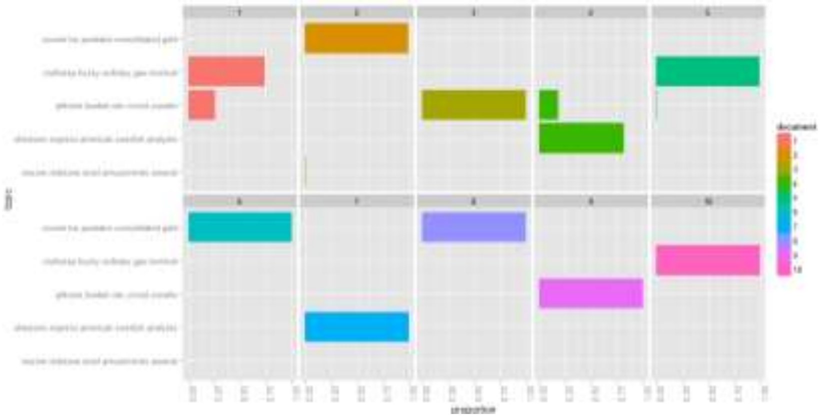
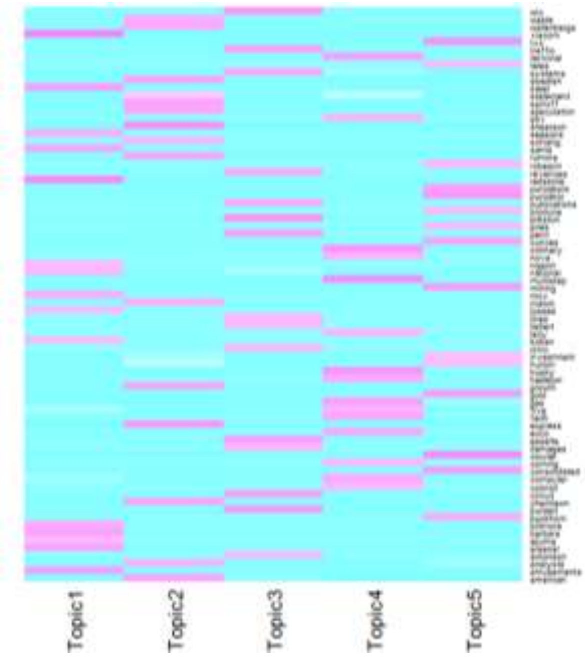
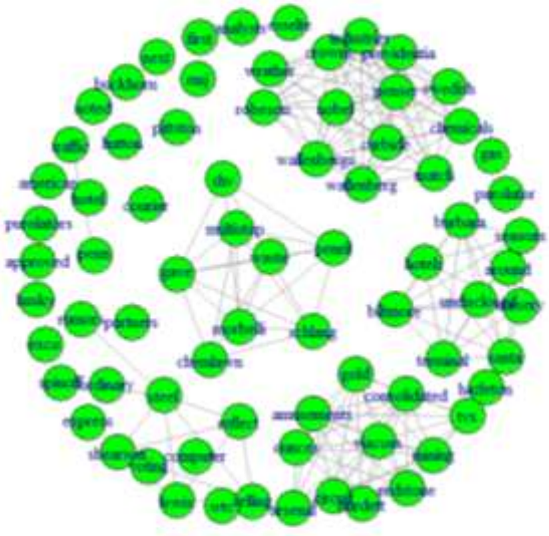
Text Mining Lab: Visualization of Topic Modelling

You will be able to create following visualization:

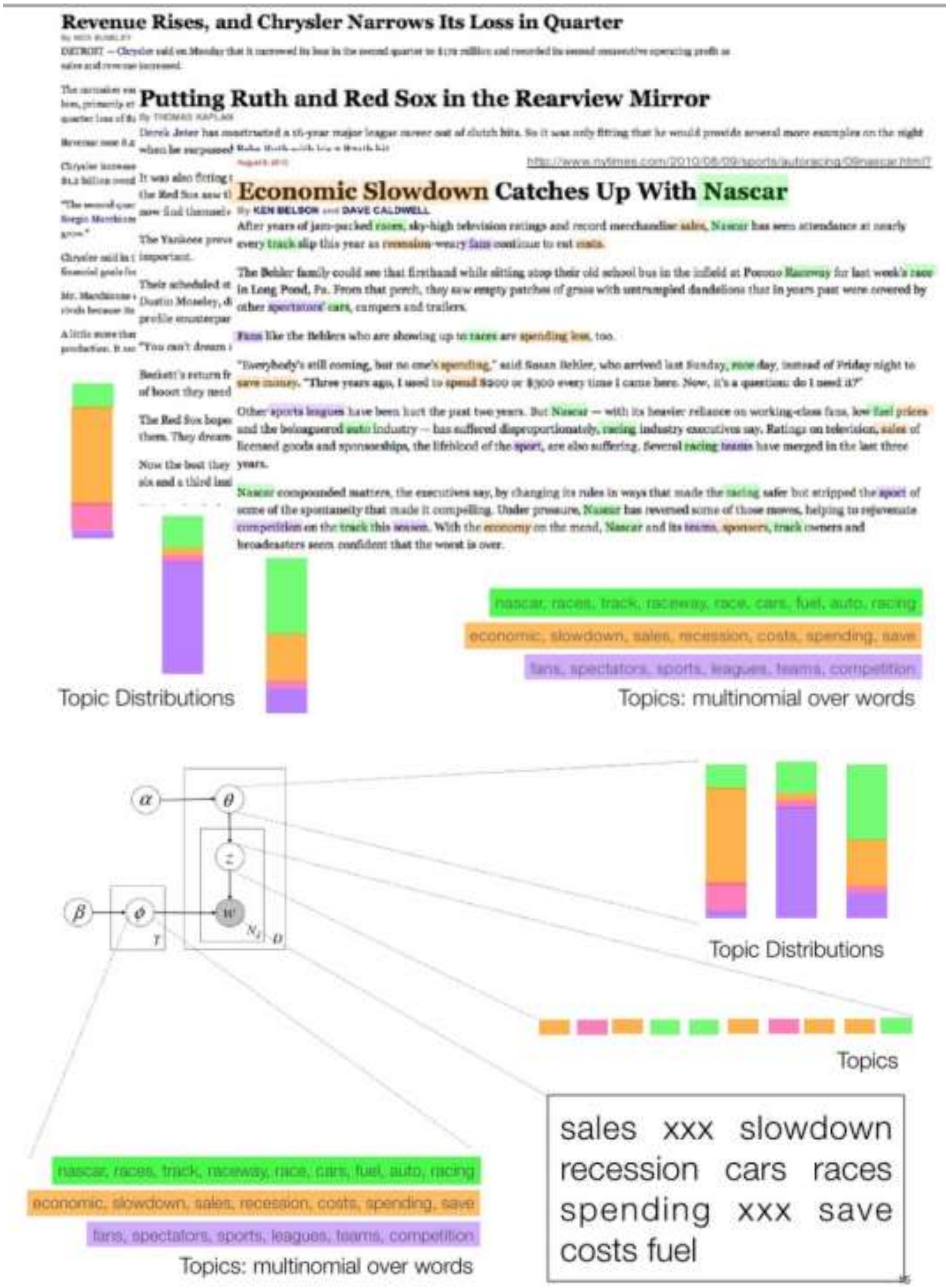






What is the "Topic Modeling"?

Probabilistic models for uncovering the underlying semantic structure of a document collection based on hierarchical Bayesian analysis of the original texts.



Revenue Rises, and Chrysler Narrows Its Loss in Quarter

DETROIT — Chrysler said on Monday that it narrowed its loss in the second quarter to $172 million and recorded its second consecutive operating profit as sales and revenue increased.

Putting Ruth and Red Sox in the Rearview Mirror

By THOMAS KAPLAN

Derek Jeter has constructed a 16-year major league career out of clutch hits. So it was only fitting that he would provide several more examples on the night when he surpassed Babe Ruth with his n Ruth hit

http://www.nytimes.com/2010/08/09/sports/autoracing/09nascar.html?

Economic Slowdown Catches Up With Nascar

By KEN BELSON and DAVE CALDWELL

After years of jam-packed races, sky-high television ratings and record merchandise sales, Nascar has seen attendance at nearly every track slip this year as recession-weary fans continue to cut costs.

The Behler family could see that firsthand while sitting atop their old school bus in the infield at Pocono Raceway for last week's race in Long Pond, Pa. From that perch, they saw empty patches of grass with untrampled dandelions that in years past were covered by other spectators' cars, campers and trailers.

Fans like the Behlers who are showing up to races are spending less, too.

"Everybody's still coming, but no one's spending," said Susan Behler, who arrived last Sunday, race day, instead of Friday night to save money. "Three years ago, I used to spend $200 or $300 every time I came here. Now, it's a question: do I need it?"

Other sports leagues have been hurt the past two years. But Nascar — with its heavier reliance on working-class fans, low fuel prices and the beleaguered auto industry — has suffered disproportionately, racing industry executives say. Ratings on television, sales of licensed goods and sponsorships, the lifeblood of the sport, are also suffering. Several racing teams have merged in the last three years.

Nascar compounded matters, the executives say, by changing its rules in ways that made the racing safer but stripped the sport of some of the spontaneity that made it compelling. Under pressure, Nascar has reversed some of those moves, helping to rejuvenate competition on the track this season. With the economy on the mend, Nascar and its teams, sponsors, track owners and broadcasters seem confident that the worst is over.

nascar, races, track, raceway, race, cars, fuel, auto, racing
economic, slowdown, sales, recession, costs, spending, save
fans, spectators, sports, leagues, teams, competition
Topic Distributions
Topics: multinomial over words
α
θ
z
β
ϕ
w
T
N_d
D
Topic Distributions
Topics
sales xxx slowdown
recession cars races
spending xxx save
costs fuel
nascar, races, track, raceway, race, cars, fuel, auto, racing
economic, slowdown, sales, recession, costs, spending, save
fans, spectators, sports, leagues, teams, competition
Topics: multinomial over words

LDA Illustration (Yohan&Alice, 2010)

Tools:

“Gensim” ,”Tethne” and “Mallet” on Python: http://devo-evo.lab.asu.edu/?q=node/474

“lda” and “topicmodels” on R: http://cran.r-project.org/web/packages/lda/ (or /topicmodels)

Data :

data “acq” from “tm” package: a list of 50 Reuters articles

Preprocess data for topic modeling

```
#Pre-process data:

>install.packages("tm")

> require("tm")

# Load the corpus of Reuters articles from the package of “tm”

>data(acq)

>reuters<-Corpus(VectorSource(as.vector(acq)))

#Convert to lower case

> reuters <- tm_map(reuters, tolower)
#Remove stop words

> reuters <- tm_map(reuters, removeWords, stopwords("english"))

# Creat an term-document matrix

> tdm <- TermDocumentMatrix(reuters, control = list(removePunctuation = TRUE, removeNumbers
= TRUE))

#Inspect() to check part of the term-document matrix

#Further Thought: How to stem words and typos?

# reuters <- tm_map(reuters,stemDocument) #May need to install “SnowballC”
```

Construct LDA model and visualize topics in documents

```
#LDA model

> install.packages("topicmodels")

> require("topicmodels")

#Choose number of topics (we arbitrarily choose 5 here) and type of methods(VEM or Gibbs)

>lda_model <- LDA(tdm, method="VEM", control = list(alpha = 0.1), k = 5)

>lda_inf <- posterior(lda_model, tdm)
```

```
# lda_inf contains two matrices: topics and terms

#Choose top five possible words to represent each topic

#Use top.topic.words()function from lda package to do so

>install.packages("lda")

>require("lda")

> top.words <- top.topic.words(t(lda_inf$topics), 5, by.score=TRUE)

#Get topic proportion in each documents

>topic.proportions <- t(lda_inf$terms) / colSums(lda_inf$terms)

#Assign topic as the top five words of each topic

> colnames(topic.proportions) <- apply(top.words, 2, paste, collapse=" ")

#Use melt() function from “reshape2” to transform data for the plot purpose

>install.packages("reshape2")

>require(reshape2)

> topic.proportions.df <- melt(cbind(data.frame(topic.proportions), document=factor(1:50)),
variable.name="topic",id.vars ="document")

#Visualize topics in documents:

>require(ggplot2)

> qplot(topic, value, fill=document, ylab="proportion", data=topic.proportions.df,
geom="bar")+opts(axis.text.x = theme_text(angle=90, hjust=1))+coord_flip()+facet_wrap(~
document, ncol=5)
```

Visualize words in topics

```
#Visualize words in topics

#Visualize top 15 words in each topic

>top.words <- top.topic.words(t(lda_inf$topics), 15, by.score=TRUE)

#Search for the top 15 words

> index<-c()

> for(element in top.words)   index<-c(index,which(rownames(lda_inf$topics)==element))

#Use heatmap() to plot

>hm<- heatmap(lda_inf$topics[index,], Rowv=NA, Colv=NA, col = cm.colors(256), scale="column", margins=c(5,10))

#Create another heatmap sorted by alphabetic order

> sub_matrix<-lda_inf$topics[index,]

#Use order() function to sort the matrix

>sorted_matrix<-sub_matrix[order(rownames(sub_matrix)),]

#Assign column names

>colnames(sorted_matrix)<-c("Topic1","Topic2","Topic3","Topic4","Topic5")

> hm2 <- heatmap(sorted_matrix, Rowv=NA, Colv=NA, col = cm.colors(256), scale="column", margins=c(5,10))
```

Visualize a word-network

```
#Visualize a network of words  :

> correlation<-cor(t(lda_inf$topics))

#Choose top 15 words of each topics from above to get a subset of correlation matrix

> index<-c()

> for(element in top.words)   index<-c(index,which(rownames(correlation)==element))

#Get a subset of correlation matrix

> adjacency<-correlation[index,index]

>diag(adjacency)<-0

#Use functions from "igraph" package to construct a network object and then plot it.

>install.packages("igraph")

> require("igraph")

#Convert adjacency matrix to a network object

>g<-graph.adjacency(adjacency,mode="undirected")

> plot(g,layout=layout.kamada.kawai,vertex.color="green")
```

After Class Exercise:

Use the “get.edgelist()” function from the “igraph” package and import edgelist to a .csv file using write.csv() function. Reconstruct a nicer word-network on Gephi.

Reference:

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993-1022.

Yohan Jo& Alice Oh (2010)[Slides Show], Aspect and Sentiment Unification Model, *AMC Websearch and Data Mining* https://vialogues.com/vialogues/play/14351