CS 6320 Data Mining Data Collection Report

1 How we obtained the data?

Since this is a project motivated by Kaggle competition we just download their data as our data.

2 How large is the data?

The data scraped over 17,000 tweets from 100+ pro-ISIS fanboys from all over the world since the November 2015 Paris Attacks. The size of data is 4.5 MB.

3 Format we are storing our data in

We stored our data in csv file, and there are eight features for the data. The dataset includes the following fields:

Fields	**Data type**
Name	string type
Username	string type
Description	string type
Location	string type
Number of followers at the time the tweet was downloaded	int type
Number of statuses by the user when the tweet was downloaded	int type
Date and timestamp of the tweet	float type
The tweet itself	text

Table 1: dataset format

4 Processing of original data

The data is only 6MB, also the format is already clear for analyzing. Hence we don't need to do process the original data into other format, like the extra compress work. However, for data analysis, since there might be some missing in the collected data, for example, missing in the description, we have to do some text pre-processing to make the following analysis easier.

5 Simulate similar data

Since the data is based on the text on twitter, we can simulate the similar data by adding more user(node) in the first column. And use the existed users' data which share a relationship with the added node in the first column. In this way, we could simulate the similar data.