|  |
| --- |
| ***Big Data Analytics - Team Project Proposal - Summer 2019*** |
| ***Part 1. General Project Information*** |
| **Names of team members:** Yijun Tian, Fan Xie, Rui Jin.  **Project Title:** What’s the next hot area in Computer Science field?  **Project Description: *(Write one paragraph to describe what this analytic will do.)***  Computer Science field have so many branch: Artificial Intelligence, Hardware Architecture, Databases, Graphics, Multimedia, etc. Every year, there are question like “What are currently the hot topics in computer science research?” appears on Quora. People are enthusiastic to ask and answer these questions. Students who will choose their major want to know which subject is the most promising. Researchers want to know which subject attract the most attention. In this case, newspapers, magazines and surveys publish trend ranking yearly. However, most of their predictions are based on some experts and there always existing some difference between each one. Thus, in this project, we will use the subject information and the paper information within arXiv, to analyze the next hot area.  **Who is a typical user of your application:**  Computer Science field researchers.  Newcomers who want to step into the computer field.  Venture capital.  **Describe who will benefit from your analytic:**  Computer Science field researchers are able to study the changing trend of the whole industry. Newcomers who want to step into the computer field can get a taste what’s the most attractive area currently within the field. Especially for Venture capital, what’s the next hot area is the most attractive question. The hot area point out the research interests of researchers, which is also a significant indicators to the strength of attention received from scientific communities. Therefore, Venture capitalist can easily track hot topics and make appropriate assessments and investments.  **What insight will you derive from the data?**  Idealy, we can develop a project to indicate the hot area trend within the past years, and the future trends of each area within computer science field.  **Describe how you will check the goodness of the analytic, i.e., how will you prove the results are accurate and can be trusted:**   |  | | --- | | We compare it with the past reality data. We can do cross validation on the dataset. Also, the features analysis will help us determine the confidence of the prediction result. | |

|  |  |  |
| --- | --- | --- |
| ***Big Data Analytics - Team Project Proposal - Summer 2019*** | | |
| ***Part 2. Data Source Information*** | | |
| ***Name of Data Source 1:*** **arXiv**  ***Data Source Description:***  Arxiv is a document submission and retrieval system that is heavily used by computer science communities. It has become the primary means of communicating cutting-edge manuscripts on current and ongoing research. Almost all scientific papers are self-archived on the arXiv repository. Arrive API allows application developers to access all of the arXiv data, search and linking facilities with an easy-to-use programmatic interface.  API: [*https://arxiv.org/help/api/user-manual*](https://arxiv.org/help/api/user-manual) | | |
| **Data Collection Frequency**   * Are you collecting the data in realtime? Or are you collecting it periodically? * Are you collecting static data? (e.g. historic data that you load once) | **Data Size**   * *Estimate size of the data you will store, e.g. MB, GB, TB, PB* | **Data Frequency**   * If realtime data, what is the frequency and volume of data (how often and how much data will you collect at a time)? * If batch data, how often will you collect it? |
| ☐ Realtime (ongoing near-realtime collection)  ☐ Batch (multiple non near-realtimecollections)  ☐ Static (one time collection) | ☐ MB  ☐ 1-10 GB  ☐ 10-100 GB  ☐ 100-300 GB  ☐ 300-500 GB  ☐ > 500 GB | **If realtime data:**   * How often will you collect data?   ☐ Every second, or every few seconds  ☐ Every minute, or every few minutes   * What is the size of data you will collect at each interval?   \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **If *not* realtime data:**   * Will you collect a batch of data periodically or just once (static)?   ☐ Just once  ☐ Every hour, or every few hours  ☐ Every day, or every few days, or every week, or every month   * How much data that will be collected at each interval?   All of the data. In our case, it will be all the paper appeared in the arXiv.  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

|  |  |  |
| --- | --- | --- |
| ***Big Data Analytics - Team Project Proposal - Summer 2019*** | | |
| ***Part 2. Data Source Information*** | | |
| ***Name of Data Source 2:***  ***Data Source Description:*** | | |
| **Data Collection Frequency**   * Are you collecting the data in realtime? Or are you collecting it periodically? * Are you collecting static data? (e.g. historic data that you load once) | **Data Size**   * *Estimate size of the data you will store, e.g. MB, GB, TB, PB* | **Data Frequency**   * If realtime data, what is the frequency and volume of data (how often and how much data will you collect at a time)? * If batch data, how often will you collect it? |
| ☐ Realtime (ongoing near-realtime collection)  ☐ Batch (multiple non near-realtimecollections)  ☐ Static (one time collection) | ☐ MB  ☐ 1-10 GB  ☐ 10-100 GB  ☐ 100-300 GB  ☐ 300-500 GB  ☐ > 500 GB | **If realtime data:**   * How often will you collect data?   ☐ Every second, or every few seconds  ☐ Every minute, or every few minutes   * What is the size of data you will collect at each interval?   \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **If *not* realtime data:**   * Will you collect a batch of data periodically or just once (static)?   ☐ Just once  ☐ Every hour, or every few hours  ☐ Every day, or every few days, or every week, or every month   * How much data that will be collected at each interval?   \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

|  |  |  |
| --- | --- | --- |
| ***Big Data Analytics - Team Project Proposal - Summer 2019*** | | |
| ***Part 2. Data Source Information*** | | |
| ***Name of Data Source n:***  ***Data Source Description:*** | | |
| **Data Collection Frequency**   * Are you collecting the data in realtime? Or are you collecting it periodically? * Are you collecting static data? (e.g. historic data that you load once) | **Data Size**   * *Estimate size of the data you will store, e.g. MB, GB, TB, PB* | **Data Frequency**   * If realtime data, what is the frequency and volume of data (how often and how much data will you collect at a time)? * If batch data, how often will you collect it? |
| ☐ Realtime (ongoing near-realtime collection)  ☐ Batch (multiple non near-realtimecollections)  ☐ Static (one time collection) | ☐ MB  ☐ 1-10 GB  ☐ 10-100 GB  ☐ 100-300 GB  ☐ 300-500 GB  ☐ > 500 GB | **If realtime data:**   * How often will you collect data?   ☐ Every second, or every few seconds  ☐ Every minute, or every few minutes   * What is the size of data you will collect at each interval?   \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **If *not* realtime data:**   * Will you collect a batch of data periodically or just once (static)?   ☐ Just once  ☐ Every hour, or every few hours  ☐ Every day, or every few days, or every week, or every month   * How much data that will be collected at each interval?   \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |