Proposal for Collecting, Analyzing, and Visualizing Reported Mining Fatalities

Team Information

Team Name: Text Miners
Team Members:
Alexander Antonison (***ada4@illinois.edu***) (Team Leader)
Sai Rao (***sairao2@illinois.edu***)
Amartya Chowdhury (***amartya4@illinois.edu***)

What is the function of the tool?

Our project aims to create a user friendly display of MSHA (Mining Safety and Health Administration) fatality reports in a manner that allows the user to quickly assess information in a manner that does not require them to peruse hundreds of reports spanning dozens of pages to identify if a particular mining area or type of mine sees higher incidence of issues over time.
We envision scraping data available on the MSHA fatality reports site and use the classification metadata to collate and present on a heat map of the USA, where users can specify time frames that will appropriately display reported fatalities for that time period.

Who will benefit from such a tool?

The target audience is expected to be twofold - first, it would appeal to miners who work in surface or underground mines and before they take on any assignments can check if the area that they are planning to work in has higher fatality trends than other areas. This would hopefully encourage them to perform a more detailed analysis of the circumstances surrounding the incident(s) and would help them make informed decisions.
The second target audience set would be the government agency itself(MSHA folks) as they do not seem to currently have any dashboards for public consumption that would let folks view consolidated data over time. They only have canned reports that run per year and it would require a deep dive into multiple pdfs to truly get a bigger picture or identify any trends that might eventually require additional legislation.

Does this kind of tools already exist? If similar tools exist, how is your tool different from them? Would people care about the difference?

There is no as such tool present accordingly to our knowledge.
There are some white papers available related to this topic for example “Machine-related injuries in the US mining industry and priorities for safety research” but our tool will be different as there will be a user friendly display that help the miners to understand the potential risk based on the analysis of the data.
People will care about this tool as it will help the insurance company to check the whole statistics analyzed in one page and determine the life insurance cost. This will also help the miners to take better safety precautions on their daily mining activity to prevent casualties.

What existing resources can you use?

The main resource we will be using will be from US department of Labor Fatality report from MSHA fatality reports site
https://www.msha.gov/data-reports/fatality-reports/search
We will be getting all the metadata like Category, Year, Territory, Location of Accident, Accident Classification, Mined Material from this site.

We will be referring the investigation report from the following link from year 1995 till 2007
https://arlweb.msha.gov/fatals/fabmc-1995-2007.htm

The daily fatality report we will collect from the below link for both underground mines and surface mines
https://arlweb.msha.gov/stats/charts/combined.php

For the Coal Fatalities for 1900 Through 2018 we will be using the following resources
https://arlweb.msha.gov/stats/centurystats/coalstats.asp

For Metal/Nonmetal Fatalities for 1900 through 2018 we will refer the following resources
https://arlweb.msha.gov/stats/centurystats/mnmstats.asp

The other resources we will be using are from
https://www.mining-technology.com/mining-safety/us-mining-accidents-fatalities-2018
https://www.mining-technology.com/mining-safety/msha-announces-first-us-mining-fatality-of-2019
https://www.riskope.com/2017/01/25/mining-deaths-injuries/
We also referring some of the white paper available on these topics as below, these will help us to analyses what user can look for which will be different from what already exists.
https://www.cdc.gov/NIOSH/Mining/UserFiles/works/pdfs/mriit.pdf
https://www.bls.gov/opub/mlr/cwc/transportation-fatalities-in-the-mining-sector-20042008.pdf

What techniques/algorithms will you use to develop the tool?

Our project will be broken into three different phases.

1. The first phase of the project is data collection. This will involve web scraping the fatality reports from the https://www.msha.gov/data-reports/fatality-reports/search website. This will involve using a combination of either Scrapy or Selenium to pull the text and then using beautiful soup to process and store the data of interest into a json text file. This will focus on pulling in as much raw text as possible with minimal pre-processing. This approach allows us to be more flexible in the pre-processing text phase without having to re-scrape data.
2. The second phase will be the text pre-processing. This involves building a text pre-processing pipeline that will pull the documents from the json text file and pre-process the documents into CSV files that will later be used to power the dashboard. Additionally, will use some text analysis tools such as Topic Analysis on the long text reports.
3. The third phase involves building an interactive dashboard. In order to present the results, an interactive dashboard will be created to allow an end-user to explore the data extracted from the mining

fatality reports. Planning to use either plot.ly in Python to create a dashboard or may consider using a free business intelligence tool like Tableau depending on available time.

How will you demonstrate the usefulness of your tool?

Given that the mining industry(rock/ore mining, not data mining) is vastly underserved in terms of innovative technology trends in comparison to most other domains, this tool might be a step up in helping visualize to the general public or miners or MSHA itself - the information that is out there but not intuitively usable to identify trends with issues of a certain category or incidence of multiple events of the same nature in a specific type of mine.
We hope that this tool will help drive discussion around additional safeguards that might be needed in terms of government oversight or regulation - while allowing the mining community to understand where incidents of a similar nature had occurred/are occurring, which might, in turn, help them make better safety decisions in their day to day work resulting in fewer fatalities over time.

A very rough timeline to show when you expect to finish what.

With a total of 9 weeks from the submission of the proposal to the due date of the final project.

Phase 1 - Data Collection (2.5 weeks)

As the fatality reports all follow a consistent format, we expect it will take about two weeks to build the web scraper to extract the fatality reports and store them in a JSON text file.

Phase 2 - Text Data Processing (3.5 weeks)

In this phase, the first step will be to establish the types of data visualizations we are wanting to create as this will inform the data we extract from the fatality reports. Once the desired data is defined, will need to create a series of scripts capable of extracting the data into CSV files in preparation to be ingested by the dashboard platform.

Phase 3 - Build an Interactive Dashboard (3 weeks)

In this final phase, we will be focusing on finalizing the data visualizations and building initial prototypes. Once finalized, we will integrate data visualizations into the interactive dashboard.