Statistical Analysis of ISMB Coverage at Twitter 2012

Neil Saunders

August 15, 2012

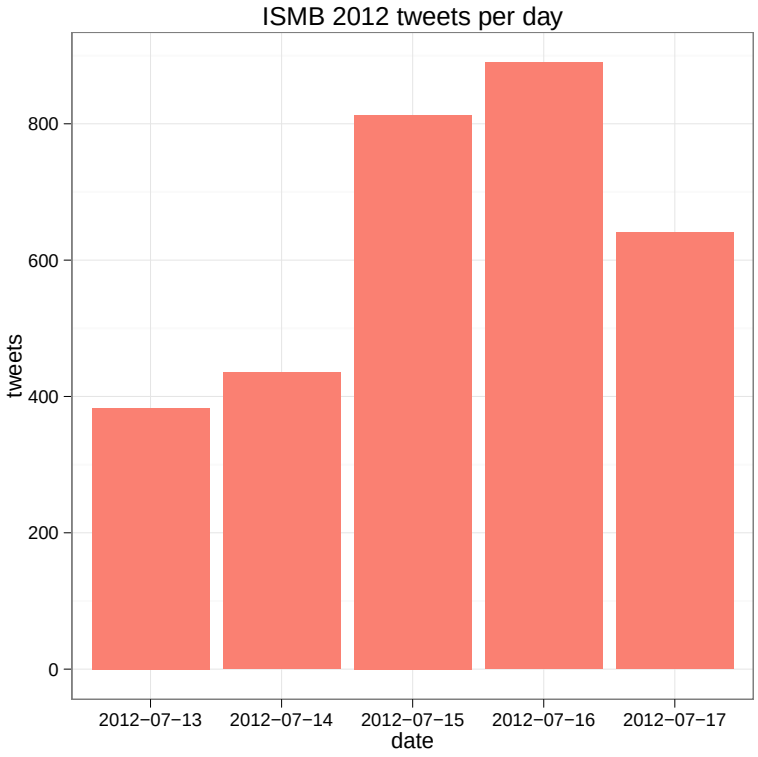
1 Preamble

Load required libraries and data.

```
> library(ggplot2)
> library(xtable)
> library(RColorBrewer)
> library(tm)
> library(wordcloud)
> library(sentiment)
> load("~/Dropbox/projects/twitter/ismb/data/ismb.RData")
```

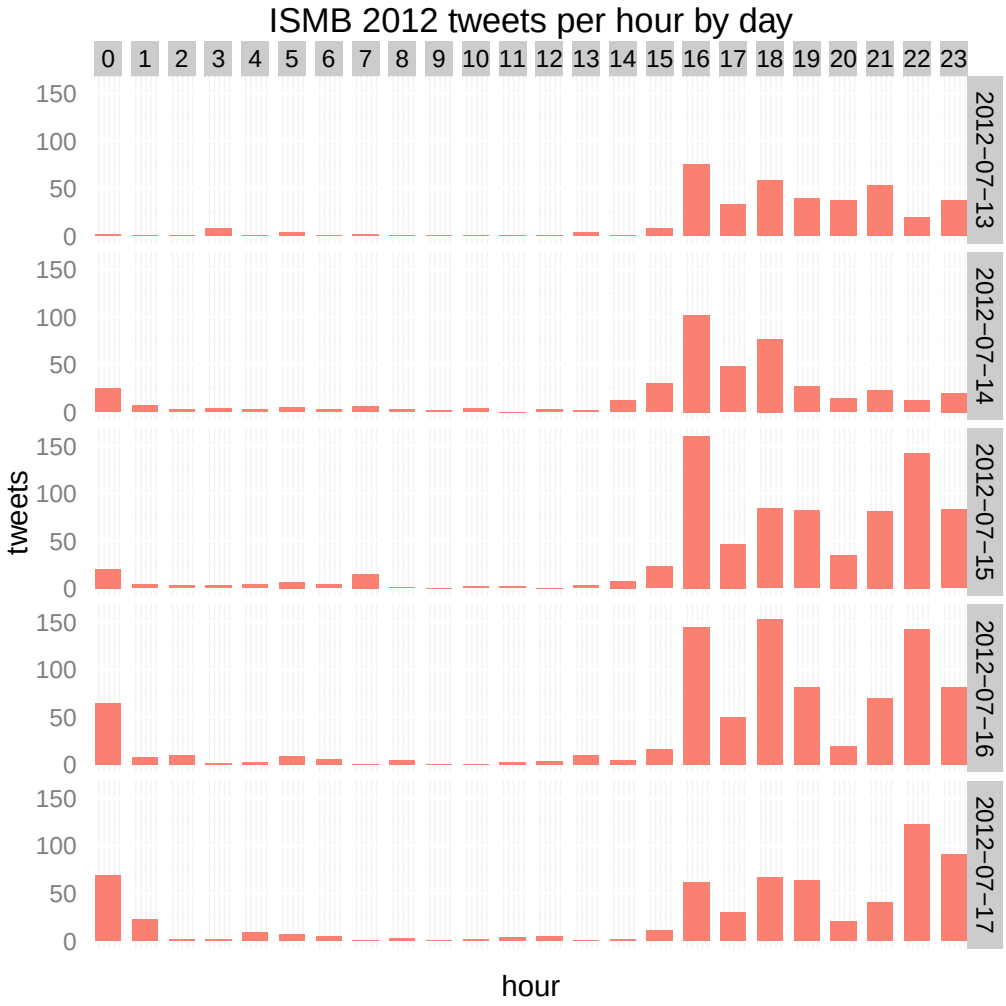
2 Tweets per day

```
> ismb$date <- as.Date(ismb$created)
> byDay <- as.data.frame(table(ismb$date))
> colnames(byDay) <- c("date", "tweets")
> print(ggplot(byDay) + geom_bar(aes(date, tweets), fill = "salmon") + theme_bw()
+       + opts(title = "ISMB 2012 tweets per day"))
```



3 Tweets per hour by day

```
> ismb$hour <- as.POSIXlt(ismb$created)$hour
> byDayHour <- as.data.frame(table(ismb$date, ismb$hour))
> colnames(byDayHour) <- c("date", "hour", "tweets")
> byDayHour$hour <- as.numeric(as.character(byDayHour$hour))
> print(ggplot(byDayHour) + geom_bar(aes(hour, tweets), fill = "salmon", binwidth = 1) +
+   facet_grid(date ~ hour) +
+   opts(axis.text.x = theme_blank(),
+        axis.ticks  = theme_blank(),
+        panel.background = theme_blank(),
+        title = "ISMB 2012 tweets per hour by day"))
```



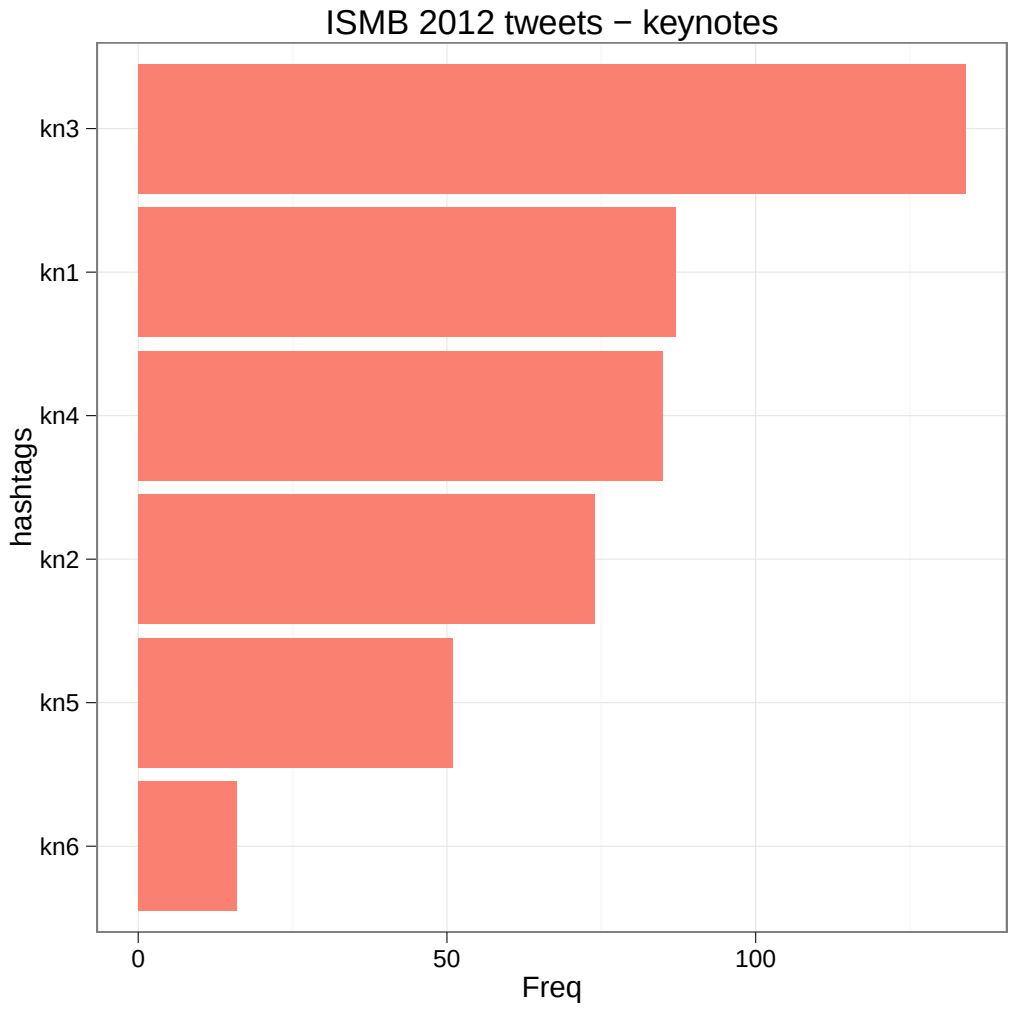
4 Popular talks

First, get the hashtags:

```
> words <- strsplit(ismb$text, " ")
> hashtags <- lapply(words, function(x) x[grep("^#", x)])
> hashtags <- unlist(hashtags)
> hashtags <- tolower(hashtags)
> hashtags <- gsub("[^A-Za-z0-9]", "", hashtags)
> ht <- as.data.frame(table(hashtags))
> ht <- ht[sort.list(ht$Freq, decreasing=F),]
```

4.1 Keynotes

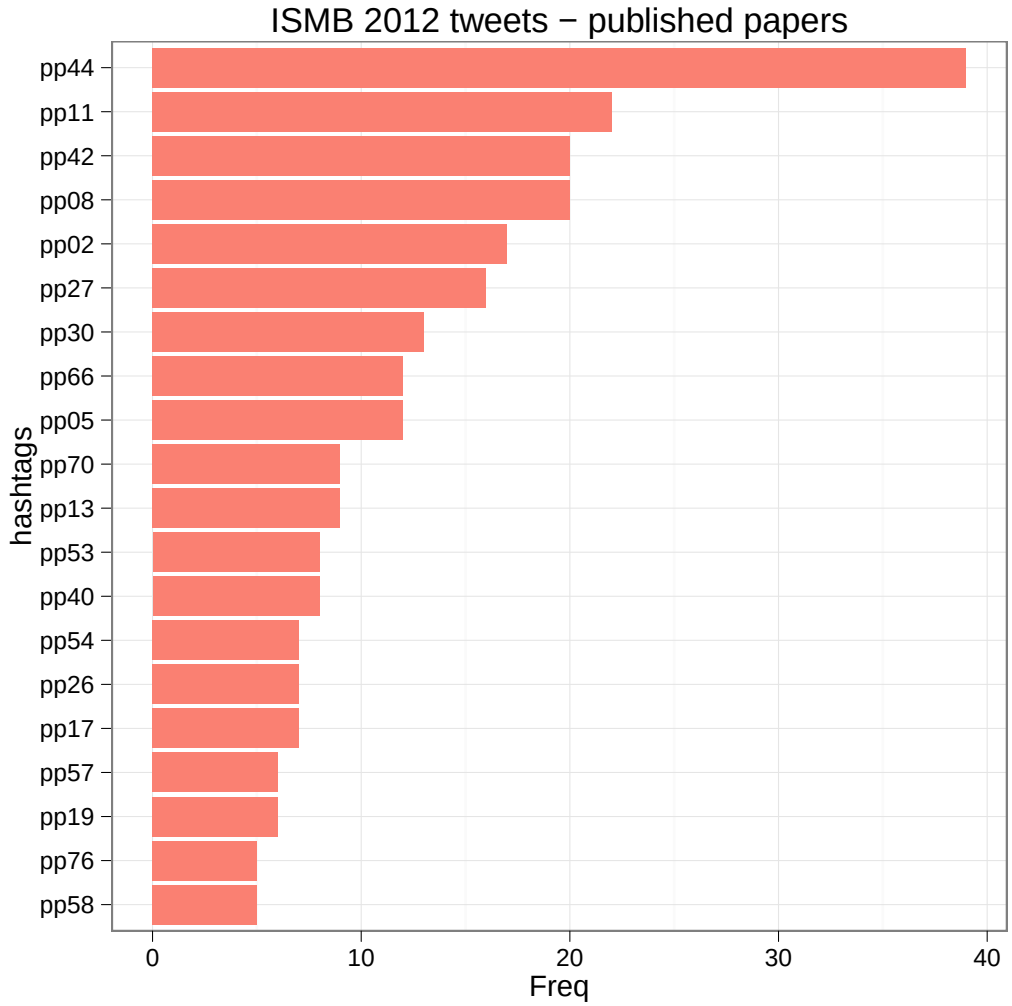
```
> kn <- ht[grep("^kn", ht$hashtags),]
> kn$hashtags <- factor(kn$hashtags, levels = as.character(kn$hashtags))
> print(ggplot(tail(kn)) + geom_bar(aes(hashtags, Freq), fill = "salmon") +
+   coord_flip() + theme_bw() + opts(title = "ISMB 2012 tweets - keynotes"))
```



KN 3: Analysis of transcriptome structure and chromatin landscapes (Barbara Wold).

4.2 Published Papers

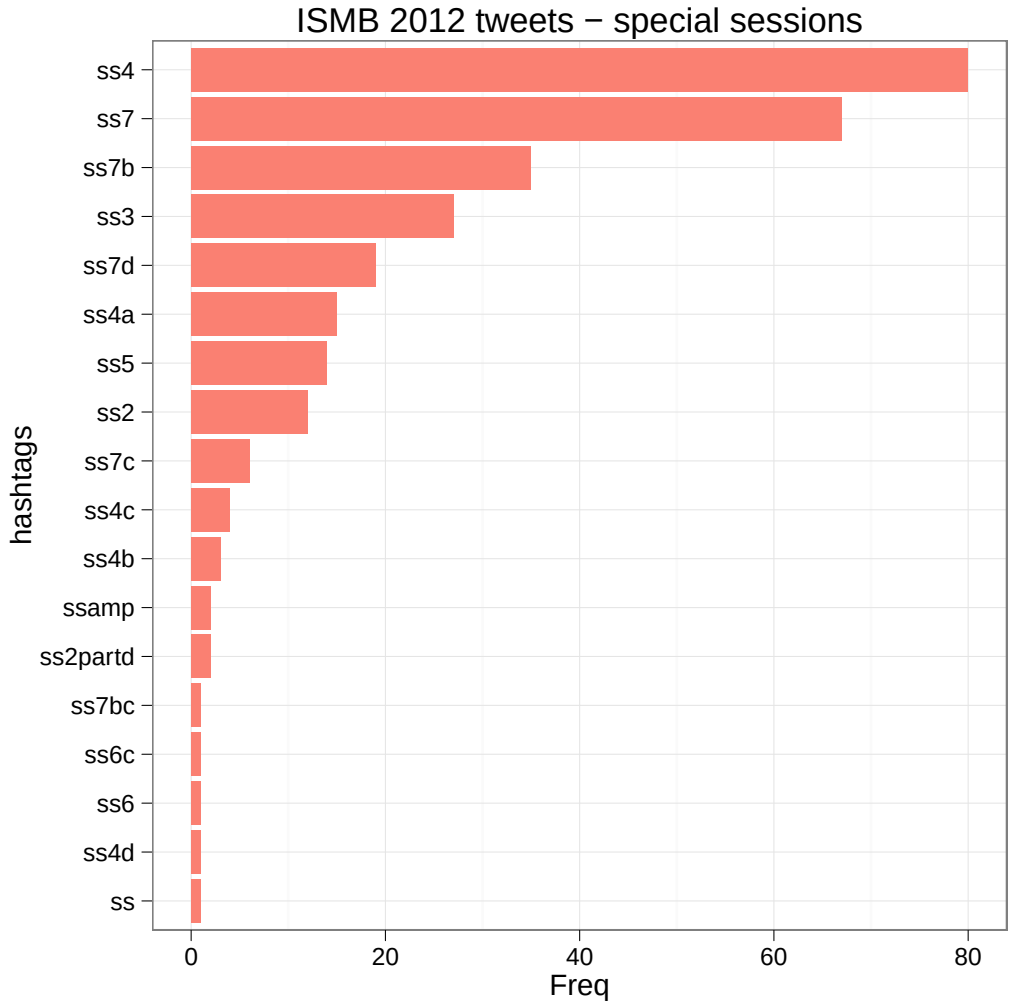
```
> pp <- ht[grep("^pp", ht$hashtags),]
> pp[47,2] <- 39
> pp <- pp[-5,]
> pp$hashtags <- factor(pp$hashtags, levels = as.character(pp$hashtags))
> print(ggplot(tail(pp, 20)) + geom_bar(aes(hashtags, Freq), fill = "salmon") +
+   coord_flip() + theme_bw() + opts(title = "ISMB 2012 tweets - published papers"))
```



PP 44: Toward interoperable bioscience data (Susanna-Assunta Sansone).

4.3 Special Sessions

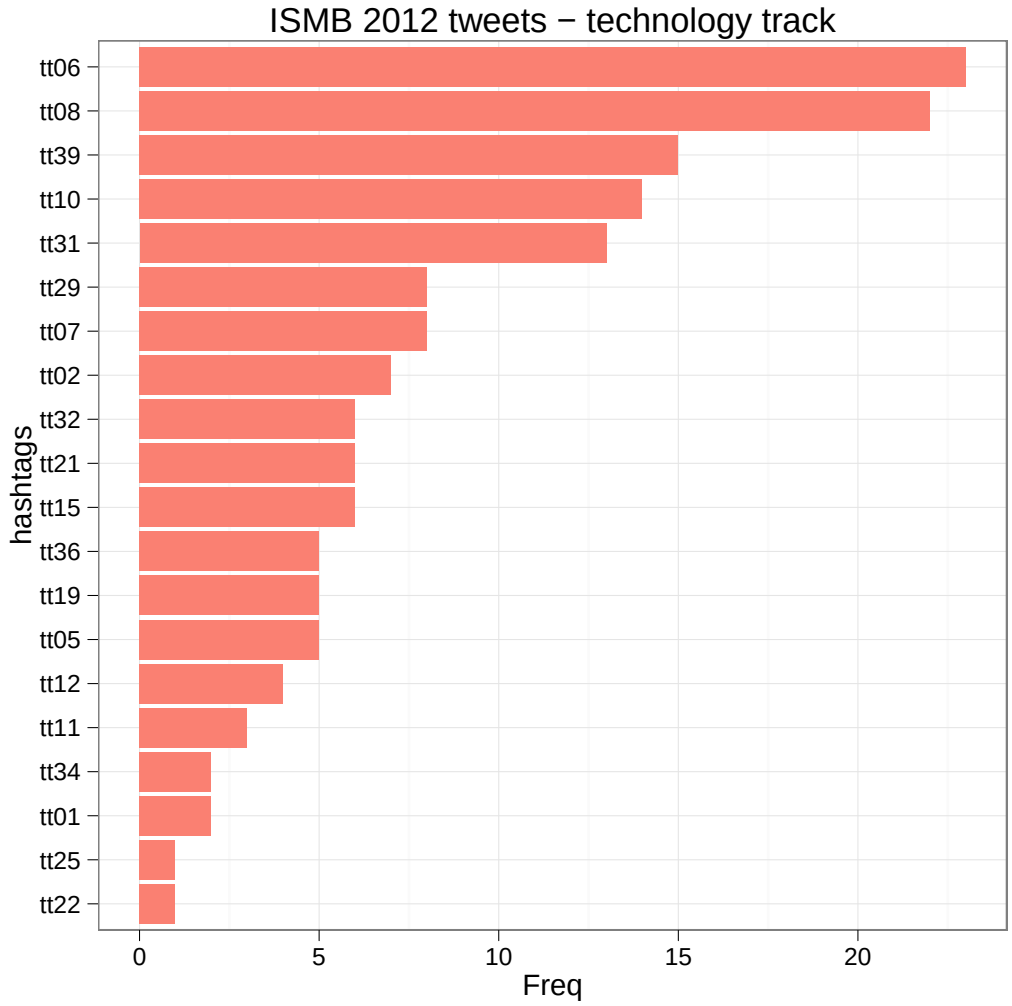
```
> ss <- ht[grep("^ss", ht$hashtags),]
> ss$hashtags <- factor(ss$hashtags, levels = as.character(ss$hashtags))
> print(ggplot(ss) + geom_bar(aes(hashtags, Freq), fill = "salmon") +
+   coord_flip() + theme_bw() + opts(title = "ISMB 2012 tweets - special sessions"))
```



SS 4: Bioinformatic Integration of Diverse Experimental Data Sources (Kyle Ellrott, David Haussler, Artem Sokolov, Josh Stuart).

4.4 Technology Track

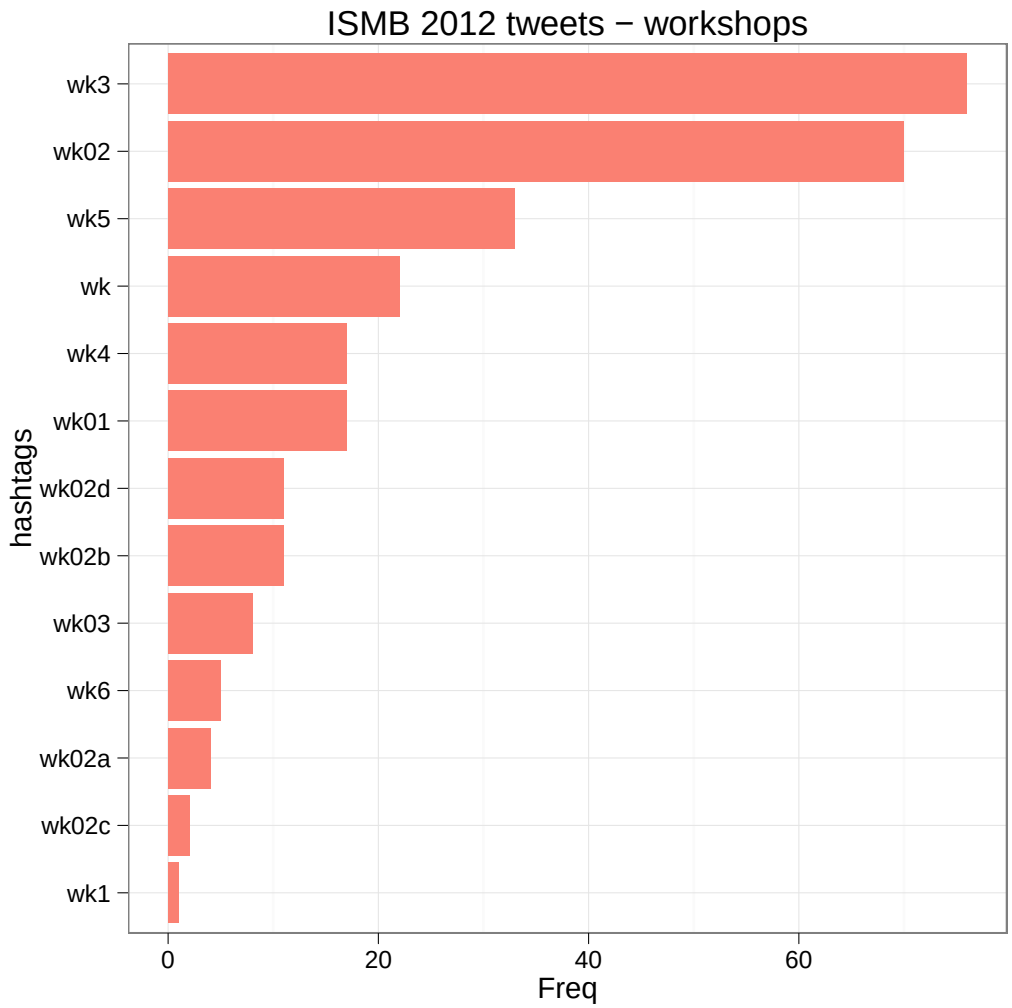
```
> tt <- ht[grep("^tt", ht$hashtags),]
> tt$hashtags <- factor(tt$hashtags, levels = as.character(tt$hashtags))
> print(ggplot(tail(tt, 20)) + geom_bar(aes(hashtags, Freq), fill = "salmon") +
+   coord_flip() + theme_bw() + opts(title = "ISMB 2012 tweets - technology track"))
```



TT06: The Taverna Server - Executing Scientific Workflows Remotely (Katy Wolstencroft).

4.5 Workshops

```
> wk <- ht[grep("^wk", ht$hashtags),]
> wk$hashtags <- factor(wk$hashtags, levels = as.character(wk$hashtags))
> print(ggplot(wk) + geom_bar(aes(hashtags, Freq), fill = "salmon") +
+   coord_flip() + theme_bw() + opts(title = "ISMB 2012 tweets - workshops"))
```

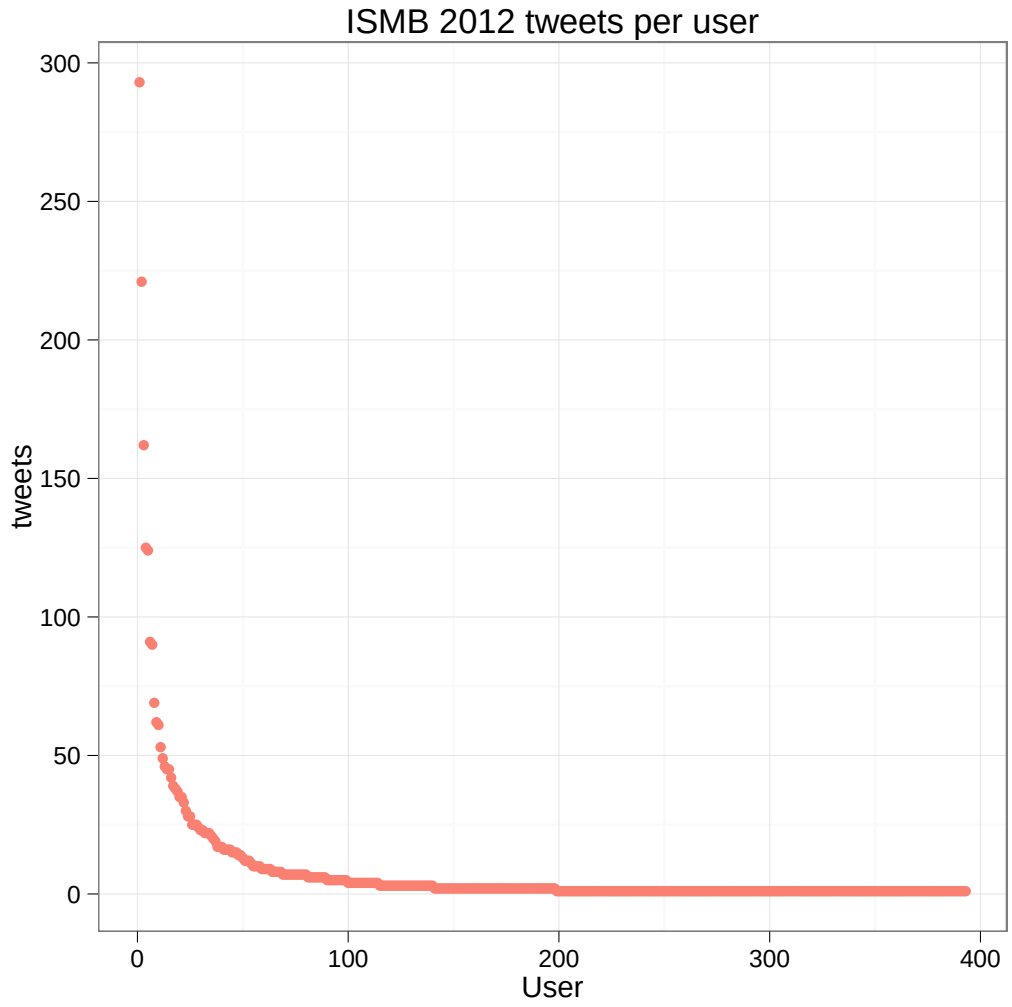


WK 3: Bioinformatics Core Facilities (Simon Andrews, Fran Lewitter, Brent Richter, David Sexton).

5 Users

5.1 The long tail

```
> users <- as.data.frame(table(ismb$screenName))
> colnames(users) <- c("user", "tweets")
> users <- users[sort.list(users$tweets, decreasing = T),]
> print(ggplot(users) + geom_point(aes(1:nrow(users), tweets), color = "salmon") + theme_bw() +
+   opts(title = "ISMB 2012 tweets per user") + xlab("User"))
```



5.2 The top 10

user	tweets
Chris_Evelo	293
genetics_blog	221
tladeras	162
iGenomics	125
WonderMixTape	124
alexishkin	91
bffo	90
spitshine	69
Albertagael	62
andrewsu	61

Table 1: Most tweets - top 10 users

6 Text mining

6.1 Word frequency

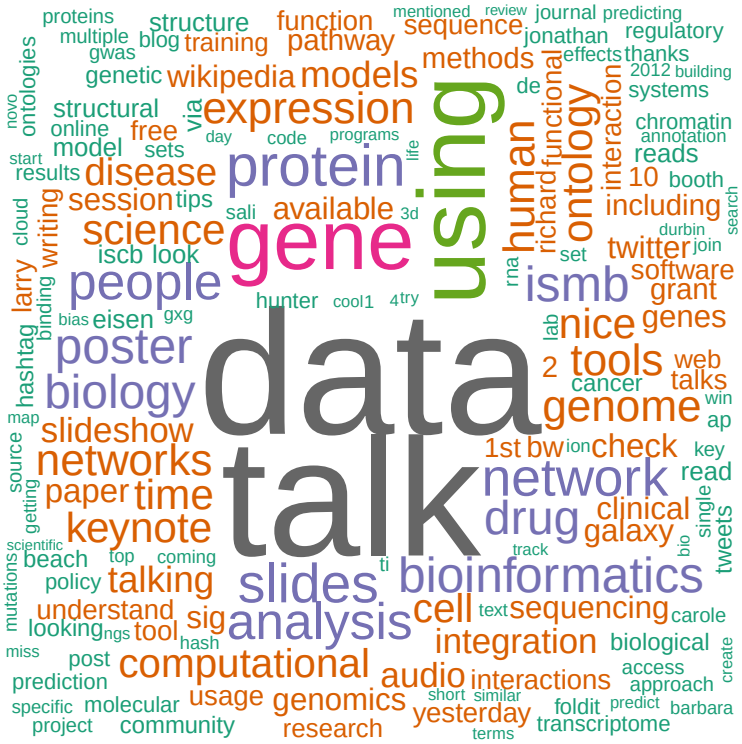
First, get all words composed only of alphanumeric characters.

```
> sw    <- stopwords("en")
> words <- lapply(words, function(x) x[grep("^[A-Za-z0-9]+$", x)])
> words <- unlist(words)
> words <- tolower(words)
> words <- words[-grep("^[rm]t$", words)]
> words <- words[!words %in% sw]
> words.t <- as.data.frame(table(words))
> words.t <- words.t[sort.list(words.t$Freq, decreasing = T),]
> print(xtable(head(words.t, 10), caption = "Top 10 words in tweets"), include.rownames = FALSE)
```

words	Freq
data	251
talk	238
using	132
gene	117
protein	80
slides	73
network	71
people	69
analysis	68
ismb	66

Table 2: Top 10 words in tweets

```
> pal2 <- brewer.pal(8, "Dark2")
> wordcloud(words.t$words, words.t$Freq, scale = c(8, .2), min.freq = 3,
+     max.words = 200, random.order = FALSE, rot.per = .15, colors = pal2)
```



6.2 Sentiment analysis

6.2.1 Emotions

```
> em <- classify_emotion(ismb$text, algorithm = "bayes")
> print(xtable(table(em[, "BEST_FIT"]), caption = "Tweet emotion"))
```

	V1
anger	24
disgust	4
fear	11
joy	149
sadness	44
surprise	37

Table 3: Tweet emotion

Shall we look at a "joyous" tweet?

Hooray! First #comicsans talk of the conference. Didn't take long. It's like an old friend that refuses to die. #ismb

6.2.2 Polarity

```
> po <- classify_polarity(ismb$text, algorithm = "bayes")
> print(xtable(table(po[, "BEST_FIT"]), caption = "Tweet polarity"))
```

	V1
negative	573
neutral	197
positive	2392

Table 4: Tweet polarity

A "positive" tweet:

Often forgotten: CS Optimizing (network) models only makes sense when you keep the model simple to not overfit #ISMB #netbiosiig

A "negative" tweet:

Chris Sander in Network Biology SIG and how translational medicine: tries to explain how complicated it all is in cancer biology #ISMB