BERTopic

먼저, 예전에 비해 뭐가 바뀌였냐
=> 엑셀 데이터를 정제하면서 수가 줄어들면서
=> 알아서 BERTopic이 불용어 제거해주던거에 조금 문제가 생김
(문서 수가 부족해서 그런지 확실하게 제거를 못해줌)
=> 그래서 LDA에서 했던 것과 같이 데이터 전처리를 해주었는데
=> LDA 처럼 수백개 불용어 제거 안했는데도 나름 준수한 토픽들을 보여줌.
=> 우선 알아서 얘 내부적으로 알고리즘 돌려서 정해진 **토픽 수(K)**대로 뽑았음.

뽑은 토픽은 아래와 같구요

```
Topic -1: data network system device information include method communication vehicle provide
Topic 0: fluid gas vessel liquid chamber pressure flow sample heat temperature
Topic 1: container door lock cargo seal security sensor monitor monitoring open
Topic 2: container conveyor assembly end cargo support rail handle move arm
Topic 3: image object camera pixel capture generate medium neural digital processor
Topic 4: security endpoint network attack traffic malicious policy application packet program
Topic 5: vehicle traffic control transportation system information plan passenger method signal
Topic 6: chain supply inventory order fulfillment item quantity good supplier logistics
Topic 7: memory instruction process operation simulation program representation control execute data
Topic 8: storage data cloud database backup distribute transfer store process compute
Topic 9: rfid tag reader rf radio identification frequency sensor antenna read
Topic 10: shipping ship shipment package carrier label delivery shipper international item
Topic 11: device key access identity server authentication guest authenticate request certificate
Topic 12: vessel ship marine maritime sail voyage route information hazard position
Topic 13: marine propulsion vessel control sail trim helm position steer thrust
Topic 14: wireless access device communication point network connection ap station interface
Topic 15: power charge battery electrical energy signal electric soc receptacle analog
Topic 16: optical network reflectometry link wavelength dwdm node optic signal fiber
Topic 17: switch packet protocol port fabric tcp host network server message
Topic 18: item delivery sortation inventory sort container order package station deliver
Topic 19: travel route transportation delay road leg location vehicle least transport
Topic 20: call telephone voice conversational user communication protocol party network server
Topic 21: aspect schedule disclosure entity ue resource beam reference slot transmission
Topic 22: sensor monitor analyte condition monitoring signal device environmental data measurement
Topic 23: architectures network move thing communication disclosure support static node aspect
Topic 24: payment transaction user gift purchase merchant information account game product
Topic 25: terminal mobile display information fixture communication icon unit message command
Topic 26: sonar transducer acoustic signal underwater element array hydrophone pulse sound
Topic 27: track gps asset location device wireless receiver acceleration motion position
Topic 28: satellite beam constellation antenna earth communication relay terminal spot access
Topic 29: radiation detector detection neutron source quasistatic dosimeter radioactive detect container
Topic 30: seismic cable rov tow marine survey auv ocean source vessel
Topic 31: basis rrc index station terminal resource transmission channel signal pr
Topic 32: iot device smart thing internet environmental functionality provision data sensor
Topic 33: mesh node network packet mobile wireless message driver device prop
Topic 34: ledger blockchain product distribute block digital edt certificate transaction possession
Topic 35: packet traffic flow network filter application policy scheme receive queue
Topic 36: resonator magnetic power electromagnetic source coil waveguide impedance rf match
Topic 37: blockchain block hash asset data example chain operation node store
Topic 38: electronic connector electrical band audio antenna electrically signal circuit connect
Topic 39: moor hull line mast vessel yoke turret spar element connector
Topic 40: network service connectivity appliance provision accessory skid virtual path client
Topic 41: display screen video power device text content circuit remote control
Topic 42: secure master communication sm key encryption wireless network security controller
Topic 43: node multicast packet rout subnet destination port router path give
Topic 44: restoration path graph network link node inferential mean primary cost
Topic 45: power electric energy supply watercraft motor control ship vessel wind
Topic 46: location strength radio geographical user signal polygon ms mobile event
Topic 47: center humidity cool rack air compute aisle data portable vibration
Topic 48: multipath wireless tsci channel venue series object signal motion receiver
Topic 49: crane container gantry hoist lift handle component transfer elastic tractor
Topic 50: smart technology car vocabulary disclosure education dmrs city resource communication
Topic 51: plan constraint trip set mission order consolidation subcomponent computer transportation
Topic 52: track package movement carrier body td product wan indicia main
Topic 53: serviceable product pick launch embodiment item delivery quality point package
Topic 54: screen object manifest screening share tenant scan character content deployment
Topic 55: radar target mount receiver light cognitive cantilever camera member dispose
Topic 56: circuit collection collector ecu structure noise data plurality pattern collect
```

근데, 이렇게 하면 GTM돌리기에 너무나도 많은 토픽 수에요.
관리하기도 어렵고 분석하기도 복잡.

그래서 `nr_topics` 파라미터를 조정하여 임의로 20개로 만들었습니다.

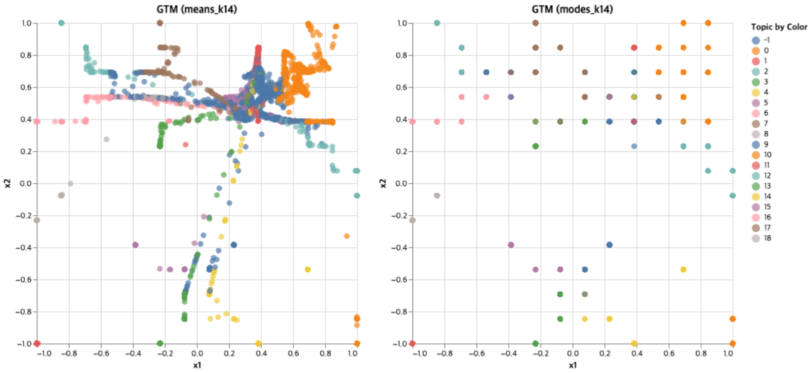
그렇게 해서 나온 토픽은 아래와 같습니다.

```
Topic -1: device system data network include information method communication provide receiv
e
Topic 0: fluid gas vessel liquid chamber pressure flow include sample system
Topic 1: vehicle item order system transportation method include information delivery supply
Topic 2: network node packet optical switch traffic path device data include
Topic 3: wireless signal rfid tag satellite communication device location antenna base
Topic 4: image object display data include system process control generate screen
Topic 5: container door lock cargo sensor radiation security seal monitor system
Topic 6: vessel marine sonar control transducer position sail acoustic signal system
Topic 7: network device iot security key user request endpoint secure access
Topic 8: container assembly support end cargo conveyor member handle rail include
Topic 9: call communication network disclosure aspect resource example telephone architectur
es node
Topic 10: storage data database compute cloud backup system server store process
Topic 11: sensor signal data filter system circuit noise monitor plurality device
Topic 12: power charge energy electrical electric battery system supply control load
Topic 13: blockchain block ledger product hash data distribute node transaction chain
Topic 14: payment user transaction account reward information product gift purchase data
Topic 15: transistor electronic layer connector signal circuit antenna audio band device
Topic 16: resonator power magnetic source electromagnetic coil couple loop least device
Topic 17: terminal mobile information display area unit ue command control challenge
Topic 18: broadcast signal mobile service data digital content video transmission ofdm
```

그런데, 이렇게 하면 GTM에서 문제가 좀 생깁니다.

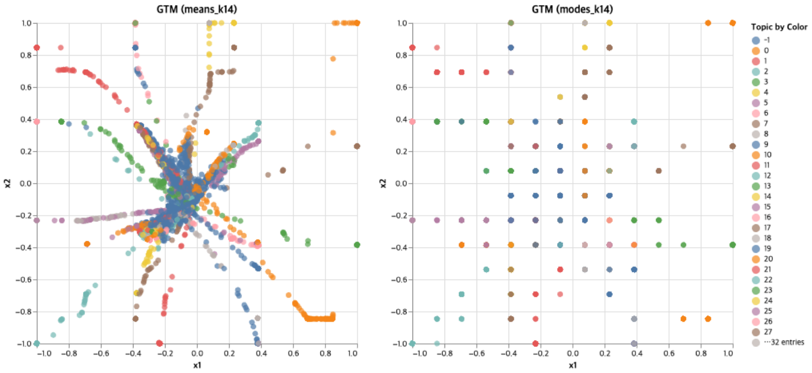
GTM

아래 그래프가 토픽 수 20으로 했을 때의 GTM.

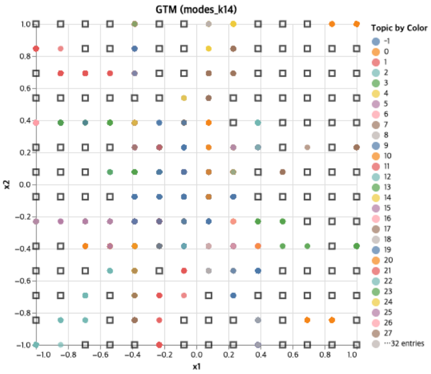


LDA가 단어 빈도를 기반한 토픽 분포를 사용하여 GTM을 실행하지만
BERTopic은 **문서, 문장의 임베딩을 기반으로 실행함. 때문에 좀 더 군집화된 형태**를 띄는데
이로인해 데이터가 무진장 많지 않다면 **공백 영역이 너무나도 많이 생겨나게 되버립니다.**

우선, 그래도 BERTopic을 통해 얻은 확률 분포를 GTM에 넣어 시험삼아 해보려고 BERTopic 자체 알고리즘대로 정한 **토픽 수(57개)** 기준으로 해서 돌려봤습니다.

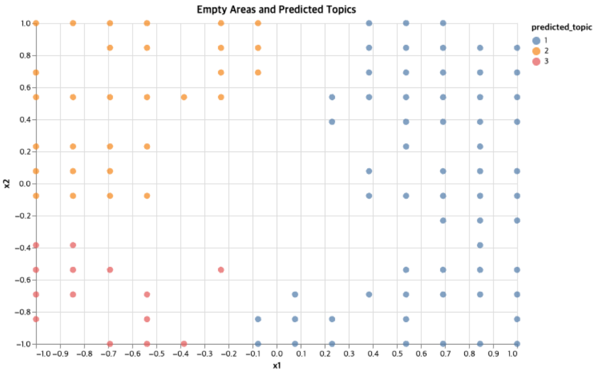


그렇게해서 나온 GTM 그래프이구요. 아까 LDA 때의 설명과 마찬가지로



빈 영역 따왔구, 이번엔 그냥 네모로 표시했어요.

그렇게 해서 가우시안 함수(LDA 해봤을때 이게 가장 잘나와서) 통해서 역매핑했고
나온 결과는 이렇게 나왔습니다.



보기쉬우려고 제일 확률 높은 Topic만 찝어서 시각화 했는데 두드러지게 1, 2, 3 토픽 영역별로
나왔구여

근데, 역시 공백 영역이 너무 많기도 하고, 토픽이 너무 많기도하고 공백 기술을 잘설명한다고 보기에 어려움이 있을 것 같습니다.

BERTopic을 통해서 GTM 하려면 데이터 수가 훨씬 더 많은 경우이어야 적용하기 적합할 것 같다는 생각이 듭니다.

이러한 결과를 봤을 때, 확실히 토픽 자체로만 보면 BERTopic이 LDA 보다 좋은 것 같은데, GTM 때문에 LDA를 택해야되지 않나싶습니다.
-> 이렇게 발표하고 교수님 의견 듣는게 좋을 듯 합니다(?)