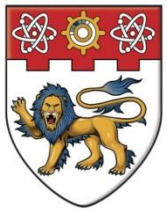


NANYANG TECHNOLOGICAL UNIVERSITY

BC2406 Business Analytics I: Predictive Techniques

Seminars 10 & 11
Text-mining and Applications

Instructor: Prof. Lee Gun-woong

Nanyang Business School

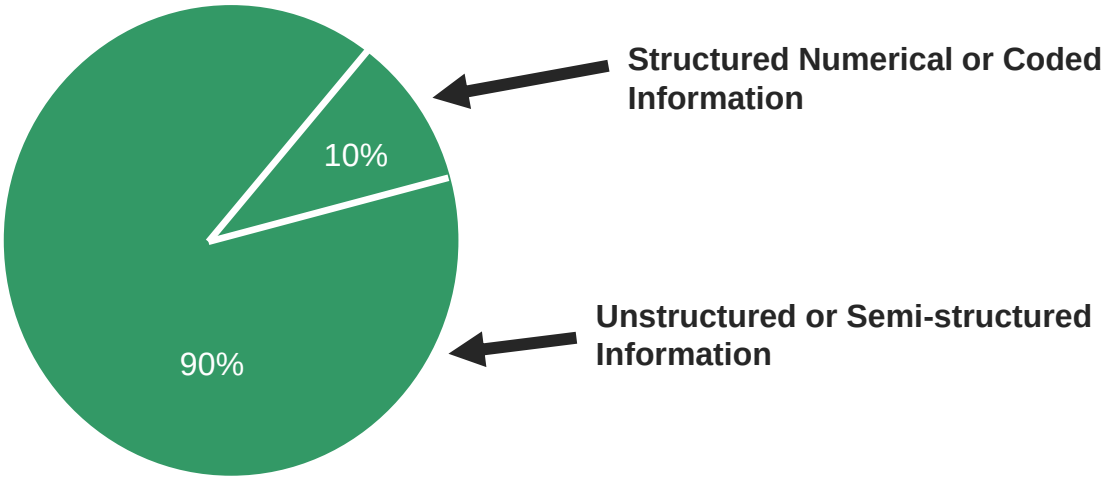
Outline

- Overview of Text-mining
- Mobile App Analytics with TM
  - Business Problem
  - Data Structure in Text-mining
    - Data Selection & Creating Corpus
  - Preparation for Text-mining
    - Parsing & Stemming
  - Mining Textual Information
    - Document Term Matrix (DTM)
  - Evaluation
    - Frequency & Associations
    - Visualization
  - Applications
    - Clustering
    - Regressions

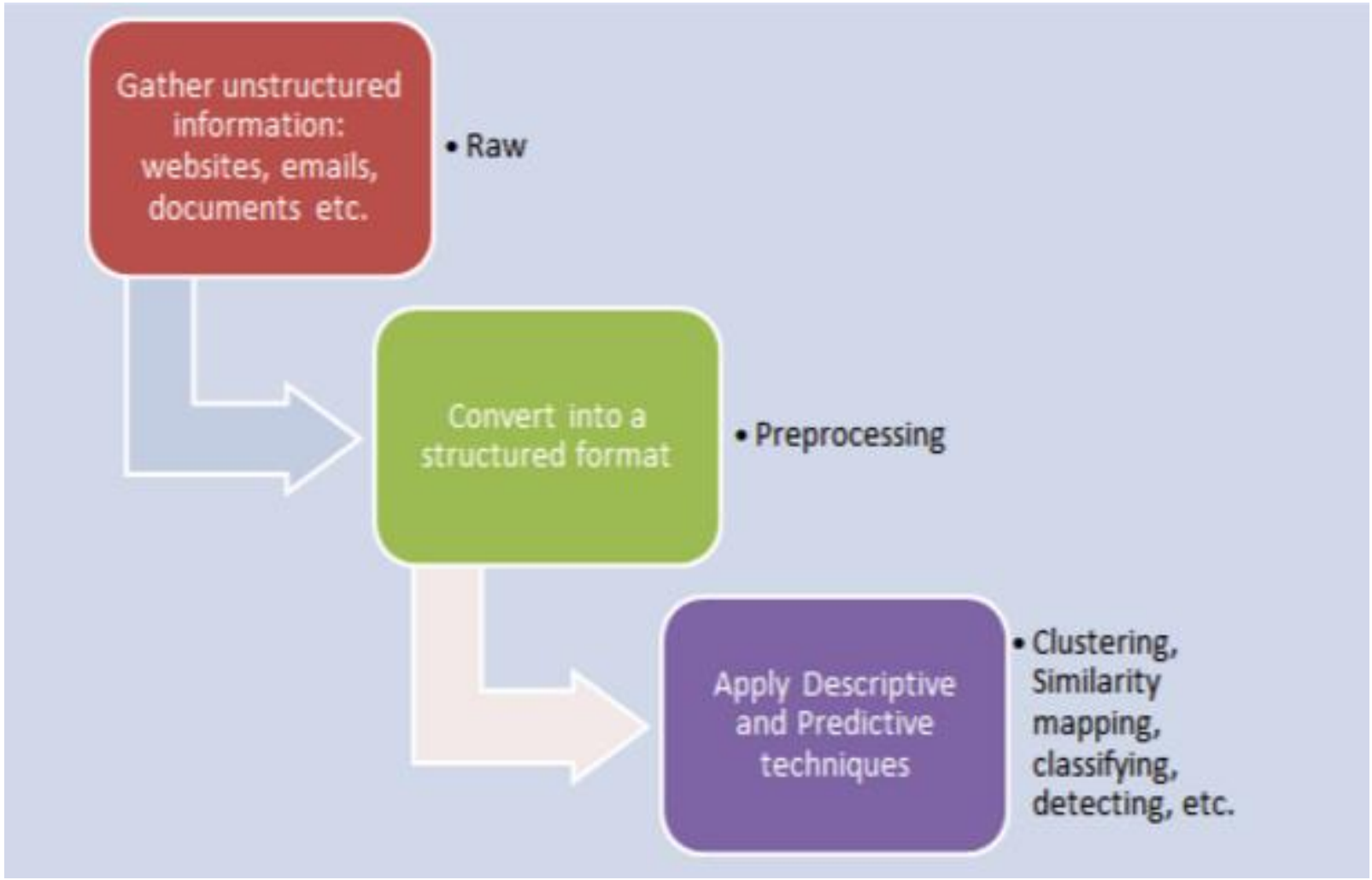
Overview of Text-mining

Text-Mining (TM)

- Definition
  - A process of identifying novel information from a collection of texts
  - Transform from unstructured textual data into structured data
- Motivation
  - Approximately 90% of the world's data is held in unstructured formats (Oracle, 2015)
  - Information intensive business processes demand that we transcend from simple document retrieval to "knowledge"



Text Mining Process



Source: www.simafore.com, 2016

Natural Language Processing

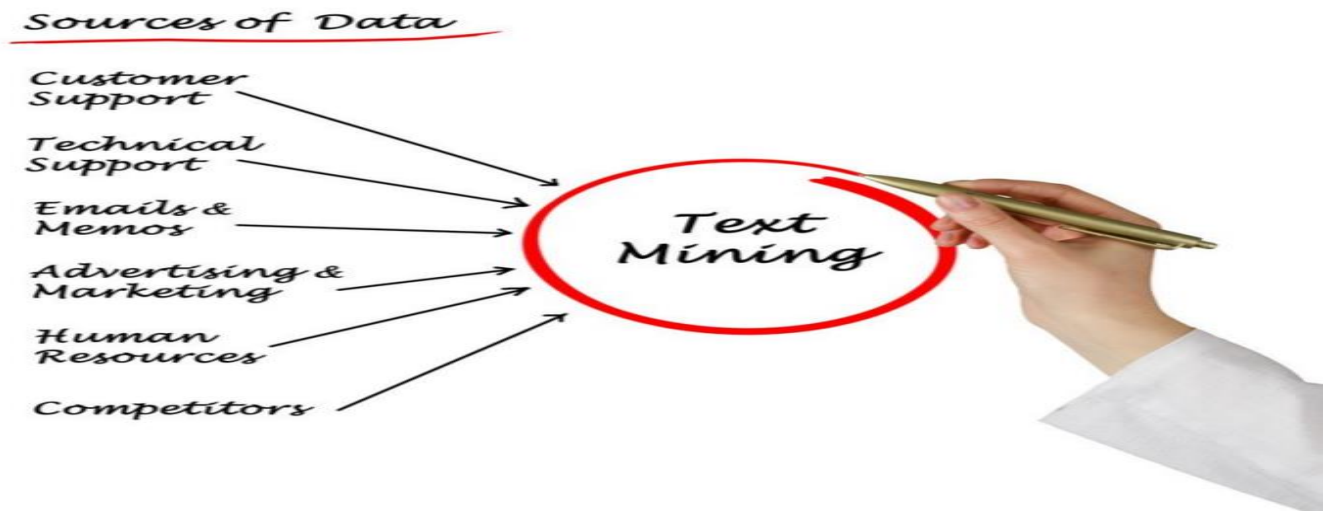
- **Natural language processing (NLP)** is a subfield of artificial intelligence and linguistics. It studies the problems of automated generation and understanding of natural human languages.

- Natural language processing uses probabilistic, statistical, and data-mining methods to resolve some of the difficulties
  - e.g., text segmentation

Business Applications

- **Decision Support in CRM**
  - What are customers' typical complaints?
  - What is the trend in the number of satisfied customers?

- **Personalization in e-Commerce (Recommendations)**
  - Suggest products that fit a user's interest profile (even based on personal info).



Text-mining Tools

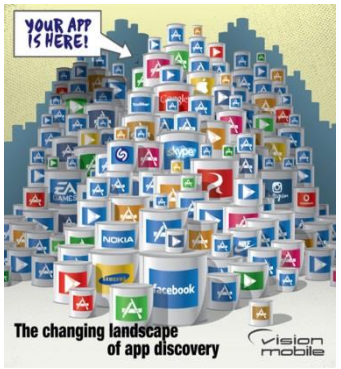
- Commercial SW Packages
  - IBM Intelligent Miner for Text
  - SAS Text Miner
  - Semio Map
  - InXight LinguistX / ThingFinder
  - LexiQuest
  - ClearForest
  - Teragram
  - SRA NetOwl Extractor
  - Autonomy

- Open-Source SW
  - R, Python, Pearl …

Mobile App Analytics with TM

Mobile App Markets

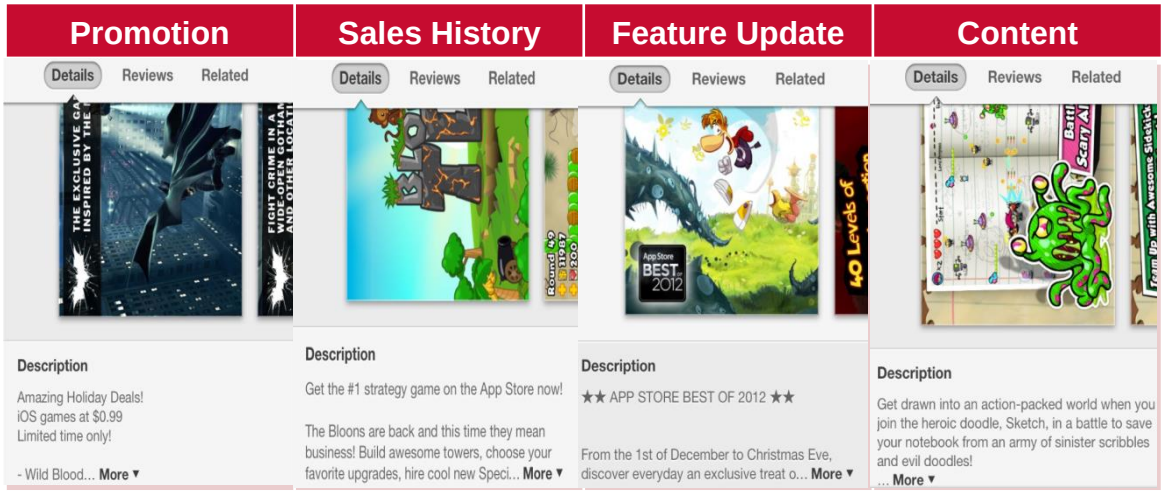
- **Opportunities**
  - **Users**: Broaden customers' access to a large database of Apps
  - **Sellers**: Create a large network of customers and accommodate heterogeneous customer preferences
- **Challenges**
  - **Users**: Many Similar Apps
    - Increased cognitive load searching for the right App
    - High transaction uncertainties related to the value/quality of an App
  - **Sellers**: Severe Competition
    - 942 new Apps appeared every single day (October 2013, 148Apps.com)
    - "*Over 30% Apps are never downloaded*" (July 2012, PCMac.com
    - "*Top 25 US developers account for half of App revenue*" (December 2012, TechCrunch)



NANYANG TECHNOLOGICAL UNIVERSITY

Motivation

- Marketing Messages in App Product Descriptions



- Developers can deliver App information and marketing messages in the head of **App product descriptions**.
  - **Price Promotion**: " *Now 50% OFF for a Limited Time*"
  - **Sales Performance**: "*#1 Downloaded Game App in 2013!*"
  - **Feature Update**: "*Feature Updates! Retina/HD Support! …*"
  - **Content**: "**Minecraft is about placing blocks to build things…**"

NANYANG TECHNOLOGICAL UNIVERSITY

How to Write a Successful App Description ?

- **"Write your Application Description with a focus on what makes the functionality of design of your App unique (should be less than 4,000 characters" (**Apple App Store**)**
- **Product Description Marketing** (SEO Service, 2012)
  - List Benefits, not Features!
  - Let Customer Describe it!
  - Make your Content Unique!
  - Provide Additional Value for the Price Customers Pay!

Step1: Identify Business Problem(s)

- **Business Problem: How to write successful App product descriptions?**
- **Task1:** Which terms/words are frequently appeared in App product descriptions?
  - **Text-Mining**
- **Task2:** What are the common patterns/styles in presenting App product descriptions?
  - **Clustering Analysis**
- **Task3:** What types of messages in the descriptions are associated with better sales performance?
  - **Regression Analysis**

Step1: Identify Business Problem(s)

- **Text-mining Procedure**
  - **Extracting Keywords from App Descriptions**
    - 250 descriptions over 20 weeks
    - Only first two sentences will be used for TM
  - **Preprocessing** (Parsing and Stemming)
    - Converts Unstructured Data to Structured Data
  - **Keywords Selection (Task 1)**
    - Selects the most frequently appearing key terms
  - **Clustering (Task 2)**
    - Groups frequently used terms together
    - Identify the common patterns/styles in presenting App product descriptions
  - **Regression Analysis (Task 3)**
    - Identify relationships between the messages in product descriptions and App sales

Required R Packages for TM

- To conduct Text-mining in R, we need the following packages:

```
##  Required Packages  ##
install.packages("NLP") #NLP Functions
install.packages("tm")  #Text-mining
install.packages("SnowballC") #Stemming
install.packages("wordcloud") #WordCloud
install.packages("RColorBrewer")  #Colours for WordCloud

##  Activate Packages ##
library(NLP)
library(tm)
library(SnowballC)
library(wordcloud)
library(RColorBrewer)
library(stats)      # Clustering
```

Step2: Understand the Data

- **Data**
  - 250 App product descriptions in the Top Paid Games Apps from Apple App Store
  - Games Category
    - Games Apps makes up 17.2 percent of the Apps on the Apple App Store in 2013*
    - It accounts for 40% of store downloads and covered 70% of App Store revenue in 2013**
  - Paid Charts
    - No dominant zero-price effects (c.f., free and grossing charts)
    - Less inclusion of in-app-purchase options
  - 20 weeks (04/18/2012 ~ 08/17/2012)
    - 250 App product descriptions were randomly selected

Sources: **Apple App Store Metrics(October 2013, 148Apps.com)*
***App Market Report Q1 2013: iOS App Store Revenue (August 2013, AppAnnie.com)*

Step2: Understand the Data

- Load textual data
  - Download App_Games_Desc.csv from the course site

```
## Data Loading ##
Apps_data <- read.csv("App_Games_Desc.csv") #The original dataset is stored in your project folder.
Apps_desc <- Apps_data[,10]                 #Descriptions are recorded in the 10th column.
n_desc <- length(Apps_desc)                 #Check the number of app descriptions
n_desc
```

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	AppID	Rank	Name	ReleaseDate	Category	Price	Developer	Screenshot	Size	Description	UpdatedDate	StarCurrentVersion	RatingCurrentVersion	TopInAppPurchases	DataDate
2	394716128	232	Urban Ninja	10,01,2010	Games	0.99	Donut Games	7	9.6	[u'-- #1 DOWNLOAD IN 28 COUNTRIES --', u"	06,14,2011	4	4426	0	04,06,2012
3	535747605	183	Master of Dungeon	07,19,2012	Games	0.99	PLAYBEAN	10	44.2	[u'\u2605\u2606 LIMITED TIME OFFER : SPEC	08,07,2012	4.5	139	5	08,10,2012
4	465072566	292	Galaxy on Fire 2??? HD	10,07,2011	Games	4.99	FISHLABS	10	622	[u'\u2729 50% OFF \u2013 SPECIAL SALE: Bu	07,03,2012	4.5	122	7	07,27,2012
5	322852954	26	Words With Friends	07,19,2009	Games	2.99	Zynga	5	13.3	[u'Play the AD-FREE version of Words With	03,15,2012	4.5	13657	8	04,06,2012
6	337402021	211	Harry Potter: Spells	11,15,2009	Games	0.99	Warner Bros.	5	102	[u'Now available at $0.99 for a limited time	11,04,2010	4	2715	2	04,06,2012
7	404086528	56	LEGO Harry Potter: Years 1-4	11,19,2010	Games	0.99	Warner Bros.	10	453	[u'Now available at $0.99 for a limited time	01,18,2011	4	2749	0	04,06,2012
8	342842881	109	Where's Waldo???? The Fantastic Journe	12,09,2009	Games	0.99	Ludia	5	169	[u'***March Madness Limited Time Promo -	04,05,2011	4	4631	0	04,06,2012
9	414723566	293	Smoothie Moves	01,19,2011	Games	0.99	Gameloft	10	834	[u'New iOS game sale! Order & Chaos\	05,31,2012	4.5	2750	10	08,17,2012
10	441242571	118	RPG - Fantasy Chronicle	06,07,2011	Games	0.99	Kotobuki Sol	5	28.4	[u'** LIMITED OFFER: SPECIAL PRICE !!! 85%	01,19,2012	4.5	20	0	05,04,2012
11	489454710	294	Adventure Bar Story	03,08,2012	Games	0.99	RideonJapan	5	19.4	[u'\u2605\u2605 Summer Sale! \u2605\u260	06,22,2012	4.5	66	7	07,20,2012
12	500428935	67	Brandnew Boy	03,08,2012	Games	0.99	Oozoo Inc.	10	121	[u'**** $3.99 -> $0.99 **** HOT SALE FOR	04,20,2012	4.5	107	5	04,20,2012
13	307132353	133	Sally's Spa	04,08,2009	Games	0.99	Games Cafe I	5	26.7	[u'Apple\u2019s Rewind 2009! Apple featur	07,07,2011	4.5	1499	0	04,06,2012
14	344186162	31	Grand Theft Auto: Chinatown Wars	01,18,2010	Games	0.99	Rockstar Gam	5	232	[u'**LIMITED TIME OFFER**', u'We\u2019re	03,28,2010	4	7922	0	05,25,2012
15	365678365	97	CHAOS RINGS	04,20,2010	Games	3.99	SQUARE ENIX	5	700	[u'---', u'The CHAOS RINGS series summer s	07,26,2012	4	3908	0	07,27,2012
16	373849969	288	Crash Bandicoot Nitro Kart 2	05,27,2010	Games	0.99	Activision Pu	5	71.8	[u'** To celebrate the release of the new Sk	06,15,2010	3.5	1288	0	04,27,2012
17	396669943	292	Death Worm	11,05,2010	Games	0.99	PlayCreek	10	25.4	[u'* LIMITED TIME OFFER - $0.99 * ', u"Over	10,04,2011	4.5	909	0	04,13,2012
18	399045373	208	Aralon: Sword and Shadow HD	12,16,2010	Games	0.99	Crescent Moo	10	311	[u'---- For a limited time only, on sale for ov	01,09,2012	4.5	716	6	04,27,2012
19	414664715	269	Order & Chaos?? Online	04,27,2011	Games	6.99	Gameloft	10	775	[u'Finally a true real-time, full-3D MMORPG	03,15,2012	4.5	1253	10	04,13,2012
20	418520600	255	RPG???Eve of the Genesis	02,18,2011	Games	0.99	Kotobuki Sol	5	16.2	[u'** LIMITED OFFER: SPECIAL PRICE!!! 89% (	12,09,2011	4	40	0	05,25,2012
21	429000513	204	Sally's Studio	06,16,2011	Games	0.99	Games Cafe I	5	42.1	[u'\u222a EASTER WEEKEND SALE!!! \u222a	06,23,2011	4.5	844	0	04,06,2012
22	433127774	230	DrawRace 2	09,01,2011	Games	0.99	Chillingo Ltd	5	300	[u'\u2605\u2605\u2605\u2605\u2605 MOTC	03,23,2012	4.5	80	6	05,18,2012
23	454316134	155	Sonic CD	12,15,2011	Games	2.99	SEGA	10	258	[u'Don\u2019t miss out on the Sonic Sale of	12,15,2011	4.5	7383	0	06,22,2012
24	454679700	243	Mega Mall Story	08,08,2011	Games	0.99	Kairosoft Co.	5	10.8	[u'--LIMITED TIME OFFER! 75% OFF!--', u'Des	08,09,2011	4.5	504	0	06,01,2012
25	456244246	112	Blueprint 3D	11,10,2011	Games	0.99	FDG Entertair	5	55.6	[u'AN ENTIRELY NEW GAMING EXPERIENCE '	02,28,2012	5	7034	2	04,06,2012
26	463928420	12	Are You Smarter Than a 5th Grader??? &a	06,11,2012	Games	0.99	Ludia	10	20.6	[u'--LIMITED TIME SALE - GET THE AD FREE V	06,11,2012	4	188	6	06,16,2012
27	464308703	119	Prince of Persia?? Classic	02,29,2012	Games	0.99	Ubisoft	5	194	[u'LIMITED TIME OFFER!', u'Prince Of Persia	04,12,2012	4	104	1	04,13,2012
28	465918332	36	RPG - Symphony of Eternity	09,26,2011	Games	0.99	Kotobuki Sol	5	33.5	[u'** LIMITED OFFER: SPECIAL PRICE !!! 89%	10,25,2011	4.5	71	0	04,06,2012
29	473645314	161	RPG Alphadia	10,25,2011	Games	0.99	Kotobuki Sol	5	43.5	[u'** LIMITED OFFER: SPECIAL PRICE!!! 86%	10,26,2011	4	66	0	05,18,2012
30	479799035	141	MapleStory Live Deluxe	02,23,2012	Games	0.99	NEXON MOBI	5	35.2	[u'\u2605\u2605\u2605 Big Sale for 7 days! '	07,13,2012	4	86	6	07,27,2012
31	488627858	1	Draw Something by OMGPOP	02,01,2012	Games	0.99	OMGPOP	10	12.9	[u'**LIMITED TIME ONLY SALE - UPGRADE TO	03,19,2012	5	105054	4	04,06,2012

Step3: Data Preparation

- Use the first two sentences
  - App developers are likely to deliver the key marketing messages in the head of App product descriptions.

```
## Use the First Two Sentences ##
# Test with the first Description #
Test_sent <- unlist(strsplit(as.character(Apps_desc[1]), split = "[.!?]+")) #Split a document into sentences
Test_sent
```

```
 [1] "[u'-- #1 DOWNLOAD IN 28 COUNTRIES --', u\"What's even better than Frenchmen, ninjas and heroes"

 [2] "\", u'A chubby Ninja Hero with a French accent of course"

 [3] "', u'Join Francois in a wall climbing, roof jumping adventure as a secret agent"

 [4] "', u\"Wrapped in a tight latex suit, he's ready to take on a list of classified missions entrusted him by his agency"

 [5] "\", u'Can you handle the dangerous rope swinging, pole slidning and acrobatics required to survive 40 breath-taking missions"

 [6] "', u'* * * * * * * * * * * * * * * * * * * * * * * * * * * *', u'WHAT REVIEWERS SAY:', u'- iPild"

 [7] "de: \"Gameplay 20 / 20\"', u'- Apple: Rated \"New &amp;amp; Noteworthy\" in iTunes', u'- Apple: Rated \"Hot New Game\" in App Store', u'- Touch Arcade: \"It wouldn\
\'t surprise me if I went back and ran the numbers that Donut Games is responsible for more favorable game reviews than any other developer\"', u'- GamePro: Score 4"
 [8] "5 / 5"

 [9] "0 in \"App Store Games of the Week\"', u'WHAT FANS SAY:', u'- \"Donut Games hits another one outta the park"

[10] "\"', u'- \"This is an awesome game"

[11] " Very addictive and challenging"

[12] " My 6yo loves it too"

[13] "\"', u'- \"Another great game from donut games"

[14] " Fun, addictive, and challenging to get three stars starting at level 3"

[15] "\"', u'* * * * * * * * * * * * * * * * * * * * * * * * * * * *', u'GAME FEATURES:', u'- Jump, sneak, climb and slide your way through various locations', u'- 40 classif
ied missions', u'- BONUS GAME: Ninja Slide', u'- BONUS GAME: Bubble Ride', u\"- TRAINER mode: Ability to play any level in Trainer mode -- Perfect for inexperienced g
amers or if you'd like to practice"
[16] "\", u\"- Donut Games' famous 3-star ranking system: Increased replay value"

[17] "\", u'- Achievements to unlock', u'- Global High Scores: Submit your scores online', u'- Collectors Icon #23', u'- EXCLUSIVE: Not available on any other platfor
m than iDevices (iPhone, iPad, iPod Touch)', u'- And so much more"

[18] "', u'TECH FEATURES:', u'- Retina / HD support', u'- Game Center support', u'- Universal App (iPad, iPhone, iPod Touch)', u'* * * * * * * * * * * * * * * * * * * *
 * * * * *', u'Enjoy another Donut Games release"

[19] "']"
```

Step3: Data Preparation

- Use the first two sentences
  - App users are also likely to read the first few sentences to evaluate Apps before downloading.

```
> Test_sent[1:2]
[1] "[u'-- #1 DOWNLOAD IN 28 COUNTRIES --', u\"What's even better than Frenchmen, ninjas and heroes"
[2] "\", u'A chubby Ninja Hero with a French accent of course"
>
> # Select the first two sentences #
> Apps_Sent <- list()
>
> for (i in 1:n_desc) {
+     temp <- unlist(strsplit(as.character(Apps_desc[i]), split = "[.!?]+"))
+     Apps_Sent[[i]] <- temp[1:2]           #Use the first 2 sentences
+ }
>
> Apps_Sent[[1]]
[1] "[u'-- #1 DOWNLOAD IN 28 COUNTRIES --', u\"What's even better than Frenchmen, ninjas and heroes"
[2] "\", u'A chubby Ninja Hero with a French accent of course"
> Apps_Sent[[2]]
[1] "[u'\\u2605\\u2606 LIMITED TIME OFFER : SPECIAL PRICE"
[2] "  \\u2606\\u2605', u'Appreciate customers event for only limited period on sale"
```

Step3: Data Preparation

- Create a Corpus
  - Corpus is a collection of documents in linguistics
    - A list of text documents (e.g., news article, user reviews, books)
    - The Corpus function reads each text document as a vector

```
> ## Convert Vector/List to Corpus ##
> Apps   <- Corpus(VectorSource(Apps_Sent))
>
> Apps
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 250
> Apps[[1]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 148
> as.character(Apps[[1]])
[1] "[u'-- #1 DOWNLOAD IN 28 COUNTRIES --', u\"What's even better than Frenchmen, ninjas and heroes"
[2] "\", u'A chubby Ninja Hero with a French accent of course"
```

Step3: Data Preparation - Parsing

- The process of structuring the input text and deriving patterns within the structured data
  - Convert upper-case letters to lower-case letters
  - Remove unnecessary/less-important terms
  - Remove stop words
  - Remove numbers
  - Remove punctuations and symbols
  - Remove white space

- The tm_map function applies these NLP rules (mappings) to texts

Step3: Parsing

- Convert Upper-case letters to Lower-case versions
  - e.g., from COUNTIRES to countries

```
> ## Upper-case Letters to Lower-case Letters ##
> Apps  <- tm_map(Apps, content_transformer(tolower))
> as.character(Apps[[1]])
[1] "[u'-- #1 download in 28 countries --', u\"what's even better than frenchmen, ninjas and heroes"
[2] "\", u'a chubby ninja hero with a french accent of course"
```

Step3: Parsing

- Delete unnecessary/less-important terms
  - HTML operators

```
> # Delete HTML Tags #
> for (j in 1:n_desc) Apps[[j]] <- gsub("u2019", " ", Apps[[j]]) #delete "u2019"
> for (j in 1:n_desc) Apps[[j]] <- gsub("u'", " ", Apps[[j]])  #delete u'
> for (j in 1:n_desc) Apps[[j]] <- gsub("u\"", " ", Apps[[j]]) #delete u"
> as.character(Apps[[1]])
[1] "[ -- #1 download in 28 countries --',  what's even better than frenchmen, ninjas and heroes"
[2] "\",  a chubby ninja hero with a french accent of course"
> inspect(Apps[1:3])
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 3

[[1]]
[1] [ -- #1 download in 28 countries --',  what's even better than frenchmen, ninjas and heroes
[2] ",  a chubby ninja hero with a french accent of course

[[2]]
[1] [ \\u2605\\u2606 limited time offer : special price                    \\u2606\\u2605',  appreciate customers event for only limited period on sale

[[3]]
[1] [ \\u2729 50% off \\u2013 special sale: buy galaxy on fire 2 hd for spectacular $ 4,99 / \\u20ac 3,99 / \\xa3 2,99 for a limited time only

[2]  \\u2729',  galaxy on fire 2\\u2122 hd is the completely remodeled hd-version of the award-winning saga that set the standard for sci-fi on the app store, now ava
ilable for iphone 4s, ipad 2 and retina display of the new ipad
```

Step3: Parsing

- Delete unnecessary/less-important terms
  - Additional HTML operators

```
> # Delete HTML Tags #
> for (j in 1:n_desc) Apps[[j]] <- gsub("u2605", " ", Apps[[j]]) #delete "u2605"
> for (j in 1:n_desc) Apps[[j]] <- gsub("u2606", " ", Apps[[j]]) #delete "u2606"
> for (j in 1:n_desc) Apps[[j]] <- gsub("u201c", " ", Apps[[j]]) #delete "u201c"
> for (j in 1:n_desc) Apps[[j]] <- gsub("u201d", " ", Apps[[j]]) #delete "u201d"
> for (j in 1:n_desc) Apps[[j]] <- gsub("u2011", " ", Apps[[j]]) #delete "u2011"
> for (j in 1:n_desc) Apps[[j]] <- gsub("u2013", " ", Apps[[j]]) #delete "u2013"
> for (j in 1:n_desc) Apps[[j]] <- gsub("u2014", " ", Apps[[j]]) #delete "u2014"
> for (j in 1:n_desc) Apps[[j]] <- gsub("u2022", " ", Apps[[j]]) #delete "u2022"
> for (j in 1:n_desc) Apps[[j]] <- gsub("u2122", " ", Apps[[j]]) #delete "u2122"
> for (j in 1:n_desc) Apps[[j]] <- gsub("u2026", " ", Apps[[j]]) #delete "u2026
> for (j in 1:n_desc) Apps[[j]] <- gsub("u2028", " ", Apps[[j]]) #delete "u2028"
> for (j in 1:n_desc) Apps[[j]] <- gsub("u2729", " ", Apps[[j]]) #delete "u2729"
> for (j in 1:n_desc) Apps[[j]] <- gsub("u20ac", " ", Apps[[j]]) #delete "u20ac"
> for (j in 1:n_desc) Apps[[j]] <- gsub("amp", " ", Apps[[j]]) #amp
> for (j in 1:n_desc) Apps[[j]] <- gsub("xae", " ", Apps[[j]]) #xae
> for (j in 1:n_desc) Apps[[j]] <- gsub("xa0", " ", Apps[[j]]) #xa0
> for (j in 1:n_desc) Apps[[j]] <- gsub("xa3", " ", Apps[[j]]) #xa3
> inspect(Apps[1:3])
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 3

[[1]]
[1] [ -- #1 download in 28 countries --',  what's even better than frenchmen, ninjas and heroes
[2] ",  a chubby ninja hero with a french accent of course

[[2]]
[1] [ \\ \\  limited time offer : special price                                \\ \\ ',  appreciate customers event for only limited period on sale

[[3]]
[1] [ \\  50% off \\  special sale: buy galaxy on fire 2 hd for spectacular $ 4,99 / \\  3,99 / \\  2,99 for a limited time only

[2]  \\ ',  galaxy on fire 2\\  hd is the completely remodeled hd-version of the award-winning saga that set the standard for sci-fi on the app store, now available f
or iphone 4s, ipad 2 and retina display of the new ipad
```

Step3: Parsing

- Delete unnecessary/less-important terms
  - Device-related terms: Frequently appeared but less-informative

```
> # Remove less important terms: Device names #
> for (j in 1:n_desc) Apps[[j]] <- gsub("apple", " ", Apps[[j]]) #Apple
> for (j in 1:n_desc) Apps[[j]] <- gsub("iphone", " ", Apps[[j]]) #iphone
> for (j in 1:n_desc) Apps[[j]] <- gsub("touch", " ", Apps[[j]]) #touch
> for (j in 1:n_desc) Apps[[j]] <- gsub("ipod", " ", Apps[[j]]) #ipod
> for (j in 1:n_desc) Apps[[j]] <- gsub("ipad", " ", Apps[[j]]) #ipad
> for (j in 1:n_desc) Apps[[j]] <- gsub("3gs", " ", Apps[[j]]) #iPohne 3GS
> for (j in 1:n_desc) Apps[[j]] <- gsub("3rd", " ", Apps[[j]]) #3rd Gen. iPod
> for (j in 1:n_desc) Apps[[j]] <- gsub("2nd", " ", Apps[[j]]) #2nd Gen. iPod
> for (j in 1:n_desc) Apps[[j]] <- gsub("4th", " ", Apps[[j]]) #4th Gen. iPod
> inspect(Apps[1:3])
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 3

[[1]]
[1] [ -- #1 download in 28 countries --',  what's even better than frenchmen, ninjas and heroes
[2] ",  a chubby ninja hero with a french accent of course

[[2]]
[1] [ \\ \\  limited time offer : special price                          \\ \\ ',  appreciate customers event for only limited period on sale

[[3]]
[1] [ \\  50% off \\  special sale: buy galaxy on fire 2 hd for spectacular $ 4,99 / \\  3,99 / \\  2,99 for a limited time only

[2]  \\ ',  galaxy on fire 2\\  hd is the completely remodeled hd-version of the award-winning saga that set the standard for sci-fi on the app store, now available f
or   4s,   2 and retina display of the new
```

Step3: Parsing

- Delete unnecessary/less-important terms
  - App Store-related terms: Frequently appeared but less-informative

```
> # Remove less important terms: App, Game, Play #
> for (j in 1:n_desc) Apps[[j]] <- gsub("app", " ", Apps[[j]]) #app
> for (j in 1:n_desc) Apps[[j]] <- gsub("store", " ", Apps[[j]]) #store
> for (j in 1:n_desc) Apps[[j]] <- gsub("game", " ", Apps[[j]]) #game
> for (j in 1:n_desc) Apps[[j]] <- gsub("play", " ", Apps[[j]]) #play
> for (j in 1:n_desc) Apps[[j]] <- gsub("mobile", " ", Apps[[j]]) #mobile
> for (j in 1:n_desc) Apps[[j]] <- gsub("free", " ", Apps[[j]]) #Free
> for (j in 1:n_desc) Apps[[j]] <- gsub("new", " ", Apps[[j]]) #new
> for (j in 1:n_desc) Apps[[j]] <- gsub("world", " ", Apps[[j]]) #world
> inspect(Apps[1:3])
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 3

[[1]]
[1] [ -- #1 download in 28 countries --',  what's even better than frenchmen, ninjas and heroes
[2] ",  a chubby ninja hero with a french accent of course

[[2]]
[1] [ \\ \\  limited time offer : special price                          \\ \\ ',   reciate customers event for only limited period on sale

[[3]]
[1] [ \\  50% off \\  special sale: buy galaxy on fire 2 hd for spectacular $ 4,99 / \\  3,99 / \\  2,99 for a limited time only

[2]  \\ ',  galaxy on fire 2\\  hd is the completely remodeled hd-version of the award-winning saga that set the standard for sci-fi on the    , now available for  4
s,  2 and retina dis  of the
```

Step3: Parsing

- Convert important/meaningful numbers to characters (terms)
  - #1 App in 30 courtiers
  - $0.99 for a limited time only
  - Limited Time Sale! 50% off!

```
> # Convert Important numbers to words/terms #
> for (j in 1:n_desc) Apps[[j]] <- gsub("#1", "numberone", Apps[[j]]) #1
> for (j in 1:n_desc) Apps[[j]] <- gsub("99", "nintyninecent", Apps[[j]]) #$0.99
> for (j in 1:n_desc) Apps[[j]] <- gsub("%", "percent", Apps[[j]]) #percent
> inspect(Apps[1:3])
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 3

[[1]]
[1] [ -- numberone download in 28 countries --',  what's even better than frenchmen, ninjas and heroes
[2] ",  a chubby ninja hero with a french accent of course

[[2]]
[1] [ \\ \\  limited time offer : special price                    \\ \\ ',  reciate customers event for only limited period on sale

[[3]]
[1] [ \\  50percent off \\  special sale: buy galaxy on fire 2 hd for spectacular $ 4,nintyninecent / \\  3,nintyninecent / \\  2,nintyninecent for a limited time onl
y
[2]  \\ ',  galaxy on fire 2\\  hd is the completely remodeled hd-version of the award-winning saga that set the standard for sci-fi on the    , now available for  4
s,  2 and retina dis  of the
```

Step3: Parsing

- Delete Stop words
  - By removing the words that are very commonly used (but less informative) in a given language, we can focus on the important words instead.
  - Articles (the, a, an…)
  - Prepositions (for, after, above, across, before, under…)
  - Conjunctions (and, but, nor, yet, so, than…)
  - Pronouns (she, he, I, you, they, them…)
  - Auxiliary Verbs (can, will, could, would, must) and Linking Verbs (is, are, am)
  - When, where, how, what, which

```
> # Remove stopwords #
> Apps <- tm_map(Apps, removeWords, stopwords("english"))
> inspect(Apps[1:3])
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 3

[[1]]
[1] [ -- numberone download  28 countries --',   even better  frenchmen, ninjas  heroes
[2] ",   chubby ninja hero   french accent  course

[[2]]
[1] [ \\ \\  limited time offer : special price                \\ \\ ',   reciate customers event   limited period  sale

[[3]]
[1] [ \\  50percent  \\  special sale: buy galaxy  fire 2 hd  spectacular $ 4,nintyninecent / \\  3,nintyninecent / \\  2,nintyninecent   limited time
[2]  \\ ',  galaxy  fire 2\\  hd   completely remodeled hd-version   award-winning saga  set  standard  sci-fi      , now available    4s,   2  retina dis
```

Step3: Parsing

- Delete Stop words

```
> # You can define stopwords #
> newstopwords <-c("and", "for", "the", "to", "in", "when", "then", "he", "she", "than", "can", "get", "one");
> Apps <- tm_map(Apps, removewords, newstopwords)
>
> for (j in 1:n_desc) Apps[[j]] <- gsub("don", " ", Apps[[j]]) #don
> for (j in 1:n_desc) Apps[[j]] <- gsub("won", " ", Apps[[j]]) #won't
> for (j in 1:n_desc) Apps[[j]] <- gsub("ing", " ", Apps[[j]]) # -ing
> for (j in 1:n_desc) Apps[[j]] <- gsub("http", " ", Apps[[j]]) # http
> for (j in 1:n_desc) Apps[[j]] <- gsub("'ll", " ", Apps[[j]]) #'ll
> for (j in 1:n_desc) Apps[[j]] <- gsub("www", " ", Apps[[j]]) # www
> for (j in 1:n_desc) Apps[[j]] <- gsub("com", " ", Apps[[j]]) # com
> inspect(Apps[1:3])
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 3

[[1]]
[1] [ -- numberone download  28 countries --',   even better  frenchmen, ninjas  heroes
[2] ",   chubby ninja hero   french accent  course

[[2]]
[1] [ \\ \\  limited time offer : special price                \\ \\ ',   reciate customers event   limited period  sale

[[3]]
[1] [ \\  50percent  \\  special sale: buy galaxy  fire 2 hd  spectacular $ 4,nintyninecent / \\  3,nintyninecent / \\  2,nintyninecent   limited time
[2]  \\ ',  galaxy  fire 2\\  hd    pletely remodeled hd-version   award-winn  saga  set  standard  sci-fi      , now available    4s,   2  retina dis
```

Step3: Parsing

- Delete Numbers
  - if numeric values are not important.

```
> # Remove Numbers #
> Apps <- tm_map(Apps, removeNumbers)
> inspect(Apps[1:3])
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 3

[[1]]
[1] [ -- numberone download   countries --',   even better  frenchmen, ninjas  heroes
[2] ",   chubby ninja hero   french accent  course

[[2]]
[1] [ \\ \\  limited time offer : special price                \\ \\ ',   reciate customers event   limited period  sale

[[3]]
[1] [ \\  percent  \\  special sale: buy galaxy  fire  hd  spectacular $ ,nintyninecent / \\  ,nintyninecent / \\  ,nintyninecent   limited time
[2]  \\ ',  galaxy  fire \\  hd    pletely remodeled hd-version   award-winn  saga  set  standard  sci-fi      , now available    s,     retina dis
```

Step3: Parsing

- Delete Punctuations and Symbols
  - Punctuations and Symbols (!, ?, $, @, . ;, , :, “”, [,], ( )…)

```
> # Remove Punctuations and Symbols #
> Apps <- tm_map(Apps, removePunctuation)
> inspect(Apps[1:3])
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 3

[[1]]
[1]   numberone download   countries    even better  frenchmen ninjas  heroes    chubby ninja hero   french accent  course

[[2]]
[1]    limited time offer  special price                    reciate customers event   limited period  sale

[[3]]
[1]   percent    special sale buy galaxy  fire  hd  spectacular  nintyninecent    nintyninecent    nintyninecent   limited time
[2]    galaxy  fire   hd    pletely remodeled hdversion   awardwinn  saga  set  standard  scifi      now available    s     retina dis


> # You can also manually delete non-characters #
> for (j in 1:n_disc) Apps[[j]] <- gsub("['*|&|-|/|\\|()|\\.,!-_]", " ", Apps[[j]]) # remove non-characters
```

Step 3: Parsing

- Remove White Space(s)

```
> # Remove Extra White Space #
> Apps <- tm_map(Apps, stripWhitespace)
> inspect(Apps[1:3])
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 3

[[1]]
[1]  numberone download countries even better frenchmen ninjas heroes  chubby ninja hero french accent course

[[2]]
[1]  limited time offer special price            reciate customers event limited period sale

[[3]]
[1]  percent special sale buy galaxy fire hd spectacular nintyninecent nintyninecent nintyninecent limited time
[2]  galaxy fire hd pletely remodeled hdversion awardwinn saga set standard scifi now available s retina dis
```

Step3: Stemming

- Reduce inflected (or sometimes derived) words to their word stem (i.e., root form)
  - work: works, worked, working, worker
  - countri: country, countries,

```
> ## Stemming ##
> Apps <- tm_map(Apps, PlainTextDocument)  # Remove common word endings ("es", "ed", "s", "ing")
> Apps <- tm_map(Apps, stemDocument)
>
> as.character(Apps[[1]])
[1] " numberon download countri even better frenchmen ninja hero" " chubbi ninja hero french accent cours"
> as.character(Apps[[3]])
[1] " percent special sale buy galaxi fire hd spectacular nintyninec nintyninec nintyninec limit time"
[2] " galaxi fire hd plete remodel hdversion awardwinn saga set standard scifi now avail s retina dis"
```

Step3: Convert Texts to Numbers

- Create a Document Term Matrix (DTM) from a corpus
  - DTM represents the relationship between terms and documents
    - Rows: Documents, Columns: Terms
    - An entry is the number of occurrences of the term in the document.
  - Unstructured data is converted into structured/quantitative data for further data-mining analyses.

```
> # Original DTM #
> dtm_Apps <- DocumentTermMatrix(Apps)
> dtm_Apps
<<DocumentTermMatrix (documents: 250, terms: 1440)>>
Non-/sparse entries: 3130/356870
Sparsity           : 99%
Maximal term length: 21
Weighting          : term frequency (tf)
> inspect(dtm_Apps)
<<DocumentTermMatrix (documents: 250, terms: 1440)>>
Non-/sparse entries: 3130/356870
Sparsity           : 99%
Maximal term length: 21
Weighting          : term frequency (tf)
```

```
> as.character(Apps[[1]])
[1] " numberon download countri even better frenchmen ninja hero" " chubbi ninja hero french accent cours"
```

Docs	Terms abil	abl	absolut	absorb	academi	accent	access	accessori	acclaim	accuraci	acdc	achiev	across	action	actionpack	activis	adapt	aday	add	addict	addit
character(0)	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
character(0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
character(0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
character(0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
character(0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
character(0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Step3: Convert Texts to Numbers

- Create a Document Term Matrix (DTM) from a corpus
  - To remove less-important and sparse terms,
    - Enforce lower and upper limits to the length of the words included (between 3 and 20 characters)
    - Select the terms occurred in 10 to 200 app descriptions

```
> # DTM with Controls #
> dtm_Apps_Ctrl <- DocumentTermMatrix(Apps, control=list(wordLength=c(3,20), bounds=list(global=c(10,200))))
> dtm_Apps_Ctrl
<<DocumentTermMatrix (documents: 250, terms: 44)>>
Non-/sparse entries: 746/10254
Sparsity           : 93%
Maximal term length: 10
Weighting          : term frequency (tf)
> inspect(dtm_Apps_Ctrl) # Display DTM for the descriptions
<<DocumentTermMatrix (documents: 250, terms: 44)>>
Non-/sparse entries: 746/10254
Sparsity           : 93%
Maximal term length: 10
Weighting          : term frequency (tf)

              Terms
Docs           action addict adventur arcad avail award best challeng countri creat download ever experi featur first fun gam graphic hit launch level like limit
  character(0)      0      0        0     0     0     0    0        0       1     0        1    0      0      0     0   0   0       0   0      0     0    0     0
  character(0)      0      0        0     0     0     0    0        0       0     0        0    0      0      0     0   0   0       0   0      0     0    0     2
  character(0)      0      0        0     0     1     0    0        0       0     0        0    0      0      0     0   0   0       0   0      0     0    0     1
  character(0)      0      1        0     0     0     0    0        0       0     0        0    0      1      0     0   0   0       0   0      0     0    0     0
  character(0)      0      0        0     0     1     0    0        0       0     0        0    0      0      0     0   0   0       0   0      0     0    0     1
  character(0)      0      0        0     0     1     0    0        0       0     0        0    0      0      0     0   0   0       0   0      0     0    0     1
  character(0)      0      0        0     0     0     0    0        0       0     0        0    0      0      0     0   0   0       0   0      0     0    0     1
  character(0)      0      0        0     0     0     0    0        0       0     0        0    0      0      0     0   0   0       0   0      0     0    0     0
  character(0)      0      0        0     0     0     0    0        0       0     0        0    0      0      0     0   0   0       0   0      0     0    0     1
```

Step3: Convert Texts to Numbers

- Create a Document Term Matrix (DTM) from a corpus

```
> as.character(Apps[[1]])
[1] " numberon download countri even better frenchmen ninja hero" " chubbi ninja hero french accent cours"
```

```
> inspect(dtm_Apps_Ctrl[1:5,1:15]) # DTM for the first 5 descriptions with the first 15 terms
<<DocumentTermMatrix (documents: 5, terms: 15)>>
Non-/sparse entries: 6/69
Sparsity           : 92%
Maximal term length: 8
Weighting          : term frequency (tf)

              Terms
Docs           action addict adventur arcad avail award best challeng countri creat download ever experi featur first
  character(0)      0      0        0     0     0     0    0        0       1     0        1    0      0      0     0
  character(0)      0      0        0     0     0     0    0        0       0     0        0    0      0      0     0
  character(0)      0      0        0     0     1     0    0        0       0     0        0    0      0      0     0
  character(0)      0      1        0     0     0     0    0        0       0     0        0    0      1      0     0
  character(0)      0      0        0     0     1     0    0        0       0     0        0    0      0      0     0
```

NANYANG TECHNOLOGICAL UNIVERSITY

Task 1:
Identify the most frequently used terms in the descriptions

Step4: Evaluate the Outcomes

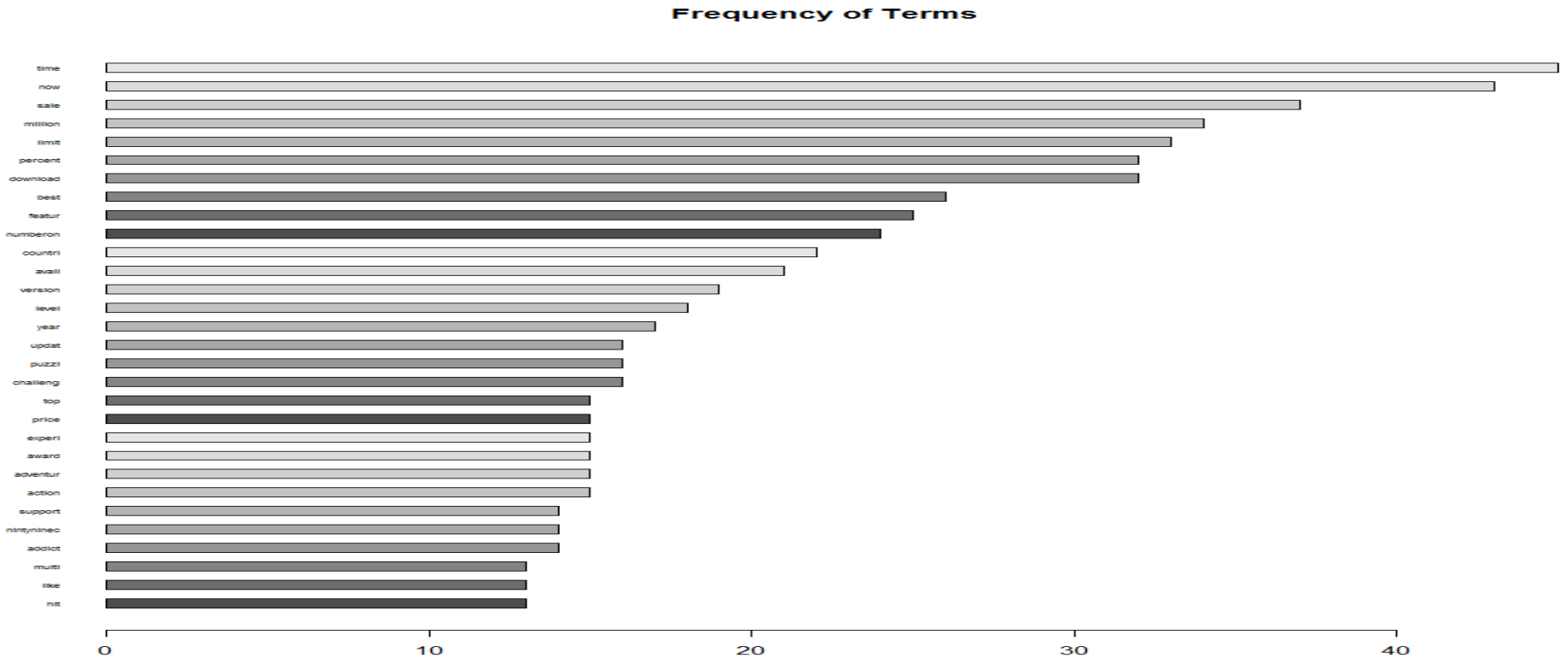
- Create a Frequency Table for the terms

```
> # Find the terms that occur at least 30 times #
> findFreqTerms(dtm_Apps_Ctrl, 30)
[1] "download" "limit"    "million"  "now"      "percent"  "sale"     "time"
>
> # Frequency of terms #
> Freq_term <-colSums(as.matrix(dtm_Apps_Ctrl))
> Order_Freq_term <- order(Freq_term, decreasing = TRUE)
> Freq_term[Order_Freq_term]
     time       now      sale   million     limit  download   percent      best    featur  numberon  countri     avail   version     level
       45        43        37        34        33        32        32        26        25        24        22        21        19        18
     year   challeng     puzzl     updat    action  adventur     award    experi     price       top   addict nintyninec   support       hit
       17        16        16        16        15        15        15        15        15        15        14        14        14        13
     like     multi     arcad     creat    launch   special      ever     offer      real    retina     first       fun       gam   graphic
       13        13        12        12        12        12        11        11        11        11        10        10        10        10
     week      will
       10        10
```

Step4: Evaluate the Outcomes

- Create a Frequency Diagram for the terms

```
# Frequency Diagram #
library(grDevices); # for colours
Apps_DTM_DF = as.data.frame(as.matrix(dtm_Apps_Ctrl))
numwords <- 30; # the most frequent 30 terms

# sum each column and sort by descending order #
Terms_Freq <- as.matrix(sort(sapply(Apps_DTM_DF, FUN=sum),decreasing=TRUE)[1:numwords], colnames=count)
x <- sort(Terms_Freq[1:numwords,], decreasing=FALSE)
barplot(x, horiz=TRUE, cex.names=0.5, space=1, las=1, col=grey.colors(10), main="Frequency of Terms")
```



Step4: Evaluate the Outcomes

- Create a Word Cloud
  - The larger the word in the visual the more common the word was in the document(s).

```
# For Original DTM #
 set.seed(2406) #set the same seed each time ensures consistent look across clouds
 m <- as.matrix(t(dtm_Apps)) # Convert it to a matrix
 v <- sort(rowSums(m),decreasing=TRUE) # Sort the terms in a descending order
 w <- data.frame(word = names(v),freq=v) # Create a data frame indicating the name&frequency of terms

 WC_color <- brewer.pal(8,"Set2")
 wordcloud(w$word,w$freq, scale=c(3,.1),min.freq=1, max.words=200, random.order=F, rot.per=.3, colors=WC_color)
```

NANYANG TECHNOLOGICAL UNIVERSITY

Step4: Evaluate the Outcomes

- Create a Word Cloud

```
# For DTM with Controls #
 dev.off()
 set.seed(2406) #set the same seed each time ensures consistent look across clouds
 m <- as.matrix(t(dtm_Apps_Ctrl)) # Convert it to a matrix
 v <- sort(rowSums(m),decreasing=TRUE) # Sort the terms in a descending order
 w <- data.frame(word = names(v),freq=v) # Create a data frame indicating the name&frequency of terms

 WC_color <- brewer.pal(8,"Set2")
 wordcloud(w$word,w$freq, scale=c(3,.1),min.freq=1, max.words=200, random.order=F, rot.per=.3, colors=WC_color)
```

NANYANG TECHNOLOGICAL UNIVERSITY

Step4: Evaluate the Outcomes

- Find associated terms
  - Check for correlations between terms
  - Measure the co-occurrence of terms in multiple documents
    - "time" is used together with "limit" and "sale" at a correlation of 64% and 34% respectively
      - Limited time sales!

```
> ## Find Associated Terms ##
> findAssocs(dtm_Apps_Ctrl, "time", .3);
$time
limit  sale
 0.64  0.34

> findAssocs(dtm_Apps_Ctrl, "now", .2);
$now
avail   hit
 0.29  0.22

> findAssocs(dtm_Apps_Ctrl, "best", .2);
$best
award  ever
 0.31  0.25
```

NANYANG TECHNOLOGICAL UNIVERSITY

Task 2:

Identify the common patterns/styles in presenting App product descriptions?

K-means Clustering

Step3: Data Preparation

- Select the number of terms for clustering
  - Remove sparse terms

```
## Select the number of Terms for Clustering ##
dtm_Apps_Sparse <- removeSparseTerms(dtm_Apps_Ctrl, 0.92)
#remove the terms which have at least a 89 percentage of sparse

nrow(dtm_Apps_Sparse); ncol(dtm_Apps_Sparse)
inspect(dtm_Apps_Sparse)
```

```
<<DocumentTermMatrix (documents: 250, terms: 12)>>
Non-/sparse entries: 354/2646
Sparsity           : 88%
Maximal term length: 8
Weighting          : term frequency (tf)

              Terms
Docs           avail best countri download featur limit million now numberon percent sale time
  character(0)     0    0       1        1      0     0       0   0        1       0    0    0
  character(0)     0    0       0        0      0     2       0   0        0       0    1    1
  character(0)     1    0       0        0      0     1       0   1        0       1    1    1
  character(0)     0    0       0        0      0     0       1   0        0       0    0    0
  character(0)     1    0       0        0      0     1       0   1        0       0    0    1
  character(0)     1    0       0        0      0     1       0   1        0       0    0    1
  character(0)     0    0       0        0      0     1       0   0        0       0    0    1
  character(0)     0    0       0        0      0     0       0   1        0       0    1    0
  character(0)     0    0       0        0      0     1       0   0        0       1    0    0
  character(0)     0    0       0        0      0     0       0   0        0       0    1    0
  character(0)     0    0       0        0      0     0       0   0        0       0    0    0
  character(0)     0    0       0        0      1     0       0   0        0       0    0    0
  character(0)     1    0       0        0      0     1       0   1        0       1    1    1
```

Step4: Build a Clustering Model

- Problem: Identify the common patterns/styles in presenting App product descriptions?
- Variables
  - 12 terms
    - *time, nintyninec, sales, limit, percent*
    - *featur, level, best, avail, now*
    - *download, million, countri, numberon*
- Expected Clusters
  - Price Promotion
  - Update Information
  - Sales Performance

Step5: Train a Model on the Data

- K-means Clustering (K=3)

```
## K-means Clustering ##
dtm_Apps_cluster <-as.matrix(dtm_Apps_Sparse)
library(stats)      # Clustering

set.seed(2406)
Apps_KM <- kmeans(t(dtm_Apps_cluster), 3)
```

Step6: Validate Clusters

- K-means Clustering Output (K=3)

```
> Apps_KM
K-means clustering with 3 clusters of sizes 2, 4, 6

Cluster means:
  character(0) character(0) character(0) character(0) character(0) character(0) character(0)
1    0.5000000            0    0.0000000          0.5    0.0000000    0.0000000          0.0
2    0.0000000            1    1.0000000          0.0    0.5000000    0.5000000          0.5
3    0.3333333            0    0.3333333          0.0    0.3333333    0.3333333          0.0
  character(0) character(0) character(0) character(0) character(0) character(0) character(0)
1    0.0000000          0.0         0.00            0    0.0000000    0.0000000    0.0000000
2    0.2500000          0.5         0.25            0    0.0000000    1.0000000    1.0000000
3    0.1666667          0.0         0.00            0    0.1666667    0.3333333    0.1666667
  character(0) character(0) character(0) character(0) character(0) character(0) character(0)
1            0          1.0            0            0          0.0          0.0            0
2            1          0.5            1            0          0.5          0.5            1
3            0          0.0            0            0          0.0          0.0            0
  character(0) character(0) character(0) character(0) character(0) character(0) character(0)
1          0.0         0.00    0.0000000    0.0000000    0.0000000          0.0          0.0
2          0.5         0.75    1.0000000    0.7500000    0.5000000          0.5          0.5
3          0.0         0.00    0.3333333    0.1666667    0.3333333          0.0          0.0
  character(0) character(0) character(0) character(0) character(0) character(0) character(0)
1         0.00         0.00    0.0000000            0            0         0.00          0.0
2         0.25         0.75    1.0000000            1            1         0.75          0.5
3         0.00         0.00    0.1666667            0            0         0.00          0.0

Clustering vector:
   avail     best  countri download   featur    limit  million      now numberon  percent     sale
       3        3        3        1        3        2        1        3        3        2        2
    time
       2

Within cluster sum of squares by cluster:
[1]  15.0000  53.7500 132.1667
 (between_SS / total_SS =  36.9 %)
```

NANYANG TECHNOLOGICAL UNIVERSITY

Step6: Validate Clusters

- Look at the size of the clusters (i.e., #terms for each cluster)
- Look at the memberships of 12 terms

```
> Apps_KM$size
[1] 2 4 6
> sort(Apps_KM$cluster)
download  million    limit  percent     sale     time    avail     best  countri   featur      now numberon
       1        1        2        2        2        2        3        3        3        3        3        3
```

- Characterize the Clusters
  - Expected Clusters: Price Promotion, Sales Performance, and Update Info.

	Cluster 1	Cluster 2	Cluster 3
N	2	4	6
Terms	download million	limit percent (%) sale time	avail best countri feature now numberon
Example	"Downloaded over 10 million times in 2012"	"Limited Time SALE! 50% OFF!" "$ 0.99 for a limited time only"	"Screen rotation and zoom are now available "'NEW FEATURES: Retina / HD support', "Play all 60 levels, then take on 60 tests in Challenge mode" "#1 App in 30 countries"
Label	Sales Performance	Price Promotion	Update Info. + Sales Performance

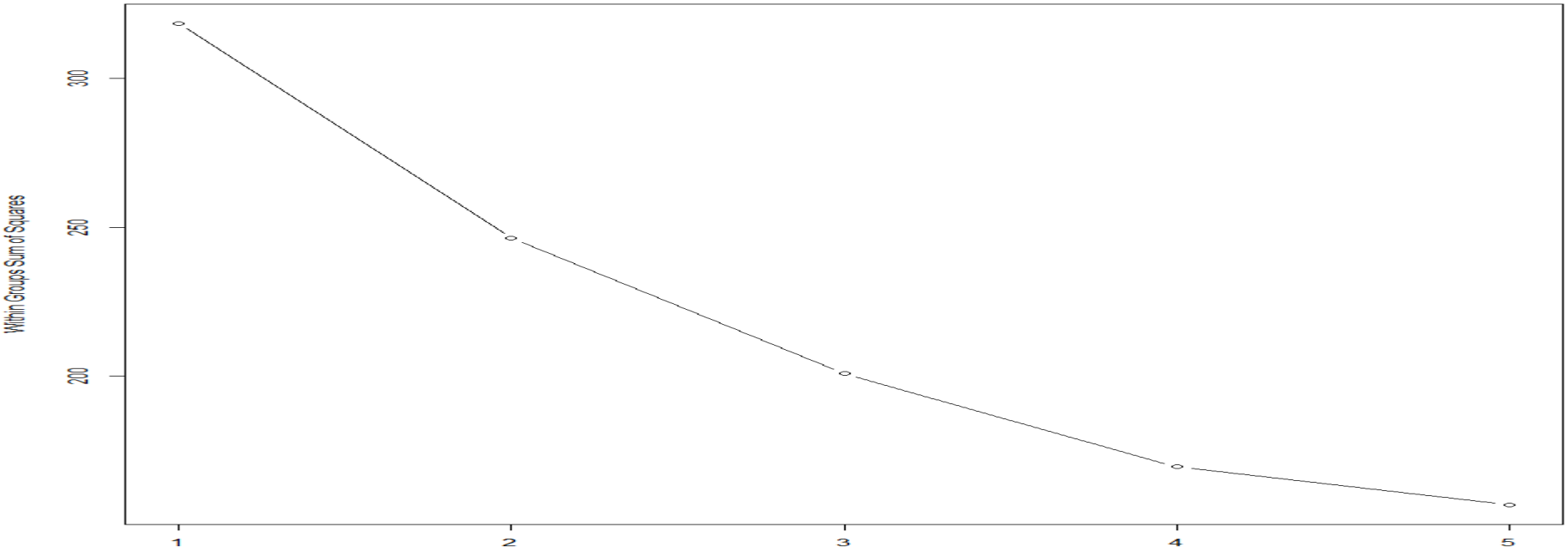
Step6: Validate Clusters

- Evaluate the number of K (Elbow Method)

```
## Identify the optimal k ##
#Within Groups Sum of Squares#
wss <- 1:5
for(i in 1:5) {wss[i] <- sum(kmeans(t(dtm_Apps_cluster),i)$withinss)}
plot(1:5, wss[1:5], type="b", xlab="Number of Clusters", ylab="Within Groups Sum of Squares")
# type="b" creates a plot with lines between points #
```

```
> WSS
[1] 318.5000 246.3750 200.9167 169.5833 156.7500
```

Difference: 72.12 45.59 31.33 12.83



Hierarchical Clustering

Step5: Train a Model on the Data

- Hierarchical Clustering
  - Measure distance between records

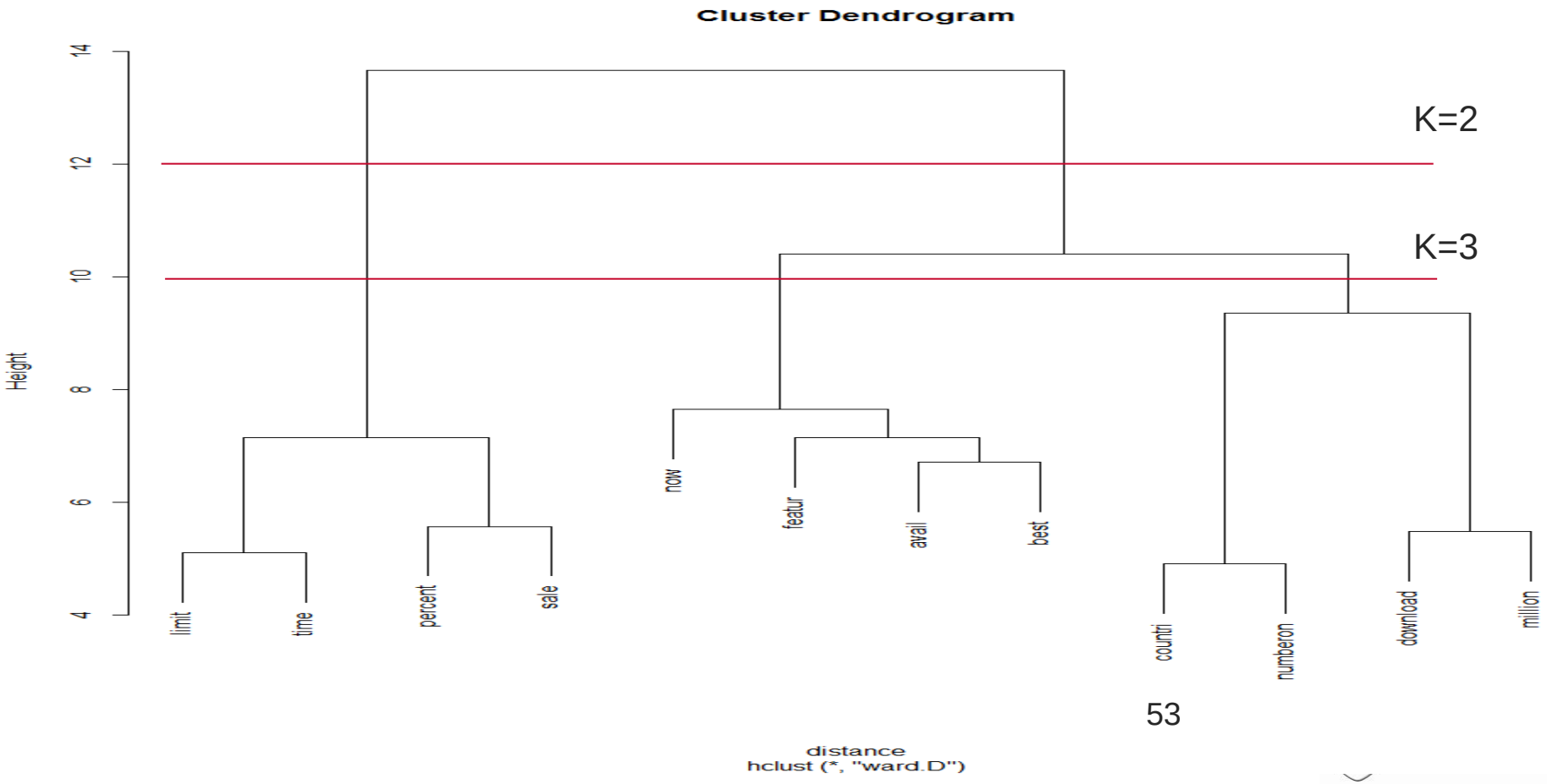
```
## Hierarchical Clustering ##
dtm_Apps_cluster <-as.matrix(dtm_Apps_Sparse)

# Calculate the Distance between Terms #
distance <- dist(t(dtm_Apps_cluster), method="euclidean")
distance
```

```
          avail     best  countri download   featur    limit  million      now numberon  percent     sale
best     6.708204
countri  7.141428 7.483315
download 7.416198 7.071068 6.928203
featur   6.782330 7.280110 7.416198 7.681146
limit    6.633250 8.185353 7.810250 8.062258 7.615773
million  8.062258 8.124038 7.615773 5.477226 8.426150 8.774964
now      6.782330 7.549834 8.426150 8.062258 7.874008 8.246211 8.660254
numberon 7.141428 7.211103 4.898979 7.071068 7.416198 8.062258 7.483315 8.306624
percent  6.855655 7.615773 7.615773 8.000000 7.681146 5.196152 8.717798 8.306624 7.874008
sale     7.211103 8.306624 7.937254 8.426150 7.615773 5.830952 9.110434 8.485281 8.185353 5.567764
time     7.071068 7.937254 7.937254 8.426150 8.246211 5.099020 9.000000 8.246211 8.306624 7.000000 6.928203
```

Step6: Validate Clusters

- Display the clustering output in a dendrogram

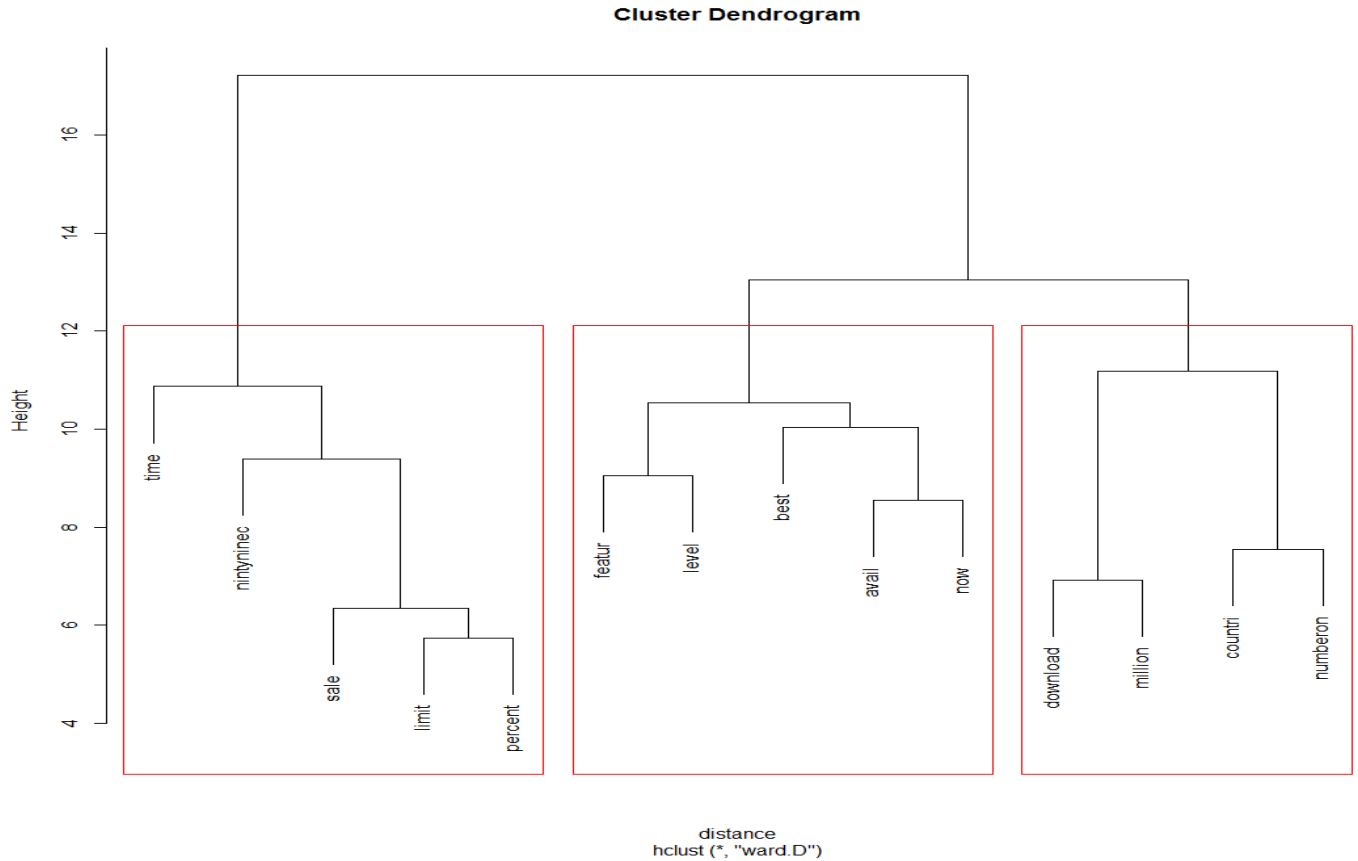
```
Apps_HC <- hclust(distance, method = "ward.D")
plot(Apps_HC)
```



Step6: Validate Clusters

- With 3 Clusters

```
# Draw dendrogram with red borders around the 3 clusters #
rect.hclust(Apps_HC, k=3, border="red")
```



Step6: Validate Clusters

- Cut off the tree at the desired number of clusters (k=3)
  - Examine the clusters for 12 terms

```
> Apps_HC_Cut <- cutree(Apps_HC, k=3) # cut tree into 3 clusters
> Apps_HC_Cut
   avail     best  countri download   featur    limit  million      now numberon  percent     sale     time
       1        1        2        2        1        3        2        1        2        3        3        3
>
> sort(Apps_HC_Cut)
   avail     best   featur      now  countri download  million numberon    limit  percent     sale     time
       1        1        1        1        2        2        2        2        3        3        3        3
```

- Characterize the Clusters

	Cluster 1	Cluster 2	Cluster 3
N	4	4	4
Terms	avail best feature now	countri download million numberon (#1)	limit percent (%) sale time
Example	"Screen rotation and zoom are now available "'NEW FEATURES: Retina / HD support', "Play all 60 levels, then take on 60 tests in Challenge mode"	"#1 App in 30 countries" "Downloaded over 10 million times in 2012"	"Limited Time SALE! 50% OFF!" "$ 0.99 for a limited time only"
Label	Update Info.	Sales Performance	Price Promotion

Task 3:

Identify what types of messages in the descriptions are associated with better sales performance

Linear Regression

Build a Regression Model from TM

- Problem: Identify what types of messages in the descriptions are associated with better sales performance
- Dependent Variable
  - *Sales*: Continuous variable
- Independent Variables
  - Key predictors from app product descriptions (from HC)
    - *Update_Info* (# terms in Cluster1)
    - *Sales_Performance* (#terms in Cluster 2)
    - *Price_Promotion* (#terms in Cluster 3)
    - *Content* (# terms not appeared in the clusters 1,2, and 3)
  - Other predictors
    - *Price*, *#Screenshots*, *Size*, *StarRating*, *#Ratings*, *#IAP*

Cluster Scores: Example

Cluster(s)	App	Description	Cluster 1	Cluster 2	Cluster 3
1. Update Info.	Batman Arkham City	UPDATED WITH NEW LEVELS & FEATURES - OPTIMIZED FOR THE NEW iPAD! The inmates have escaped and Batman has his hands full defeating an army of henchmen and some of his most iconic villains.	2	0	0
2. Sales Performance	Bridge Constructor	# 1 in the game charts for iPad & iPhone in 24 countries. #1 in the overall app charts for iPad in 19 countries Already #1 in the overall app charts for iPhone in 4 countries	0	6	0
3. Price Promotion	Infinity Blade	Super Summer Sale! INFINITY BLADE IS NOW ONLY $0.99. FOR A LIMITED TIME! The critically acclaimed best seller is celebrating a limited time $0.99 sale.	8	0	0
4. Content	Epic Astro Story	Ready to test your mettle against the final frontier? Pioneer an untamed planet, building roads and houses for your fellow denizens of the future.	0	0	0

- Cluster Score
  - The number of terms used for a cluster in an App's product description
- Common Patterns: commonly appearing terms in hit Apps:
  - Terms used for '*Update Info.*', '*Sales Performance*', and '*Price Promotion*'
- Content: characterizes an App's key feature/uniqueness of '*Content*'
  - Terms that did not appear in other three clusters but used in the description

Step3: Data Preparation

- Count the number of terms in a description
  - The number of terms appeared in the identified clusters will be used for a regression.

```
## Count the number of terms in a description ##
library(stringr) # counting the number of terms
Num_Terms <- matrix(data=0 , n_desc,1);

for (j in 1:n_desc){
str1 <- Apps[[j]]
str2 <- str_match_all(str1, "\\S+" )
Num_Terms[j] <- length(str2[[1]]) + length(str2[[2]])
}

head(Num_Terms)
```

```
     [,1]
[1,]   14
[2,]   11
[3,]   29
[4,]   21
[5,]    5
[6,]    5
```

Step3: Data Preparation

- Compute Cluster Scores from HC output

```
## Compute Cluster Scores ##
# Identify the terms for each cluster #
cluster1 <- dtm_Apps_cluster[,c("avail", "best", "featur","now")]
cluster2 <- dtm_Apps_cluster[,c("countri", "download", "million", "numberon")]
cluster3 <- dtm_Apps_cluster[,c("limit", "percent", "sale", "time")]
```

```
> head(cluster1)
              Terms
Docs           avail best featur now
  character(0)     0    0      0   0
  character(0)     0    0      0   0
  character(0)     1    0      0   1
  character(0)     0    0      0   0
  character(0)     1    0      0   1
  character(0)     1    0      0   1
```

```
> head(cluster2)
              Terms
Docs           countri download million numberon
  character(0)       1        1       0        1
  character(0)       0        0       0        0
  character(0)       0        0       0        0
  character(0)       0        0       1        0
  character(0)       0        0       0        0
  character(0)       0        0       0        0
```

```
> head(cluster3)
              Terms
Docs           limit percent sale time
  character(0)     0       0    0    0
  character(0)     2       0    1    1
  character(0)     1       1    1    1
  character(0)     0       0    0    0
  character(0)     1       0    0    1
  character(0)     1       0    0    1
```

Step3: Data Preparation

- Compute Cluster Scores

```
# Sums #
C1_Sum <- rowSums(cluster1)
C2_Sum <- rowSums(cluster2)
C3_Sum <- rowSums(cluster3)
C4_Sum <- Num_Terms - (C1_Sum + C2_Sum + C3_Sum)


# Create a Score table #
Score <- matrix(data=0 , n_desc,4);
Score[,1] <- as.matrix(C1_Sum)
Score[,2] <- as.matrix(C2_Sum)
Score[,3] <- as.matrix(C3_Sum)
Score[,4] <- as.matrix(C4_Sum)

# Name the Columns/Clusters #
colnames(Score) <- c("Cluster1", "Cluster2", "Cluster3", "Cluster4")
head(Score)
```

```
     Cluster1 Cluster2 Cluster3 Cluster4
[1,]        0        3        0       11
[2,]        0        0        4        7
[3,]        2        0        4       23
[4,]        0        1        0       20
[5,]        2        0        2        1
[6,]        2        0        2        1
```

Step3: Data Preparation

- Create a new dataset for a regression

```
## Add a Score matrix to the original Data ##
Apps_new <- cbind(Apps_data, Score)
str(Apps_new)
```

```
'data.frame':   250 obs. of  19 variables:
 $ AppID             : int  394716128 535747605 465072566 322852954 337402021 404086528 342842881 414723566 441242571 489454710 ...
 $ Rank              : int  232 183 292 26 211 56 109 293 118 294 ...
 $ Name              : Factor w/ 250 levels "???Solitaire+",..: 227 124 85 241 95 112 238 192 175 5 ...
 $ ReleaseDate       : Factor w/ 198 levels "01,14,2011","01,14,2012",..: 143 112 149 111 168 174 189 4 81 31 ...
 $ Category          : Factor w/ 1 level "Games": 1 1 1 1 1 1 1 1 1 1 ...
 $ Price             : num  0.99 0.99 4.99 2.99 0.99 0.99 0.99 0.99 0.99 0.99 ...
 $ Developer         : Factor w/ 165 levels "[adult swim]",..: 41 119 54 165 160 160 95 67 90 130 ...
 $ Screenshot        : int  7 10 10 5 5 10 5 10 5 5 ...
 $ Size              : num  9.6 44.2 622 13.3 102 453 169 834 28.4 19.4 ...
 $ Description       : Factor w/ 250 levels "[u'---', u'The CHAOS RINGS series summer sale is here for one week only!', u'Chaos Rings now 70% off for a limited time only
,..: 4 87 93 181 170 169 46 168 26 74 ...
 $ UpdatedDate       : Factor w/ 174 levels "01,03,2012","01,05,2011",..: 93 126 107 37 153 6 56 86 9 101 ...
 $ StarCurrentVersion  : num  4 4.5 4.5 4.5 4 4 4 4.5 4.5 4.5 ...
 $ RatingCurrentVersion: int  4426 139 122 13657 2715 2749 4631 2750 20 66 ...
 $ TopInAppPurchases   : int  0 5 7 8 2 0 0 10 0 7 ...
 $ DataDate          : Factor w/ 20 levels "04,06,2012","04,13,2012",..: 1 19 17 1 1 1 1 20 5 16 ...
 $ Cluster1          : num  0 0 2 0 2 2 0 1 0 0 ...
 $ Cluster2          : num  3 0 0 1 0 0 0 0 0 0 ...
 $ Cluster3          : num  0 4 4 0 2 2 2 1 2 1 ...
 $ Cluster4          : num  11 7 23 20 1 1 7 4 4 1 ...
```

Step3: Data Preparation

- A Summary of Variables

```
> summary(Apps_new[,c('Rank','Price','Screenshot','Size','StarCurrentVersion', 'RatingCurrentVersion', 'TopInAppPurchases','Cluster1','Cluster2','Cluster3','Cluster4')])
      Rank            Price          Screenshot         Size          StarCurrentVersion RatingCurrentVersion TopInAppPurchases     Cluster1         Cluster2         Cluster3         Cluster4     
 Min.   :  1.0   Min.   :0.990   Min.   : 3.000   Min.   :  0.80   Min.   :1.000      Min.   :    5.0      Min.   : 0.000   Min.   :0.00   Min.   :0.000   Min.   :0.000   Min.   : 0.00  
 1st Qu.: 74.5   1st Qu.:0.990   1st Qu.: 5.000   1st Qu.: 17.93   1st Qu.:4.000      1st Qu.:   66.0      1st Qu.: 0.000   1st Qu.:0.00   1st Qu.:0.000   1st Qu.:0.000   1st Qu.: 6.00  
 Median :171.5   Median :0.990   Median : 5.500   Median : 36.45   Median :4.500      Median :  324.5      Median : 0.000   Median :0.00   Median :0.000   Median :0.000   Median :11.00  
 Mean   :161.1   Mean   :1.838   Mean   : 7.264   Mean   :102.46   Mean   :4.322      Mean   : 3773.8      Mean   : 2.332   Mean   :0.46   Mean   :0.448   Mean   :0.588   Mean   :13.39  
 3rd Qu.:243.0   3rd Qu.:1.990   3rd Qu.:10.000   3rd Qu.: 97.47   3rd Qu.:4.500      3rd Qu.: 2259.8      3rd Qu.: 4.000   3rd Qu.:1.00   3rd Qu.:0.000   3rd Qu.:1.000   3rd Qu.:18.75  
 Max.   :300.0   Max.   :6.990   Max.   :10.000   Max.   :962.00   Max.   :5.000      Max.   :192624.0     Max.   :10.000   Max.   :4.00   Max.   :6.000   Max.   :4.000   Max.   :46.00  
```

- Variable Transformation

```
# Variable Transformation #
Sales <- -log(Apps_new$Rank)
Log_Rating_Num <- log(Apps_new$RatingCurrentVersion+1)

Apps_new <-cbind(Apps_new, Sales)
Apps_new <-cbind(Apps_new, Log_Rating_Num)
```

Step4: Build a Regression Model

- Multiple Regression Model

$$-\log(Rank) = \beta_0 + \beta_1 Price + \beta_2 Screenshot + \beta_3 Size + \beta_4 Rating\_Score + \beta_5 \log(Rating\_Num) + \beta_6 IAP + \beta_7 Cluster1 + \beta_8 Cluster2 + \beta_9 Cluster3 + \beta_{10} Cluster4 + \varepsilon$$

  - *Cluster1*: Score for Update Info.
  - *Cluster2*: Score for Sales Performance
  - *Cluster3*: Score for Price Promotion
  - *Cluster4*: Score for Content

Step5: Estimate the Regression Model

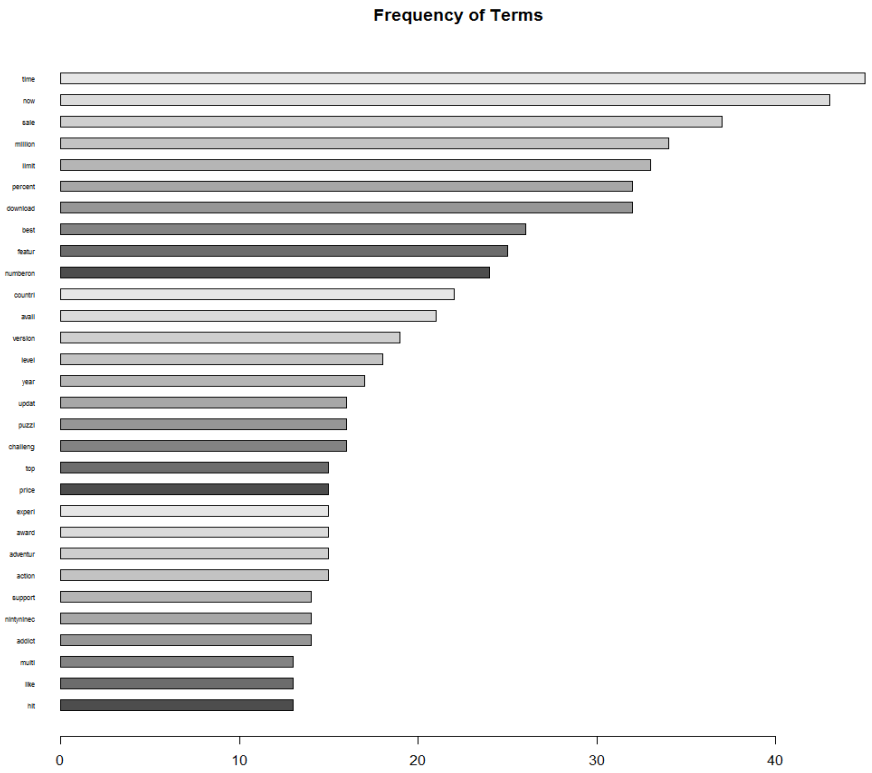
- Model Estimation by OLS

```
# Build a regression model #
Apps_Reg <- lm(Sales ~ Price + Screenshot + Size + StarCurrentVersion + Log_Rating_Num + TopInAppPurchases + Cluster1 + Cluster2 + Cluster3 + Cluster4, data=Apps_new)
summary(Apps_Reg)
```

```
Call:
lm(formula = Sales ~ Price + Screenshot + Size + StarCurrentVersion +
    Log_Rating_Num + TopInAppPurchases + Cluster1 + Cluster2 +
    Cluster3 + Cluster4, data = Apps_new)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5351 -0.6922 -0.2924  0.4786  4.8457

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -6.706e+00  5.632e-01 -11.908  < 2e-16 ***
Price             -5.019e-02  5.252e-02  -0.956  0.34022
Screenshot        -1.787e-02  2.834e-02  -0.631  0.52887
Size              -7.051e-05  4.713e-04  -0.150  0.88120
StarCurrentVersion 3.366e-01  1.300e-01   2.589  0.01021 *
Log_Rating_Num     1.073e-01  3.088e-02   3.473  0.00061 ***
TopInAppPurchases  1.000e-02  2.019e-02   0.496  0.62065
Cluster1          -8.726e-02  8.587e-02  -1.016  0.31058
Cluster2          -1.427e-01  7.135e-02  -2.000  0.04662 *
Cluster3           3.233e-02  6.277e-02   0.515  0.60698
Cluster4           1.149e-02  7.026e-03   1.635  0.10341
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.016 on 239 degrees of freedom
Multiple R-squared:   0.11,     Adjusted R-squared:  0.0728
F-statistic: 2.955 on 10 and 239 DF,  p-value: 0.001602
```

Step6: Evaluate the Business Problem

- **Problem: How to write successful App product descriptions?**
- **Task1:** Which terms/words are most frequently appeared in App product descriptions?



NANYANG TECHNOLOGICAL UNIVERSITY

Step6: Evaluate the Business Problem

- **Task2:** What are the common patterns/styles in presenting App product descriptions?

K-means Clustering

	Cluster 1	Cluster 2	Cluster 3
N	2	4	6
Terms	download million	limit percent (%) sale time	avail best countri feature now numberon
Example	"Downloaded over 10 million times in 2012"	"Limited Time SALE! 50% OFF!" "$ 0.99 for a limited time only"	"Screen rotation and zoom are now available ""NEW FEATURES: Retina / HD support', "Play all 60 levels, then take on 60 tests in Challenge mode" "#1 App in 30 countries"
Label	Sales Performance	Price Promotion	Update Info. + Sales Performance

Hierarchical Clustering

	Cluster 1	Cluster 2	Cluster 3
N	4	4	4
Terms	avail best feature now	countri download million numberon (#1)	limit percent (%) sale time
Example	"Screen rotation and zoom are now available ""NEW FEATURES: Retina / HD support', "Play all 60 levels, then take on 60 tests in Challenge mode"	"#1 App in 30 countries" "Downloaded over 10 million times in 2012"	"Limited Time SALE! 50% OFF!" "$ 0.99 for a limited time only"
Label	Update Info.	Sales Performance	Price Promotion

Step6: Evaluate the Business Problem

- **Task3:** What types of messages in the descriptions are associated with better sales performance?
  - With the current sample and model specification, the significant impacts of certain messages in App product descriptions on App sales are not identified.
  - WHY?

Step6: Evaluate the Business Problem

- Findings
  - There are commonly used product description patterns for hit Apps in App store markets.
    - Price Promotion, Sales Performance, Update Information
  - App product descriptions solely do not have significant impact on App sales
    - Low prediction for App sales

- Managerial Implications/Insights
  - Assists App developers to formulate a successful product description
  - Provides guidelines to new market entrants and enable them to improve sales performance