

Text Analytics

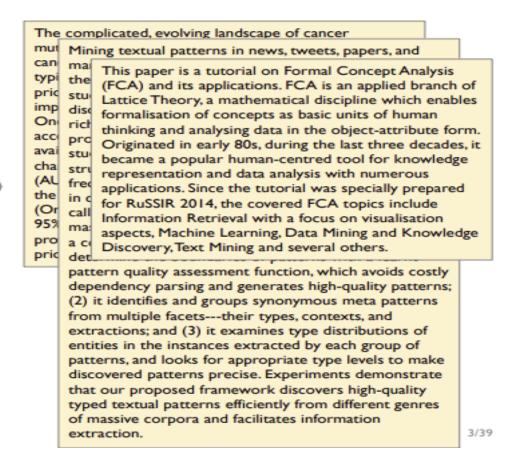
Ch4





- 텍스트 데이터를 모으면, 대부분 문서의 길이가 같지 않다.
 - 고전적 머신러닝은 기본적으로 x값인 문서의 차원수가 같아야 한다.
 - ✓ 세개의 문서 다 길이가 다름으로, 고정적 길이로 변환 해주는 과정이 필요하다.







avai



The complicated, evolving landscape of cancer

mul Mining textual patterns in news, tweets, papers, and

typi the This paper is a tutorial on Formal Concept Analysis the (FCA) and its applications. FCA is an applied branch (FCA) and its applications. FCA is an applied branch of pric stu Lattice Theory, a mathematical discipline which enables formalisation of concepts as basic units of human thinking and analysing data in the object-attribute form. Originated in early 80s, during the last three decades, it became a popular human-centred tool for knowledge representation and data analysis with numerous applications. Since the tutorial was specially prepared the in c for RuSSIR 2014, the covered FCA topics include Information Retrieval with a focus on visualisation aspects, Machine Learning, Data Mining and Knowledge pro a c Discovery, Text Mining and several others.

pattern quality assessment function, which avoids costly dependency parsing and generates high-quality patterns; (2) it identifies and groups synonymous meta patterns from multiple facets---their types, contexts, and extractions; and (3) it examines type distributions of entities in the instances extracted by each group of patterns, and looks for appropriate type levels to make discovered patterns precise. Experiments demonstrate that our proposed framework discovers high-quality typed textual patterns efficiently from different genres of massive corpora and facilitates information extraction.

the complic evolv landscap of cancer mutat pose a fori mine textual pattern in news tweet paper and mani list oth this paper is a tutori on formal concept analysi fca and it applic fca is an appli branch of lattic theori a mathemat disciplin which enabl formalis of concept as basic unit of human think and analys data in the base on objectattribut form origin in earli s dure the last three curv and decad it becam a popular humancentr tool for curat da knowledg represent and data analysi with numer applic oncosco sinc the tutori was special prepar for russir the cover oncosco fca topic includ inform retriev with a focus on visualis priorit d aspect machin learn data mine and knowledg discoveri text mine and sever other

	Var I	Var 2	 	Var P
Doc I				
Doc 2				
Doc 3				
Doc D				

4/39



Stop Words



A total of 571 stop words

- 275	-	Table 1	2002000	2010000				0.2000.002.0	2010/02/2020	Charles and	444
[1]	"a"	"a's"	"able"	"about"	"above"	"according"	"accordingly"	"across"	"actually"	"after"	"afterwards"
[12]	"again"	"against"	ain't"	"all"	"allow"	"allows"	"almost"	"allone"	"along"	"already"	"also"
[23]	"although"	"always"	an'	"among"	"amongst"	"an"	"and"	"another"	"arry"	"anybody"	"anyhou"
[34]	"arryone"	"anything"	"arryway"	"anyways"	"anywher e"	"apart"	"appear"	"appreciate"	"appropriate"	are"	"aren't"
[45]	"around"	"25"	"as1de"	"ask"	"asking"	"associated"	"at"	"available"	"away"	"awfully"	"b"
[56]	"be"	"becase"	"because"	"becone"	"becomes"	"becoming"	"been"	"before"	"beforehand"	"behtind"	"being"
[67]	"believe"	"below"	"beside"	"bestdes"	"best"	"better"	"between"	"beyond"	"both"	"brief"	"but"
[78]	"by"	"c"	"C mon"	"E'S"	"Came"	"can"	can't"	"cannot"	"Carit"	"CHUSE"	"Causes"
[89]	"certain"	"certainly"	"changes"	"clearly"	"co"	"COM"	"cone"	"comes"	"concerning"	"consequently"	"consider"
[100]	"considering"	"contain"	"containing"	"contains"	"corresponding"	"could"	"cowlidn't"	"course"	"currently"	"d"	"definitely"
[111]	"described"	"despite"	"did"	"didn't"	"different"	"do"	"does"	"doesn't"	"doing"	"don't"	"done"
[122]	"down"	"downwards"	"dur ing"	· e"	"each"	"edu"	"40"	"elight"	"either"	"else"	"elsewhere"
[133]	"enough"	"entirely"	"especially"	"et"	"etc"	"even"	ever	"every"	"everybody"	"everyone"	"everything"
[144]	"everywhere"	"ex"	"exactly"	"example"	"except"	***	"far"	"few"	"fifth"	"first"	"five"
[155]	"followed"	"fallowing"	"follows"	"for"	"former"	"formerly"	"forth"	"four"	"from"	"further"	"furthersore"
[166]	"q"	"get"	"gets"	"getting"	"given"	"gives"	"00"	"goes"	"going"	"gone"	"005"
[177]	"gotten"	"greetings"	Th*	"had"	"hadn't"	"happens"	"hardly"	"has"	"hasn"t"	"have"	"haven't"
[188]	"having"	"he"	"he's"	"hello"	"help"	"hence"	"her"	"here"	"here's"	"hereafter"	"hereby"
[199]	"herein"	"hereupon"	"hers"	"herself"	"htt	"him"	"himself"	"his"	"hither"	"hapefully"	"how"
[210]	"howbelt"	"however"	44.4	4.9.	"1'11"	"f"m"	"5've"	"ie"	1177	"ignored"	"inmediate"
£2233	"in"	"Inasmuch"	"Inc"	"Indeed"	"indicate"	"indicated"	"indicates"	"inner"	"insofar"	"Instead"	"into"
(232)	"trawerd"	"is"	"tan"t"	"it"	"it'd"	"it']]"	"15" 5"	"its"	"itself"	-4	"just"
[243]	"k"	"keep"	"keeps"	"kept"	"kmow"	"knaws"	"known"	man.	"last"	"lately"	"later"
[254]	"latter"	"latterly"	"least"	"less"	"lest"	"let"	"let's"	"15ke"	"11ked"	"Hikely"	"little"
[265]	"look"	"Tooking"	"Tooks"	"Itd"	"8"	"mainly"	"many"	"may"	"maybe"	me"	"mean"
276)	"nearwhile"	"merely"	"might"	"more"	"moreover"	"most"	"mostly"	"much"	"must"	"my"	"mynelf"
[287]	"0"	"name"	"name ty"	"nd"	"near"	"near Ty"	"necessary"	"need"	"needs"	"neither"	"never"
[298]	"nevertheless"	"new"	"next"	"mine"	"no"	"nobody"	"non"	"none"	"noone"	"mor"	"normally"
[906]	"not"	"nothing"	"nove1"	"mow"	"nowhere"	"0"	"obviously"	"of"	"off"	"often"	"oh"
[120]	"ak"	"okay"	"old"	"on"	"once"	"one"	"gnea"	"only"	"anta"	"or"	"other"
[331]	"others"	"otherwise"	"ought"	"our"	"ours"	"ourselves"	"out"	"outside"	"gyer"	"overall"	"own"
13421	"0"	"particular"	"particularly"	"per"	"perhaps"	"placed"	"please"	"plus"	"possible"	"presumably"	"probably"
[353]	"provides"	"9"	"mue"	"quite"	"gy"	***	"rather"	"rd"	're'	"really"	"reasonably"
[164]	"regarding"	"regardless"	"regards"	"relatively"	"respectively"	"right"	190	"said"	"Same"	5.00	"say"
13757	"saying"	"says"	"second"	"secondly"	"598"	"seeing"	"seem"	"seemed"	"seewing"	"sees"	"seen"
[386]	"self"	"selves"	"sensible"	"sent"	"serfous"	"seriously"	"seven"	"several"	"sha11"	"she"	"should"
1971	"shouldn't"	"since"	"six"	"50"	"some"	"somebody"	"somehow"	"someone"	"nomething"	"agnetime"	"sometimes"
[406]	"somewhat"	"s grewhere"	"soon"	"sorry"	"specified"	"specify"	"specifying"	"agill"	"sub"	"auch"	"sup"
4197	"sure"	in E in	"E "S"	"Take"	"taken"	"tell"	"Tends"	"TB"	"than"	"thank"	"thanks"
4303	"thanx"	"that"	"that 's"	"thats"	"the"	"their"	"theirs"	"then"	"themselves"	"then"	"thence"
441	"there"	"there's"	"thereafter"	"thereby"	"therefore"	"therein"	"theres"	"thereupon"	"these"	"they"	"they d"
[452]	"they"11"	"they're"	"they ve"	"think"	"third"	"this"	"thorough"	"thoroughly"	"those"	"though"	"three"
[463]	"through"	"throughout"	"thru"	"thus"	"to"	"together"	"T00"	"took"	"toward"	"towards"	"tried"
4747	"tries"	"truly"	"try"	"trying"	"twice"	"two"	"u"	"em"	"under"	"unfortunately"	"unless"
485	"unlikely"	"until"	"unto"	"up"	"upon"	"un"	"une"	"imed"	"useful"	"unen"	"using"
[496]	"usually"	"eucp"	"V"	"value"	"var fous"	"Very"	Tyle"	"viz"	"VS"	"W"	"want"
[507]	"wants"	"was"	"wasn't"	"way"	"we"	"we. q.	"we"11"	"we're"	"we" ve"	"welcome"	"we11"
[518]	"West"	"were"	weren't"	"what"	"what 's"	"whatever"	"when"	"whence"	"whenever"	"where"	"where 's"
(529)	"whereafter"	"wherean"	"whereby"	"wherein"	"wher eupon"	"wherever"	"whether"	"which"	while"	"whither"	'who'
5407	"who 's"	"whoever"	"whoTe"	"whom"	"whose"	"why"	"will"	"willing"	"wish"	"with"	"within"
5517	"without"	"won"t"	"wonder"	"would"	"would"	"wouldn't"	***	-V-	"ves"	"yet"	"you"
[562]	"you'd"	"you" 11"	"you're"	"You've"	"your"	"yours"	"yourself"	"yourselves"	75	"zero"	3-00
frant.	300 0	300 11	300 16	you re	- Jones	your o	Jon Sell	Jam Serves		20.0	

A total of 677 stop words

40	08895	801909	88E+04U2+	다시 말하자면	까닭으로	할 생각이다	즈통하여	202		년	호자
			면이아니다			43(E)(E)		7.1		너희	자기
	OHICEO F		만든 아니다			10181510	다른 발명으로	DI	Ch2:	일신	자기집
			막돈하고					이쪽	6	(동)	자신
						그렇게 짧으		0171	23	6001	우레 졸합한것고
04	286E		그치지 않다							자라리	201
				Alam Alam						함지연절	흥적으로 보면
				이상						할지라도	총적으로 말하면
			하지만	ti ti		9313				활망경	송적으로
			E210I				하는것만 못하다				08.80
	8	첫									으로서
	ECICI		논하지 않다							MTC.	8
			마지지 않다							28C	DEMOCH.
							매선			에쓰겁다	
				배도온다		교사					할때등이다
				OEC129		2670				입사람	8
									대해 말하자		99
으문	500 H									첫	20.00
크	DIOLA!	@OF	하는 편이 낫							의거하여	88
0020	31005)	Q4240H	Ch							근거하며	55
문이다	위때라	하는것도	문문하고	可						의해	영간
	위이어	29totci	\$78104	필정	여부	\$24.00E	머느록		반대로 말하		401010
근거하여	살궁	어필수 있다	\$766.AI	풍안	하기보다는	일에 통립없	어느것				OHLI
		88.1			하느니	Cf	아느해		이외 반대로		2101
	2000	8/	500	한고인인다	하면 할수록	원다면	이노 년도		바꾸어서 말		0
	동하여		高E)	0(2)(1)	88	0	간해도	와 같은 사람들	하면	버금	0101
	XIOEXE		이용하여	OLAI	0(2)0(2)6(C)	55	언젠아	부튜의 사장들	바꾸어서 한		합나
예를 듣지면								왜니라면		기타	년
				70(7)				見りおし	Elsi	첫번자로 -	문
						EtAl	301	요취	그렇지않으	나머지는	일
						CES	对英	오르지	[29]	그용에서	9
									20605	견지에서	엄
	2		2H010E					81기만 81면	-	활식으로 쓰며	92
	e.u			같이				도착하다	10	일절에서	01
	世里		한다면 불라			DHEAL			빠걸거리다	위해서	di di
	92	817(8)		DEM					보도목	단지	사
		마물건						절도에 이르다	비결거리다	의해되다	오
										하도록시키다	8
										EE/OH ISS	8
그리고 비길수 없다										반대로	6
해서는 안된		0121210								건호	8
	22H		마리사						2)	건자	5
CH											

•••



Bag of words

- 가변길이의 문서를 고정길이의 벡터로 변환하는 방식을 의미한다.
 - ✓ 문장을 숫자로 표현하는 방식 중 하나라는 의미이다.
 - ✓ 같은 차원의 벡터로 변환 하는 것을 의미 한다. (이를 document representation)이라고 한다.
- 단어들의 순서는 전혀 고려하지 않고 단어의 출현 빈도에만 집중하는 방식이다.(count 기반)





Bag of words

- 가변길이의 문서를 고정길이의 벡터로 변환하는 방식을 의미한다.
 - ✓ 문장을 숫자로 표현하는 방식 중 하나라는 의미이다
 - ✓ 같은 차원의 벡터로 변환 하는 것을 의미 한다. (이를 document representation)이라고 한다.
- 단어들의 순서는 전혀 고려하지 않고 단어의 출현 빈도에만 집중하는 방식이다.

txt1:Idon't love dog. I love cat not dog.

txt2: I love dog. dog love me too.

	I	love	dog	don't	cat	me	too	not
txt1	2	2	2	1	1	0	0	1
txt2	1	2	2	0	0	1	1	0





Bag of words

- 가변길이의 문서를 고정길이의 벡터로 변환하는 방식을 의미한다.
 - ✓ 문장을 숫자로 표현하는 방식 중 하나라는 의미이다
 - ✓ 같은 차원의 벡터로 변환 하는 것을 의미 한다. (이를 document representation)이라고 한다.
- 단어들의 순서는 전혀 고려하지 않고 단어의 출현 빈도에만 집중하는 방식이다.

```
Ex:
     five_random_documents = [
                      sentences
      'i like this movie',
      'the movie hunger games is a trilogy movie',
documents
      'jennifer lawrence is an excellent actor',
      'i would give the film an 8 out of 10',
      'you can observe some jaw-dropping cleverness'
     bag of words = [
                                   words
      documents
```





- Bag of words: Term Document Matrix
 - 두가지 표현 방식이 존재 한다.
 - Binary representation : 문서에 해당 단어가 등장 했는지 안 했는지 유무만 판별, 등장하면 1 아니면 0 으로 표현한다.
 - Frequency representation : 등장빈도를 기록한다.

S1: John likes to watch movies. Mary likes too.

S2: John also likes to watch football game.

Binary	v re	prese	enta	tion

Binary	rep	rese	ntati	ion
--------	-----	------	-------	-----

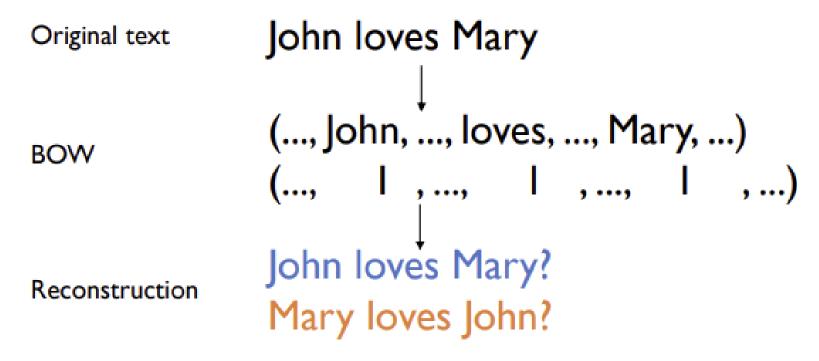
billary representa		
Word	51	S 2
John	1	1
Likes	1	1
То	1	1
Watch	1	1
Movies	1	О
Also	0	1
Football	0	1
Games	0	1
Mary	1	О
too	1	О

Frequency representation

Word	S1	S ₂
John	1	1
Likes	2	1
То	1	1
Watch	1	1
Movies	1	О
Also	О	1
Football	О	1
Games	О	1
Mary	1	О
too	1	0



- ▶ Bag of words : 한계점
 - 벡터 표현은 문서에서 단어의 순서를 고려하지 않는다.
 - ✓ 수원이는 커피를 마시고, 지훈이는 라떼를 마신다. = 지훈이는 커피를 마시고, 수원이는 라떼를 마신다.
 - 벡터로 표현하게 되면 다시 원 문장으로 돌릴 수 없다.





Word weighting

Word Weighting: Term-Frequency



- Word Weighting : Term-Frequency
 - 어떤 문서에서 어떤 단어가 중요한지 파악하기 위함이다.
 - 문서와 단어의 출현 빈도수를 확인한다.

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	o	0	0
Brutus	4	157	0	1	o	0
Caesar	232	227	o	2	1	1
Calpurnia	o	10	o	o	0	0
Cleopatra	57	0	o	o	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0



Word Weighting : Document Frequency



- Word Weighting : Document Frequency
 - 어떤 코퍼스에서 어떤 단어가 중요한지 파악하기 위함이다.
 - 특정 단어가 특정 문서에 몇 번 등장했는지 보다, 등장한 문서의 수 에만 관심을 갖는다.
 - 많이 출현할수록 덜 중요한 단어일 확률이 높다.
 - √ the, is,...

tf,,=)

Word Weighting: Inverse Document Frequency



Word Weighting: Inverse Document - Frequency

$$\checkmark idf_t = log_{10}(N/df_t)$$

- N은 코퍼스 문서의 개수를 의미한다.
- Log를 취하는 이유는 값이 기하급수적으로 증가하는 경향이 있기 때문이다.
- 밑 표는 단어가 100만개가 들어있는 코퍼스의 기준으로 예시를 들은 것이다.

term	df_t	idf _t
calpurnia	1	6
animal	100	4
sunday	1,000	3
fly	10,000	2
under	100,000	1
the	1,000,000	0



Word Weighting: TF – IDF



- Word Weighting: TF IDF
 - TF IDF 값이 낮으면 중요도가 낮고, 크면 중요도가 크다.

$$TF - IDF(w) = \underbrace{tf(w)} \times \log\left(\frac{N}{df(w)}\right)$$



Word Weighting: TF – IDF



- Word Weighting : TF IDF
 - TF IDF 값이 낮으면 중요도가 낮고, 크면 중요도가 크다.
 - ✓ 너무 고차원이 된다는 단점이 있다.
 - V-dimensional vector space(v= vocabulary)
 - ❖ 해당 문서가 쓰인 언어에서의 총 단어의 수를 의미한다.
 - » 한국어의 경우 100만개의 단어가 표현된다.
 - ✓ 희소성의 문제가 발생한다.(거의 모든 셀의 값이 0이 된다.)

.



Word Weighting: TF – IDF



- Word Weighting : TF IDF
 - TF IDF 값이 낮으면 중요도가 낮고, 크면 중요도가 크다.
 - ✓ 너무 고차원이 된다는 단점이 있다.

	Docl	Doc2	Doc3
TermI	5	0	0
Term2	1	0	0
Term3	5	5	5
Term4	3	3	3
Term5	3	0	1



Docl	TF	DF	IDF	TF-IDF
TermI	5	- 1	Log3	5log3
Term2	1	1	Log3	Hog3
Term3	5	3	LogI	0
Term4	3	3	LogI	0
Term5	3	2	Log(3/2)	3log(3/2)

Word weighting: Term I > Term 5 > Term 2 > Term 3 = Term 4



Word Weighting: TF Variants



- Word Weighting : TF Variants
 - TF값의 변형을 할 수 있다.
 - ✓ 한 논문에만 6가지가 넘고, d나 k의 값에 따라서도 다르다.

Definition 2.1 TF Variants: TF(t, d). TF(t, d) is a quantification of the within-document term frequency, tf_d . The main variants are:



Word Weighting : DF Variants



- Word Weighting : DF Variants
 - DF가 0 일 경우 문제가 발생하는 경우가 있다.
 - ✓ 스무딩을 추가한다.

Definition 2.3 DF Variants. DF(t, c) is a quantification of the document frequency, df(t, c). The main variants are:

$$df(t,c) := df_{total}(t,c) := n_D(t,c)$$
(2.18)

$$\mathrm{df}_{\mathrm{sum}}(t,c) := \frac{n_D(t,c)}{N_D(c)} \qquad \left(=\frac{\mathrm{df}(t,c)}{N_D(c)}\right) \tag{2.19}$$

$$df_{\text{sum,smooth}}(t,c) := \frac{n_D(t,c) + 0.5}{N_D(c) + 1}$$
 (2.20)

$$df_{BIR}(t,c) := \frac{n_D(t,c)}{N_D(c) - n_D(t,c)}$$
 (2.21)

$$df_{BIR,smooth}(t,c) := \frac{n_D(t,c) + 0.5}{N_D(c) - n_D(t,c) + 0.5}$$
(2.22)

Definition 2.4 IDF Variants. IDF(t, c) is the negative logarithm of a DF quantification. The main variants are:

$$idf_{total}(t,c) := -\log df_{total}(t,c)$$
 (2.23)

$$idf(t,c) := idf_{sum}(t,c) := -\log df_{sum}(t,c)$$
 (2.24)

$$idf_{sum,smooth}(t,c) := -log df_{sum,smooth}(t,c)$$
 (2.25)

$$idf_{BIR}(t,c) := -\log df_{BIR}(t,c)$$
 (2.26)

$$idf_{BIR,smooth}(t,c) := -log df_{BIR,smooth}(t,c)$$
 (2.27)



Word Weighting : DF Variants



Term frequency		Document frequency		Normalization		
n (natural)	$tf_{t,d}$	n (no)	1	n (none)	1	
I (logarithm)	$1 + \log(tf_{t,d})$	t (idf)	$\log \frac{N}{df_t}$	c (cosine)	$\frac{1}{\sqrt{w_1^2+w_2^2++w_M^2}}$	
a (augmented)	$0.5 + \frac{0.5 \times tf_{t,d}}{max_t(tf_{t,d})}$	p (prob idf)	$\max\{0,\log \frac{N-\mathrm{df}_t}{\mathrm{df}_t}\}$	u (pivoted unique)	1/ <i>u</i>	
b (boolean)	$\begin{cases} 1 & \text{if } \operatorname{tf}_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$			b (byte size)	$1/\mathit{CharLength}^{lpha}, \ lpha < 1$	
L (log ave)	$\frac{1 + \log(\operatorname{tf}_{t,d})}{1 + \log(\operatorname{ave}_{t \in d}(\operatorname{tf}_{t,d}))}$					



N-grams



- N-Gram-based Language Models in NLP
 - 앞의 n-1개의 단어로 다음 단어를 예측한다.

$$P(w_n|w_{n-1},w_{n-2},...,w_1) = \frac{P(w_n,w_{n-1},w_{n-2},...,w_1)}{P(w_{n-1},w_{n-2},...,w_1)}$$

— Q) One of the hottest topics in artificial intelligence is deep _____?

- N-Gram in Text Mining
 - 토큰화가 쉽다.
 - ✓ Six sigma, supply chain management





• Bigram example

✓ Total counts in a corpus

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0





N-Gram Research

- 20카테고리의 2만개의 신문기사를 활용하여 N-Gram을 적용 했을 때의 결과 값을 나타낸 논문을 본다.
 - ✓ Gram을 늘리면 error rate가 내려가긴 하지만, 늘린 만큼의 수고만큼 올라가지 않는다.

Pruning	n-grams	Error rate	CPU secs.	No. Features
set-of-words		47.07 ± 0.92	n.a.	71,731
	1	46.18 ± 0.94	12686.12	36,534
DF: 3	2	45.28 ± 0.51	15288.32	113,716
TF: 5	3	45.05 ± 1.22	15253.27	155,184
	4	45.18 ± 1.17	14951.17	189,933
	1	45.51 ± 0.83	12948.31	22,573
DF: 5	2	45.34 ± 0.68	13280.73	44,893
TF: 10	3	46.11 ± 0.73	12995.66	53,238
	4	46.11 ± 0.72	13063.68	59,455
	1	45.88 ± 0.89	10627.10	13,805
DF: 10	2	45.53 ± 0.86	13080.32	20,295
TF: 20	3	45.58 ± 0.87	11640.18	22,214
	4	45.74 ± 0.62	11505.92	23,565

		1		
	1	48.23 ± 0.69	10676.43	n.a.
DF: 25	2	48.97 ± 1.15	8870.05	n.a.
TF: 50	3	48.69 ± 1.04	10141.25	n.a.
	4	48.36 ± 1.01	10436.58	n.a.
	5	48.36 ± 1.01	10462.65	n.a.
	1	51.54 ± 0.60	8547.43	n.a.
DF: 50	2	49.71 ± 0.53	8164.27	n.a.
TF: 100	3	51.21 ± 1.26	8079.59	n.a.
	4	$\textbf{51.21} \pm \textbf{1.26}$	8078.55	n.a.
	5	51.21 ± 1.26	8147.75	n.a.
	1	52.59 ± 0.71	6609.05	n.a.
DF: 75	2	52.83 ± 0.25	6532.80	n.a.
TF: 150	3	52.36 ± 0.48	6128.49	n.a.
	4	$\textbf{52.36} \pm \textbf{0.48}$	6128.49	n.a.
	5	$\textbf{52.36} \pm \textbf{0.48}$	6119.27	n.a.





N-Gram Research

Pruning	n-grams	Recall	Precision	F1	Accuracy	No. Features
set-of-words		76.71	83.42	79.92	99.5140	n.a.
	1	77.22	83.55	80.26	99.5211	9,673
DF: 3	2	80.34	82.03	81.18	99.5302	28,045
TF: 5	3	77.56	82.74	80.07	99.5130	38,646
	4	78.18	82.31	80.19	99.5130	45,876
	1	77.19	83.65	80.29	99.5221	6,332
DF: 5	2	80.05	82.06	81.04	99.5278	13,598
TF: 10	3	77.96	82.29	80.07	99.5106	17,708
	4	78.21	82.13	80.12	99.5106	20,468
	1	76.92	83.99	80.30	99.5241	4,068
DF: 10	2	79.06	82.04	80.52	99.5177	7,067
TF: 20	3	77.32	82.67	79.91	99.5096	8,759
	4	76.98	82.91	79.84	99.5096	9,907

