Unsurprised leanring: PCA vs Clustering

~~wordcloud~~

~~remove stopwords “XXXX”~~

~~more clusters~~

LDA text modeling: <https://github.com/cpsievert/LDAvis>

n-gram: like tri-gram

ROC, PR, and the area under them as a metric

*\*to remove “xxxx” in the narrative*

removeNumPunct <- function(x){gsub("[^[:alpha:][:space:]]\*", "", x)}

removeXs <- function(x){gsub("[Xx]{2,}", "", x)}

clean\_corpus <- function(corpus){

corpus <- tm\_map(corpus, removePunctuation)

#corpus <- tm\_map(corpus, removeWords, c("XXXX", "XXXXXXXX", "XXXXXXXXXXXXXXXX", "XXXXXXXX"))

corpus <- tm\_map(corpus, content\_transformer(removeXs))

corpus <- tm\_map(corpus, content\_transformer(tolower))

#corpus <- tm\_map(corpus, content\_transformer(replace\_symbol))

corpus <- tm\_map(corpus, removeWords, c(stopwords("en")))

# We could add more stop words as above

corpus <- tm\_map(corpus, removeNumbers)

corpus <- tm\_map(corpus, content\_transformer(removeNumPunct))

corpus <- tm\_map(corpus, stripWhitespace)

return(corpus)

}

summary: <https://arxiv.org/abs/1704.04368>

IBM Watson Discovery

*\*\*\*dashboard*