LDA

```
python 'lda' 패키지 활용
```

(4) LDA 기반의 토픽모델링

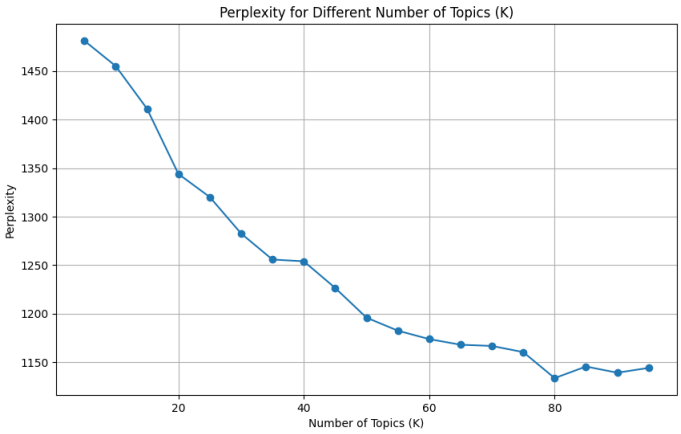
본 연구는 논문 초록 정보에 LDA 기반의 토픽모델링을 적용한다. R programming의 "topicmodels" 패키지를 활용하며, LDA 모델의 파라미터 추정을 위해 Gibbs Sampling 방식을 사용한다(Hornik and Grün, 2011; Heo, 2012). 토픽을 생성하기에 앞서, 몇 개의 토픽을 생성할지에 대한 기준이 필요하다. 토픽의 개수는 사전에 주어지지 않기 때문에 이를 고려해야 한다. 이는 언어모델의 복잡도(Perplexity) 알고리즘에 의해 최적의 K개를 도출하였다.

위 구현 내용 똑같이 python 'lda' 패키지로만 바꾸어서 구현.
(저번주는 딴 거였는데 이 방식이 훨씬 우리 데이터에 적합하길래 바꿨음.)

텍스트 전처리 - NLTK(Natural Language Toolkit)를 사용

1. **소문자 변환**:
   - 모든 텍스트를 소문자로 변환하여 일관성 유지.
2. **개행 문자 제거 및 다중 공백 처리**:
   - 텍스트 내 개행 문자를 공백으로 대체하고, 여러 개의 공백을 하나의 공백으로 변환.
3. **단어 토큰화**:
   - 텍스트를 단어 단위로 분리.
4. **알파벳 문자만 남기기**:
   - 알파벳이 아닌 문자는 제거.
5. **불용어 제거**:
   - `stopwords` 라이브러리와 추가 지정된 단어를 불용어 목록에 추가하여 제거.
6. **품사 태깅 및 표제어 추출**:
   - `pos_tag` 함수를 사용하여 단어에 품사 태그를 달고, `WordNetLemmatizer`를 사용하여 표제어 추출(lemmatization)을 수행.

Topics(K) 선정
Perplexity(복잡도) 기준으로 선정
(1000회 반복 샘플링기준) -> 이거 교수님 LDA 5000회로 하신 것 같던데 이유가 있으신가?



K에 따른 복잡도 지표(우선적으로 5단위로 뽑았음)
-> 급격하게 감소 후 완만해지는 K = 20으로 선정(애매하지만 추후 촘촘하게 해볼 예정)

그렇게 해서 나온 토픽은

Topic 0: system method embodiment present invention various use example disclosure provide
Topic 1: information device terminal location data mobile tag communication tracking identification
Topic 2: item storage system delivery package carrier station transfer location configured
Topic 3: network data packet message traffic address protocol destination information host
Topic 4: container cargo system shipping load door includes trailer within transport

Topic 5: data object image model based process set using processing input
Topic 6: power unit device port circuit control electrical output signal includes
Topic 7: data system content digital storage server application block file code
Topic 8: method least includes plurality data based receiving determining associated information
Topic 9: vessel fluid gas system flow pressure liquid chamber temperature air
Topic 10: signal transmission wireless communication channel radio base station frequency antenna
Topic 11: sensor data configured system module processor least controller control plurality
Topic 12: device network communication wireless access user server connection service computing
Topic 13: end assembly includes portion least member surface structure side body
Topic 14: event time monitoring security value condition detection state based system
Topic 15: vessel system ship marine control position water target sonar method
Topic 16: network node communication link path plurality mobile method satellite moving
Topic 17: vehicle location transportation system transport route information method time traffic
Topic 18: component system element processing mean material module radiation site substrate
Topic 19: product service user system chain information shipping shipment order supply

아직 조금 일반적인 단어들이 많아서 불용어로 처리해야될 것 같은데 어떤 단어 불용어로 뺄지 기준이 의문

,,,
그리고 나온 키워드로 Topic이 어떤 토픽이라고 하기 조금 어려움이 있음...

Table 5. The Results of Topic Modeling

No.	Keyword	Topic
Topic 1	conditions, voaltage, converter, battery, inverter, resonant, controller, charging, control, efficiency	고주파 전력변환제어 기술
Topic 2	energy, harvester, resonant, voltage, creepage, coil, electromagnetic, magnetic, levitation, transmission	전력변환시스템 효율최적화 설계
Topic 3	design, coil, receiver, transfer, magnetic, coupling, transmitter, resonant, battery, PTE	전력변환을 위한 데이터 송신
Topic 4	CPT, circuit, pad, weavingtype, pickup, inductivecoupled, transfer, core, coil, airgap	Z-source 컨버터(전자장치 토폴로지) 설계제어 및 제어기술

요렇게 Topic 1이 어떤 토픽이다가 좀 필요한데

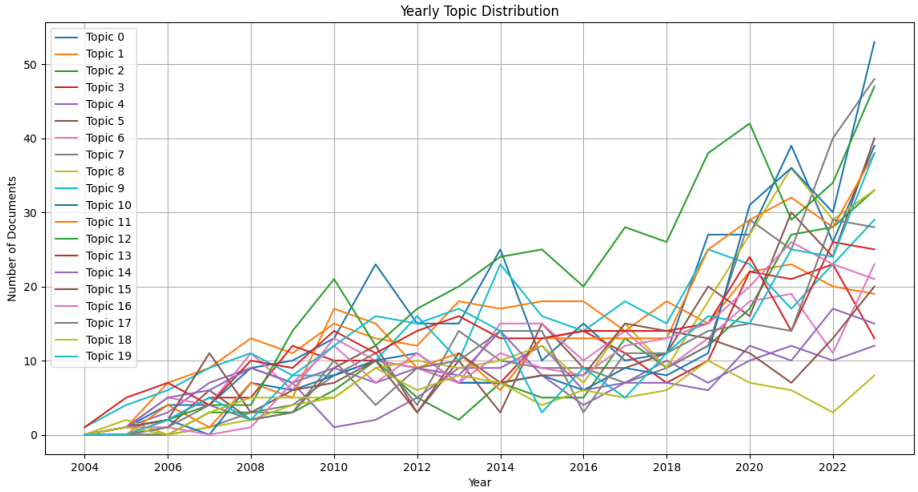
이후,

시계열 분석

과정

1. **LDA 모델로부터 문서별 토픽 할당 추출**
2. **원본 DataFrame에 토픽 할당 추가**
3. **연도별로 데이터 집계**

우선, 나온 토픽들 등록일자 기준으로 연도별 집계가 아래 그래프.



(다만, 연도별 집계하는데 2024.01 데이터가 포함되어 있는데 이걸 2024로 빼버리면 MAPE값에 이상이 생겨서 우선 뺐음.)

이후

Forecasting

교수님 논문에 나온거랑 유사하게 구현

(5) 시계열 분석

본 연구에서는 철도차량용 무선급전시스템과 관련된 연구 논문이 증가하기 시작하는 2008년부터 2018년까지 최근 10년간의 토픽별 시계열 분석을 통해 추세를 확인하였다. 각 토픽의 추이를 파악하기 위해 선형회귀분석을 포함하여 비선형 시계열 분석인 ARIMA, 지수평활법, Naive, STS,[1] BATS,[2] TBATS,[3] Neural network, STL[4] 모형에 대해 분석하였다. 학습데이터(training data)와 테스트데이터(test data)는 8:2로 나누었고 예측력이 가장 좋은 모형을 선택하였다. 토픽별 최적의 모형을 선택하기 위해 모형 적합성과 예측력을 설명할 수 있는 지표인 MAPE를 활용하였고(Lee and Park, 2018), MAPE 값이 50%를 넘는 모형은 예측력이 낮은 것으로 판단하여 제거하였다(Lewis, 1982). 각 토픽별로 MAPE 값이 가장 작은 모형을 최적의 모형으로 선정

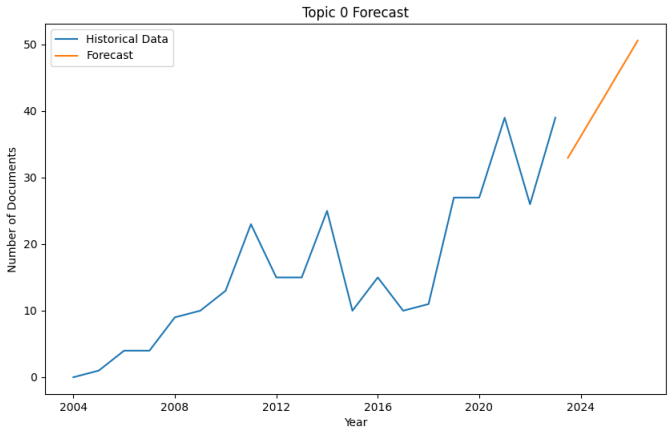
모델 : [LinearRegression, TBATS, SimpleExpSmoothing, ExponentialSmoothing, ARIMA]
(몇 개 더 추가하면 좋을듯? 지수평활 말곤 아직 모델에 대한 이해가 없어서 이것만)
각 토픽에 대해서 여러 모델 평가(MAPE < 50)하고, 가장 적합한 모델을 선택

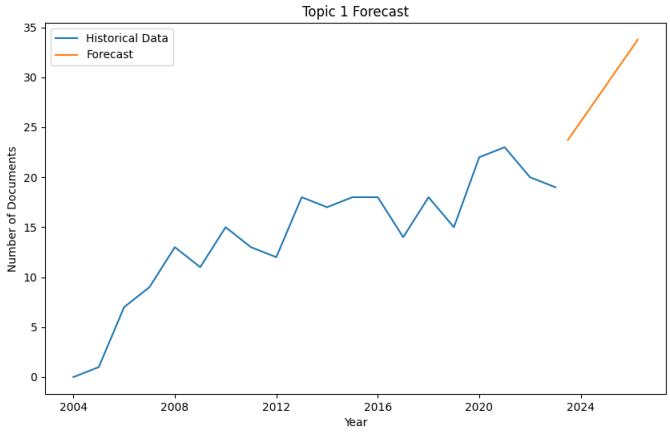
나온 결과

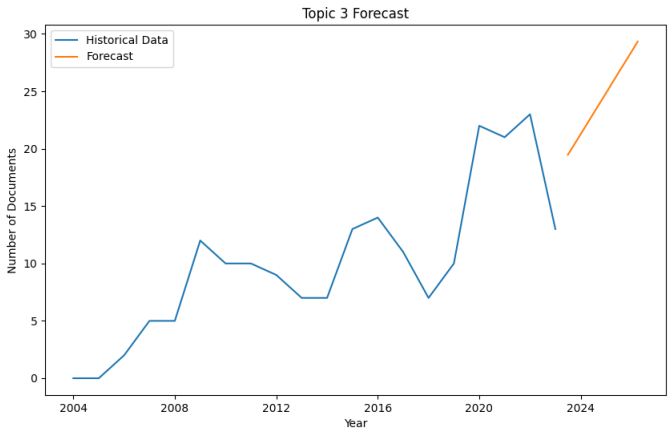
Topic 0: Best Model = LinearRegression, MAPE = 25.915975364504785%
Topic 0 model selected as optimal
Topic 1: Best Model = LinearRegression, MAPE = 12.494371718939293%
Topic 1 model selected as optimal
Topic 2: Best Model = TBATS, MAPE = 51.28630304189012%
Topic 2 model not selected due to high MAPE
Topic 3: Best Model = LinearRegression, MAPE = 31.960186552577852%
Topic 3 model selected as optimal
Topic 4: Best Model = LinearRegression, MAPE = 7.310049019607841%
Topic 4 model selected as optimal
Topic 5: Best Model = LinearRegression, MAPE = 34.12530637254901%
Topic 5 model selected as optimal
Topic 6: Best Model = LinearRegression, MAPE = 26.88938448966717%
Topic 6 model selected as optimal
Topic 7: Best Model = LinearRegression, MAPE = 58.1569620520622%
Topic 7 model not selected due to high MAPE
Topic 8: Best Model = TBATS, MAPE = 41.76347072550263%
Topic 8 model selected as optimal
Topic 9: Best Model = LinearRegression, MAPE = 14.726866670470992%
Topic 9 model selected as optimal
Topic 10: Best Model = LinearRegression, MAPE = 66.15852979962527%
Topic 10 model not selected due to high MAPE
Topic 11: Best Model = ARIMA, MAPE = 29.63158751699522%
Topic 11 model selected as optimal
Topic 12: Best Model = Naive, MAPE = 16.877359793591367%
Topic 12 model selected as optimal
Topic 13: Best Model = LinearRegression, MAPE = 25.970959249084252%
Topic 13 model selected as optimal
Topic 14: Best Model = LinearRegression, MAPE = 26.10802335640138%
Topic 14 model selected as optimal
Topic 15: Best Model = Naive, MAPE = 34.094666405664086%
Topic 15 model selected as optimal
Topic 16: Best Model = LinearRegression, MAPE = 25.57736777821498%
Topic 16 model selected as optimal
Topic 17: Best Model = LinearRegression, MAPE = 24.223727422003282%
Topic 17 model selected as optimal
Topic 18: Best Model = TBATS, MAPE = 47.484879382269696%
Topic 18 model selected as optimal

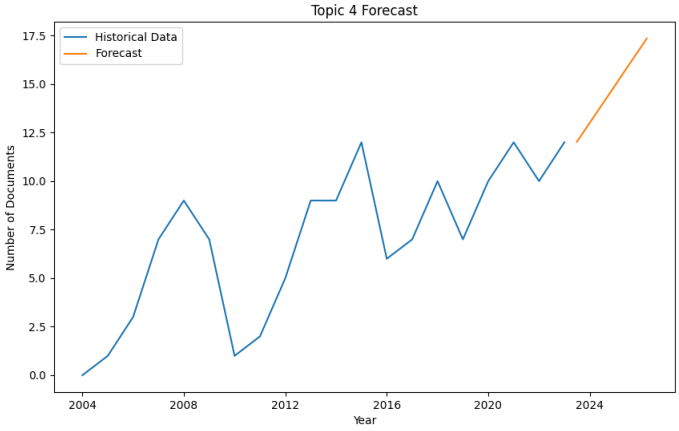
Topic 19: Best Model = LinearRegression, MAPE = 31.966586042311658%
Topic 19 model selected as optimal

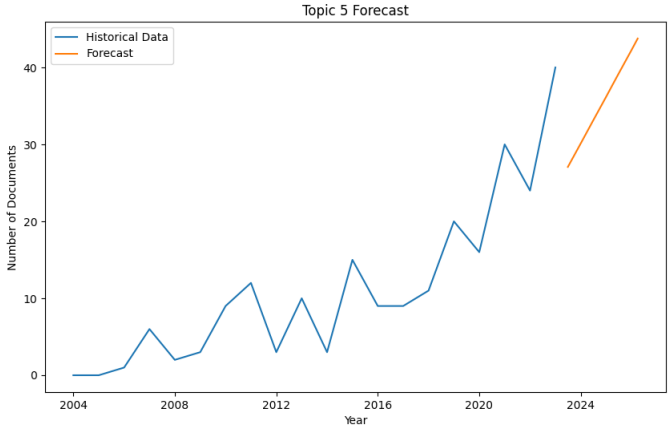
결과에 따라 그린 그림이 아래 그림 시계열 뒤에서 부터

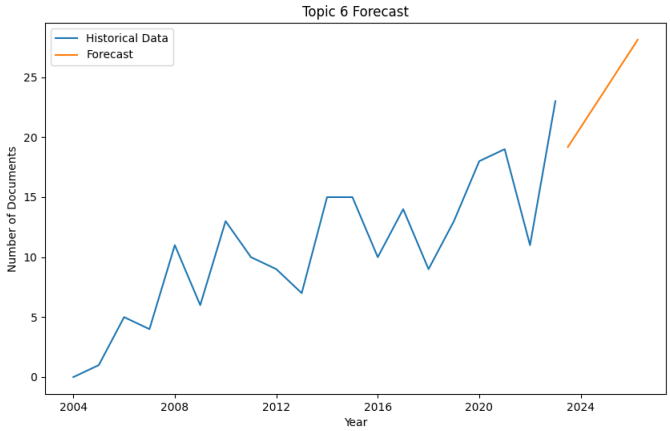


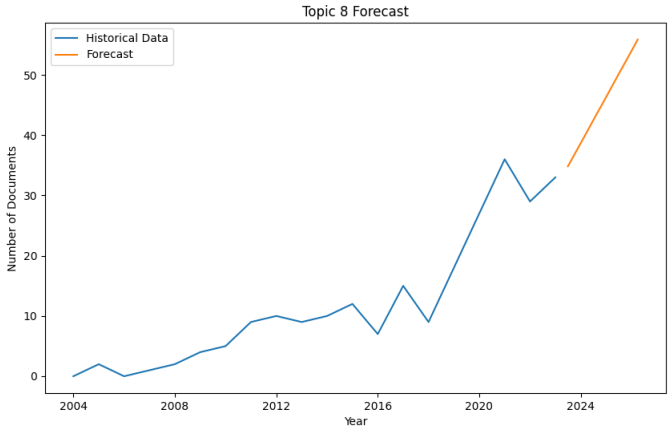


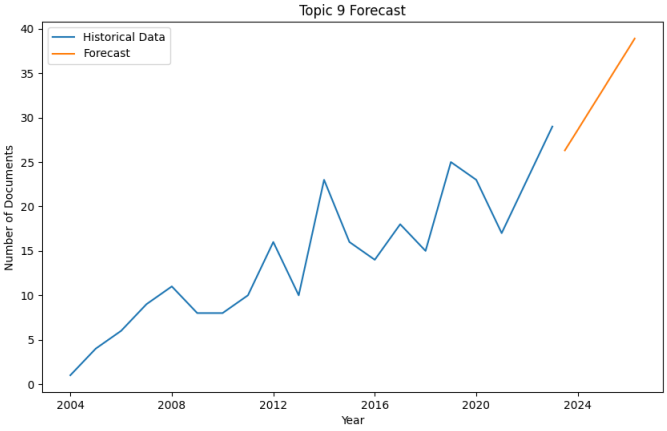


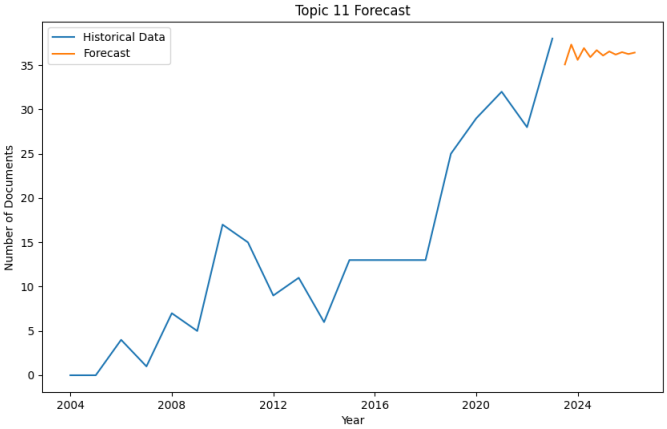

Topic 4 Forecast
Historical Data
Forecast
Number of Documents
Year
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
2004
2008
2012
2016
2020
2024

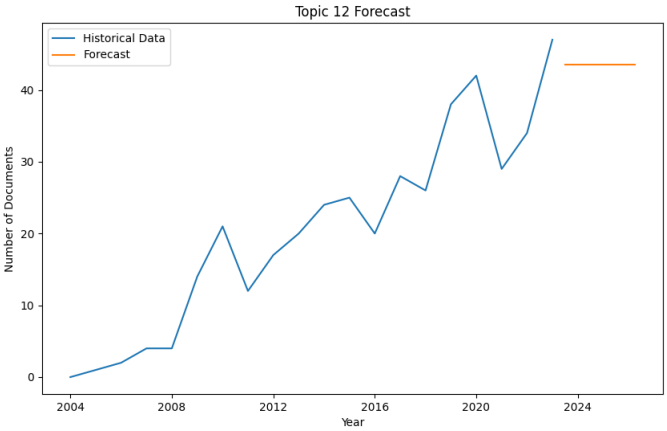

Topic 5 Forecast
Historical Data
Forecast
Number of Documents
Year
0
10
20
30
40
2004
2008
2012
2016
2020
2024

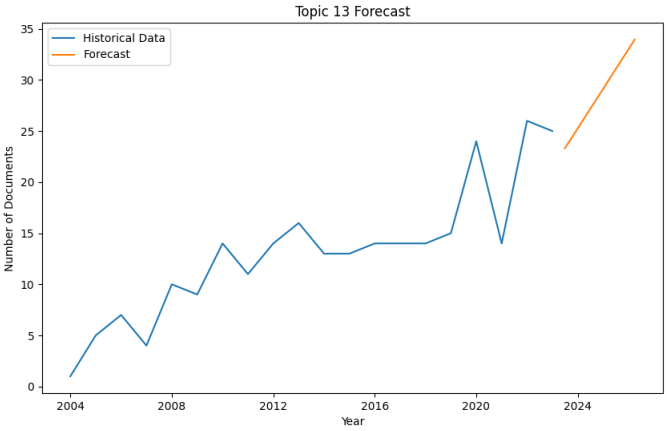

Topic 6 Forecast
Historical Data
Forecast
Number of Documents
Year
0
5
10
15
20
25
2004
2008
2012
2016
2020
2024

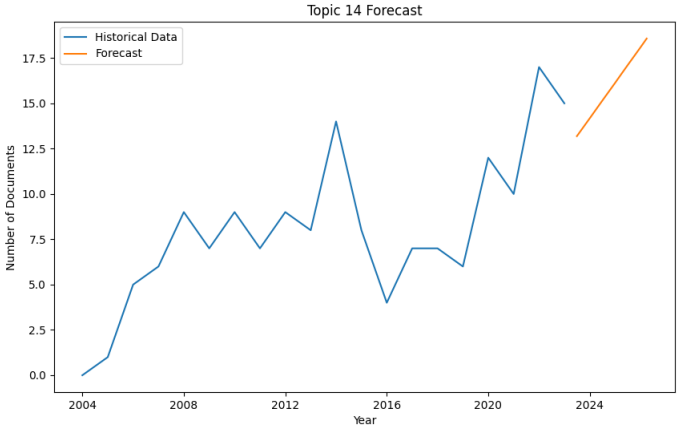

Topic 8 Forecast
Historical Data
Forecast
Number of Documents
Year
0
10
20
30
40
50
2004
2008
2012
2016
2020
2024

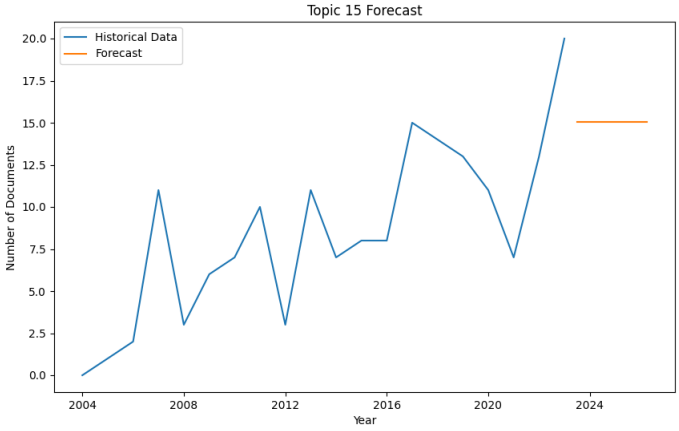

Topic 9 Forecast
Historical Data
Forecast
Number of Documents
Year
0
5
10
15
20
25
30
35
40
2004
2008
2012
2016
2020
2024

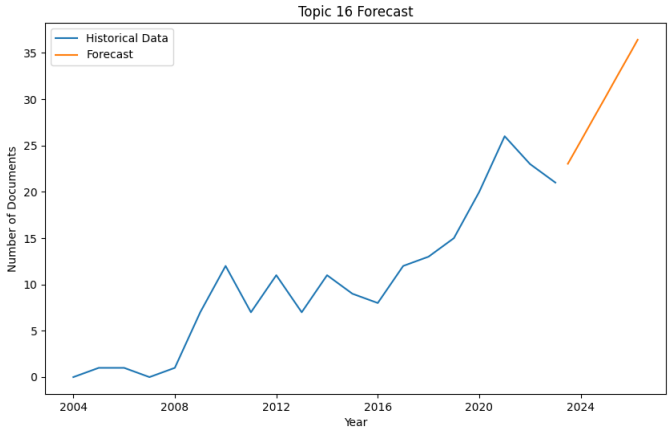

Topic 11 Forecast
Historical Data
Forecast
Number of Documents
Year
0
5
10
15
20
25
30
35
2004
2008
2012
2016
2020
2024

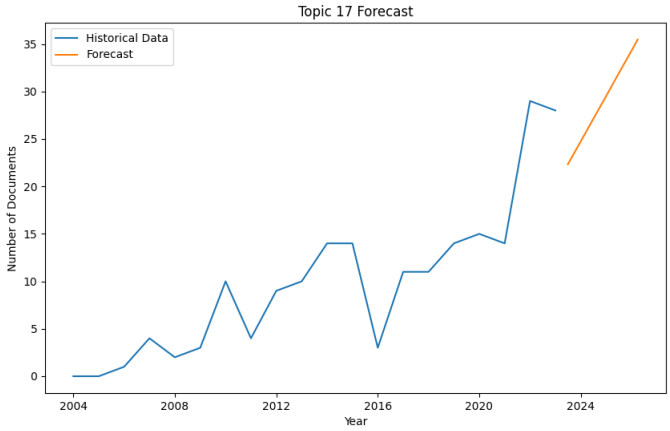

Topic 12 Forecast
Historical Data
Forecast
Number of Documents
Year
0
10
20
30
40
2004
2008
2012
2016
2020
2024

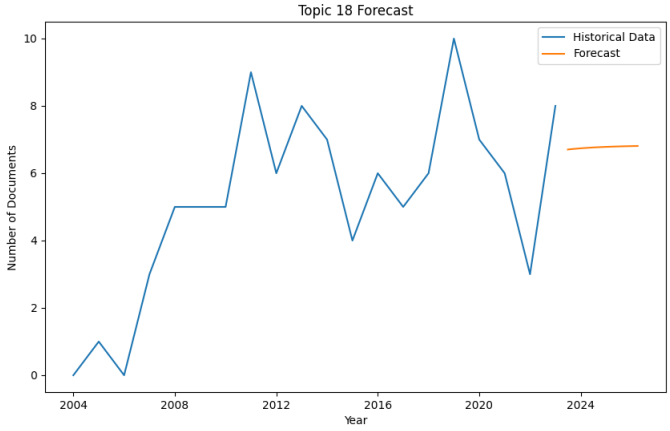

Topic 13 Forecast
Historical Data
Forecast
Number of Documents
Year
0
5
10
15
20
25
30
35
2004
2008
2012
2016
2020
2024

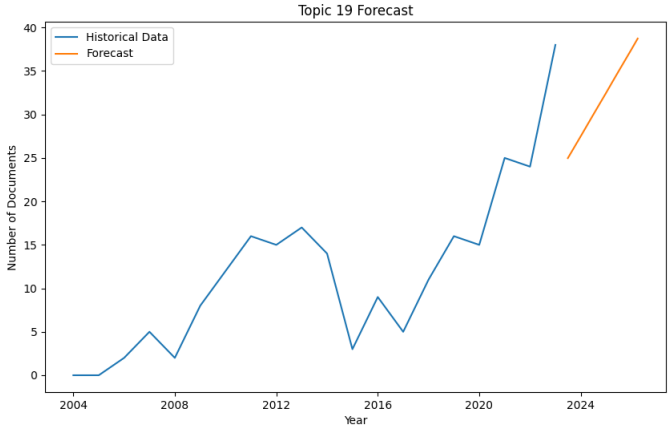

Topic 14 Forecast
Historical Data
Forecast
Number of Documents
Year
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
2004
2008
2012
2016
2020
2024


Topic 15 Forecast
Historical Data
Forecast
Number of Documents
Year
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
2004
2008
2012
2016
2020
2024


Topic 16 Forecast
Historical Data
Forecast
Number of Documents
Year
0
5
10
15
20
25
30
35
2004
2008
2012
2016
2020
2024


Topic 17 Forecast
Historical Data
Forecast
Number of Documents
Year
0
5
10
15
20
25
30
35
2004
2008
2012
2016
2020
2024



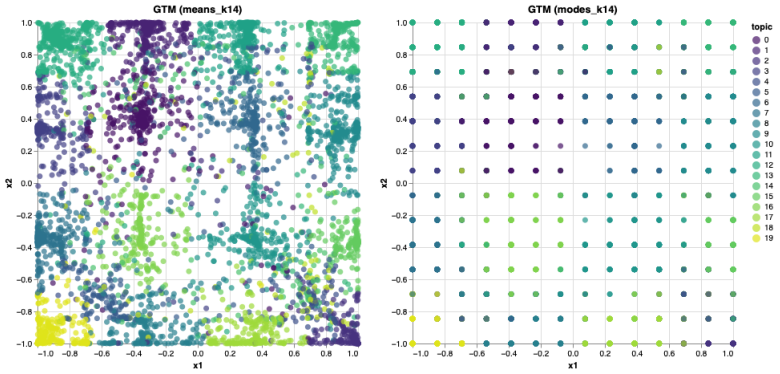


LinearRegression 가 왜 이렇게 많이 적합한 모델로 선정됐는지는 모르겠지만,
모델 더 추가해서 할 예정
그리고 MAPE 값 50%보다 낮은 애들이 생각보다 많네??

GTM

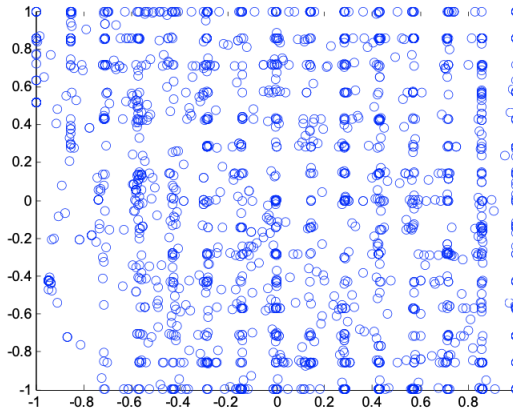
저번주까지 GTM 나온건 제대로 된게 아니였음.

과정 : LDA 분포 추출해서 -> 차원 축소 -> GTM 모델 생성 및 학습 -> Altair 사용한 시각화

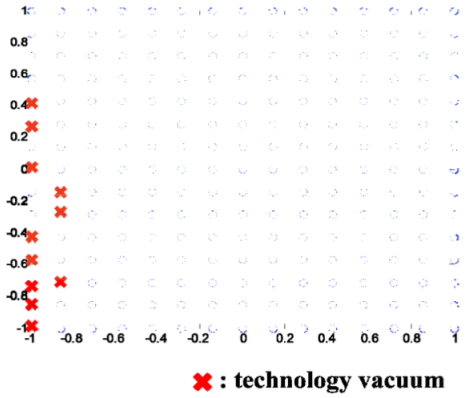


여기서 이제 시각적으로 데이터 포인터가 밀집되지 않은 영역을 찾아 내서 공백 영역 식별하면 되나??

근데 뭐 이걸 눈으로 안찾고 수치로 찾는 논문도 있긴하던데



[Figure 6] The posterior-mean projection of the data in the latent space



[Figure 7] The posterior-mode projection of the data in the latent space

5) Identifying Technology Vacuum

Patent vacuums which are expressed as "blank" areas in the map are discovered from the Fig. 6 and 7. Since all data points are mapped at each latent grid in the posterior-mode projection, the patent vacuums are discovered more clearly than the posterior-mean projection. That is, each 'o' represents a keyword vector is mapped at one of the latent points in the posterior-mode projection and the sparse latent points like a left-low part of the GTM-based patent map are the patent vacuums. In this example figures, 11 patent vacuums of 225 latent points are discovered.

As mentioned above, the main advantage of GTM is inverse mapping. The characteristics of inverse mapping, which differentiate GTM from other latent variable models which only conduct the projection from the D-dimension data space onto a two-dimension space, enables the projection from the latent space into the data space [3, 4]. Consequently, technology vacuums are identified by inversely mapping ((5) in section 2) patent vacuums in latent space into new vectors in data space as shown in Fig. 8(a). The keyword vector fields of technology vacuums are filled as '0' or '1' by threshold value, so the value '1' implies that the value is over threshold value determined by analyst as shown in Fig. 8(b).

	A	B	C	D	E	F	G	H
1		immersion	apparatus	exposure	radiation	reticle	euv	nersion liq
2	1	0.312653	0.028165	-0.03878	0.014079	0.395969	0.897664	0.07359
3	2	0.009104	0.045167	-0.15467	0.066727	0.261174	0.548218	-0.00273
4	3	-0.04357	0.054173	-0.0653	-0.0065	0.102163	0.273198	-0.02141
5	4	0.034636	0.051592	0.217153	-0.14225	0.010487	0.081197	-0.00882
6	5	0.119057	0.049556	0.595054	-0.19825	0.005953	-0.0278	0.007563
7	6	0.139787	0.060798	0.928646	-0.06245	0.048925	-0.06657	0.011834
8	7	0.100183	0.081655	1.100652	0.259019	0.080541	-0.05831	0.004095
9	8	0.047642	0.089067	1.067726	0.633398	0.065612	-0.02883	-0.00543
10	9	0.023448	0.058518	0.871254	0.878258	0.013271	0.001436	-0.00764
11	10	0.030724	-0.00833	0.604217	0.87304	-0.03444	0.020991	-0.00255
12	11	0.0413	-0.07028	0.357119	0.632185	-0.03559	0.026045	0.001983
13	12	0.029357	-0.06505	0.176695	0.29393	0.020331	0.018162	-0.00086
14	13	-0.0006	0.047316	0.059336	0.034692	0.10395	0.003387	-0.00812
15	14	-0.01622	0.247338	-0.02394	-0.0345	0.165694	-0.00855	-0.00683
16	15	0.012652	0.45781	-0.09434	0.074771	0.170045	-0.00822	0.01621
17	16	0.082661	0.590877	-0.14878	0.249025	0.12037	0.006442	0.061074

(a) The result from inverse mapping

Threshold

	A	B	C	D	E	F	G	H
1		immersion	apparatus	exposure	radiation	reticle	euv	nersion liq
2	1	0	0	0	0	0	1	0
3	2	0	0	0	0	0	1	0
4	3	0	0	0	0	0	0	0
5	4	0	0	0	0	0	0	0
6	5	0	0	1	0	0	0	0
7	6	0	0	1	0	0	0	0
8	7	0	0	1	0	0	0	0
9	8	0	0	1	1	0	0	0
10	9	0	0	1	1	0	0	0
11	10	0	0	1	1	0	0	0
12	11	0	0	0	1	0	0	0
13	12	0	0	0	0	0	0	0
14	13	0	0	0	0	0	0	0
15	14	0	0	0	0	0	0	0
16	15	0	0	0	0	0	0	0
17	16	0	1	0	0	0	0	0

(b) After applying threshold

[Figure 8] An example of inverse mapping

If a technology vacuum has two '1' values in its 2nd and 10th keyword vector fields, this means this technology vacuum is consisted of two keywords, keyword 2 and keyword 10. In fact, since there is no definitive method to determine the threshold value, it is determined depending on the purpose of research. That is, if threshold value is low, identified technology vacuums are consisted of many keywords.

Technology vacuums consisting of few keywords can be excluded since the information is not rich enough to be considered as the potential new technologies. Finally, only technology vacuums having at least 2 or more keywords are selected as shown in Table 3. Each technology vacuum including several keywords can be an alternative of future new technology development. Although technology vacuums are selected according to above processes, the result should be validated by experts in related technology area.