AI Author Prediction

여영웅, 김형기

DSS 14TH ML-6

머신 러닝 프로젝트 소개

소설 AI 작가 분류

- 5명의 작가들의 소설 문장 기반
- 문장 유사도 및 벡터화
- 모델 학습
- 작가 분류 및 예측
- LDA 토픽 모델링 & 시각화

INDEX

1. 데이터 소개 (EDA)
2. 유사도 측정
3. 모델링
4. LDA (Topic Modeling)
5. 결과 해석
6. 한계점 및 추후 과제

1

데이터 소개

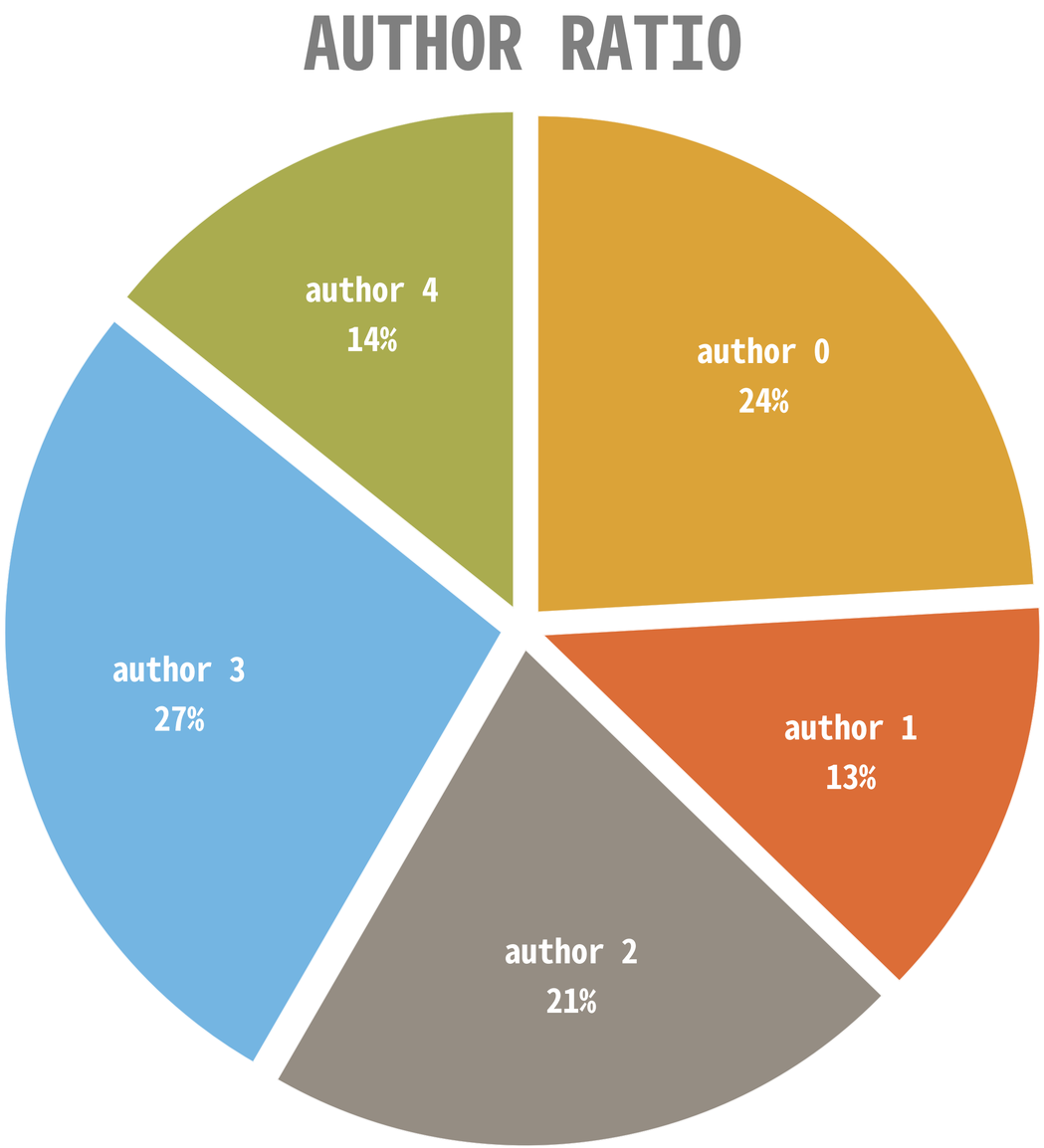
EDA

데이터 소개

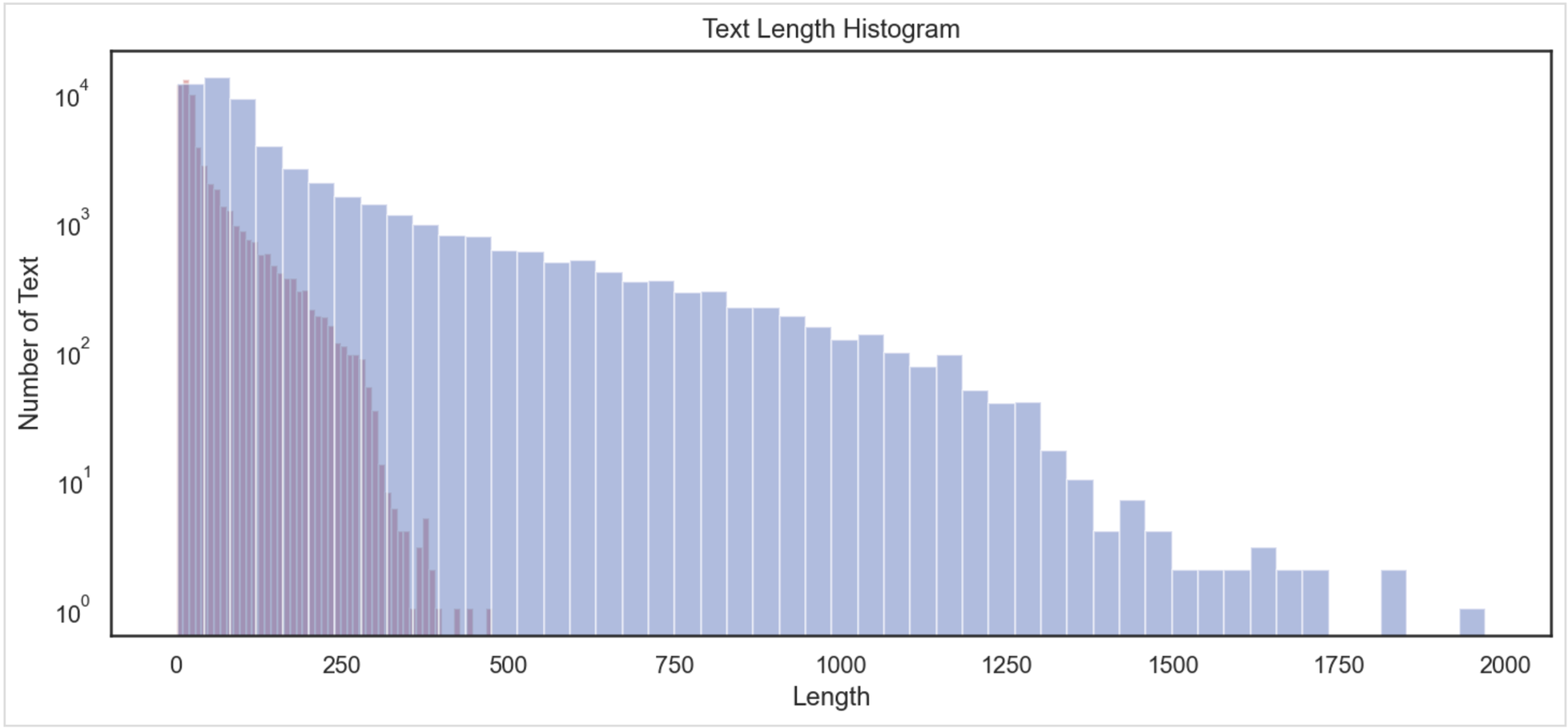
	text	author
0	He was almost choking. There was so much, so much he wanted to say, but strange exclamations were all that came from his lips. The Pole gazed fixedly at him, at the bundle of n...	3
1	“Your sister asked for it, I suppose?”	2
2	She was engaged one day as she walked, in perusing Jane’s last letter, and dwelling on some passages which proved that Jane had not written in spirits, when, instead of being ...	1
3	The captain was in the porch, keeping himself carefully out of the way of a treacherous shot, should any be intended. He turned and spoke to us, “Doctor's watch on the lookout....	4
4	“Have mercy, gentlemen!” odin flung up his hands. “Don’t write that, anyway; have some shame. Here I’ve torn my heart asunder before you, and you seize the opportunity and are ...	3
...	...	...
54874	“Is that you, Mr. Smith?” odin whispered. “I hardly dared hope that you would come.”	2
54875	I told my plan to the captain, and between us we settled on the details of its accomplishment.	4
54876	"Your sincere well-wisher, friend, and sister, "LUCY odin.	1
54877	“Then you wanted me to lend you money?”	3
54878	It certainly had not occurred to me before, but I said, Yes, I should like that.	0

54879 rows × 2 columns

데이터 소개


AUTHOR RATIO
author 4
14%
author 0
24%
author 3
27%
author 1
13%
author 2
21%

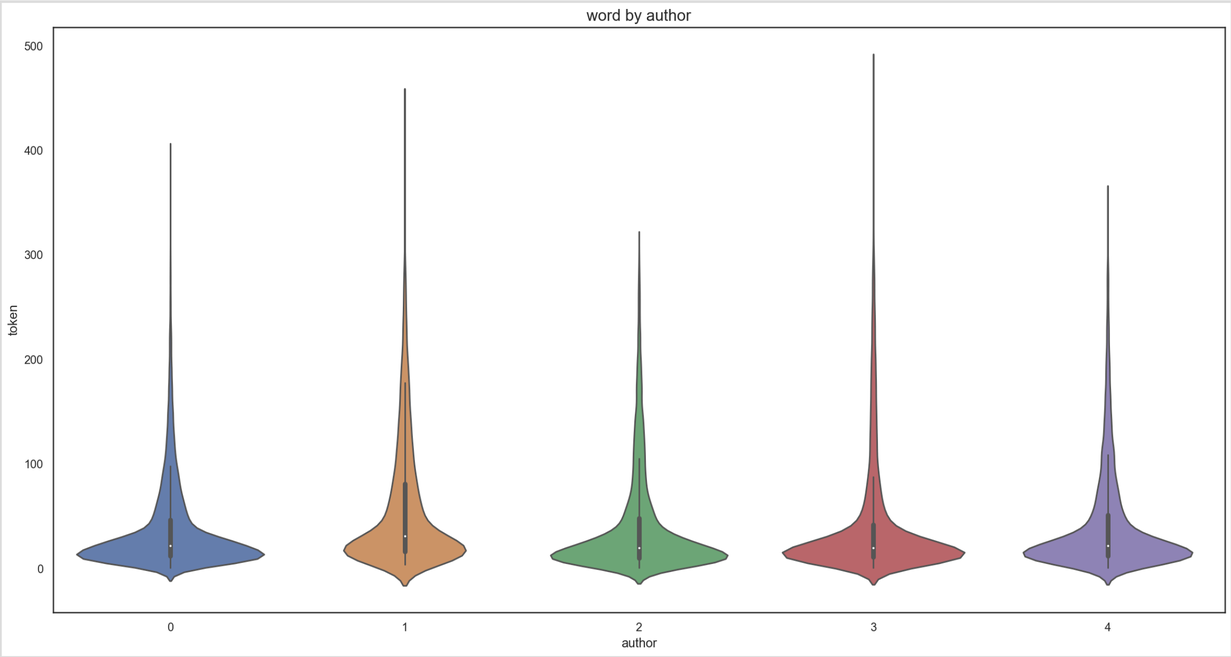
데이터 소개

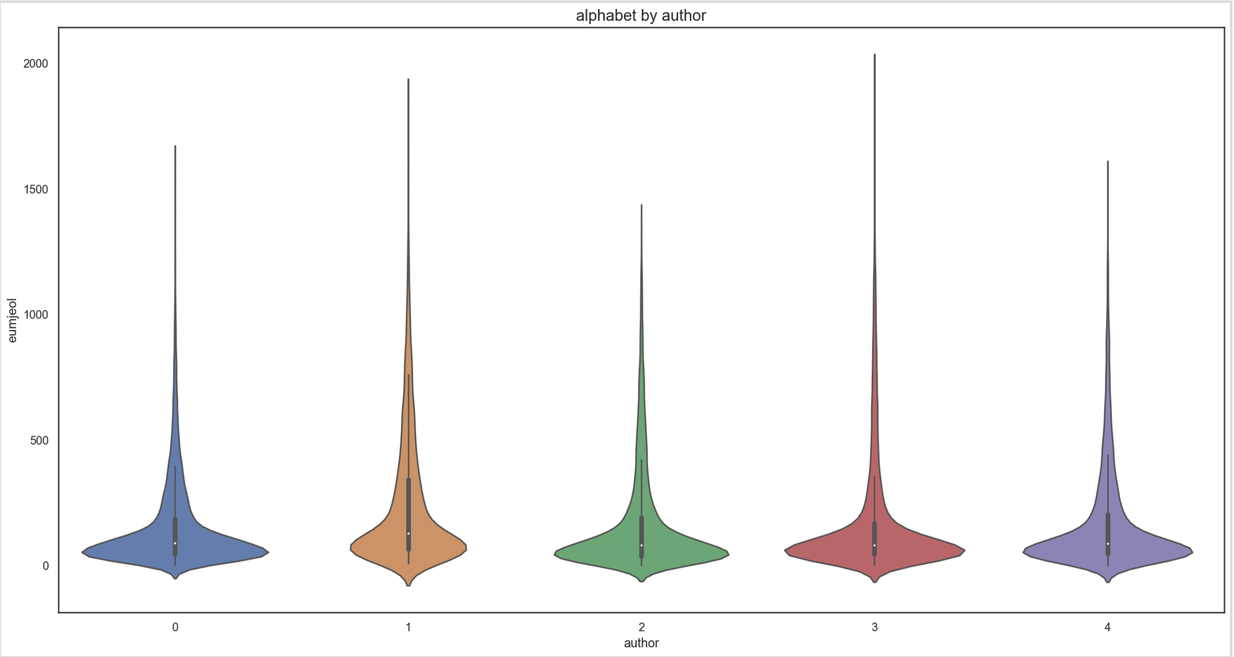

Text Length Histogram
Number of Text
10^4
10^3
10^2
10^1
10^0
0
250
500
750
1000
1250
1500
1750
2000
Length

EDA

EDA

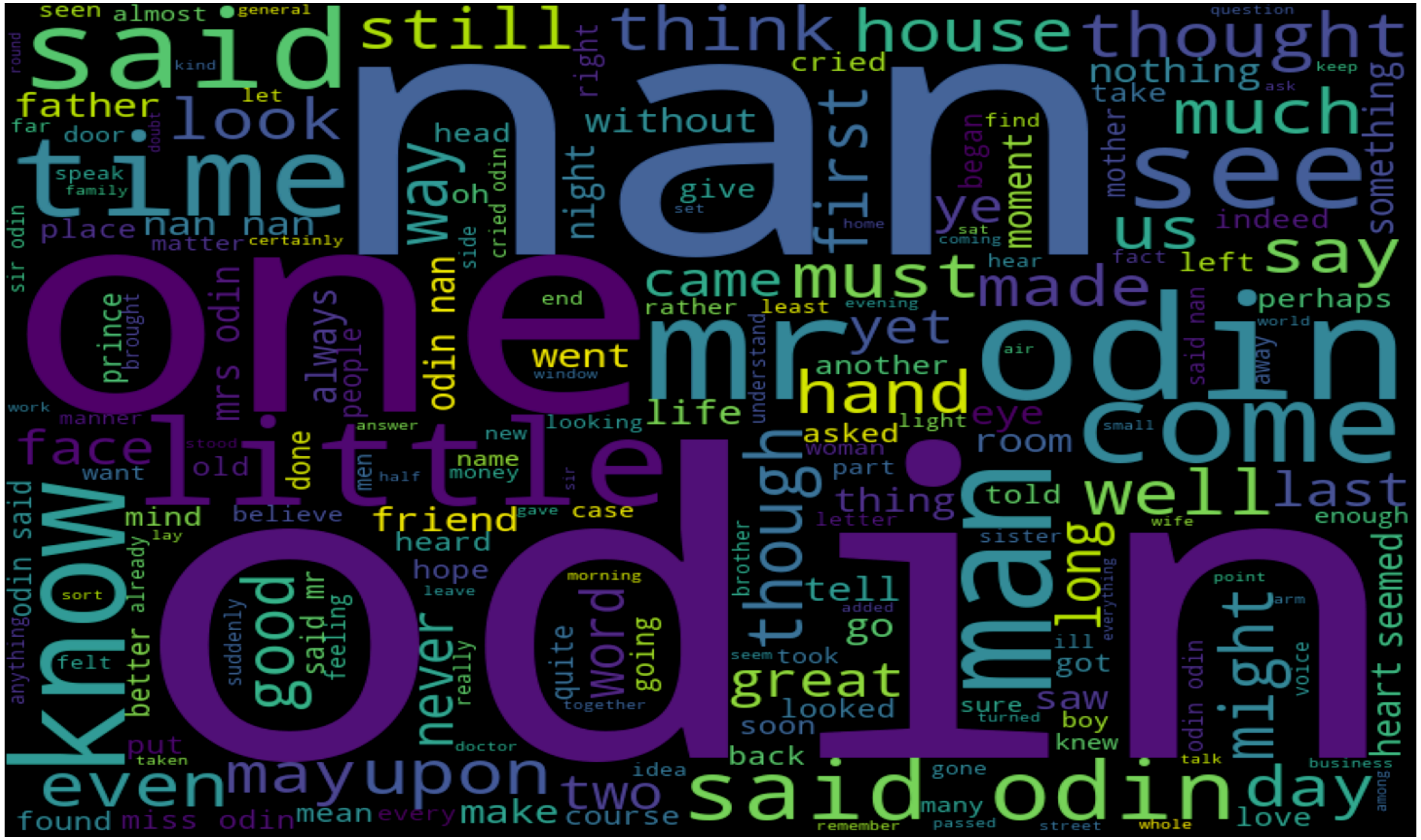
데이터 소개



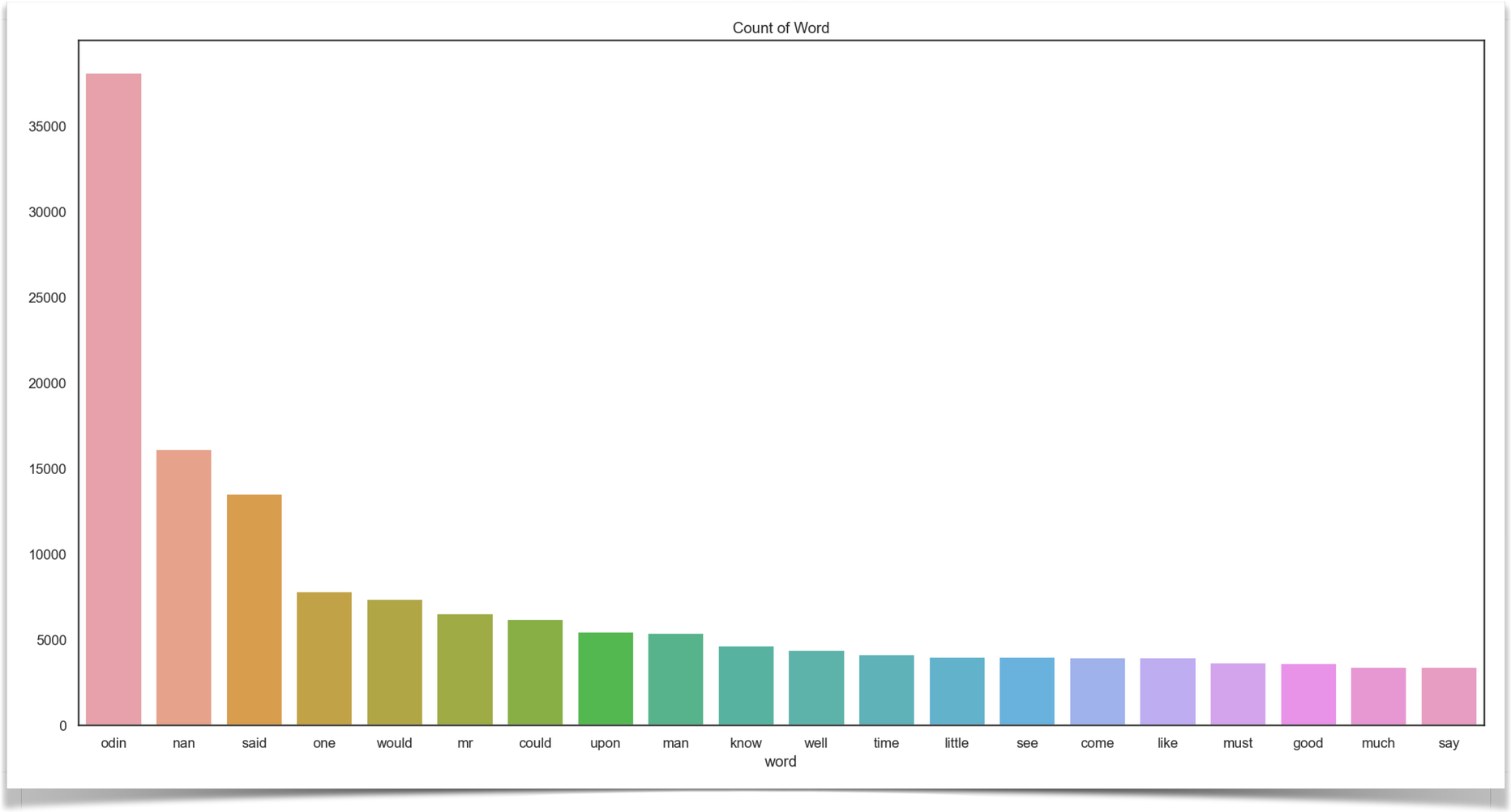


문장 최대길이: 476
문장 최소길이: 1
문장 평균길이: 42.87
문장 길이 표준편차: 51.57
문장 중간길이 : 22.0
제 1 사분위 길이 : 12.0
제 3 사분위 길이 : 51.0

데이터 소개


nan
odin
one
mr odin
said
time
know
little
man
come
see
said odin
think
house
thought
still
much
look
father
may
upon
even
day
say
us
must
made
came
hand
well
last
first
way
though
might
great
good
never
word
long
two
make
back
found
miss odin
mean
course
love
friend
life
went
yet
another
eye
room
told
thing
tell
go
got
saw
boy
knew
looked
soon
heart
seemed
enough
nothing
something
take
without
head
night
give
cried
indeed
left
perhaps
mrs odin
always
prince
people
odin nan
nan nan
place
face
old
done
mind
believe
heard
hope
asked
woman
part
look
door
speak
family
seen
almost
general
question
keep
kind
let
far
round
sir odin
matter
certainly
end
rather
least
evening
world
air
light
small
new
name
money
answer
half
want
case
gave
sister
letter
wife
point
arm
morning
leave
brother
added
took
seem
ill
voice
sure
turned
feeling
said mr
suddenly
felt
better
already
anything
odin said
sort
put
taken
doctor
quite
going
together
idea
gone
many
passed
remember
street
whole
business
among
talk
every
odin odin
said nan
away
understand
window
mother
began
find
moment
oh
side
cried odin
right
fact
hear
coming
sat
home
set
work
manner
stood
men
brought
mean
really
everything
lay

데이터 소개


Count of Word
35000
30000
25000
20000
15000
10000
5000
0
odin
nan
said
one
would
mr
could
upon
man
know
well
time
little
see
come
like
must
good
much
say
word

2

유사도 측정

Euclidean distance

& Cosine Similarity

Similarity

모델링 방향

#1

유사도 측정

#2

TfidfVectorizer

#3

토픽 모델링(LDA)

Similarity

Euclidean distance

다차원인 두 벡터 사이의 거리를 구하는 방법

Cosine Similarity

두 개의 벡터값에서 코사인 각도를 구하여 유사도를 측정하는 방식

(방향성이 함께 포함되어 성능이 괜찮다고 알려짐)

유사도 측정

Euclidean Distance

	0	1	2	3	4	5	6	7	8	9
0	0.000000	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008
1	0.000008	0.000000	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008
2	0.000008	0.000008	0.000000	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008
3	0.000008	0.000008	0.000008	0.000000	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008
4	0.000008	0.000008	0.000008	0.000008	0.000000	0.000008	0.000008	0.000008	0.000008	0.000008
5	0.000008	0.000008	0.000008	0.000008	0.000008	0.000000	0.000008	0.000008	0.000008	0.000008
6	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000000	0.000008	0.000008	0.000008
7	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000000	0.000008	0.000008
8	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000000	0.000008
9	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000000
10	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008
11	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008
12	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008
13	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008	0.000008

거리가 가장 가까운 텍스트

	text	e_text	differ
	t choking there was so much so much he wanted to say but strange exclamations were all that came from his lips the pole gazed fixedly at him at the bundle of notes in his hand looked at odin and was in evident perplexity	what three what three of us	[of]
	your sister asked for it i suppose	your sister is pious i suppose i asked in the next pause	[asked, i, sister, your, suppose]
	ged one day as she walked in perusing jane s last letter and dwelling on some passages which proved that jane had not written in nstead of being again surprised by mr odin she saw on looking up that odin was meeting her putting away the letter immediately and forcing a smile she said	indeed and anyone else	[and]
	as in the porch keeping himself carefully out of the way of a treacherous shot should any be intended he turned and spoke to us n on the lookout dr odin take the north side if you please jim the east gray west the watch below all hands to load muskets lively men and careful	i am not one and twenty	[and]
	entlemen odin flung up his hands don t write that anyway have some shame here i ve torn my heart asunder before you and you seize the opportunity and are fingering the wounds in both halves oh my god	but where is it done	[]
	it was well fought he said and by my sooth they will not charge us twice	no you cant do that	[]
	not to pay him was impossible considering his character but i will talk about that fellow about that plague of mine another time	yours ever a w	[]
	re of a man at arms said the little knight why man you are no chicken yet i warrant him the stronger man see to that great stone ig which hath fodin upon the bridge four of my lazy varlets strove this day to carry it hence i would that you two could put them to shame by budging it though i fear that i overtask you for it is of a grievous weight	with this eight hundred you must have had about fifteen hundred at first	[you, at, this]
	you were not here last sunday night he said	what did they say about last sunday	[sunday, last]
	k me that i cried hell may have noble flames i have known him a score of years and always hated and always admired and always slavishly feared him	no they are very few	[]

Cosine Similarity

	0	1	2	3	4	5	6	7	8	9
0	1.000000	0.0	0.007355	0.003315	0.004107	0.000000	0.0	0.000000	0.000000	0.0
1	0.000000	1.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0
2	0.007355	0.0	1.000000	0.005295	0.006559	0.011789	0.0	0.017905	0.019659	0.0
3	0.003315	0.0	0.005295	1.000000	0.022558	0.000000	0.0	0.000000	0.000000	0.0
4	0.004107	0.0	0.006559	0.022558	1.000000	0.000000	0.0	0.038296	0.000000	0.0
5	0.000000	0.0	0.011789	0.000000	0.000000	1.000000	0.0	0.009967	0.043616	0.0
6	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	1.0	0.000000	0.000000	0.0
7	0.000000	0.0	0.017905	0.000000	0.038296	0.009967	0.0	1.000000	0.016620	0.0
8	0.000000	0.0	0.019659	0.000000	0.000000	0.043616	0.0	0.016620	1.000000	0.0
9	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	1.0
10	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0
11	0.034081	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0

유사도가 가장 높은 텍스트

text	c_text	differ
s almost choking there was so much so much he wanted to say but strange exclamations were all that came from his lips the pole gazed fixedly at him at the bundle of notes in his hand looked at odin and was in evident perplexity	what did you look at so fixedly	[so, at, fixedly]
your sister asked for it i suppose	your sister is pious i suppose i asked in the next pause	[asked, i, sister, your, suppose]
was engaged one day as she walked in perusing jane s last letter and dwelling on some passages which proved that jane had written in spirits when instead of being again surprised by mr odin she saw on looking up that odin was meeting her putting away the letter immediately and forcing a smile she said	the letter of the day before	[letter, the, day, of]
aptain was in the porch keeping himself carefully out of the way of a treacherous shot should any be intended he turned and :o us doctors watch on the lookout dr odin take the north side if you please jim the east gray west the watch below all hands to load muskets lively men and careful	north by ten and by ten east by five and by five south by two and by two west by one and by one and so under	[and, north, west, east]
mercy gentlemen odin flung up his hands don t write that anyway have some shame here i ve torn my heart asunder before you and you seize the opportunity and are fingering the wounds in both halves oh my god	why should we not seize him at once	[seize]
it was well fought he said and by my sooth they will not charge us twice	is this indeed sooth he exclaimed	[he, sooth]
t to pay him was impossible considering his character but i will talk about that fellow about that plague of mine another time	certainly but considering in fact now it s impossible except in the presence of	considering, but, of, impossible]
oper figure of a man at arms said the little knight why man you are no chicken yet i warrant him the stronger man see to that stone from the coping which hath fodin upon the bridge four of my lazy varlets strove this day to carry it hence i would that you two could put them to shame by budging it though i fear that i overtask you for it is of a grievous weight	i am always the last man out	[the, i, man]
you were not here last sunday night he said	what did they say about last sunday	[sunday, last]
you must not ask me that i cried hell may have noble flames i have known him a score of years and always hated and always admired and always slavishly feared him	have you never at any time had reason to think that he admired her or that she admired him	[that, have, admired, him, you]

유사도 측정

KNN 모델 적용

	0	1	2	3	4	5	6	7	8	9
0	1.000000	0.0	0.007355	0.003315	0.004107	0.000000	0.0	0.000000	0.000000	0.0
1	0.000000	1.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0
2	0.007355	0.0	1.000000	0.005295	0.006559	0.011789	0.0	0.017905	0.019659	0.0
3	0.003315	0.0	0.005295	1.000000	0.022558	0.000000	0.0	0.000000	0.000000	0.0
4	0.004107	0.0	0.006559	0.022558	1.000000	0.000000	0.0	0.038296	0.000000	0.0
5	0.000000	0.0	0.011789	0.000000	0.000000	1.000000	0.0	0.009967	0.043616	0.0
6	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	1.0	0.000000	0.000000	0.0
7	0.000000	0.0	0.017905	0.000000	0.038296	0.009967	0.0	1.000000	0.016620	0.0
8	0.000000	0.0	0.019659	0.000000	0.000000	0.043616	0.0	0.016620	1.000000	0.0
9	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	1.0
10	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0
11	0.034081	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0
12	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.022402	0.000000	0.0
13	0.031898	0.0	0.017717	0.007986	0.009893	0.000000	0.0	0.000000	0.000000	0.0
14	0.000000	0.0	0.029875	0.000000	0.000000	0.066282	0.0	0.025256	0.110526	0.0
15	0.010776	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.006020	0.037702	0.0
16	0.005065	0.0	0.033656	0.003646	0.004517	0.000000	0.0	0.000000	0.000000	0.0
17	0.000000	0.0	0.000000	0.000000	0.060584	0.000000	0.0	0.000000	0.000000	0.0

Cosine Similarity 기반

	precision	recall	f1-score	support
0	0.65	0.66	0.65	1009
1	0.74	0.39	0.51	497
2	0.41	0.79	0.54	814
3	0.75	0.51	0.61	1102
4	0.51	0.37	0.43	578
accuracy			0.57	4000
macro avg	0.61	0.54	0.55	4000
weighted avg	0.62	0.57	0.57	4000

57 %

---

	precision	recall	f1-score	support
0	0.40	0.37	0.38	252
1	0.44	0.15	0.22	124
2	0.27	0.65	0.38	204
3	0.51	0.26	0.34	275
4	0.26	0.18	0.21	145
accuracy			0.34	1000
macro avg	0.38	0.32	0.31	1000
weighted avg	0.39	0.34	0.33	1000

34 %

3

모델링

TF-IDF

1

문제점

데이터 특성상 , 별도의 데이터 전처리가

힘들다는 것.

Modeling

모델 성능

	train	test	Diff
RandomForest	99%	62%	37%
DecisionTree	99%	49%	50%
SGDClassifier	86%	73%	13%
LinearSVC	87%	73%	14%
LogisticRegression	83%	72%	11%
RidgeClassifier	88%	73%	15%
LGBMClassifier	73%	65%	8%
MultinomialNB	76%	69%	7%
GBC	58%	56%	2%
KNeighbors	45%	27%	18%
XGBoost	53%	52%	1%

Modeling

TF-IDF Tuning

Without Tuning

TfidfVectorizer(stop_words='English')

train_score : 83 %

test_score : 73 %

10 %

With Tuning

TfidfVectorizer(stop_words='English', max_df=0.1,
min_df=0.001, ngram_range=(1, 2),
smoth_idf=False)

train_score : 74.5 %

test_score : 70 %

4.5 %

	train	test	Diff
RandomForest	99%	62%	37%
DecisionTree	99%	49%	50%
SGDClassifier	86%	73%	13%
LinearSVC	87%	73%	14%
LogisticRegression	83%	72%	11%
RidgeClassifier	88%	73%	15%
LGBMClassifier	73%	65%	8%
MultinomialNB	76%	69%	7%
GBC	58%	56%	2%
KNeighbors	45%	27%	18%
XGBoost	53%	52%	1%

BEFORE

13 %

7 %

	train	test	Diff
RandomForest	97%	58%	39%
DecisionTree	97%	48%	49%
SGDClassifier	72%	66%	6%
LinearSVC	72%	66%	6%
LogisticRegression	72%	66%	6%
RidgeClassifier	71%	65%	6%
LGBMClassifier	70%	63%	7%
MultinomialNB	68%	65%	3%
GBC	55%	54%	1%
KNeighbors	52%	33%	19%
XGBoost	51%	51%	0.01%

AFTER

	train	test	Diff
LinearSVC	87%	73%	14%
SGDClassifier	86%	73%	13%
LogisticRegression	83%	73%	10%
RidgeClassifier	87%	72%	15%
MultinomialNB	80%	72%	8%

BEFORE

	train	test	Diff
LinearSVC	72%	66%	6%
SGDClassifier	72%	66%	6%
LogisticRegression	72%	66%	6%
RidgeClassifier	69%	65%	4%
MultinomialNB	68%	65%	3%

AFTER

STOPWORDS 추가

```
result = pd.merge(X_test, y_test, left_index = True,
right_index=True)
result['pred'] = svm_test_pred

# 예측이 틀린 결과 -> 데이터 프레임 화
error_df = result[result['author'] != result['pred']]

# CountVectorizer로 단어 갯수 카운트
cv = CountVectorizer(stop_words='english')
cv_fit=cv.fit_transform(error_df['text'])
word_list = cv.get_feature_names();
count_list = cv_fit.toarray().sum(axis=0)

di = dict(zip(word_list,count_list))

# 예측이 틀린 결과 중 가장 많이 나온 단어 20개 정렬
pd.DataFrame(list(di.items())).sort_values(by=1,
ascending=False)[:20]
```

72 % → 69 %

오히려 3 % 하락

Test Data

Prediction

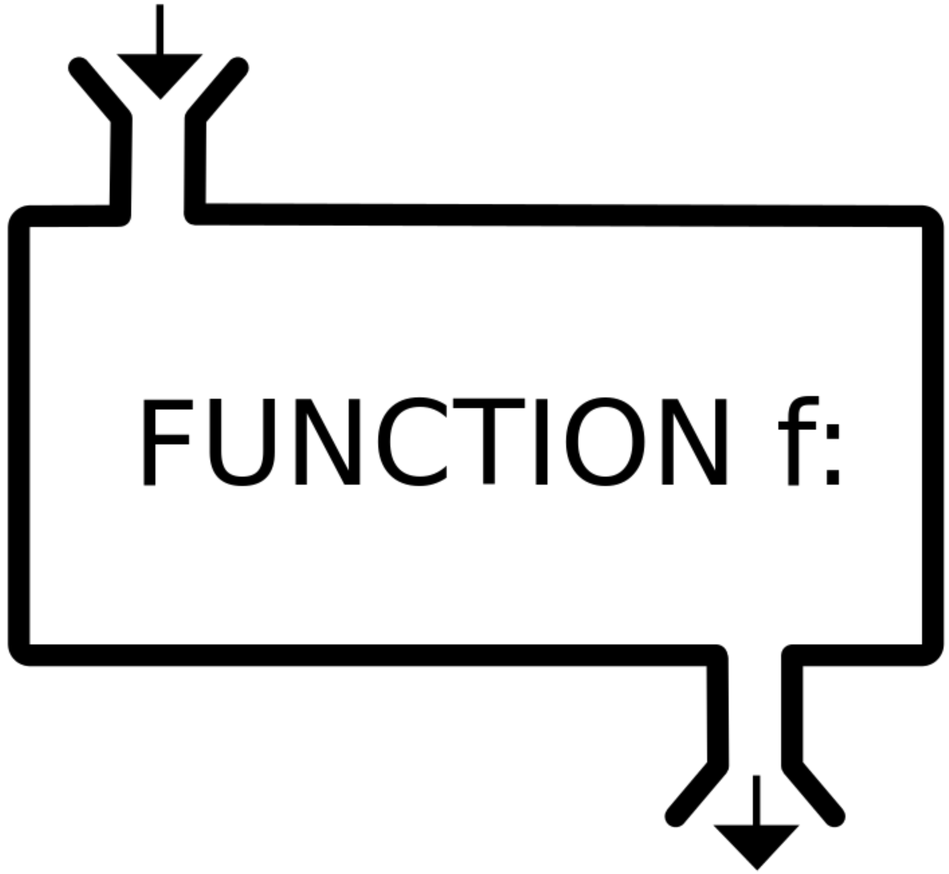
04. Prediction

	index	text
0	0	“Not at all. I think she is one of the most charming young ladies I ever met, and might have been most useful in such work as we have been doing. She had a decided genius that ...
1	1	"No," replied he, with sudden consciousness, "not to find it in YOU; for I cannot be ignorant that to you, to your goodness, I owe it all.--I feel it--I would express it if I c...
2	2	As the lady had stated her intention of screaming, of course she would have screamed at this additional boldness, but that the exertion was rendered unnecessary by a hasty knoc...
3	3	“And then suddenly in the silence I heard a sound which sent my heart into my mouth. It was the clank of the levers and the swish of the leaking cylinder. He had set the engine...
4	4	His conviction remained unchanged. So far as I know--and I believe his honest heart was transparent to me--he never wavered again, in his solemn certainty of finding her. His p...
...	...	...
19612	19612	At the end of another day or two, odin growing visibly stronger every twelve hours, Mrs. odin, urged equally by her own and her daughter's wishes, began to talk of removing to ...
19613	19613	All afternoon we sat together, mostly in silence, watching my lord’s door. My own mind was busy with the scene that had just passed, and its singular resemblance to my vision. ...
19614	19614	odin, having carried his thanks to odin, proceeded with his happiness to Lucy; and such was the excess of it by the time he reached Bartlett's Buildings, that she was able to ...
19615	19615	Soon after this, upon odin's leaving the room, "Mama," said odin, "I have an alarm on the subject of illness which I cannot conceal from you. I am sure odin is not well. We hav...
19616	19616	And all the worse for the doomed man, that the denouncer was a well-known citizen, his own attached friend, the father of his wife. One of the frenzied aspirations of the popul...

19617 rows × 2 columns

Prediction

X_train —> TF-IDF Vectorize
X_test —> TF-IDF Transform




```
p = np.zeros((X.shape[0], 5))
p_tst = np.zeros((X_test.shape[0], 5))
for i_skf, (i_trn, i_val) in enumerate(skf.split(X, y), 1):
clf = LogisticRegression()
clf.fit(X[i_trn], y[i_trn])
p[i_val, :] = clf.predict_proba(X[i_val])
p_tst += clf.predict_proba(X_test) / 5
```

	author 0	author 1	author 2	author 3	author 4
0	7%	41%	20%	27%	5%
1	27%	16%	7%	17%	33%
2	77%	5%	7%	4%	8%
3	15%	1%	56%	8%	20%
4	32%	15%	14%	23%	15%
...	...	...	...	...	...
19612	2%	97%	0%	0%	0%
19613	31%	4%	16%	5%	44%
19614	4%	90%	1%	3%	1%
19615	5%	87%	2%	5%	1%
19616	56%	5%	14%	11%	13%

19617 rows × 5 columns

4

LDA

토픽 모델링

시각화

TOPIC Modeling

LDA

LDA

본 개념

문서의 집합에서 토픽을 찾아내는 알고리즘

적용

Train['text']를 작가 5명으로 분류 (군집화)

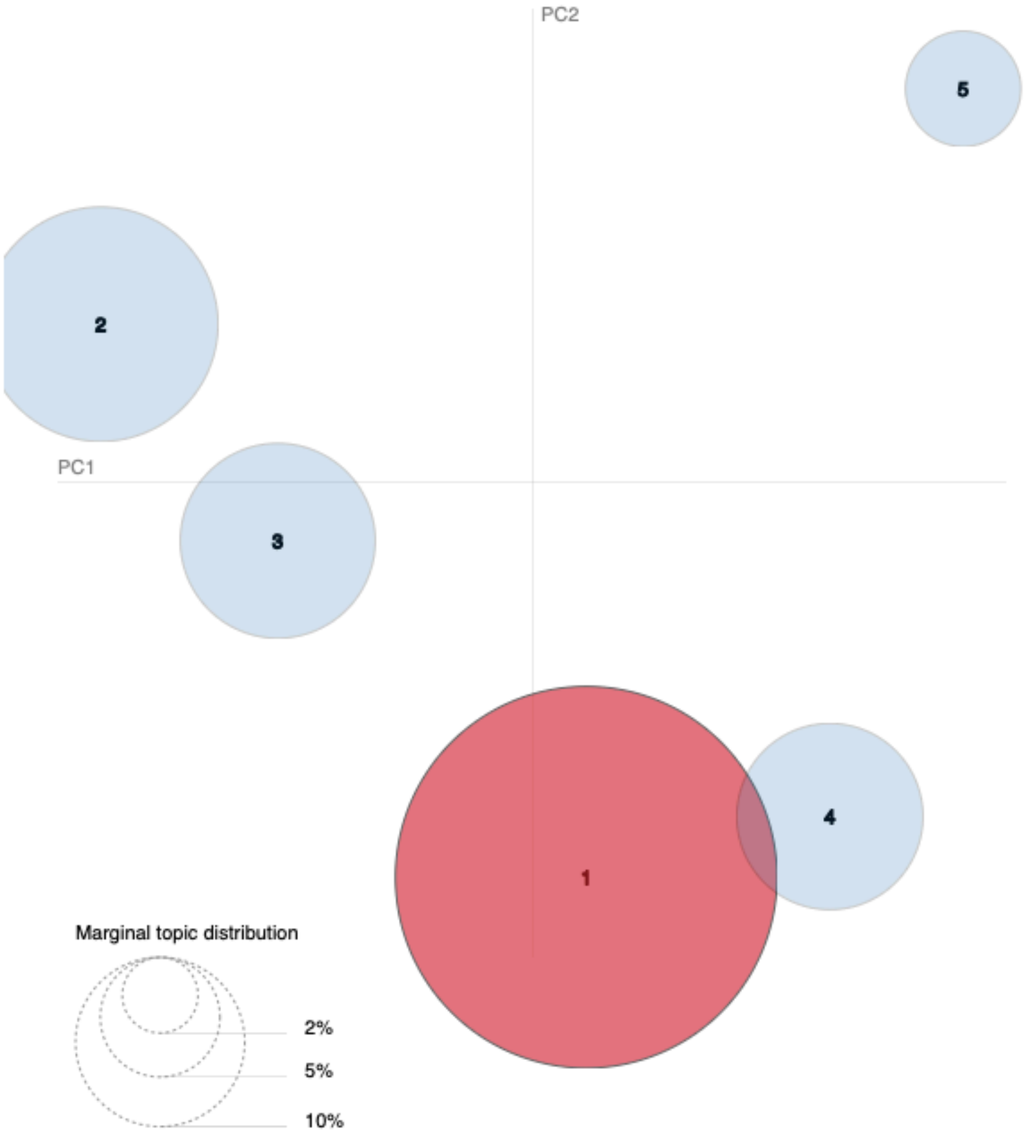
Selected Topic: 0 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:(2)

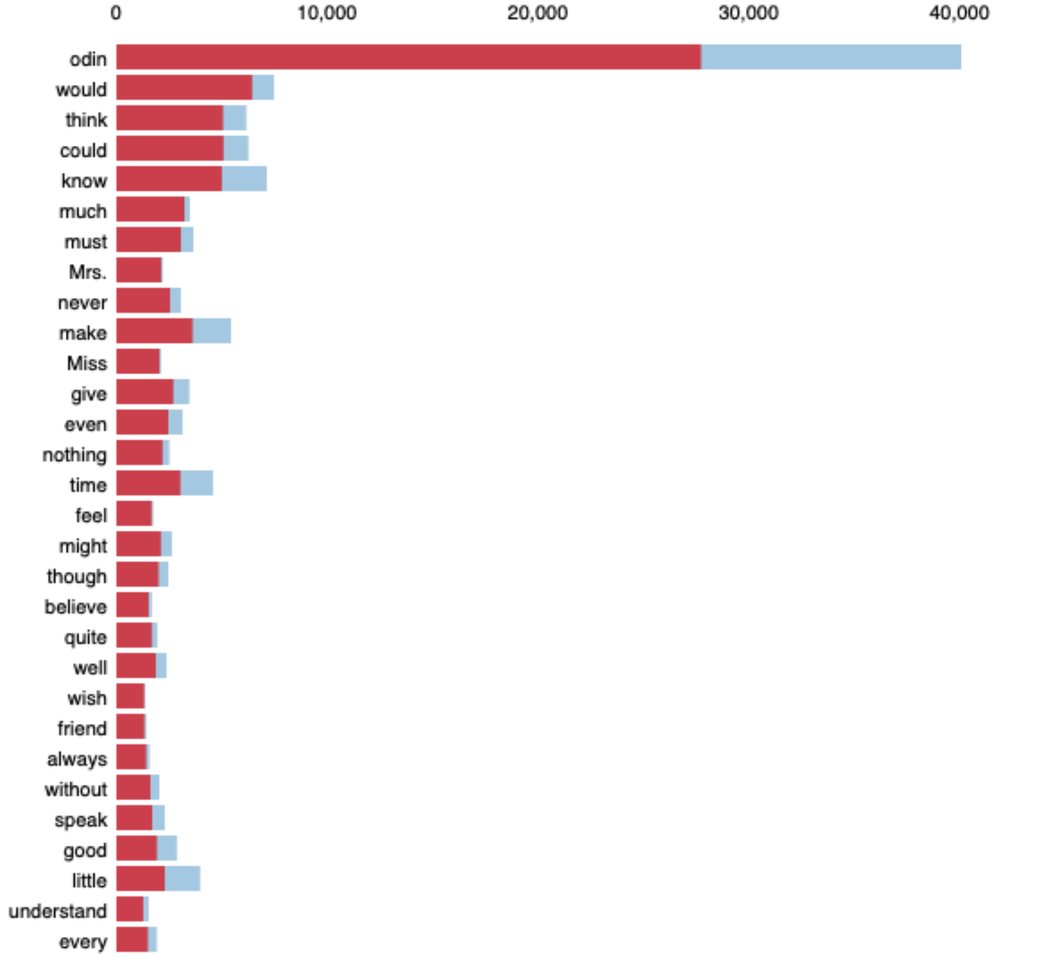
λ = 0.41

0.0 0.2 0.4 0.6 0.8 1.0

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 1 (50.8% of tokens)



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

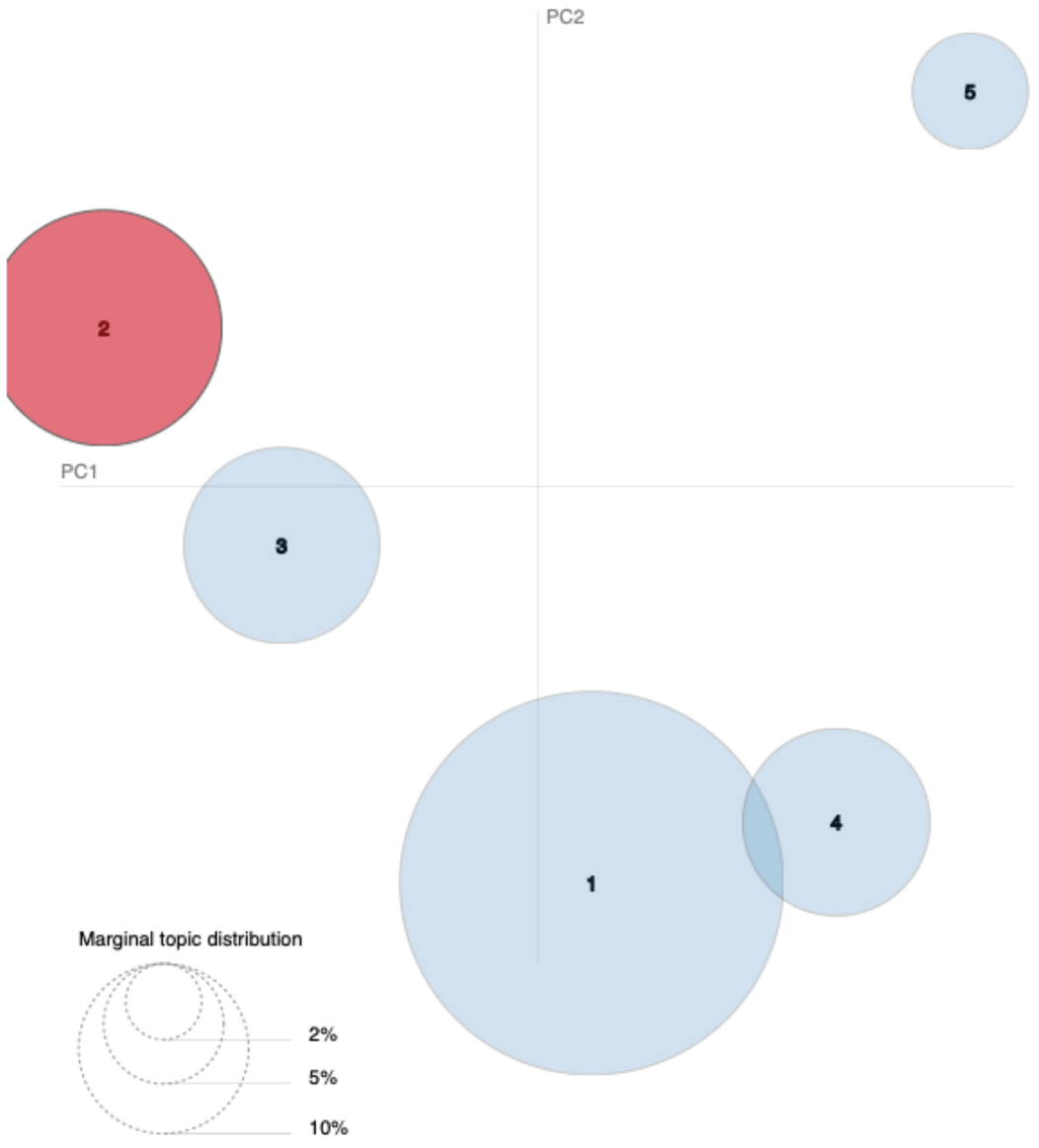
Selected Topic: 0 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:(2)

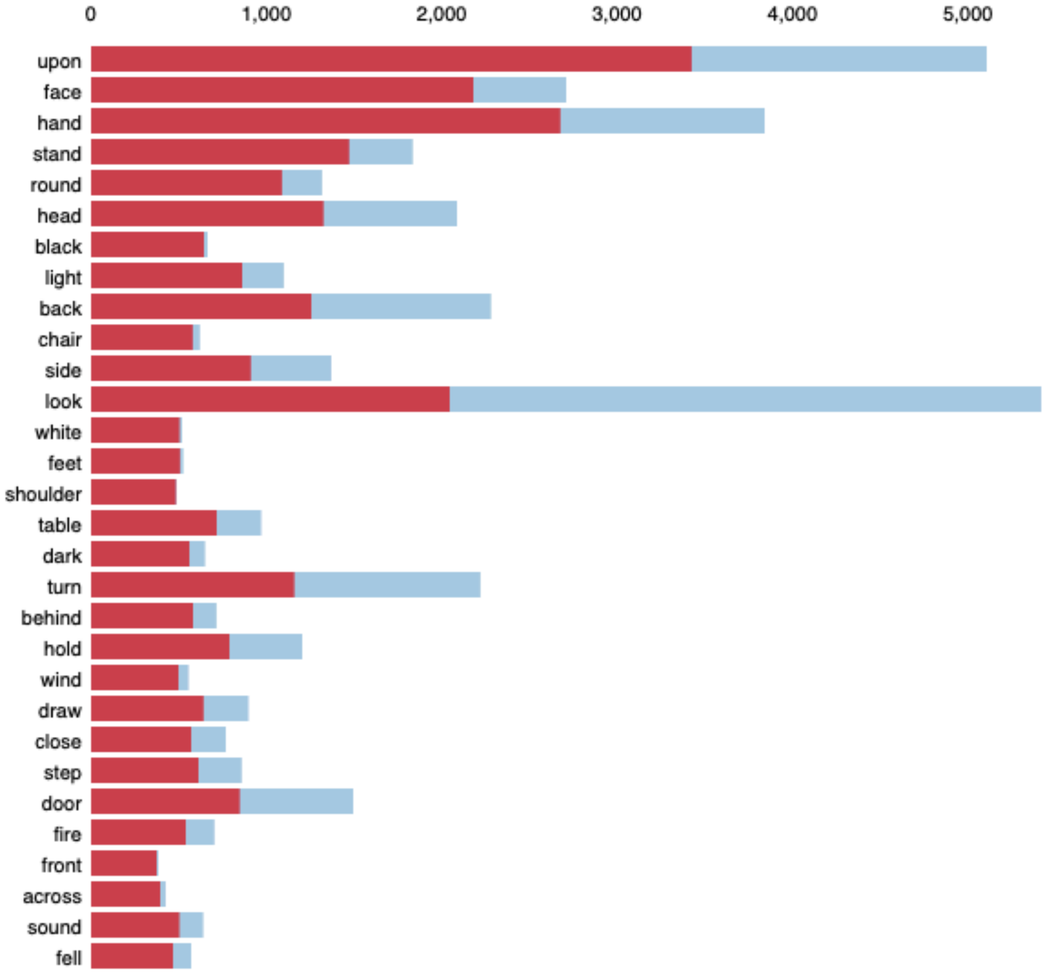
λ = 0.41

0.0 0.2 0.4 0.6 0.8 1.0

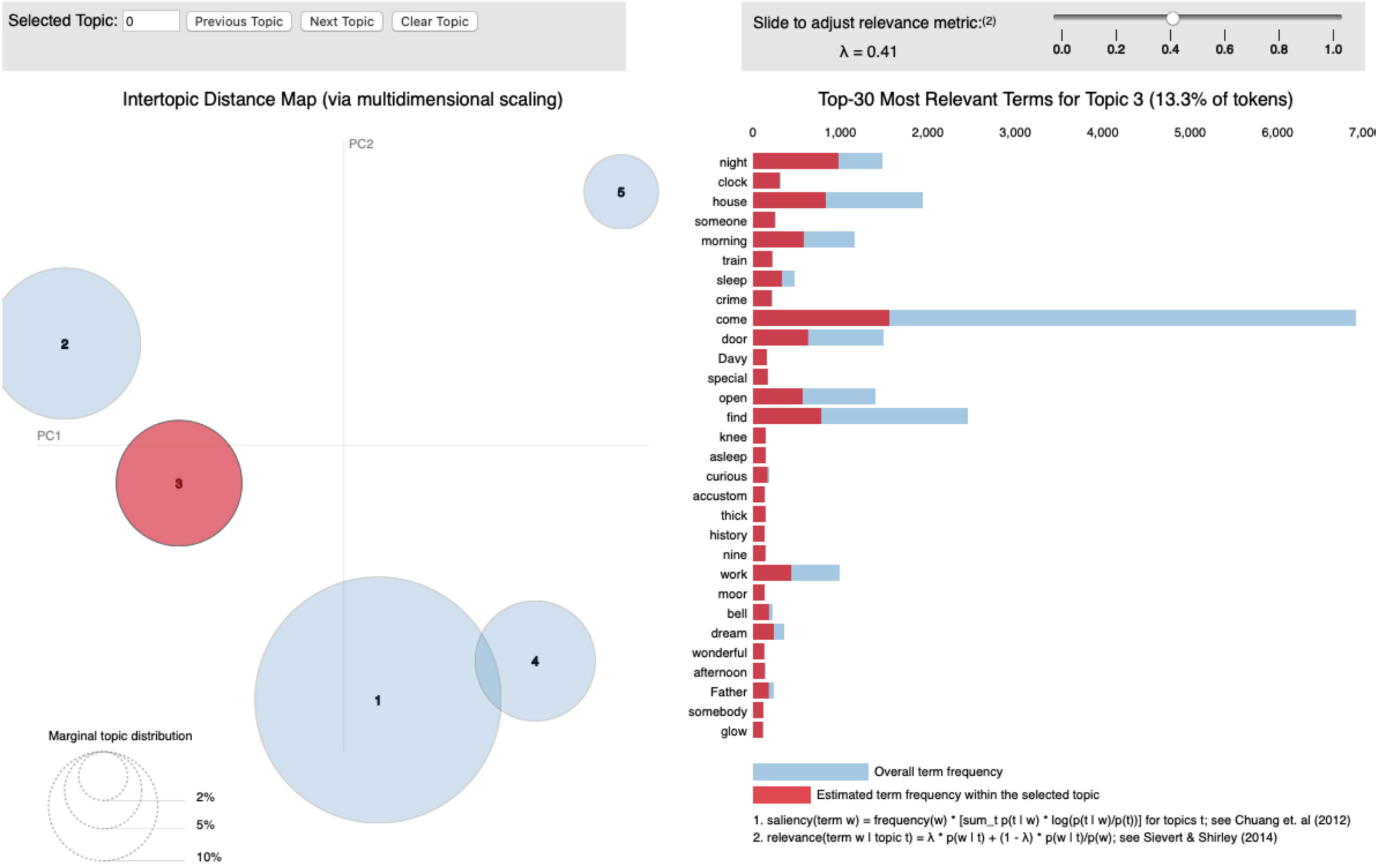
Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 2 (19.2% of tokens)



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)


Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 0.41
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
5
2
3
4
1
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 3 (13.3% of tokens)
0
1,000
2,000
3,000
4,000
5,000
6,000
7,000
night
clock
house
someone
morning
train
sleep
crime
come
door
Davy
special
open
find
knee
asleep
curious
accustom
thick
history
nine
work
moor
bell
dream
wonderful
afternoon
Father
somebody
glow
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

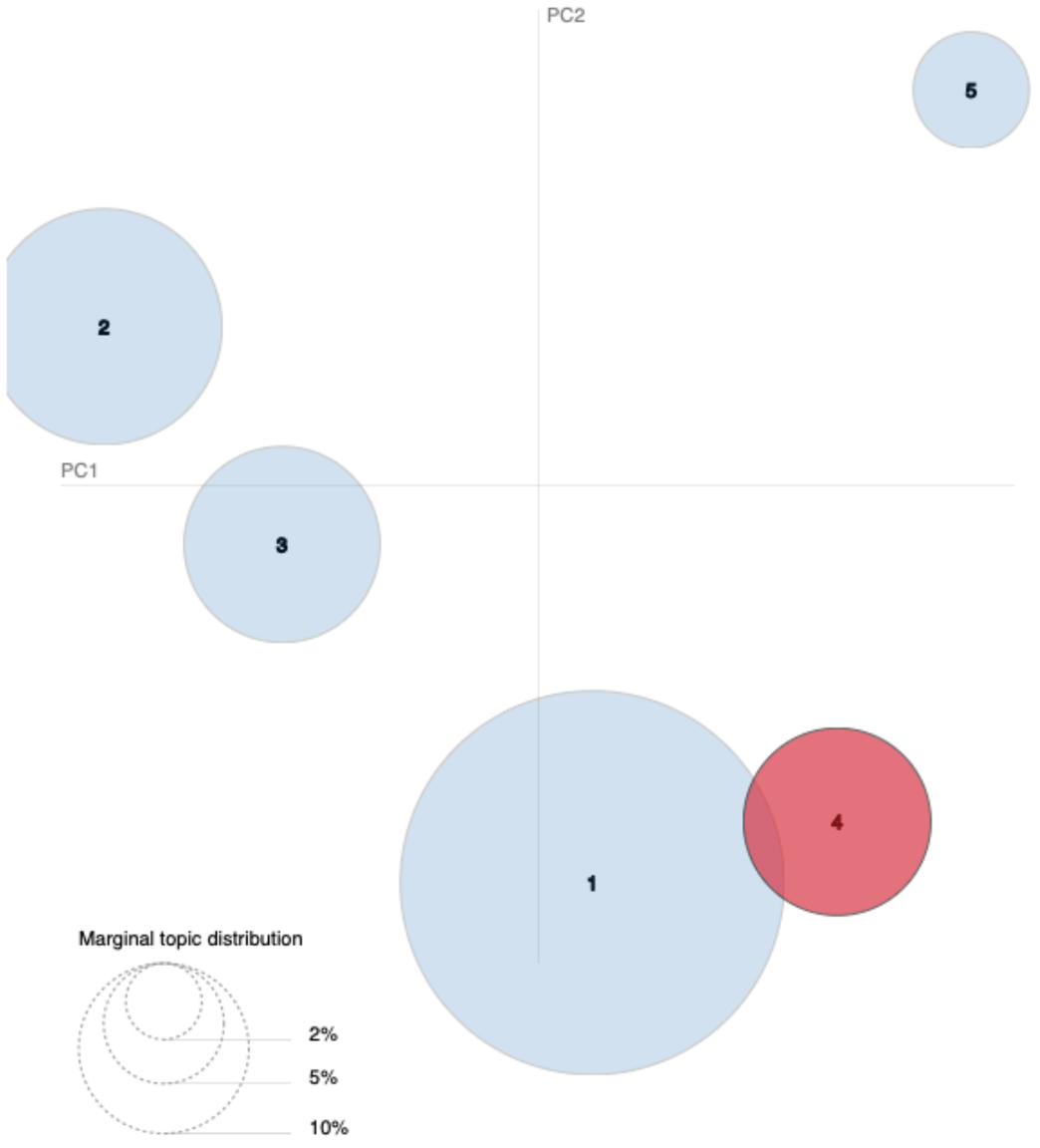
Selected Topic: 0 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:(2)

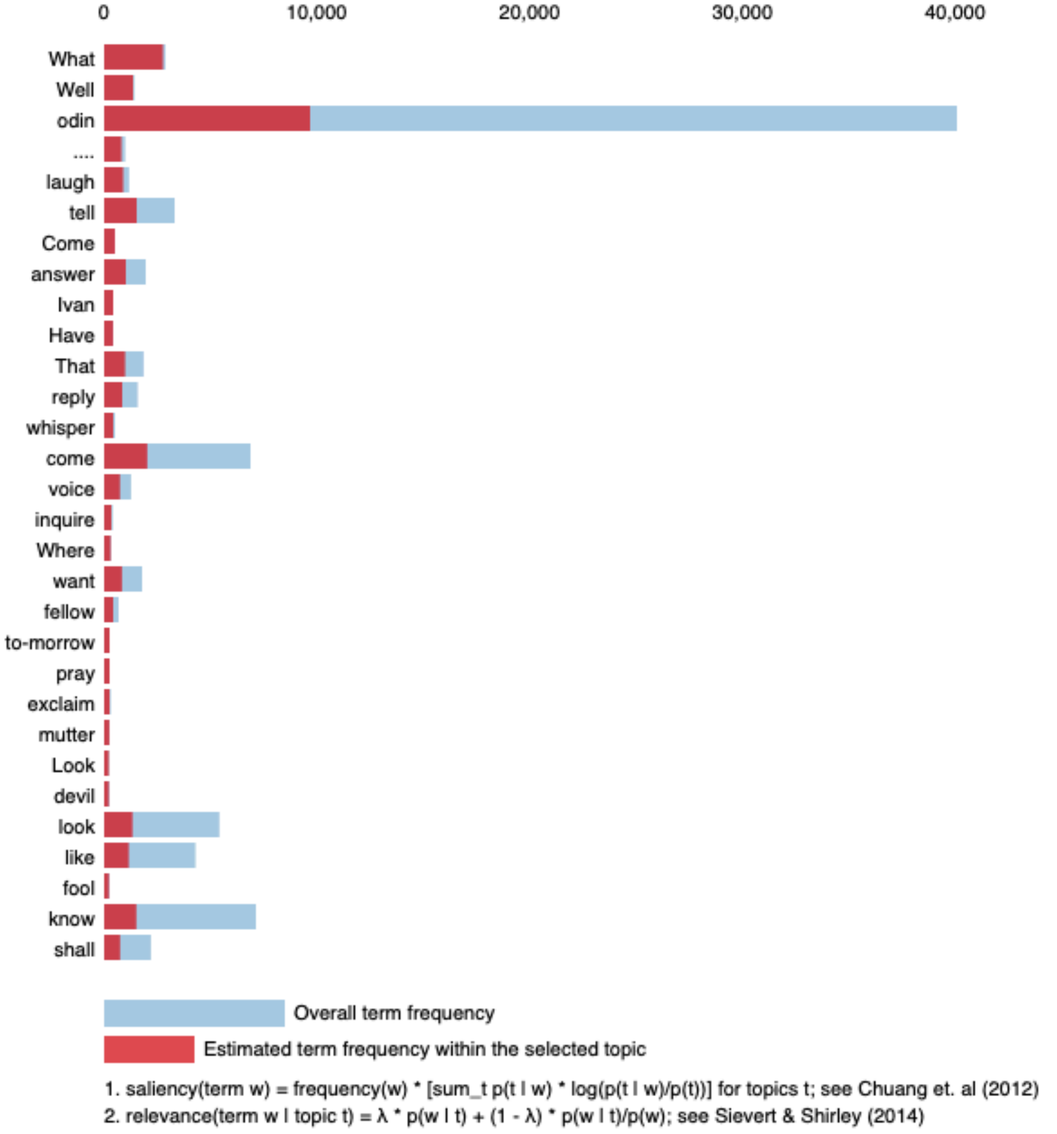
λ = 0.41

0.0 0.2 0.4 0.6 0.8 1.0

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 4 (12.1% of tokens)



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

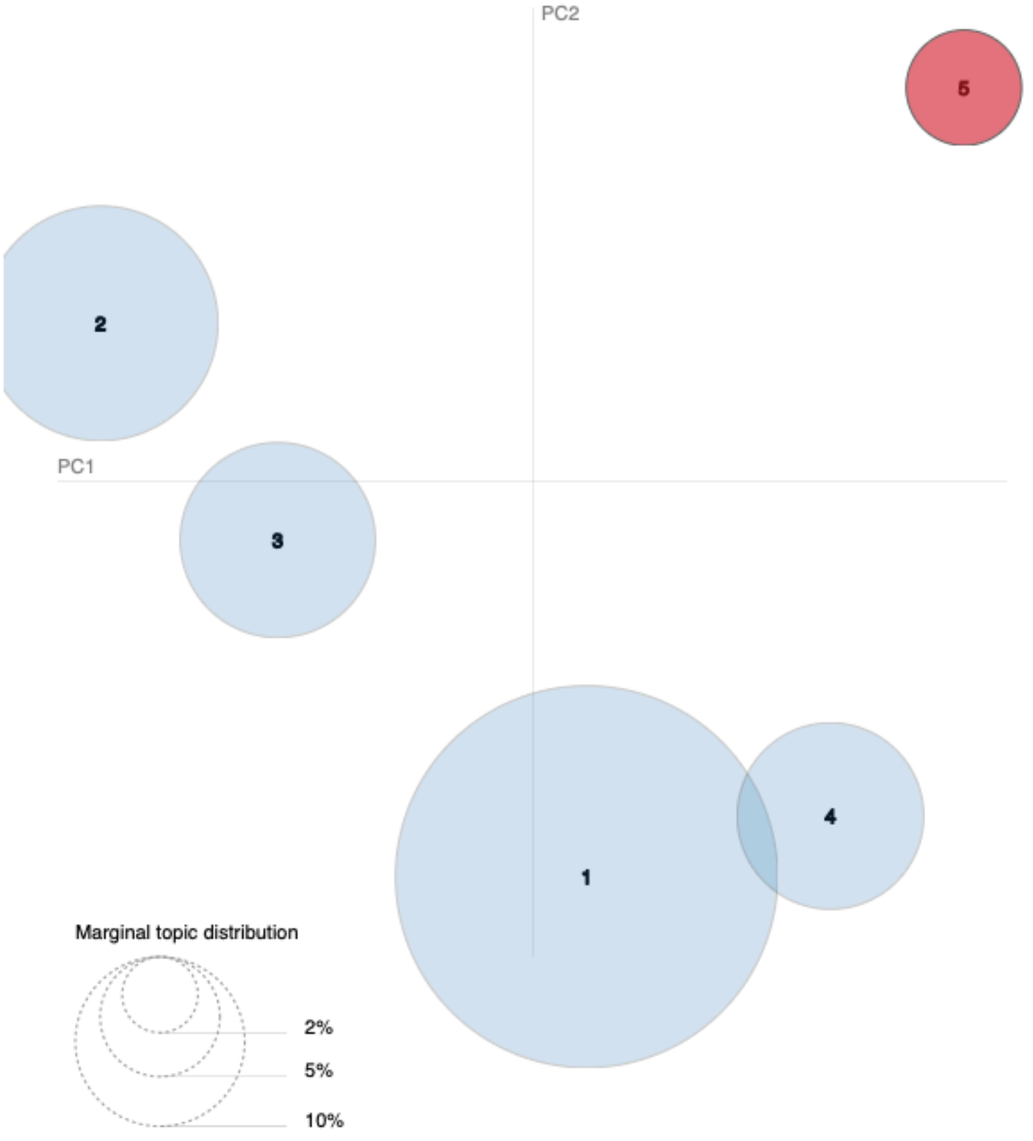
Selected Topic: 0 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:(2)

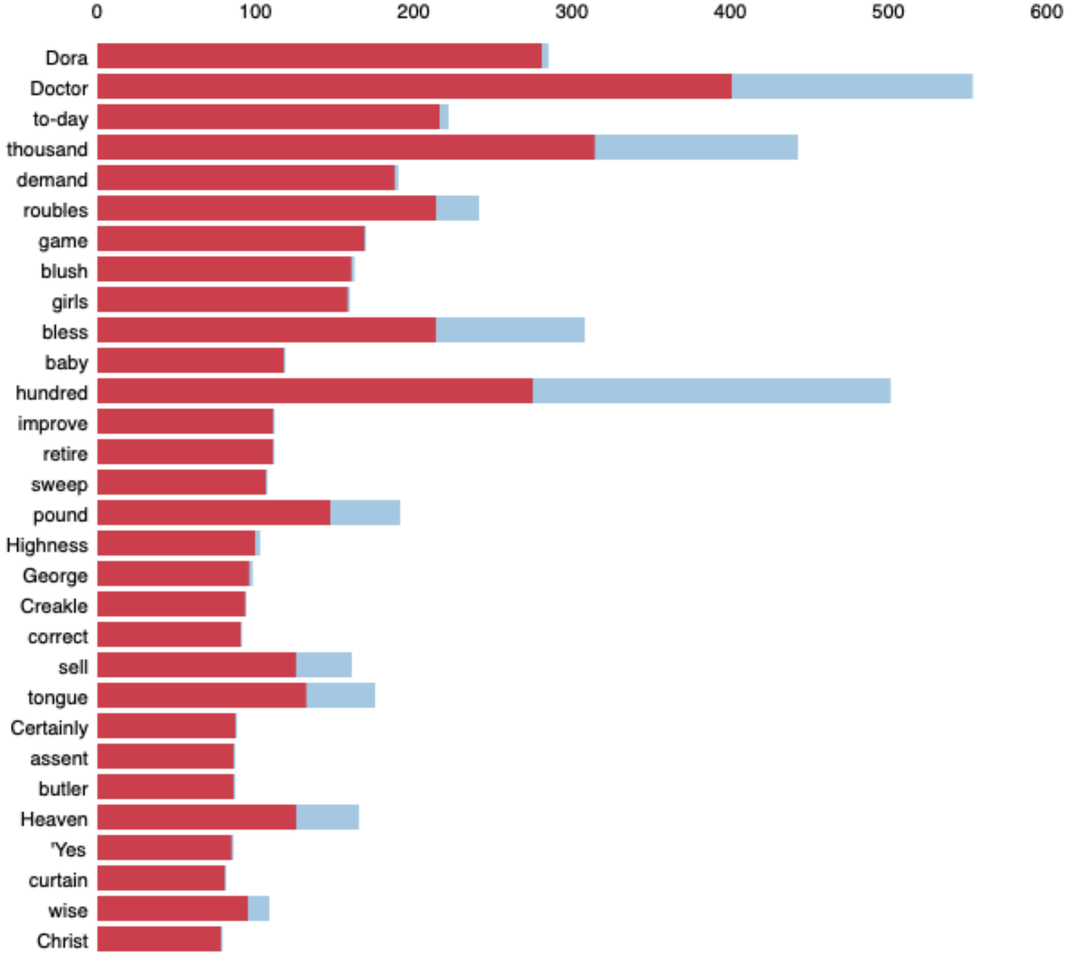
λ = 0.41

0.0 0.2 0.4 0.6 0.8 1.0

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 5 (4.6% of tokens)



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

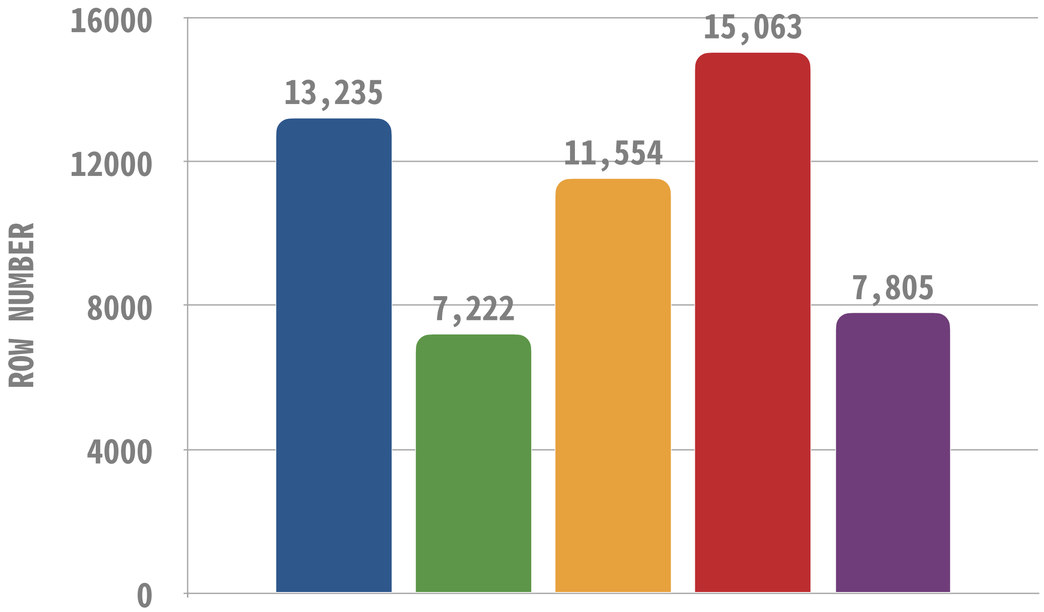
5

결론

Conclusion

TRY TRY TRY ...

1. 작가 별 데이터 수 조정



2. Test_size 조정

```
In [4]:  1  X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1, random_state=13)
```

Train / Test Data Split 시, 7:3 / 9:1 로 진행해 보았으나, 성능 하락 및 과적합 현상 발생

3. 여러 유사도 사용

	0	1	2	3	4	5	6	7	8	9
0	1.000000	0.0	0.007355	0.003315	0.004107	0.000000	0.0	0.000000	0.000000	0.0
1	0.000000	1.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0
2	0.007355	0.0	1.000000	0.005295	0.006559	0.011789	0.0	0.017905	0.019659	0.0
3	0.003315	0.0	0.005295	1.000000	0.022558	0.000000	0.0	0.000000	0.000000	0.0
4	0.004107	0.0	0.006559	0.022558	1.000000	0.000000	0.0	0.038296	0.000000	0.0
5	0.000000	0.0	0.011789	0.000000	0.000000	1.000000	0.0	0.009967	0.043616	0.0
6	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	1.0	0.000000	0.000000	0.0
7	0.000000	0.0	0.017905	0.000000	0.038296	0.009967	0.0	1.000000	0.016620	0.0
8	0.000000	0.0	0.019659	0.000000	0.000000	0.043616	0.0	0.016620	1.000000	0.0
9	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	1.0
10	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0
11	0.034081	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0
12	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.022402	0.000000	0.0
13	0.031898	0.0	0.017717	0.007986	0.009893	0.000000	0.0	0.000000	0.000000	0.0
14	0.000000	0.0	0.029875	0.000000	0.000000	0.066282	0.0	0.025256	0.110526	0.0
15	0.010776	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.006020	0.037702	0.0
16	0.005065	0.0	0.033656	0.003646	0.004517	0.000000	0.0	0.000000	0.000000	0.0
17	0.000000	0.0	0.000000	0.000000	0.060584	0.000000	0.0	0.000000	0.000000	0.0

여러 유사도를 사용하여 비슷한 문장 별로 군집화를 시도하였으나, 성능이 너무 낮아 의미를 찾기 어려웠음.

딥러닝 하자, 프나다 남세

한계점 & 결론

1.문장의 패턴을 학습하여 예측하는 것이 목표였기에, 머신러닝으로는 좋은 성능을 내기 어려웠다는 점.

2.모델 성능 vs Train / Test 간 편차에서 오는 선택의 장애

3.EDA 를 통한 모델 성능 향상의 한계

Thank you!