Evaluation of running time for NW Algorithm

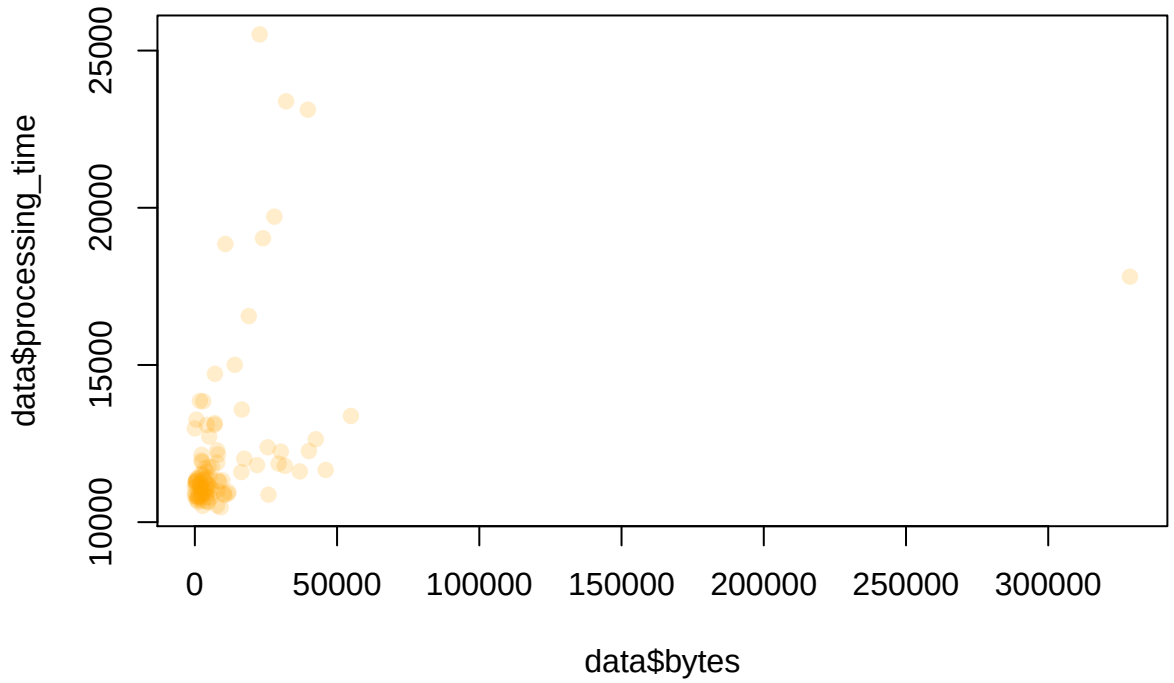
Aimee Barciauskas

May 14, 2016

Original Data:

```
setwd('~/IdeaProjects/gaceta/')
data <- read.csv('nw_split_norm_times_orig.csv', header = FALSE)
colnames(data) <- c('filename', 'bytes', 'processing_time')

plot(data$bytes, data$processing_time,
     col = add.alpha("orange", alpha=0.2),
     pch = 19)
```



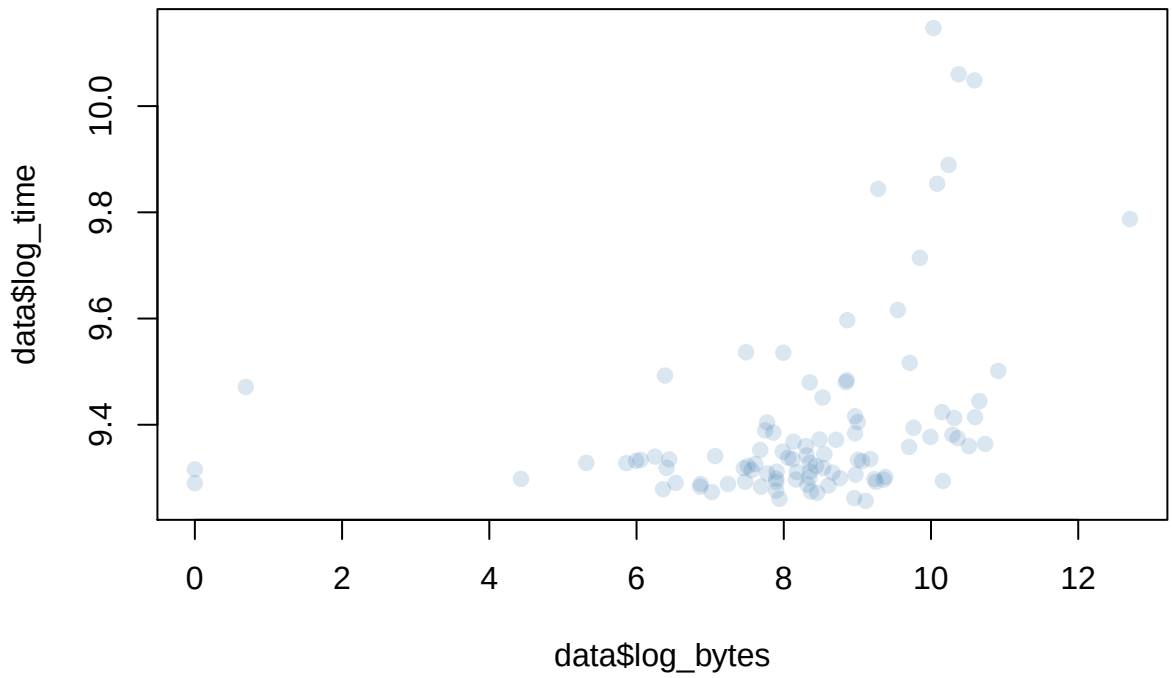
Logged data:

```
data$log_bytes = log(data$bytes)
data$log_time = log(data$processing_time)
data$log_bytes_sq = data$log_bytes**2
data$log_bytes_cub = data$log_bytes**3
plot(data$log_bytes, data$log_time,
     col = add.alpha("steelblue", alpha=0.2),
     pch = 19)#,
     #xlim = c(0,max(data$log_bytes)))


fit <- lm(log_time ~ log_bytes + log_bytes_sq + log_bytes_cub, data = data)
summary(fit)
```

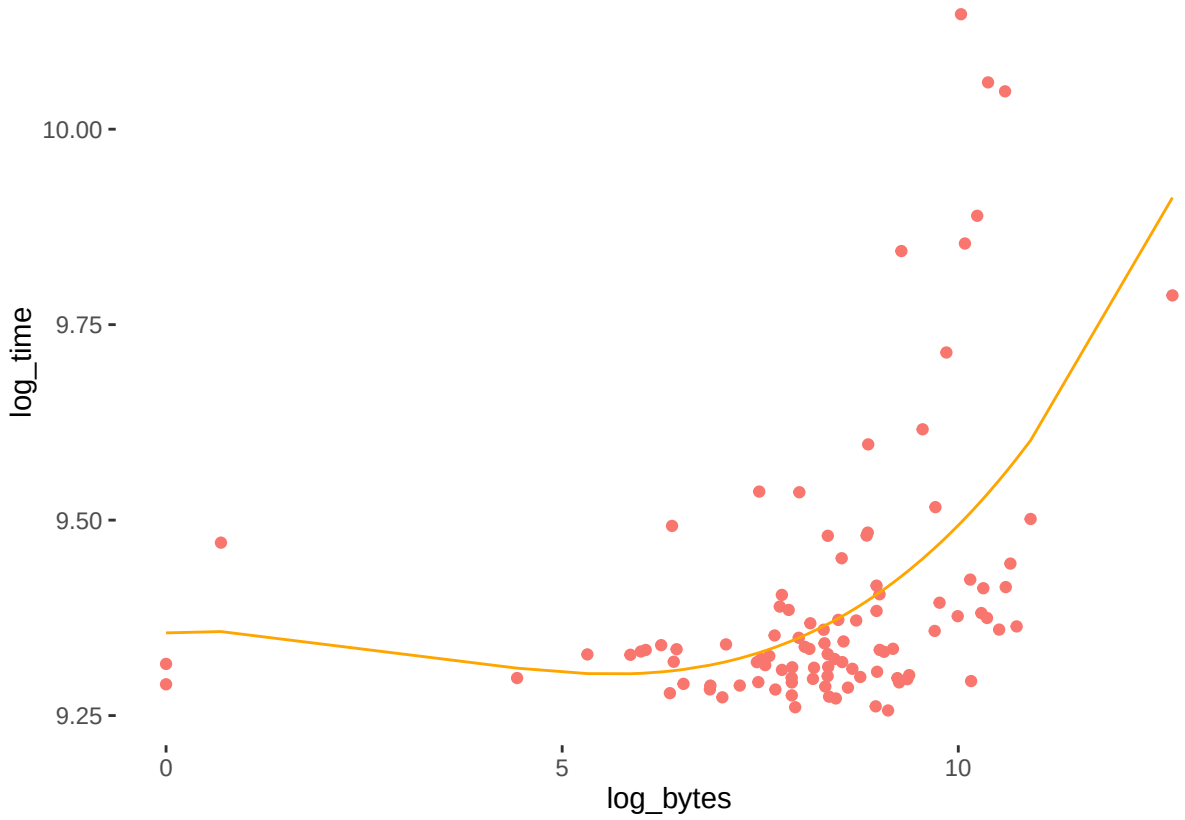
```
##
## Call:
## lm(formula = log_time ~ log_bytes + log_bytes_sq + log_bytes_cub,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21629 -0.08473 -0.03166  0.02516  0.65012
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.3557299  0.0943855  99.123   <2e-16 ***
## log_bytes      0.0073171  0.0638113   0.115    0.909
## log_bytes_sq  -0.0076027  0.0123612  -0.615    0.540
## log_bytes_cub  0.0008247  0.0006589   1.252    0.214
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1517 on 96 degrees of freedom
## Multiple R-squared:  0.2726, Adjusted R-squared:  0.2499
## F-statistic: 11.99 on 3 and 96 DF,  p-value: 9.764e-07
```

```
coeff0 <- as.numeric(fit$coefficients[1])
coeff1 <- as.numeric(fit$coefficients[2])
coeff2 <- as.numeric(fit$coefficients[3])
coeff3 <- as.numeric(fit$coefficients[4])
library(ggplot2)
```




```
p1 <- ggplot(data, aes(x=log_bytes, y = log_time, color = "grey")) +
  geom_point() +
  theme(legend.position='none', panel.background = element_blank()) +
```

```
  geom_line(aes(y=fitted.values(fit)), color='orange')
p1
```




```
# sanity check
summary(exp(coeff0 + coeff1*data$log_bytes + coeff2*data$log_bytes_sq + coeff3*data$log_bytes_cub))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10980   11390   11700   12120   12410   20180
```

```
summary(data$processing_time)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10470   10930   11330   12280   12170   25520
```

```
# total run time for all of split_norm
total_size <- 1.98e+8
ms <- exp(coeff0 + coeff1*log(total_size) +
            coeff2*(log(total_size)**2) +
            coeff3*(log(total_size)**3))
seconds <- ms/1000
minutes <- seconds/60
hours <- minutes/60
hours
```

```
## [1] 0.07240028
```

```r
# This seems like a vast underestimate

# average size of file in split_norm is 13
avg_file_size <- mean(data$bytes)
num_files <- 13116
average_expected_time_ms <- exp(coeff0 + coeff1*log(avg_file_size) +
                                  coeff2*(log(avg_file_size)**2) +
                                  coeff3*(log(avg_file_size)**3))
total_expected_time_ms <- num_files*average_expected_time_ms
seconds <- total_expected_time_ms/1000
minutes <- seconds/60
hours <- minutes/60
hours
```

```
## [1] 45.87482
```