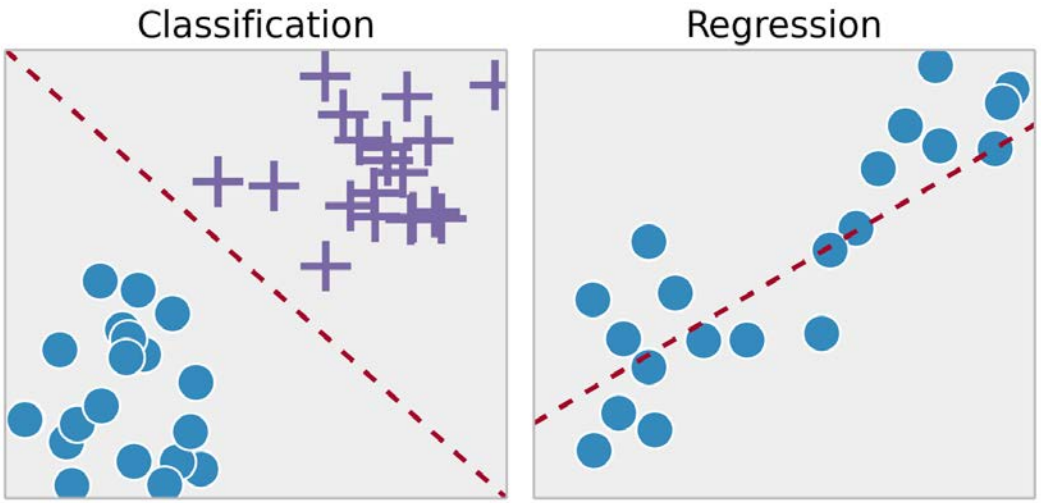
Smart Technology - AIR

Naïve Bayes

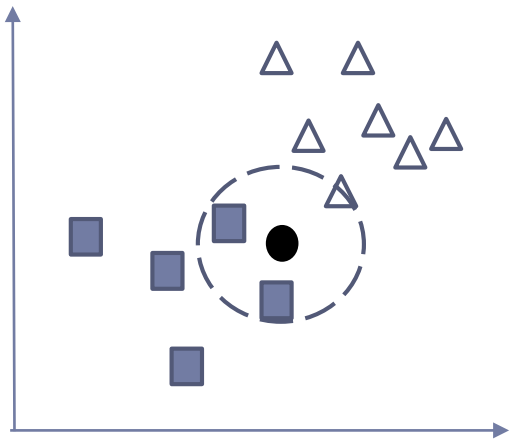
지도학습

- 지도학습 (Supervised Learning)
  - 학습데이터가 레이블을 갖고 있는 경우, 데이터로부터 형태 또는 함수를 유추해 내기 위 한 문제
  - 분류(Classification): 결과 값이 학습 데이터 세트에 포함된 값 중에 하나 도출
  - 예측(Prediction): 함수식으로 계산한 임의의 값 도출

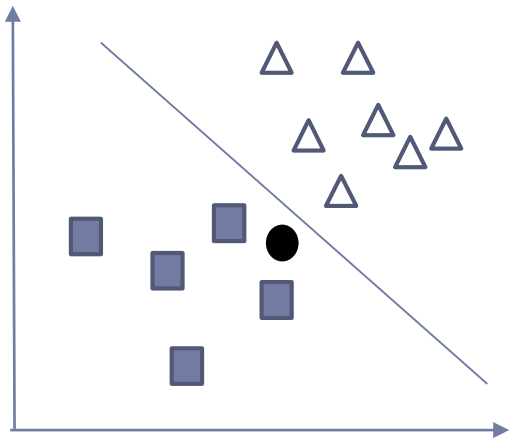


분류모델

- 분류 모델
  - 그룹명(레이블)이 적힌 학습 데이터로 학습 한 후, 새로운 데이터가 속한 그룹을 찾아내는 방법
  - 이진 분류: 분류되는 그룹의 수가 2개인 경우
  - 다항 분류: 분류되는 구룹의 수가 3개 이상인 경우
  - 나이브베이즈(Naïve bayes), kNN (k-nearest neighbor), SVM (support vector machine), 의사결정나무 (decision tree)



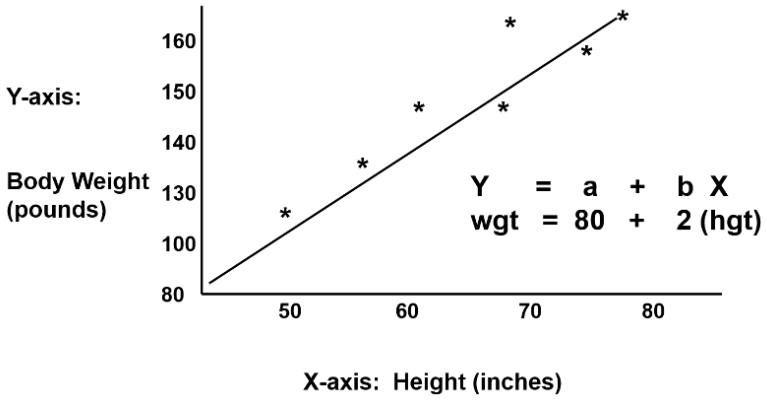
kNN (k=3)



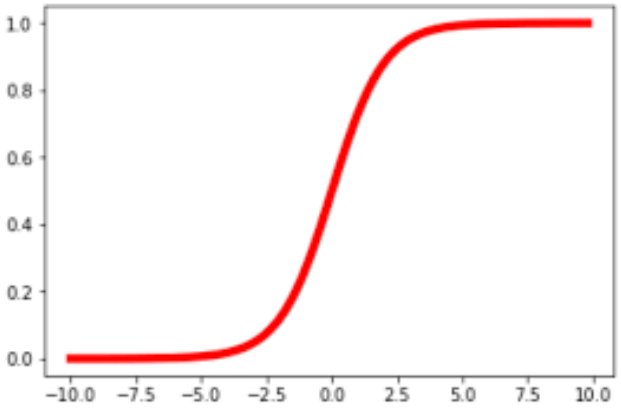
SVM

예측 모델

- 예측 모델
  - 회귀(Regression) 모델: 학습데이터를 이용하여 특성과 레이블의 관계를 함수식으로 표현하는 방법
  - 선형회귀(Linear regression) 모델: 독립변수와 종속변수의 관계를 직선 형태로 표현
    - 단순선형회귀: 독립변수가 한 개인 경우
    - 다중선형회귀: 독립변수가 2개 이상인 경우
  - 참고) 로지스틱회귀(Logistic regression) 모델: 종속 변수가 범주형 데이터로 표현



선형회귀



로지스틱회귀

빈도론 vs. 베이지안론

- 빈도론 (빈도론적 확률론)
  - John Venn (영국의 철학자):"확률은 그 사건이 일어난 횟수의 장기적인 비율이다."
  - 얼마나 빈번하게 특정 사건이 반복되는지 관찰하고 가설을 검증
  - 귀무가설, 대립가설, p-value

- 베이지안론
  - 베이즈정리(Bayes' Theorem)을 기반으로 확률을 해석해서 추론

확률 Review

- Conditional Probability
  - p(B|A) : event A 가 발생한 것을 알고 있을 때, event B가 발생할 확률

The conditional probability of B, given A, denoted by $P(B|A)$, is defined by

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \quad \text{provided} \quad P(A) > 0.$$

If in an experiment the events A and B can both occur, then

$$P(A \cap B) = P(A)P(B|A), \text{ provided } P(A) > 0.$$

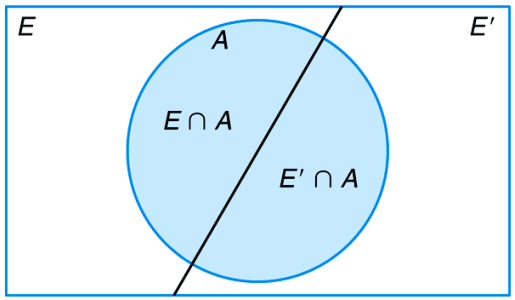
  - p(John이 채식주의자가 됨) = 0.000001
  - p(John이 채식주의자가 됨| $1B를 지불함) = 1

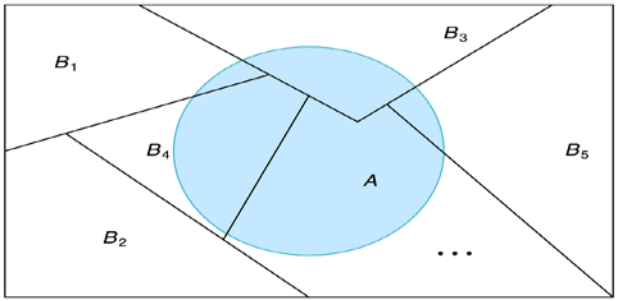
확률 Review

- Total Probability

If the events $B_1, B_2, \ldots, B_k$ constitute a partition of the sample space S such that $P(B_i) \neq 0$ for $i = 1, 2, \ldots, k$, then for any event A of S,

$$P(A) = \sum_{i=1}^{k} P(B_i \cap A) = \sum_{i=1}^{k} P(B_i)P(A|B_i).$$





- p(채식주의) = p($1B)p(채식주의| $1B) + p(not $1B)p(채식주의| not $1B)
  = 0 * 1 + 1 * 0.000001 = 0.000001

확률 Review

- Bayes Rule

(Bayes' Rule) If the events $B_1, B_2, \ldots, B_k$ constitute a partition of the sample space S such that $P(B_i) \neq 0$ for $i = 1, 2, \ldots, k$, then for any event A in S such that $P(A) \neq 0$,

$$P(B_r|A) = \frac{P(B_r \cap A)}{\sum_{i=1}^{k} P(B_i \cap A)} = \frac{P(B_r)P(A|B_r)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)} \quad \text{for } r = 1, 2, \ldots, k.$$

- p(음악듣기)=0.8
- p(K-pop 듣기) = 0.3

- p(음악듣기|K-pop 듣기) = 1
- p(K-pop 듣기|음악듣기) = p(K-pop 듣기, 음악듣기) / p(음악 듣기)
  = p(K-pop 듣기) * p(음악듣기|K-pop 듣기) / p(음악 듣기)
  = 0.3 * 1 / 0.8 = 0.375

Twitter 데이터 분석

- Mandrill.com

App

Animal

Game

- p(app | word1, word2, word3,…)
- p(other | word1, word2, word3,…)

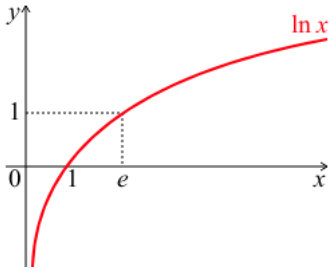
- If p(app | word1, word2, word3,…) > p(other | word1, word2, word3,…) then you have a tweet about the Mandrill app.

Twitter 데이터 분석

- Bayes Rule의 적용
  - $p(app \mid word1, word2, word3,\ldots) = \frac{p(app) * p(word1, word2, word3,\ldots \mid app)}{p(word1, word2, word3,\ldots)}$
  - $p(other \mid word1, word2, word3,\ldots) = \frac{p(other) * p(word1, word2, word3,\ldots \mid other)}{p(word1, word2, word3,\ldots)}$
  - p(word1, word2, word3,…)는 class에 따라서 변하지 않는 값
    → $p(app) * p(word1, word2, word3,\ldots \mid app)$
    vs. $p(other) * p(word1, word2, word3,\ldots \mid other)$
  - Twitter의 글이 독립적이라면..
    $p(app) * p(word1, word2,\ldots \mid app) = p(app) * p(word1 \mid app) * p(word2 \mid app)\ldots$
    $p(other) * p(word1, word2,\ldots \mid other) = p(other) * p(word1 \mid other) * p(word2 \mid other)\ldots$

Twitter 데이터 분석

- 50:50의 빈도로 가정하면 p(app) = p(others) = 0.5
  - p(word1| app) ∗ p(word2| app)… vs. p(word1|other) ∗ p(word2|other)…
- Additive smoothing
  - Rare words: 기존 학습데이터에 없던 단어가 나온다면?
  - p("Tubal-cain" | app) = 0
  - p("Tubal-cain" | app)*∗ p(word2| app) * ∗ p(word3| app)… = 0
  - 모든 단어는 한번은 나왔다고 처리
- Floating-point underflow
  - 출현 회수가 적은 단어는 확률이 매우 작음
  - 확률이 0.001인 단어 15개 : $0.001 * 0.001 * 0.001 * \ldots = 1*10^{-45}$
  - Log 함수의 활용
    - 0.2 * 0.8
    - ln(0.2 * 0.8) = ln(0.2) + ln(0.8)



Twitter 데이터의 수집

- App과 Others의 데이터 각각 150개

	A	B
1	Tweet	
2	[blog] Using Nullmailer and Mandrill for your Ub	
3	[blog] Using Postfix and free Mandrill email serv	
4	@aalbertson There are several reasons emails g	
5	@adrienneleigh I just switched it over to Mandr	
6	@ankeshk +1 to @mailchimp We use MailChimp	
7	@biggoldring That error may occur if unsuppor	
8	@BlueHayes mind sending us some details abou	
9	@cemsisman It can vary, but if sending really lo	
10	@compactcode Have you checked out Mandrill	
11	@devongovett I'm using Mandrill, but been sav	
12	@devongovett Mandrill seems pretty cheap (by	
13	@dzuelke The option to set up the Mandrill int	
14	@dzuelke You should be able to login to the M	
15	@edocr Can you send an email via http://help.m	
16	@eladlouni Et oui, Mandrill est moins cher, mais	
17	@eladlouni La raison : Mandrill = Mailchimp, et	
18	@eladlouni Mandrill = pas de list management	
19	@Elie__ @camj59 jparle de relai SMTP!1 million	
20	@Elie__ @camj59 mandrill! Sendgrid! Sans parle	
21	@EricCandino They're unfortunately not for sale	

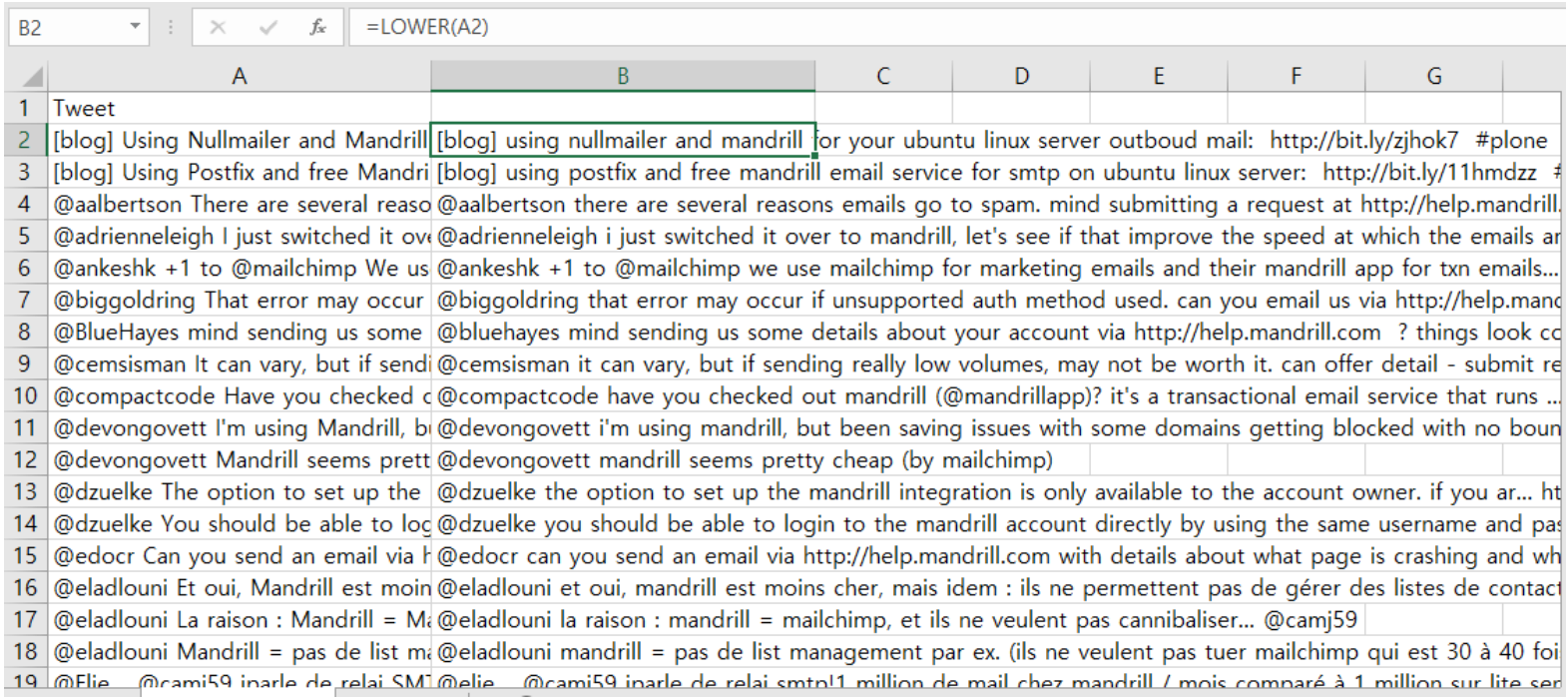
AboutMandrillApp | AboutOther

	A
1	Tweet
2	¿En donde esta su remontada Mandrill?
3	.@Katie_PhD Alternate, 'reproachful mandrill' co
4	.@theophani can i get "drill" in there? it would b
5	"@ChrisJBoyland: Baby Mandrill Paignton Zoo 29
6	"@MISSMYA #NameAnAmazingBand MANDRILL!
7	"Fat City Strut" by Mandrill is my new jam. http:/
8	【SOUL TRAIN #22】1973年 MANDRILL 中古盤屋て
9	@alicegreennn_ but how come you didn't have
10	@As_TomasRoncero a la mierda el mandrill toca
11	@Burnziey @sjsharkfinatic I have zach mandrill c
12	@charlie29598 #GreatJob #Mandrill
13	@JustDewYou @isma_longo tu no ables sapo qu
14	@Khamili_1015 @espsmile t'es vraiment idiot cc
15	@mandrill "The venue should be opening once
16	@mandrill @CCP_Manifest @CCPGames don't g
17	@mandrill @CCPGames I hope they record the
18	@mandrill @ccpgames Plus, we have a faint CCF
19	@mandrill @Freebooted probably try to corner
20	@mandrill @Freebooted. I want to say Heinlein
21	@mandrill @HilmarVeigar ZOMG Want that jack

AboutMandrillApp | AboutOther

데이터 전처리하기

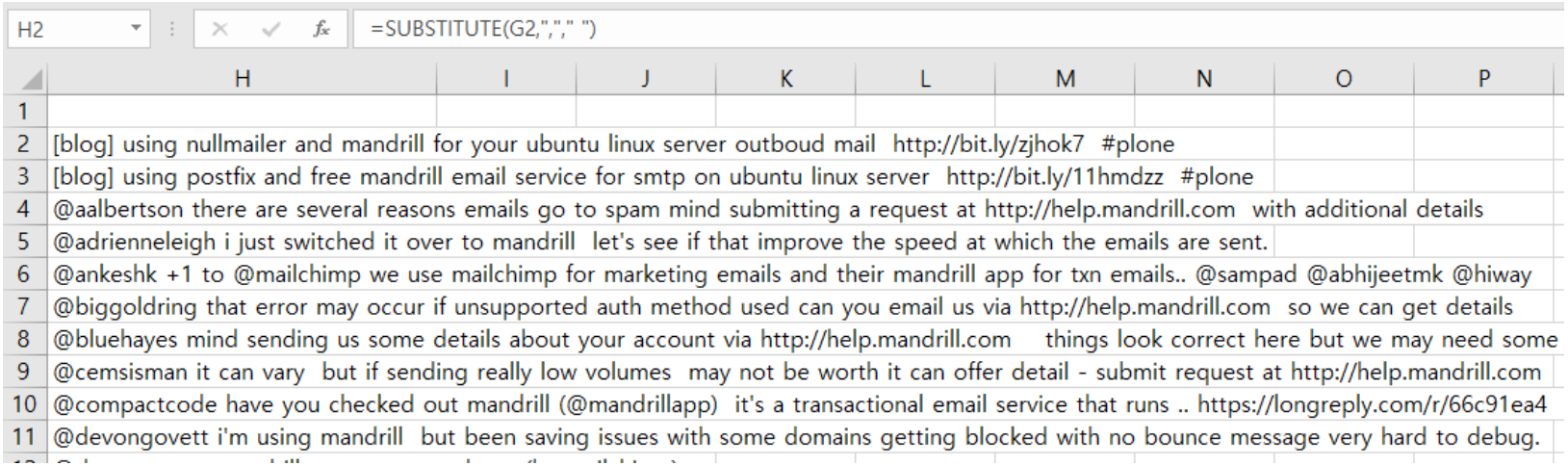
- 소문자로 변환하기
  - E-mail과 e-mail은 동일한 의미
  - B2=LOWER(A2)



	A	B
1	Tweet	
2	[blog] Using Nullmailer and Mandrill	[blog] using nullmailer and mandrill for your ubuntu linux server outboud mail: http://bit.ly/zjhok7 #plone
3	[blog] Using Postfix and free Mandri	[blog] using postfix and free mandrill email service for smtp on ubuntu linux server: http://bit.ly/11hmdzz #
4	@aalbertson There are several reaso	@aalbertson there are several reasons emails go to spam. mind submitting a request at http://help.mandrill.
5	@adrienneleigh I just switched it ov	@adrienneleigh i just switched it over to mandrill, let's see if that improve the speed at which the emails ar
6	@ankeshk +1 to @mailchimp We us	@ankeshk +1 to @mailchimp we use mailchimp for marketing emails and their mandrill app for txn emails...
7	@biggoldring That error may occur	@biggoldring that error may occur if unsupported auth method used. can you email us via http://help.manc
8	@BlueHayes mind sending us some	@bluehayes mind sending us some details about your account via http://help.mandrill.com ? things look cc
9	@cemsisman It can vary, but if sendi	@cemsisman it can vary, but if sending really low volumes, may not be worth it. can offer detail - submit re
10	@compactcode Have you checked c	@compactcode have you checked out mandrill (@mandrillapp)? it's a transactional email service that runs ...
11	@devongovett I'm using Mandrill, b	@devongovett i'm using mandrill, but been saving issues with some domains getting blocked with no boun
12	@devongovett Mandrill seems prett	@devongovett mandrill seems pretty cheap (by mailchimp)
13	@dzuelke The option to set up the	@dzuelke the option to set up the mandrill integration is only available to the account owner. if you ar... ht
14	@dzuelke You should be able to log	@dzuelke you should be able to login to the mandrill account directly by using the same username and pas
15	@edocr Can you send an email via h	@edocr can you send an email via http://help.mandrill.com with details about what page is crashing and wh
16	@eladlouni Et oui, Mandrill est moin	@eladlouni et oui, mandrill est moins cher, mais idem : ils ne permettent pas de gérer des listes de contact
17	@eladlouni La raison : Mandrill = Ma	@eladlouni la raison : mandrill = mailchimp, et ils ne veulent pas cannibaliser... @camj59
18	@eladlouni Mandrill = pas de list ma	@eladlouni mandrill = pas de list management par ex. (ils ne veulent pas tuer mailchimp qui est 30 à 40 foi
19	@Elie @camj59 jparle de relai SM	@elie @camj59 jparle de relai smtp!1 million de mail chez mandrill / mois comparé à 1 million sur lite ser

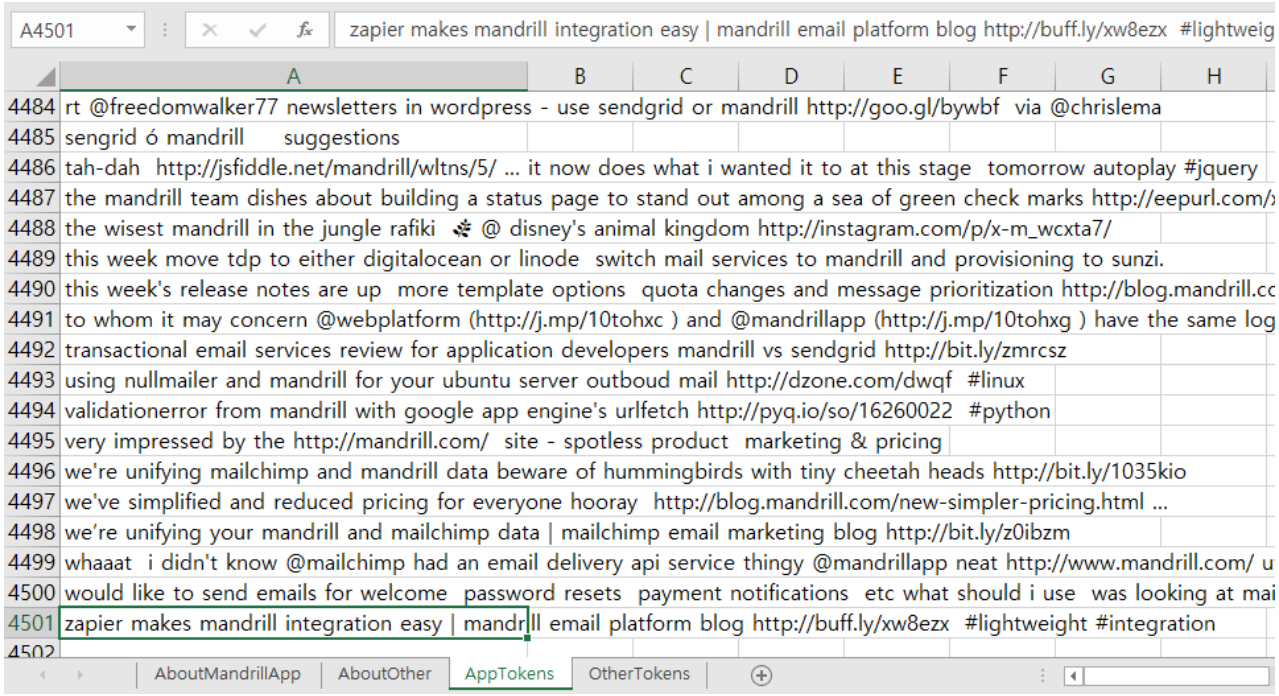
데이터 전처리하기

- 불필요한 기호 없애기
  - 마침표, 쉼표 등 의미 없는 기호는 공백으로 변환
  - C2=SUBSTITUTE(B2,". "," ")
  - D2=SUBSTITUTE(C2,": "," ")
  - E2=SUBSTITUTE(D2,"?"," ")
  - F2=SUBSTITUTE(E2,"!"," ")
  - G2=SUBSTITUTE(F2,";"," ")
  - H2=SUBSTITUTE(G2,","," ")



데이터 전처리하기

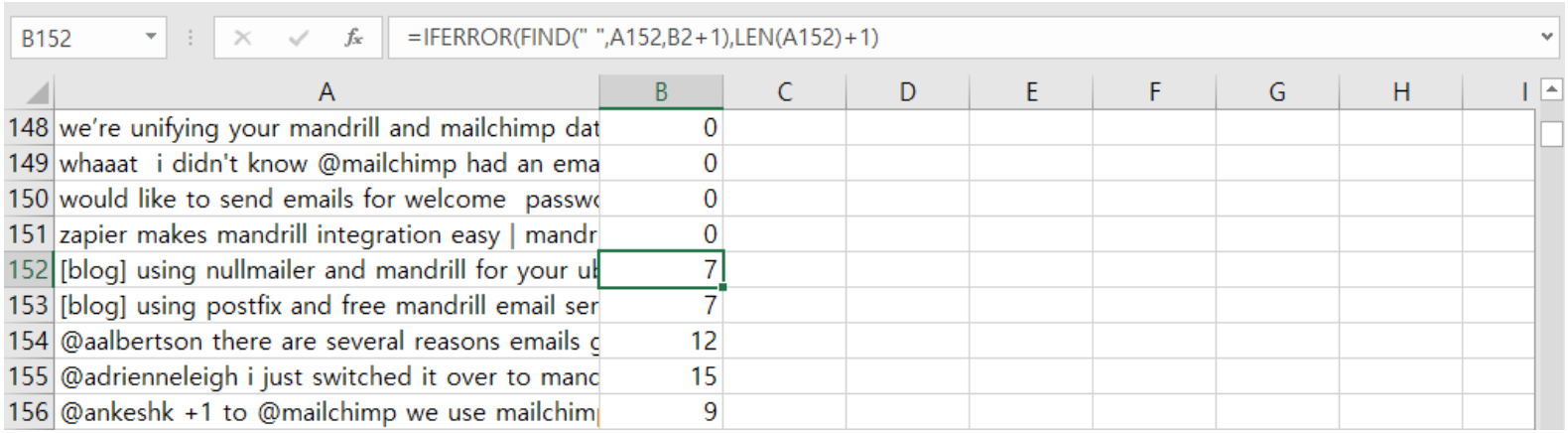
- 단어로 구분하기 (Tokenizing)
  - AppTokens, OtherTokens Tap 만들기
  - A1에 Tweet이라는 이름 작성
  - AboutMandrillApp과 AboutOther의 전처리된 Data (H2~H150)를 AppTokens와 OtherTokens의 A2~A4501에 복사.. 동일한 글이 150회 반복 복사됨



데이터 전처리하기

- 단어로 구분하기 (Tokenizing)
  - B1에 Space Position이라는 이름 작성
  - B2~B151에 0 입력: 첫 단어의 시작지점
  - 다음 공백문자 위치 찾기 (A2,A152,A302는 같은 내용의 글 반복)
  - B152=FIND(" " ,A152,B2+1)

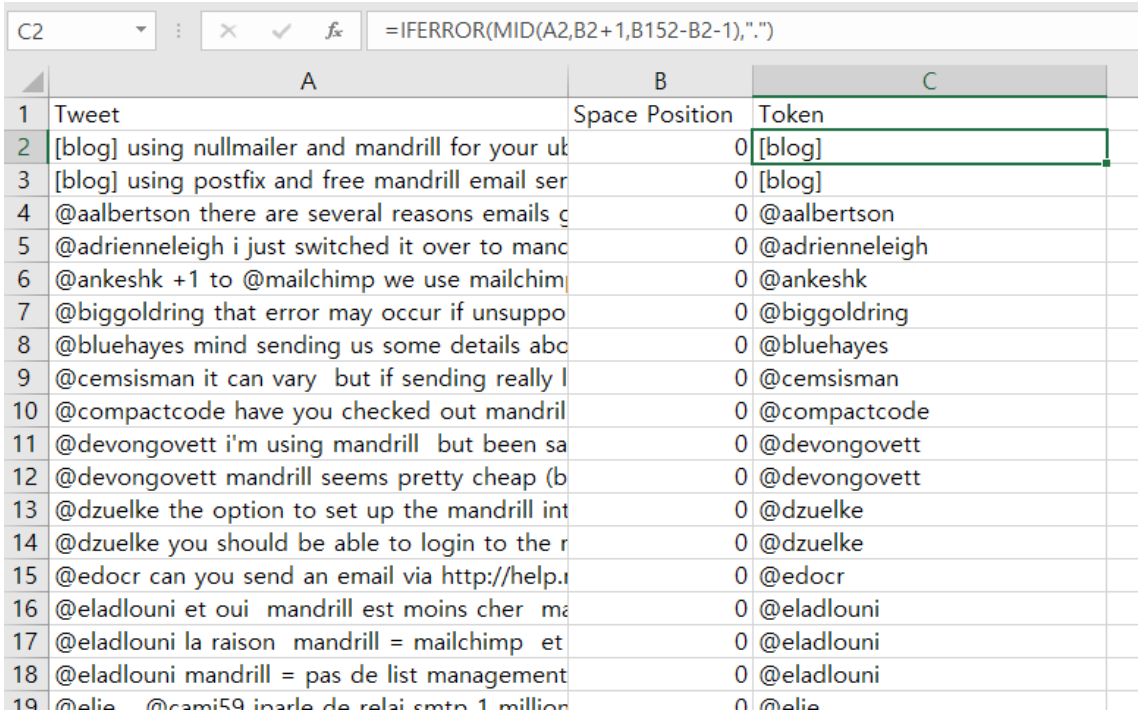
  - 만일 30개의 단어보다 적다면 Error 발생, 글자수 + 1의 값 return
  - B152=IFERROR(FIND(" ",A152,B2+1),LEN(A152)+1)



	A	B	C	D	E	F	G	H	I
148	we're unifying your mandrill and mailchimp dat	0							
149	whaaat i didn't know @mailchimp had an ema	0							
150	would like to send emails for welcome passwo	0							
151	zapier makes mandrill integration easy | mandr	0							
152	[blog] using nullmailer and mandrill for your ul	7							
153	[blog] using postfix and free mandrill email ser	7							
154	@aalbertson there are several reasons emails g	12							
155	@adrienneleigh i just switched it over to manc	15							
156	@ankeshk +1 to @mailchimp we use mailchim	9							

데이터 전처리하기

- 단어로 구분하기 (Tokenizing)
  - C1에 Token이라는 이름 작성
  - 공백문자 위치 정보로 Token 생성하기
  - C2=IFERROR(MID(A2,B2+1,B152-B2-1),".")

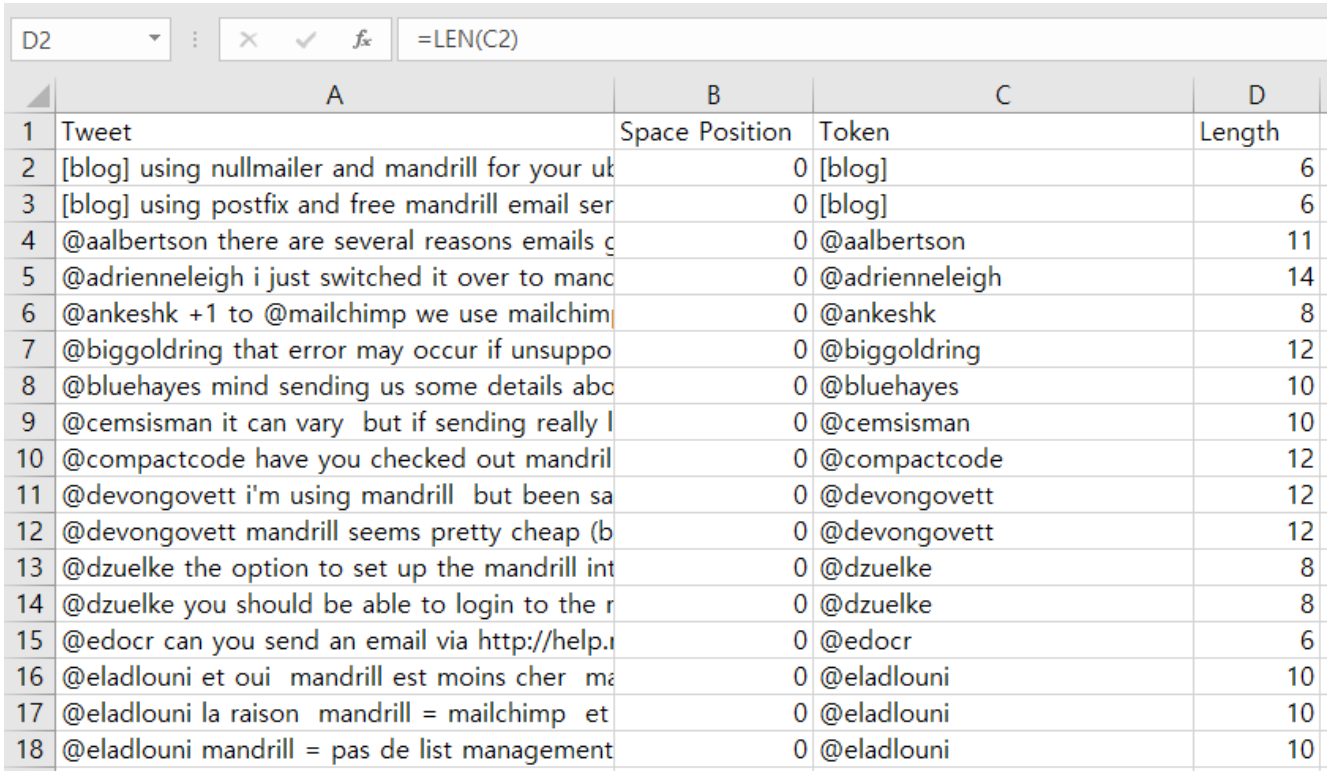


C2 =IFERROR(MID(A2,B2+1,B152-B2-1),".")

	A	B	C
1	Tweet	Space Position	Token
2	[blog] using nullmailer and mandrill for your ub	0	[blog]
3	[blog] using postfix and free mandrill email ser	0	[blog]
4	@aalbertson there are several reasons emails g	0	@aalbertson
5	@adrienneleigh i just switched it over to manc	0	@adrienneleigh
6	@ankeshk +1 to @mailchimp we use mailchim	0	@ankeshk
7	@biggoldring that error may occur if unsuppo	0	@biggoldring
8	@bluehayes mind sending us some details abc	0	@bluehayes
9	@cemsisman it can vary but if sending really l	0	@cemsisman
10	@compactcode have you checked out mandril	0	@compactcode
11	@devongovett i'm using mandrill but been sa	0	@devongovett
12	@devongovett mandrill seems pretty cheap (b	0	@devongovett
13	@dzuelke the option to set up the mandrill int	0	@dzuelke
14	@dzuelke you should be able to login to the r	0	@dzuelke
15	@edocr can you send an email via http://help.r	0	@edocr
16	@eladlouni et oui mandrill est moins cher ma	0	@eladlouni
17	@eladlouni la raison mandrill = mailchimp et	0	@eladlouni
18	@eladlouni mandrill = pas de list management	0	@eladlouni
19	@elie @cami59 inparle de relai smtp 1 million	0	@elie

데이터 전처리하기

- 단어로 구분하기 (Tokenizing)
  - D1에 Length이라는 이름 작성
  - 글자수 카운트하기
  - D2=LEN(C2)

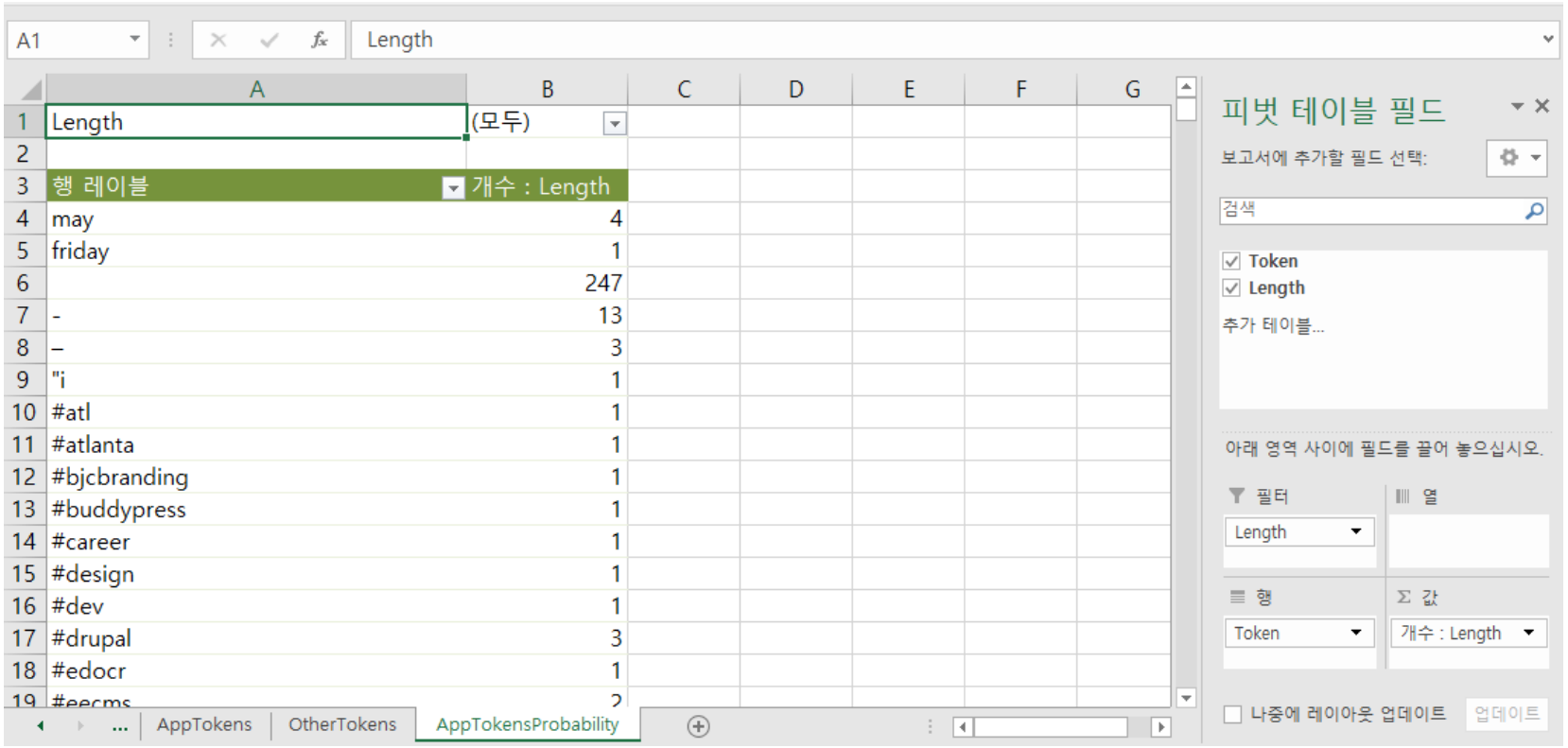


D2 =LEN(C2)

	A	B	C	D
1	Tweet	Space Position	Token	Length
2	[blog] using nullmailer and mandrill for your ub	0	[blog]	6
3	[blog] using postfix and free mandrill email ser	0	[blog]	6
4	@aalbertson there are several reasons emails g	0	@aalbertson	11
5	@adrienneleigh i just switched it over to manc	0	@adrienneleigh	14
6	@ankeshk +1 to @mailchimp we use mailchim	0	@ankeshk	8
7	@biggoldring that error may occur if unsuppo	0	@biggoldring	12
8	@bluehayes mind sending us some details abc	0	@bluehayes	10
9	@cemsisman it can vary but if sending really l	0	@cemsisman	10
10	@compactcode have you checked out mandril	0	@compactcode	12
11	@devongovett i'm using mandrill but been sa	0	@devongovett	12
12	@devongovett mandrill seems pretty cheap (b	0	@devongovett	12
13	@dzuelke the option to set up the mandrill int	0	@dzuelke	8
14	@dzuelke you should be able to login to the r	0	@dzuelke	8
15	@edocr can you send an email via http://help.r	0	@edocr	6
16	@eladlouni et oui mandrill est moins cher ma	0	@eladlouni	10
17	@eladlouni la raison mandrill = mailchimp et	0	@eladlouni	10
18	@eladlouni mandrill = pas de list management	0	@eladlouni	10

확률 계산하기

- Token 확인하기
  - Token과 Length data로 Pivot 테이블 생성 (AppTokensProbability)



A1 | Length

	A	B
1	Length	(모두)
2		
3	행 레이블	개수 : Length
4	may	4
5	friday	1
6		247
7	-	13
8	–	3
9	"i	1
10	#atl	1
11	#atlanta	1
12	#bjcbranding	1
13	#buddypress	1
14	#career	1
15	#design	1
16	#dev	1
17	#drupal	3
18	#edocr	1
19	#eecms	2

AppTokens | OtherTokens | AppTokensProbability

피벗 테이블 필드

보고서에 추가할 필드 선택:

검색

Token
Length

추가 테이블...

아래 영역 사이에 필드를 끌어 놓으십시오.

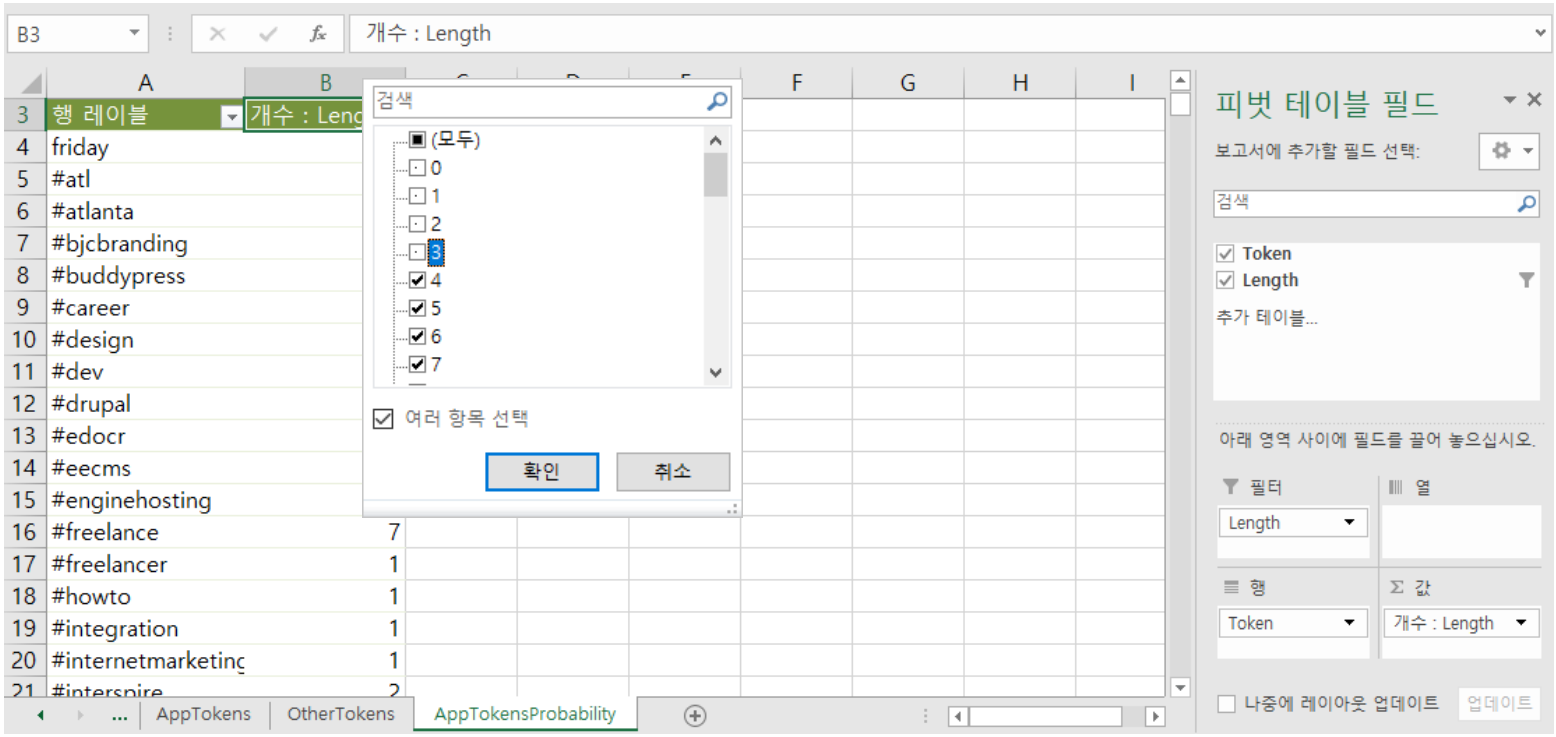
필터: Length | 열

행: Token | Σ 값: 개수 : Length

나중에 레이아웃 업데이트 | 업데이트

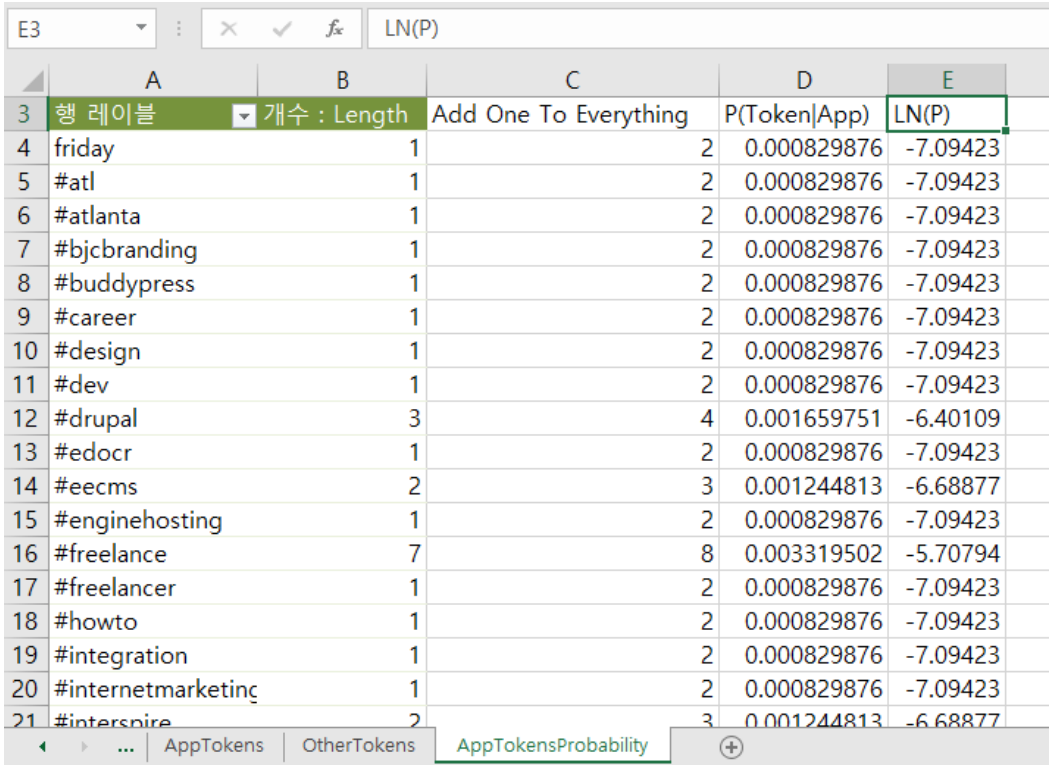
확률 계산하기

- Token 확인하기
  - 불필요한 Token은 제외하기



확률 계산하기

- Token 확률 계산하기
  - C3에 Add One To Everything 작성
  - Length에 1 더하기(additive smoothing)
  - C4=B4+1
  - 합계 계산하기
  - C827=SUM(C4:C826) = 2411
  - D3에 P(Token|App) 작성
  - Token 발생 확률 계산하기
  - D4=C4/$C$827
  - E3에 LN(P) 작성
  - Log 값으로 변환하기
  - E4=LN(D4)



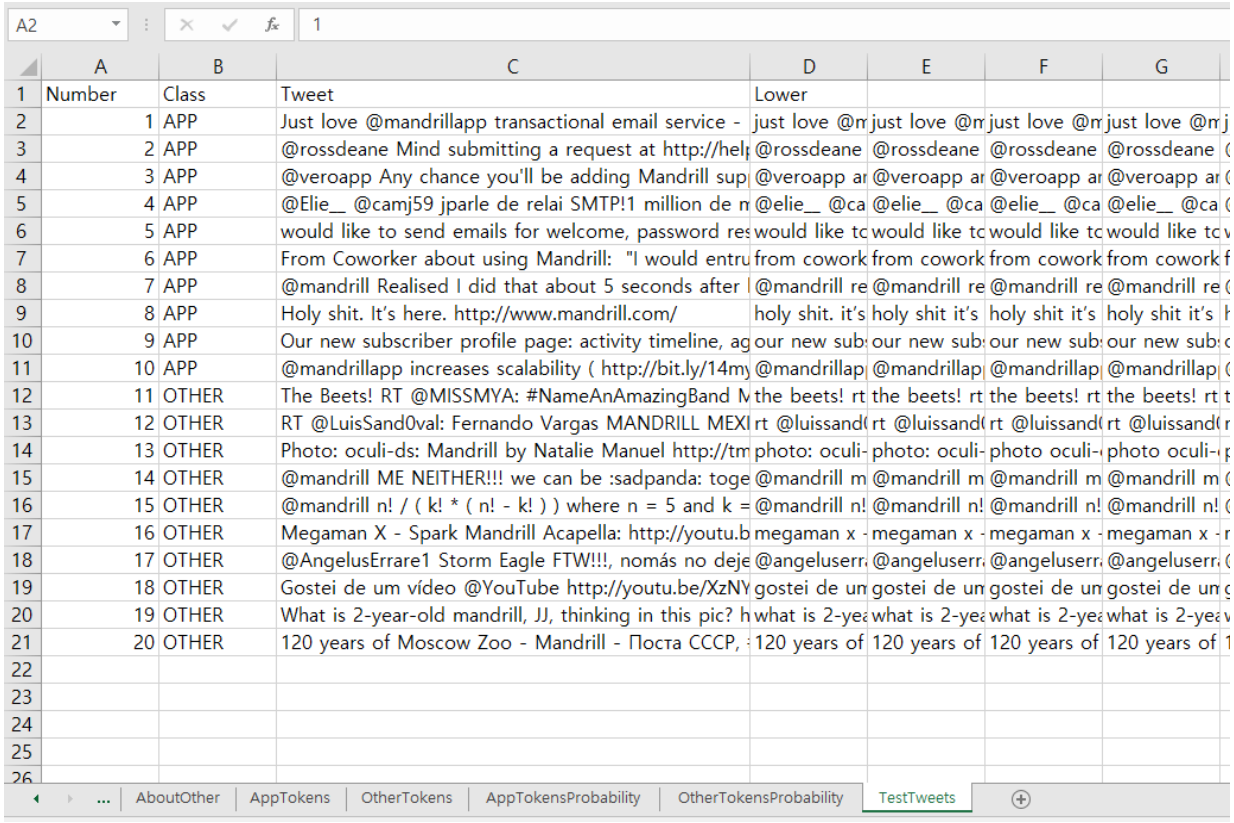
E3 | LN(P)

	A	B	C	D	E
3	행 레이블	개수 : Length	Add One To Everything	P(Token|App)	LN(P)
4	friday	1	2	0.000829876	-7.09423
5	#atl	1	2	0.000829876	-7.09423
6	#atlanta	1	2	0.000829876	-7.09423
7	#bjcbranding	1	2	0.000829876	-7.09423
8	#buddypress	1	2	0.000829876	-7.09423
9	#career	1	2	0.000829876	-7.09423
10	#design	1	2	0.000829876	-7.09423
11	#dev	1	2	0.000829876	-7.09423
12	#drupal	3	4	0.001659751	-6.40109
13	#edocr	1	2	0.000829876	-7.09423
14	#eecms	2	3	0.001244813	-6.68877
15	#enginehosting	1	2	0.000829876	-7.09423
16	#freelance	7	8	0.003319502	-5.70794
17	#freelancer	1	2	0.000829876	-7.09423
18	#howto	1	2	0.000829876	-7.09423
19	#integration	1	2	0.000829876	-7.09423
20	#internetmarketinc	1	2	0.000829876	-7.09423
21	#interspire	2	3	0.001244813	-6.68877

... | AppTokens | OtherTokens | AppTokensProbability

테스트데이터

- TestTweet의 문장으로 확인
  - APP 데이터, Other 데이터 각각 10개로 구성



	A	B	C	D	E	F	G
1	Number	Class	Tweet	Lower			
2	1	APP	Just love @mandrillapp transactional email service -	just love @m	just love @m	just love @m	just love @m
3	2	APP	@rossdeane Mind submitting a request at http://hel	@rossdeane	@rossdeane	@rossdeane	@rossdeane
4	3	APP	@veroapp Any chance you'll be adding Mandrill sup	@veroapp ar	@veroapp ar	@veroapp ar	@veroapp ar
5	4	APP	@Elie__ @camj59 jparle de relai SMTP!1 million de m	@elie__ @ca	@elie__ @ca	@elie__ @ca	@elie__ @ca
6	5	APP	would like to send emails for welcome, password res	would like to	would like to	would like to	would like to
7	6	APP	From Coworker about using Mandrill: "I would entru	from cowork	from cowork	from cowork	from cowork
8	7	APP	@mandrill Realised I did that about 5 seconds after	@mandrill re	@mandrill re	@mandrill re	@mandrill re
9	8	APP	Holy shit. It's here. http://www.mandrill.com/	holy shit. it's	holy shit it's	holy shit it's	holy shit it's
10	9	APP	Our new subscriber profile page: activity timeline, ag	our new sub	our new sub	our new sub	our new sub
11	10	APP	@mandrillapp increases scalability (http://bit.ly/14m	@mandrillap	@mandrillap	@mandrillap	@mandrillap
12	11	OTHER	The Beets! RT @MISSMYA: #NameAnAmazingBand M	the beets! rt	the beets! rt	the beets! rt	the beets! rt
13	12	OTHER	RT @LuisSand0val: Fernando Vargas MANDRILL MEXI	rt @luissand	rt @luissand	rt @luissand	rt @luissand
14	13	OTHER	Photo: oculi-ds: Mandrill by Natalie Manuel http://tm	photo: oculi-	photo: oculi-	photo oculi-	photo oculi-
15	14	OTHER	@mandrill ME NEITHER!!! we can be :sadpanda: toge	@mandrill m	@mandrill m	@mandrill m	@mandrill m
16	15	OTHER	@mandrill n! / (k! * (n! - k!)) where n = 5 and k =	@mandrill n!	@mandrill n!	@mandrill n!	@mandrill n!
17	16	OTHER	Megaman X - Spark Mandrill Acapella: http://youtu.b	megaman x -	megaman x -	megaman x -	megaman x -
18	17	OTHER	@AngelusErrare1 Storm Eagle FTW!!!, nomás no deje	@angeluserr	@angeluserr	@angeluserr	@angeluserr
19	18	OTHER	Gostei de um vídeo @YouTube http://youtu.be/XzNY	gostei de un	gostei de un	gostei de un	gostei de un
20	19	OTHER	What is 2-year-old mandrill, JJ, thinking in this pic? h	what is 2-yea	what is 2-yea	what is 2-yea	what is 2-yea
21	20	OTHER	120 years of Moscow Zoo - Mandrill - Поста СССР,	120 years of	120 years of	120 years of	120 years of
22							
23							
24							
25							
26							

AboutOther | AppTokens | OtherTokens | AppTokensProbability | OtherTokensProbability | TestTweets

테스트데이터

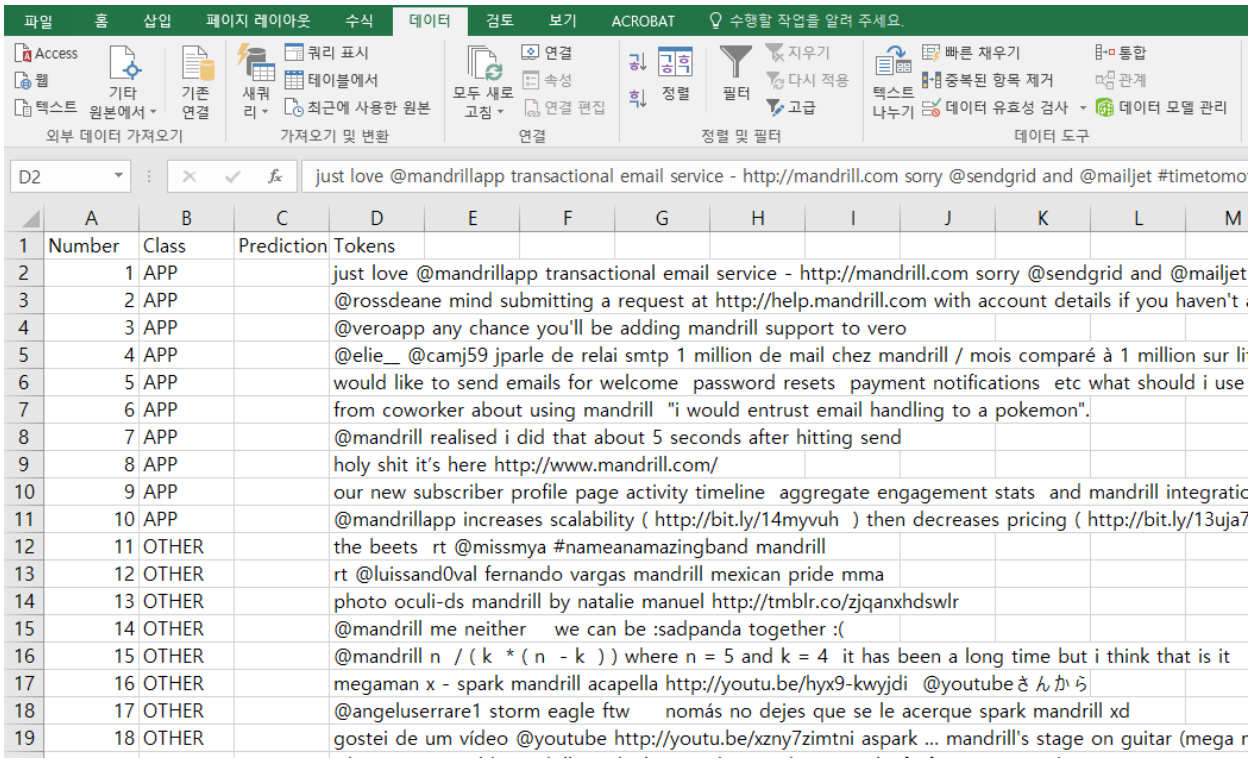
- 테스트 데이터 전처리
  - 새로운 Tap 생성(TestPredictions)
  - TestTweets의 Column A(Number)와 B(Class)를 복사
  - C1에 Prediction, D1에 Tokens 작성
  - TestTweets의 J2~J21의 값을 TestPredictions D2~D21에 복사 (선택하여 붙여넣기)

D2 | just love @mandrillapp transactional email service - http://mandrill.com sorry @sendgrid and @mailjet #timetomoveo

	A	B	C	D
1	Number	Class	Prediction	Tokens
2	1	APP		just love @mandrillapp transactional email service - http://mandrill.com sorry @sendgrid and @mailjet #ti
3	2	APP		@rossdeane mind submitting a request at http://help.mandrill.com with account details if you haven't alre
4	3	APP		@veroapp any chance you'll be adding mandrill support to vero
5	4	APP		@elie_ @camj59 jparle de relai smtp 1 million de mail chez mandrill / mois comparé à 1 million sur lite s
6	5	APP		would like to send emails for welcome password resets payment notifications etc what should i use wa
7	6	APP		from coworker about using mandrill "i would entrust email handling to a pokemon".
8	7	APP		@mandrill realised i did that about 5 seconds after hitting send
9	8	APP		holy shit it's here http://www.mandrill.com/
10	9	APP		our new subscriber profile page activity timeline aggregate engagement stats and mandrill integratio #b
11	10	APP		@mandrillapp increases scalability (http://bit.ly/14myvuh) then decreases pricing (http://bit.ly/13uja7s)
12	11	OTHER		the beets rt @missmya #nameanamazingband mandrill

테스트데이터

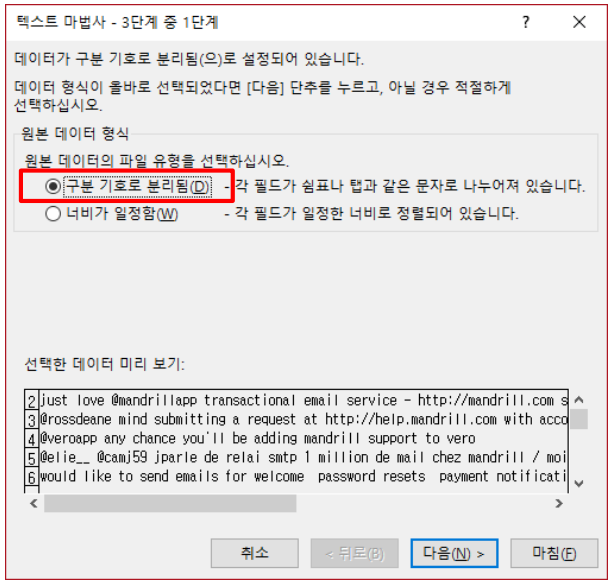
- 테스트 데이터 전처리
  - 공백 문자를 기준으로 토큰화 하기
  - D2~D21 하일라이트, 데이터>텍스트나누기

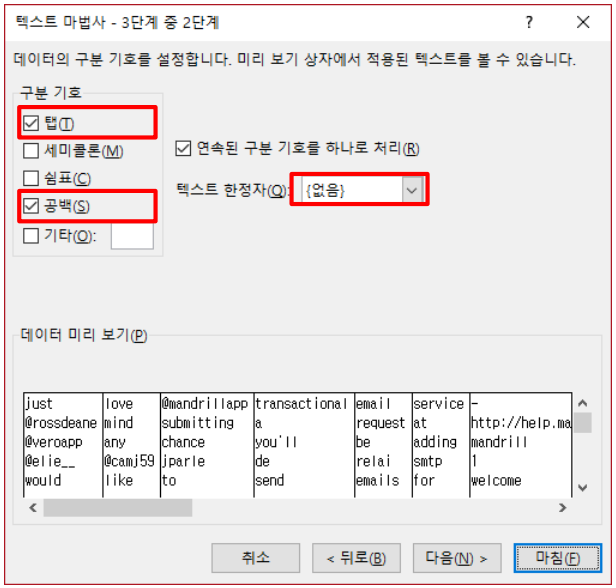


	A	B	C	D
1	Number	Class	Prediction	Tokens
2	1	APP		just love @mandrillapp transactional email service - http://mandrill.com sorry @sendgrid and @mailjet
3	2	APP		@rossdeane mind submitting a request at http://help.mandrill.com with account details if you haven't
4	3	APP		@veroapp any chance you'll be adding mandrill support to vero
5	4	APP		@elie_ @camj59 jparle de relai smtp 1 million de mail chez mandrill / mois comparé à 1 million sur li
6	5	APP		would like to send emails for welcome password resets payment notifications etc what should i use
7	6	APP		from coworker about using mandrill "i would entrust email handling to a pokemon".
8	7	APP		@mandrill realised i did that about 5 seconds after hitting send
9	8	APP		holy shit it's here http://www.mandrill.com/
10	9	APP		our new subscriber profile page activity timeline aggregate engagement stats and mandrill integratic
11	10	APP		@mandrillapp increases scalability (http://bit.ly/14myvuh) then decreases pricing (http://bit.ly/13uja7
12	11	OTHER		the beets rt @missmya #nameanamazingband mandrill
13	12	OTHER		rt @luissand0val fernando vargas mandrill mexican pride mma
14	13	OTHER		photo oculi-ds mandrill by natalie manuel http://tmblr.co/zjqanxhdswlr
15	14	OTHER		@mandrill me neither we can be :sadpanda together :(
16	15	OTHER		@mandrill n / (k * (n - k)) where n = 5 and k = 4 it has been a long time but i think that is it
17	16	OTHER		megaman x - spark mandrill acapella http://youtu.be/hyx9-kwyjdi @youtubeさんから
18	17	OTHER		@angeluserrare1 storm eagle ftw nomás no dejes que se le acerque spark mandrill xd
19	18	OTHER		gostei de um vídeo @youtube http://youtu.be/xzny7zimtni aspark ... mandrill's stage on guitar (mega r

테스트데이터

- 테스트 데이터 전처리
  - 공백 문자를 기준으로 토큰화 하기
  - 구분 기호로 분리
  - 탭, 공백으로 나누기 & 텍스트 한정자 없음





테스트데이터

- 테스트 데이터 전처리
  - 공백 문자를 기준으로 토큰화 하기 결과

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Number	Class	Prediction	Tokens											
2	1	APP		just	love	@mandrill	transactior	email	service	-	http://mar	sorry	@sendgric	and	@mailjet
3	2	APP		@rossdea	mind	submitting	a	request	at	http://help	with	account	details	if	you
4	3	APP		@veroapp	any	chance	you'll	be	adding	mandrill	support	to	vero		
5	4	APP		@elie__	@camj59	jparle	de	relai	smtp	1	million	de	mail	chez	mandrill
6	5	APP		would	like	to	send	emails	for	welcome	password	resets	payment	notificatio	etc
7	6	APP		from	coworker	about	using	mandrill	"i	would	entrust	email	handling	to	a
8	7	APP		@mandrill	realised	i	did	that	about	5	seconds	after	hitting	send	
9	8	APP		holy	shit	it's	here	http://www.mandrill.com/							
10	9	APP		our	new	subscriber	profile	page	activity	timeline	aggregate	engageme	stats	and	mandrill
11	10	APP		@mandrill	increases	scalability	(	http://bit.l	)	then	decreases	pricing	(	http://bit.l	)
12	11	OTHER		the	beets	rt	@missmya	#nameana	mandrill						
13	12	OTHER		rt	@luissand	fernando	vargas	mandrill	mexican	pride	mma				
14	13	OTHER		photo	oculi-ds	mandrill	by	natalie	manuel	http://tmblr.co/zjqanxhdswlr					
15	14	OTHER		@mandrill	me	neither	we	can	be	:sadpanda	together	:(			
16	15	OTHER		@mandrill	n	/	(	k	*	(	n	-	k	)	)
17	16	OTHER		megaman	x	-	spark	mandrill	acapella	http://you	@youtube	さんから			
18	17	OTHER		@angeluse	storm	eagle	ftw	nomás	no	dejes	que	se	le	acerque	spark
19	18	OTHER		gostei	de	um	vídeo	@youtube	http://you	aspark	...	mandrill's	stage	on	guitar
20	19	OTHER		what	is	2-year-old	mandrill	jj	thinking	in	this	pic	http://ow.l	re-tweet	with

테스트데이터

- APP에 대한 조건부 확률 계산하기
  - A25~44에 1~20까지 number 작성
  - D25에 D2단어에 대한 확률 계산하기
  - D25=VLOOKUP(D2,AppTokensProbability!$A$4:$E$826,5,FALSE)

  - 학습데이터에 없던 단어 처리하기
  - =IF(ISNA(VLOOKUP(D2,AppTokensProbability!$A$4:$E$826,5,FALSE)),LN(1/AppTokensProbability!$C$827),VLOOKUP(D2,AppTokensProbability!$A$4: $E$826,5,FALSE))
    만일 VLOOKUP으로 찾은 확률값이 에러를 리턴한다면…

  - 3글자 이하의 단어 제외하기
  - =IF(LEN(D2)<=3,0,IF(ISNA(VLOOKUP(D2,AppTokensProbability!$A$4:$E$826,5,FALSE)),LN(1/AppTokensProbability!$C$827),VLOOKUP(D2,AppTokensProbability!$A$4:$E$826,5,FALSE)))
    만일 3글자 이하의 단어라면…

테스트데이터

- APP에 대한 조건부 확률 계산하기
  - 첫번째 테스트 데이터의 토큰의 확률 계산: D25를 AI25까지 복사
  - 20개 데이터 토큰의 확률 계산:AI44까지 복사

D25 =IF(LEN(D2)<=3,0,IF(ISNA(VLOOKUP(D2,AppTokensProbability!$A$4:$E$826,5,FALSE)),LN(1/AppTokensProbability!$C$827),VLOOKUP(D2,AppTokensProbability!$A$4:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
22																	
23																	
24																	
25	1			-5.30248	-6.68877	-5.14832	-5.30248	-4.49155	-5.30248	0	-5.84147	-7.09423	-6.17794	0	-7.09423	-7.09423	0
26	2			-7.09423	-5.30248	-5.38949	0	-4.95417	0	-4.65189	-4.17646	-5.59016	-5.14832	0	0	-6.68877	-6.40109
27	3			-7.09423	0	-7.09423	-7.09423	0	-7.09423	-3.23351	-6.17794	0	-7.09423	0	0	0	0
28	4			-6.68877	-5.99562	-7.09423	0	-7.09423	-5.70794	0	-6.68877	0	-6.17794	-7.09423	-3.23351	0	-7.09423
29	5			-5.99562	-5.99562	0	-5.14832	-5.38949	0	-7.09423	-7.09423	-7.09423	-7.09423	-7.09423	0	-5.38949	-6.68877
30	6			-5.59016	-7.09423	-5.4848	-5.30248	-3.23351	0	-5.99562	-7.09423	-4.49155	-7.09423	0	0	-7.09423	0
31	7			-6.68877	-7.09423	0	0	-5.59016	-5.4848	0	-7.09423	-6.68877	-7.09423	-5.14832	0	0	0
32	8			-7.09423	-7.09423	-7.09423	-6.17794	-6.68877	0	0	0	0	0	0	0	0	0
33	9			0	0	-7.09423	-7.09423	-6.17794	-7.09423	-7.09423	-7.09423	-7.09423	-7.09423	0	-3.23351	-7.09423	-7.09423
34	10			-5.14832	-7.09423	-7.09423	0	-7.09423	0	-7.09423	-7.09423	-5.99562	0	-7.09423	0	-7.09423	0
35	11			0	-7.78738	0	-7.78738	-7.78738	-3.23351	0	0	0	0	0	0	0	0
36	12			0	-7.78738	-7.78738	-7.78738	-3.23351	-7.78738	-7.78738	0	0	0	0	0	0	0
37	13			-7.09423	-7.78738	-3.23351	0	-7.78738	-7.78738	-7.78738	0	0	0	0	0	0	0
38	14			-6.68877	0	-7.78738	0	0	0	-7.78738	-7.78738	0	0	0	0	0	0
39	15			-6.68877	0	0	0	0	0	0	0	0	0	0	0	-6.40109	0
40	16			-7.78738	0	0	-7.78738	-3.23351	-7.78738	-7.78738	-7.78738	0	0	0	0	0	0
41	17			-7.78738	-7.78738	-7.78738	0	-7.78738	0	-7.78738	0	0	0	-7.78738	-7.78738	-3.23351	0
42	18			-7.78738	0	0	-7.78738	-7.78738	-7.78738	-7.78738	0	-7.78738	-7.09423	0	-7.78738	-7.78738	0
43	19			-5.38949	0	-7.78738	-3.23351	0	-7.09423	0	-5.4848	0	-7.78738	-7.78738	-4.17646	-5.30248	-7.78738
44	20			0	-7.78738	0	-7.78738	0	0	-3.23351	0	-7.78738	-7.78738	-7.78738	-7.78738	-7.78738	-5.59016

테스트데이터

- APP에 대한 조건부 확률 계산하기
  - C24에 Sum of conditional probabilities 작성
  - 각 토큰에 대한 LN(P)값을 더하기
  - C25=SUM(D25:AI25)

C25 =SUM(D25:AI25)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
22																
23																
24			Sum of conditional probabilities													
25	1		-65.5382	-5.30248	-6.68877	-5.14832	-5.30248	-4.49155	-5.30248	0	-5.84147	-7.09423	-6.17794	0	-7.09423	-7.09423
26	2		-74.4825	-7.09423	-5.30248	-5.38949	0	-4.95417	0	-4.65189	-4.17646	-5.59016	-5.14832	0	0	-6.68877
27	3		-44.8826	-7.09423	0	-7.09423	-7.09423	0	-7.09423	-3.23351	-6.17794	0	-7.09423	0	0	0
28	4		-109.772	-6.68877	-5.99562	-7.09423	0	-7.09423	-5.70794	0	-6.68877	0	-6.17794	-7.09423	-3.23351	0
29	5		-82.7629	-5.99562	-5.99562	0	-5.14832	-5.38949	0	-7.09423	-7.09423	-7.09423	-7.09423	-7.09423	0	-5.38949
30	6		-58.475	-5.59016	-7.09423	-5.4848	-5.30248	-3.23351	0	-5.99562	-7.09423	-4.49155	-7.09423	0	0	-7.09423
31	7		-50.8835	-6.68877	-7.09423	0	0	-5.59016	-5.4848	0	-7.09423	-6.68877	-7.09423	-5.14832	0	0
32	8		-34.1494	-7.09423	-7.09423	-7.09423	-6.17794	-6.68877	0	0	0	0	0	0	0	0
33	9		-80.3538	0	0	-7.09423	-7.09423	-6.17794	-7.09423	-7.09423	-7.09423	-7.09423	-7.09423	0	-3.23351	-7.09423
34	10		-60.8036	-5.14832	-7.09423	-7.09423	0	-7.09423	0	-7.09423	-7.09423	-5.99562	0	-7.09423	0	-7.09423
35	11		-26.5957	0	-7.78738	0	-7.78738	-7.78738	-3.23351	0	0	0	0	0	0	0
36	12		-42.1704	0	-7.78738	-7.78738	-7.78738	-3.23351	-7.78738	-7.78738	0	0	0	0	0	0
37	13		-41.4773	-7.09423	-7.78738	-3.23351	0	-7.78738	-7.78738	-7.78738	0	0	0	0	0	0
38	14		-30.0509	-6.68877	0	-7.78738	0	0	0	-7.78738	-7.78738	0	0	0	0	0
39	15		-48.4432	-6.68877	0	0	0	0	0	0	0	0	0	0	0	-6.40109
40	16		-42.1704	-7.78738	0	0	-7.78738	-3.23351	-7.78738	-7.78738	-7.78738	0	0	0	0	0
41	17		-57.7452	-7.78738	-7.78738	-7.78738	0	-7.78738	0	-7.78738	0	0	0	-7.78738	-7.78738	-3.23351
42	18		-69.3933	-7.78738	0	0	-7.78738	-7.78738	-7.78738	-7.78738	0	-7.78738	-7.09423	0	-7.78738	-7.78738
43	19		-61.8305	-5.38949	0	-7.78738	-3.23351	0	-7.09423	0	-5.4848	0	-7.78738	-7.78738	-4.17646	-5.30248
44	20		-102.272	0	-7.78738	0	-7.78738	0	0	-3.23351	0	-7.78738	-7.78738	-7.78738	-7.78738	-7.78738

테스트데이터

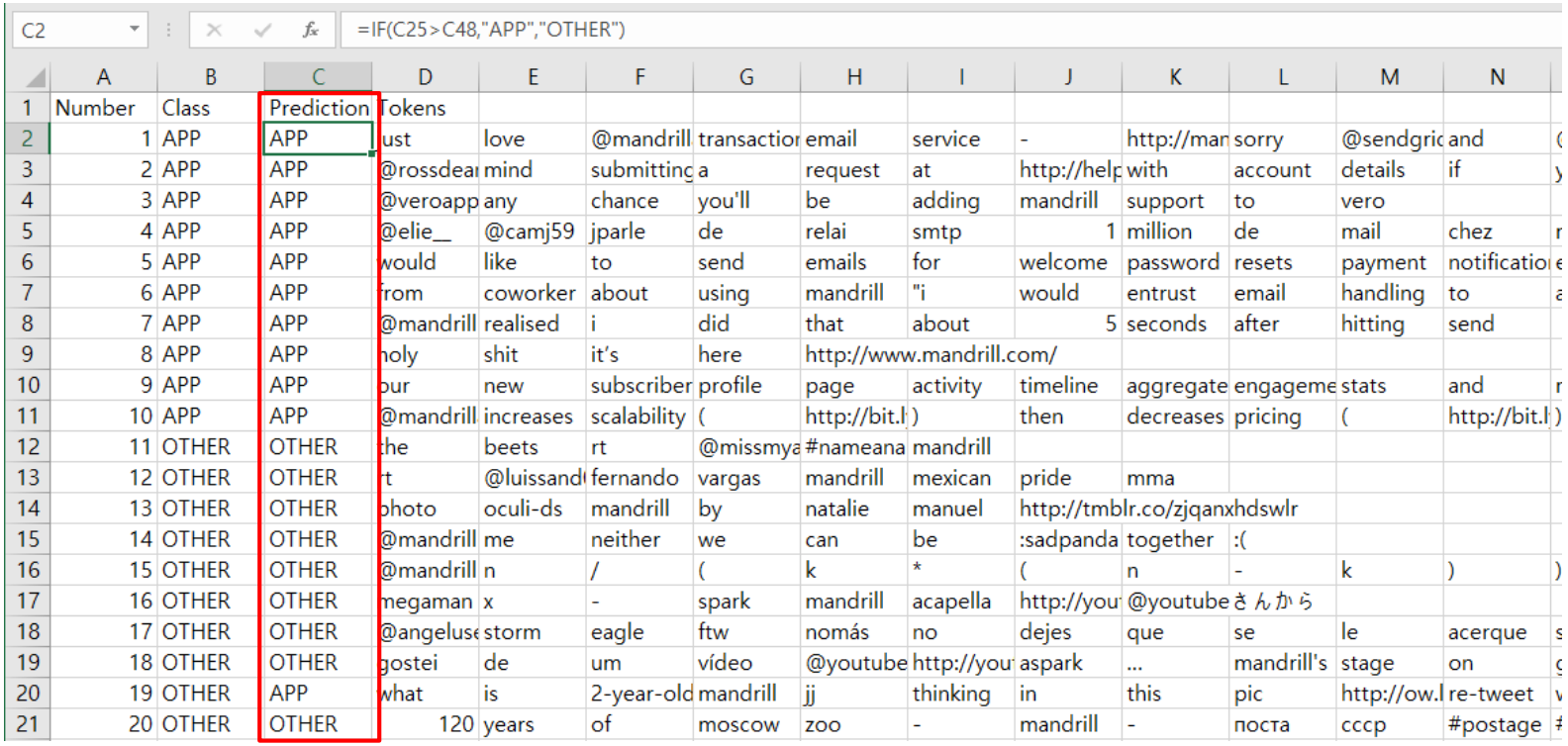
- Other에 대한 조건부 확률 계산하기
  - D48=IF(LEN(D2)<=3,0,IF(ISNA(VLOOKUP(D2,OtherTokensProbability!$A$4:$E$809,5,FALSE)),LN(1/OtherTokensProbability!$C$810),VLOOKUP(D2,OtherTokensProbability!$A$4:$E$809,5,FALSE)))

C48 =SUM(D48:AI48)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
47			Sum of conditional probabilities													
48	1		-80.0198	-5.6684	-6.5157	-7.61431	-7.61431	-7.61431	-6.92116	0	-7.61431	-7.61431	-7.61431	0	-7.61431	-7.61431
49	2		-93.7286	-7.61431	-6.92116	-7.61431	0	-7.61431	0	-7.61431	-5.12941	-6.92116	-7.61431	0	0	-7.61431
50	3		-48.0962	-7.61431	0	-7.61431	-7.61431	0	-7.61431	-3.10345	-6.92116	0	-7.61431	0	0	0
51	4		-124.239	-7.61431	-7.61431	-7.61431	0	-7.61431	-7.61431	0	-7.61431	0	-7.61431	-7.61431	-3.10345	0
52	5		-93.7286	-6.22802	-5.82255	0	-6.92116	-7.61431	0	-7.61431	-7.61431	-7.61431	-7.61431	-7.61431	0	-6.92116
53	6		-66.8448	-5.82255	-7.61431	-6.00487	-7.61431	-3.10345	0	-6.22802	-7.61431	-7.61431	-7.61431	0	0	-7.61431
54	7		-52.1017	-3.87664	-7.61431	0	0	-5.53487	-6.00487	0	-7.61431	-6.92116	-7.61431	-6.92116	0	0
55	8		-34.488	-6.92116	-6.5157	-7.61431	-5.82255	-7.61431	0	0	0	0	0	0	0	0
56	9		-86.1677	0	0	-7.61431	-7.61431	-6.92116	-7.61431	-7.61431	-7.61431	-7.61431	-7.61431	0	-3.10345	-7.61431
57	10		-68.5288	-7.61431	-7.61431	-7.61431	0	-7.61431	0	-7.61431	-7.61431	-7.61431	0	-7.61431	0	-7.61431
58	11		-23.4615	0	-7.61431	0	-6.5157	-6.22802	-3.10345	0	0	0	0	0	0	0
59	12		-38.4024	0	-7.61431	-6.92116	-6.92116	-3.10345	-6.92116	-6.92116	0	0	0	0	0	0
60	13		-41.175	-7.61431	-7.61431	-3.10345	0	-7.61431	-7.61431	-7.61431	0	0	0	0	0	0
61	14		-26.7196	-3.87664	0	-7.61431	0	0	0	-7.61431	-7.61431	0	0	0	0	0
62	15		-44.5927	-3.87664	0	0	0	0	0	0	0	0	0	0	0	-6.5157
63	16		-28.7378	-4.84172	0	0	-4.39544	-3.10345	-4.90626	-4.97525	-6.5157	0	0	0	0	0
64	17		-53.1848	-7.61431	-7.61431	-7.61431	0	-7.61431	0	-7.61431	0	0	0	-7.61431	-4.39544	-3.10345
65	18		-60.8146	-5.53487	0	0	-5.31173	-4.97525	-7.61431	-7.61431	0	-6.92116	-7.61431	0	-7.61431	-7.61431
66	19		-64.4469	-6.92116	0	-7.61431	-3.10345	0	-7.61431	0	-5.21642	0	-7.61431	-7.61431	-5.12941	-6.00487
67	20		-99.6046	0	-6.92116	0	-7.61431	0	0	-3.10345	0	-7.61431	-7.61431	-7.61431	-7.61431	-7.61431

테스트데이터

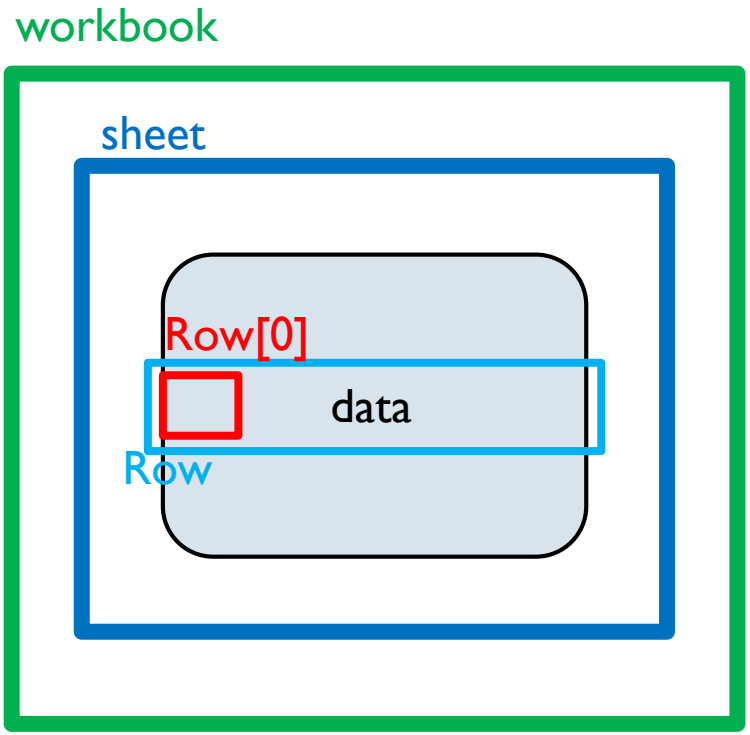
- Class 예측하기
  - C2~C21에 예측결과 표시하기
  - C2=IF(C25>C48,"APP","OTHER")



C2 =IF(C25>C48,"APP","OTHER")

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Number	Class	Prediction	Tokens										
2	1	APP	APP	ust	love	@mandrill	transactior	email	service	-	http://mar	sorry	@sendgric	and
3	2	APP	APP	@rossdear	mind	submitting	a	request	at	http://help	with	account	details	if
4	3	APP	APP	@veroapp	any	chance	you'll	be	adding	mandrill	support	to	vero	
5	4	APP	APP	@elie__	@camj59	jparle	de	relai	smtp	1	million	de	mail	chez
6	5	APP	APP	would	like	to	send	emails	for	welcome	password	resets	payment	notificatio
7	6	APP	APP	from	coworker	about	using	mandrill	"i	would	entrust	email	handling	to
8	7	APP	APP	@mandrill	realised	i	did	that	about	5	seconds	after	hitting	send
9	8	APP	APP	holy	shit	it's	here	http://www.mandrill.com/						
10	9	APP	APP	our	new	subscriber	profile	page	activity	timeline	aggregate	engageme	stats	and
11	10	APP	APP	@mandrill	increases	scalability	(	http://bit.l	)	then	decreases	pricing	(	http://bit.l
12	11	OTHER	OTHER	the	beets	rt	@missmya	#nameana	mandrill					
13	12	OTHER	OTHER	rt	@luissand	fernando	vargas	mandrill	mexican	pride	mma			
14	13	OTHER	OTHER	photo	oculi-ds	mandrill	by	natalie	manuel	http://tmblr.co/zjqanxhdswlr				
15	14	OTHER	OTHER	@mandrill	me	neither	we	can	be	:sadpanda	together	:(		
16	15	OTHER	OTHER	@mandrill	n	/	(	k	*	(	n	-	k	)
17	16	OTHER	OTHER	megaman	x	-	spark	mandrill	acapella	http://you	@youtube	さんから		
18	17	OTHER	OTHER	@angeluse	storm	eagle	ftw	nomás	no	dejes	que	se	le	acerque
19	18	OTHER	OTHER	gostei	de	um	vídeo	@youtube	http://you	aspark	...	mandrill's	stage	on
20	19	OTHER	APP	what	is	2-year-old	mandrill	jj	thinking	in	this	pic	http://ow.l	re-tweet
21	20	OTHER	OTHER	120	years	of	moscow	zoo	-	mandrill	-	носта	cccp	#postage

엑셀에서 데이터 가져오기




```
book=xlrd.open_workbook('E:/Datasmart/ch03/Mandrill.xlsx')
sheet=book.sheet_by_index(0)
data=[]
for i in range(sheet.nrows) :
    data.append(sheet.row_values(i)[0])
```

참고하면 좋은 함수들

- String.lower() : String의 모든 대문자를 소문자로 바꿈
- String.replace("A","B") : A를 B로 교체함
- String.split(' ') : 빈칸으로 단어를 구분함 (,.;:'"!@|)
- Apply(lambda x: fuction(x),axis=1) :
  x는 DataFrame 의 each row
- app_prob=[[] for cols in range(len(test_data_split))] :
  2차원 List 선언 (초기화)
- Series.isin('A') : A가 있으면 True 를 반환

파이썬 나이브 베이즈 흐름도