**Topic Extraction** is an integral part of IE (Information Extraction) from Corpus of Text to understand what are all the key things the corpus is talking about. While this can be achieved naively using unigrams and bigrams, a more intelligent way of doing it with an algorithm called RAKE is what we’re going to see in this post.

**Udpipe**

udpipe is an NLP-focused R package created and opensourced by this organization bnosac. Thanks to them, udpipe is the R package that many a times solves the pain of not having native spacy for R.

**Udpipe – Installation**

install.packages("udpipe")

**Udpipe – Loading**

library("udpipe")

**Udpipe – Language Model**

An NLP library is as good as its Language Model because the Language model contains the recipe of how to annotate your text corpus. So, before we proceed further, we need to download the language model for us to use. In this case, We’ll download English Language model as we’re going to do Topic Extraction for English Reviews (Text).

en <- udpipe::udpipe\_download\_model("english")

Language model, once downloaded can be used later on without requiring to be redownloaded for every session.

**Customer Reviews – Extraction**

We’ll use itunesr package to extract reviews of Amazon US App from Apple App Store.

library(itunesr)

reviews1 <- getReviews("297606951", "us", 1)

reviews2 <- getReviews("297606951", "us", 2)

reviews <- rbind(reviews1, reviews2)

head(reviews)

## Title

## 1 Fine Anything Easy, Good Policies

## 2 Customer support

## 3 Uh oh, something went wrong on our end

## 4 Connection Lost

## 5 Add this app to the I-Pads

## 6 Wish lists!

## URL Name

## 1 <https://itunes.apple.com/us/reviews/id899889795> KeithAppProgrammer

## 2 <https://itunes.apple.com/us/reviews/id978296731> Stormdoll

## 3 <https://itunes.apple.com/us/reviews/id33953389> Joker1138

## 4 <https://itunes.apple.com/us/reviews/id8865955> Loquacious lair

## 5 <https://itunes.apple.com/us/reviews/id43459956> MattC4U

## 6 <https://itunes.apple.com/us/reviews/id389452759> Best update ever12345

## App\_Version Rating

## 1 13.15.0 5

## 2 13.15.0 5

## 3 13.15.0 1

## 4 13.15.0 2

## 5 13.15.0 1

## 6 13.15.0 1

## Review

## 1 We’ve been quite blessed to work with Amazon. Searching for odd items, the App also has some compatibility safeguards. If I need to return something, it really couldn’t be easier.

## 2 I love not having to call if there is an issue. The mobile app has great automated features to reach someone and when there is a problem it’s resolved quickly and in the manner I request instead of just a refund . - meaning I was able to get half of my order refunded and the other half mailed again as my first package was listed lost. The items I needed more quickly than could arrive were swiftly refunded and the other items mailed again without a problem this time - super convenient!

## 3 Constantly getting the above error message combined with random pictures of dogs. Hasn’t been fixed for a couple weeks. Pretty frustrating.

## 4 The app is constantly crashing and telling me that the network connection has been lost even if I have full access to WiFi or data.

## 5 This makes me so mad.

## 6 What did you do Amazon? Changing the way we saved wish list items was a horrible idea. Whoever came up with this heart update instead of holding and dropping needs to be demoted immediately. Please fix this. We also need Amazon smile ability in the app as well.

## Date

## 1 2019-08-21 13:54:37

## 2 2019-08-21 11:39:40

## 3 2019-08-21 10:21:20

## 4 2019-08-21 07:11:33

## 5 2019-08-21 05:25:44

## 6 2019-08-21 05:20:25

At this point, We’ve about 98 Reviews (Text) of Amazon iOS App from US Apple Store.

**Customer Reviews – Only Negative (1 & 2-star)**

We’ll pick only the negative reviews (1 & 2-star) to understand what pain points are customers talking about while rating Amazon bad.

reviews\_neg <- reviews[reviews$Rating %in% c('1','2'),]

nrow(reviews\_neg)

## [1] 68

**Customer Reviews – Annotation**

We’re going to do Topic Extraction from the above extracted 70 Reviews. But before we can proceed with Topic Analysis, We need to annotate the text with the language model that we downloaded above.

model <- udpipe\_load\_model("english-ewt-ud-2.3-181115.udpipe")

doc <- udpipe::udpipe\_annotate(model, reviews\_neg$Review)

Let’s look at the object doc to see what’s there in it.

names(as.data.frame(doc))

## [1] "doc\_id" "paragraph\_id" "sentence\_id" "sentence"

## [5] "token\_id" "token" "lemma" "upos"

## [9] "xpos" "feats" "head\_token\_id" "dep\_rel"

## [13] "deps" "misc"

Considering the scope of this post is **Topic Analysis**, I’ll leave out the basics of NLP (to understand the above terms, if you’re not familiar) for another post.

**Topic Extraction using RAKE**

**RAKE** stands for Rapid Automatic Keyword Extraction.

doc\_df <- as.data.frame(doc)

topics <- keywords\_rake(x = doc\_df, term = "lemma", group = "doc\_id",

relevant = doc\_df$upos %in% c("NOUN", "ADJ"))

head(topics)

## keyword ngram freq rake

## 1 error message 2 2 2.375000

## 2 new layout 2 2 2.000000

## 3 promo pricing 2 2 2.000000

## 4 latest update 2 2 1.857143

## 5 same app 2 2 1.674242

## 6 multiple item 2 3 1.666667

Voila! Topics (or as technically it goes, Keywords) have been extracted using RAKE. As the output above states, we also get to see few metrics like ngram, freq and rake score against those Topics.

**Topic Analysis**

Let’s load up tidyverse to kickstart our Analysis

library(tidyverse)

and make a bar chart of the top 10 topics based on the rake score.

topics %>%

head() %>%

ggplot() + geom\_bar(aes(x = keyword,

y = rake), stat = "identity",

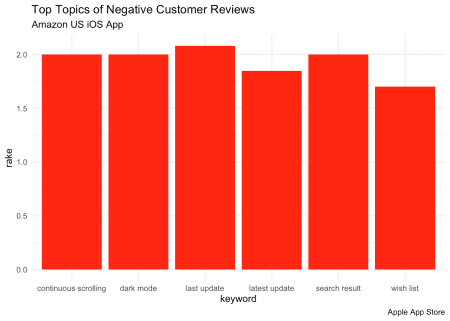
fill = "#ff2211") +

theme\_minimal() +

labs(title = "Top Topics of Negative Customer Reviews",

subtitle = "Amazon US iOS App",

caption = "Apple App Store")



That’s a nice plot indicating the top customer pain points. Seems the latest update and its error messages didn’t go well with the Customers. This is a simple bar plot but the output of RAKE could also be used to make a correlation plot between rake score and freq to add extra dimension in understanding More frequently occuring topics.

**Summary**

udpipe is a very handy package if you are in the business of NLP and Text Analytics. It also supports multiple other Languages like German, French other than English.