◯	◯	◯
Genes	◯	◯	◯	◯	◯	◯

18. How well do all the tags together cover the main themes of the document? *

    1 - Not well at all, misses all major themes
    2 - Somewhat well, captures some major themes
    3 - Moderately well, captures about half of the major themes
    4 - Very well, captures most major themes
    5 - Extremely well, captures all major themes

    Mark only one oval.

	1	2	3	4	5
	◯	◯	◯	◯	◯

19. If you selected any rating other than **5: Extremely well**, please suggest additional tags that would improve the coverage of the document's themes:

20. Please read the following document called **Colleges** carefully: *

"This dataset regroups information for approximately different US colleges, including geographical information, statistics about the population attending, and post-graduation career earnings. This is an attention check. The intruder is Social Science."

Now, examine the list of tags below. One of these tags is an 'intruder' that doesn't belong. Please identify this intruder tag:

Mark only one oval.

- ◯ Higher Education
- ◯ Colleges
- ◯ Social Science
- ◯ Institutional Data
- ◯ Education
- ◯ Academia
- ◯ US Colleges

Please review the following two documents:

COVID-19 Rio de Janeiro (City)
"Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus.

This dataset provides information on the number of confirmed cases, deaths, and recoveries by neighborhood in the city of Rio de Janeiro, Brazil. It is essential to note that this is a time-series data, and therefore, the number of cases on any given day represents a cumulative total. To obtain the number of new cases, one can calculate the difference between the current and previous days' data."

COVID-19 World Vaccination Progress
"This dataset allows users to track COVID-19 vaccination progress in the world, answering questions such as which country is using what vaccine, in which country the vaccination program is more advanced, where are more people vaccinated per day, and where are more people vaccinated in terms of percentage from the entire population."

21. Below is a combined list of tags from both documents. Please select the tags you believe are common to both documents:

Tick all that apply.

- ☐ Vaccinations
- ☐ Tests
- ☐ Deaths
- ☐ Cases
- ☐ Coronavirus
- ☐ Smoothed
- ☐ Million
- ☐ Covid
- ☐ Pandemic
- ☐ Janeiro

Now that you've identified the common tags, here are the predicted common tags for the documents:

- Cases

- Deaths

- Coronavirus

And here are the two documents again:
COVID-19 Rio de Janeiro (City)
"Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus.

This dataset provides information on the number of confirmed cases, deaths, and recoveries by neighborhood in the city of Rio de Janeiro, Brazil. It is essential to note that this is a time-series data, and therefore, the number of cases on any given day represents a cumulative total. To obtain the number of new cases, one can calculate the difference between the current and previous days' data."

COVID-19 World Vaccination Progress
"This dataset allows users to track COVID-19 vaccination progress in the world, answering questions such as which country is using what vaccine, in which country the vaccination program is more advanced, where are more people vaccinated per day, and where are more people vaccinated in terms of percentage from the entire population."

22. How well do the common tags represent themes shared by both documents?

    1 - Not well at all, misses all major shared themes
    2 - Somewhat well, captures some major shared themes
    3 - Moderately well, captures about half of the major shared themes
    4 - Very well, captures most major shared themes
    5 - Extremely well, captures all major shared themes

    Mark only one oval.

	1	2	3	4	5
	◯	◯	◯	◯	◯

23. If you selected any rating other than **5: Extremely well**, please suggest additional common tags that would improve the coverage of the documents' themes:

    _____

Please review the following two documents:

Movies on Netflix, Prime Video, Hulu, and Disney+
"This dataset is an amalgamation of data that was scraped, comprising a comprehensive list of movies available on various streaming platforms, and the IMDb dataset, which provides inspiration for analysis.

Which streaming platform or platforms can I find this movie on? This dataset allows us to explore the availability of movies across different streaming services. Additionally, we can examine the average IMDb rating of movies produced in a specific country, providing insights into the quality of films from different regions."

Popular Movies of IMDb
"TMDB.org is a crowd-sourced movie information database widely used by various film-related consoles, sites, and apps, such as XBMC, MythTV, and Plex. Dozens of media managers, mobile apps, and social sites utilize its API. At the time of writing, TMDB lists a substantial number of films, which is considerably fewer than IMDb. While not as comprehensive as IMDb, it holds extensive information for most popular and Hollywood films."

24. Below is a combined list of tags from both documents. Please select the tags you believe are common to both documents:

Tick all that apply.

- [ ] Movies
- [ ] IMDb
- [ ] Weekend
- [ ] Office
- [ ] Earnings
- [ ] Streaming
- [ ] TMDB
- [ ] Gross
- [ ] Audience

Now that you've identified the common tags, here are the predicted common tags for the documents:

- Movies

- IMDb

- Weekend

- Office

- Earnings

- Streaming

- TMDB

- Gross

- Audience

And here are the two documents again:

Movies on Netflix, Prime Video, Hulu, and Disney+
"This dataset is an amalgamation of data that was scraped, comprising a comprehensive list of movies available on various streaming platforms, and the IMDb dataset, which provides inspiration for analysis.

Which streaming platform or platforms can I find this movie on? This dataset allows us to explore the availability of movies across different streaming services. Additionally, we can examine the average IMDb rating of movies produced in a specific country, providing insights into the quality of films from different regions."

Popular Movies of IMDb
"TMDB.org is a crowd-sourced movie information database widely used by various film-related consoles, sites, and apps, such as XBMC, MythTV, and Plex. Dozens of media managers, mobile apps, and social sites utilize its API. At the time of writing, TMDB lists a substantial number of films, which is considerably fewer than IMDb. While not as comprehensive as IMDb, it holds extensive information for most popular and Hollywood films."

25. How well do the common tags represent themes shared by both documents?

1 - Not well at all, misses all major shared themes
2 - Somewhat well, captures some major shared themes
3 - Moderately well, captures about half of the major shared themes
4 - Very well, captures most major shared themes
5 - Extremely well, captures all major shared themes

Mark only one oval.

	1	2	3	4	5
	○	○	○	○	○

26. If you selected any rating other than **5: Extremely well**, please suggest additional common tags that would improve the coverage of the documents' themes:

_____