Text Mining for Daily Reports of Construction

# Introduction

Currently, for the traditional convention, most reports of construction are saved as hard copy or PDF file. Although those reports recorded very detailed information about project of construction, because of the traditional way of information storage and process, the information of those reports are hard to be efficiently extracted and exploited. Therefore, we apply the technique of text mining and machine learning to analyze those reports and try to get find a more convenient and efficient way to fully exploit the information from reports.

The daily reports we used, the CONTRACTORS QUALITY CONTROL REPORT (QCR) DAILY LOG OF CONSTRUCTION - MILITARY, including the daily logs from 01 Nov 2004 to 17 Oct 2007, recorded the information about construction activities and events for every workday. However, the records are very detailed and fractional, and information is recorded in natural language, which is relatively unstructured. So, it is not easy for human readers to quickly capture the very information they want in from such large volume of data.

To reduce those difficulties, we develop several Python programs to extract some features from the data. At first, we clean the textual data by removing the stop words and some relatively meaningless words. Then all the reports are converted to a matrix of word frequency or TF-IDF. Based on the matrices, K-means algorithm is applied to cluster both words and reports. By analyzing and visualizing the clusters of words and reports, some features of the reports can be easily observed and more information might be revealed by further research.

# Data cleaning and information retrieval

## Extract whole textual file to daily contents

Our textual data set has 2027 pages of reports. A report usually only has one to three pages. So the first step is to do is to reorganize the pages and group them into corresponding report numbers. The program groups those page by recognize the report numbers and page number. The grouping results show there are 881 reports and the report numbers are from 91 to 1212.

## Remove stop word

Since the reports are written in natural language, punctuations and many words which are relatively meaningless, namely stop words like “the”, “is” and “that”, can be eliminated from our analysis. Then we have a dictionary containing the remaining words.

## Convert reports into a matrix of word frequency. This step results in a matrix of 6782 words by 881 reports. Save to words\_counts\_doc.csv

To simplify the problem, we use every single word as a token. So every report can be converted to n-dimension vector, while n is the number of word in our dictionary, and the whole data can be converted into a matrix(*i*,*j*) where *i* represents report and *j* represents a word. Part of the matrix is shown in the Table 1.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Report  word | Report 1 | Report 2 | Report 2 | Report 2 | Report 2 | Report 2 | … |
| aai | 2 | 1 | 1 | 1 | 1 | 0 | … |
| aat | 0 | 0 | 0 | 0 | 0 | 0 |
| ab | 0 | 0 | 0 | 0 | 0 | 0 |
| abate | 0 | 0 | 0 | 0 | 0 | 0 |
| abatement | 1 | 1 | 1 | 1 | 1 | 1 |
| absence | 0 | 0 | 0 | 0 | 0 | 0 |
| ac | 0 | 0 | 0 | 0 | 0 | 0 |
| accent | 0 | 0 | 0 | 0 | 0 | 0 |
| accept | 0 | 0 | 0 | 0 | 0 | 0 |
| … | … | … | … | … | … | … | … |

Table 1: Part of the matrix of word frequency

## Re-represent reports with a matrix with TF-IDF

TF-IDF is short for term frequency-inverse & inverse document frequency, which can reflect the importance of a word in a document.

### Term frequency(TF)

Term frequency tf(d,t) is define as the number of times that term t occurs in document d. For example, “abatement” occurs in the first reports once, so the TF of “abatement” in the first report is 1. The high value of tf may imply the higher importance of the word in the report.

### Inverse document frequency(IDF)

Since some words in language occur so common that they exist almost everywhere. Obvious that kind of words are usually not so important and provide less information comparing to their frequency. So IDF is designed to diminish that effect. The inverse document frequency(IDF) is a measure of how much information the word provides, that is, whether the term is common or rare across all documents. Inverse document frequency idf(d, t) = log[1+n/1+df(d, t)]+1, where df(d, t) is the document frequency, defined as the number of documents d that contain term t.

For every word in a document, we can compute its TF-IDF, which is the product of its TF and IDF. Therefore, by computing the TF-IDF, a matrix of TF-IDF can generated, too. Part of the matrix of TF-IDF is shown in Table 2.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Report word | Report 2 | Report 2 | Report 2 | Report 2 | Report 2 | Report 2 | … |
| aai | 0.153782421 | 0.059501269 | 0.084724144 | 0.084510682 | 0.085376961 | 0 | … |
| aat | 0 | 0 | 0 | 0 | 0 | 0 |
| ab | 0 | 0 | 0 | 0 | 0 | 0 |
| abate | 0 | 0 | 0 | 0 | 0 | 0 |
| abatement | 0.07502449 | 0.058056731 | 0.08266726 | 0.082458981 | 0.083304228 | 0.107773421 |
| absence | 0 | 0 | 0 | 0 | 0 | 0 |
| ac | 0 | 0 | 0 | 0 | 0 | 0 |
| … | … | … | … | … | … | … | … |

Table 2: Part of the matrix of TF-IDF

# Clustering: input: TF-IDF matrix; output: the cluster labels of each word and report.

K-means clustering is an unsupervised algorithm for clustering aiming to partition items into k clusters in which each item belongs to the cluster with the nearest mean. Based on the TF-IDF matrix obtained above, K-means clustering is applied on both words and reports.

For the words clustering, because most of words only occur for several times or even once in such many reports, the matrix of word frequency is sparse. So, we can find in the result of words clustering that such words are grouped in the same cluster. We think that it is not easy to extract more specific information from a such large cluster, so we exclude that cluster from the following analysis. Meanwhile, some clusters contain few words, which are also excluded from the analysis. In the remaining four clusters of words, we find some words in the same cluster share some similarities. For example, many words in one cluster are related to transportation and loading.

For the clustering of reports, we found the results are apparently related to the date of the reports. For example, most reports from 15 Nov 2004 to 03 Mar 2005 are grouped into a same group. So we infer that the activities in that period recorded in the reports are probably similar.

|  |  |
| --- | --- |
| cluster | words |
| 1 | na roman trucking quantum crusher loads se sl  cat hole dec hauling services debris driver trenching update  group rfp status jd rfi demo link belt samsung aai  main abatement asbestos |
| 2 | survey poured foundation mains narrative layout patio excavation  genuine slab backfill jun foundations written ir lot arauco  footings strip porch tie footing density forms roller sewer rebar  dust received spaces unit jcb axelsen form pour place  craw grade meeting layer garage pipe attached walls  mti water |
| 3 | overhead bldg roughin sand vanities ceiling phon blow  appliances firew oct shingle hvac walls distributed jan nov  passed nailing street mold cable complete kx kobelco marble  discussed west insulate texture installation tape steer hang  drain walk skid ways kabota misc generators review interior  nutek pick porches temp installing trusses sideway cabinets  storm check electrical pressure custom rock kitchen plumbing  specific wall piping gas heaters repairs shingles soil  driveway critical floors test way observed mcgahey jack  contra carpentry base conversations backhoe apprentice stone  door sheetrock dry umc buildings dingo dump mason hanging  roof painting taping taper issues set gradall forklift west roofer  installer soffit fascia |
| 4 | forrest lebaron apr jul sep aug mar common  work reports repair saturday inspect maintance touch loops  landscaping monday friday tuesday wednesday Thursday  hardware carpet units coe doors paint performed |

Table 3: Words Clusters

# Visualization with Excel

The results obtained above are save in CSV and XLSX files for further research. We sum up the frequency of words in a same cluster and plot them with temporal order (Fig 1), then plot the reports cluster as well (Fig 2). The figures show that the peak periods of words frequencies are obviously corresponding to certain clusters of reports.

Fig 1: words frequency for different cluster of words

Fig 2: the clustering results of reports

A similar plot of TF-IDF is shown below (Fig 3). It still needs further research for the importance of the words in reports.

Fig 3: TF-IDF for different cluster of words

# Current Conclusion

According the four clusters of words and the plot shown previous, we find that the reports can be divided into four periods. The first period is from 15 Nov 2004 to 03 Mar 2005; the second one is from 04 Mar 2005 to 10 Aug 2005; the third one is from 11 Aug 2005 to 09 Feb 2007; the forth one is from 10 Feb 2007 to 27 Nov 2007. The first three periods can be represented by the first three clusters of words. These periods can still be sub-divided.