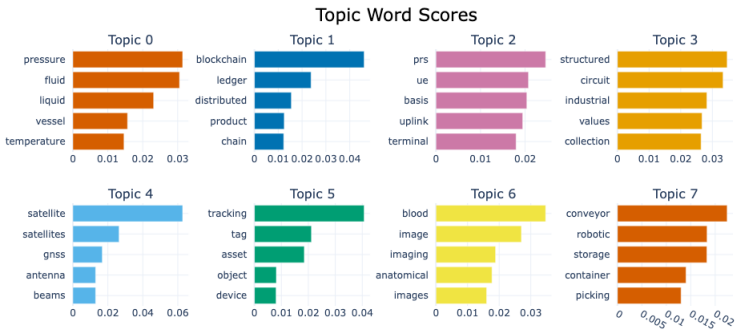
- **상황**
  - 인영이가 올려준 abc.xlsx 사용 (IPC분류만 된 12000여개)
  - '요약'열 사용함.
  - 문서 임베딩 과정에 `sentence-transformers/all-MiniLM-L6-v2` 추가. (좀 더 효율 + 최적화)
  - 데이터 전처리(조사, 관사, 불용어 제거)는 안했음.

- **결과**

```
     Topic  Count                                          Name  \
0       -1   4630                                 -1_the_to_and_of
1        0    201                  0_pressure_fluid_liquid_vessel
2        1    136       1_blockchain_ledger_distributed_product
3        2    131                           2_prs_ue_basis_uplink
4        3    126        3_structured_circuit_industrial_values
..     ...    ...                                            ...
264    263     10           263_tracking_load_tracker_connector
265    264     10  264_terminal_sources_comparison_multiplicity
266    265     10             265_replenishment_store_spoke_pick
267    266     10          266_sensor_gateway_dongle_addressable
268    267     10          267_crop_crops_harvesting_cultivation

                                        Representation  \
0      [the, to, and, of, for, is, or, an, device, in]
1    [pressure, fluid, liquid, vessel, temperature,...
2    [blockchain, ledger, distributed, product, cha...
3    [prs, ue, basis, uplink, terminal, sidelink, r...
4    [structured, circuit, industrial, values, coll...
..                                                 ...
264  [tracking, load, tracker, connector, seal, att...
265  [terminal, sources, comparison, multiplicity, ...
266  [replenishment, store, spoke, pick, fulfilled,...
267  [sensor, gateway, dongle, addressable, portion...
268  [crop, crops, harvesting, cultivation, chamber...

                                    Representative_Docs
0    [A computationally implemented system and meth...
1    [A hydronic system and method of use that will...
2    [A vehicle for use with a blockchain-based tra...
3    [Disclosed are a method and apparatus for tran...
4    [Systems and methods for data collection in an...
..                                                 ...
264  [Methods, apparatus, and systems are provided ...
265  [A network system is provided with a plurality...
266  [Systems, methods, and machine readable medium...
267  [An RF addressable sensor network includes one...
268  [A harvesting supply chain logistics system fo...

[269 rows x 5 columns]
```

(Topic -1 '잡음'으로 **the, to, and, of** 등등 다 잘 지워졌음.)

Topic 0 ~ 7까지 그래프로 보면 (이게 **유사한 문서끼리 묶어서 토픽 뽑아낸게 각각 이렇게 나왔다**는 뜻)

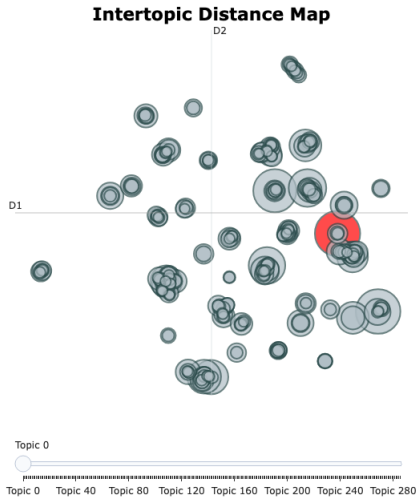


.

.

그리고 아래 **Distance Map**은 토픽 간의 유사성을 시각적으로 표현한 것. (가까울수록 토픽이 유사. 멀리 떨어져 있으면 차이 큼.)

- 빨간 원이 Topic 0



Topic 0만 자세히 보면

- **추출된 토픽**
  ('fluid'), ('pressure'), ('liquid'), ('vessel'), ('temperature'), ('valve'), ('gas'), ('conduit'), ('flow'), ('dispensing',)
- 그리고 해당 토픽 관련 문서(이 '요약'으로부터 추출된 토픽일 확률이 높은 '요약'을 선별한 것)
  - (번역)
    - 1. **로봇 기반 증기 장치에 관한 방법**: 공기를 로봇 증기 장치로 끌어들여, 끌어들인 공기를 센서에 노출시켜 공기 중 하나 이상의 성분을 감지합니다. 센서를 통해 그 성분들의 첫 번째 측정 데이터를 결정하고, 이 데이터를 네트워크를 통해 컴퓨팅 장치로 전송합니다. 네트워크를 통해 컴퓨팅 장치로부터 두 번째 측정 데이터를 받고, 첫 번째 및 두 번째 측정 데이터를 기반으로 증기화할 하나 이상의 증기화 가능 물질을 결정한 다음, 하나 이상의 증기화 가능 물질로 구성된 증기를 분사합니다.
    - 2. **휴대용 수화 시스템**: 휴대용 컨테이너 내에 액체 또는 다른 용질에 첨가물을 분사하는 기계적 또는 전기기계적 메커니즘을 포함하는 휴대용 수화 시스템입니다. 이러한 첨가물에는 고체, 액체, 분말, 가스가 포함되며, 비타민, 미네랄, 영양 보충제, 의약품 및 기타 소비 가능한 제품이 포함됩니다. 첨가물은 RFID 태그 또는 유사한 것이 장착된 밀폐 용기를 통해 수화 장치에 도입됩니다. 분사는 사용자의 직접 작동, 장치에 의한 자동, 및/또는 사용자 장치의 관련 애플리케이션을 통한 외부 작동에 의해 시작됩니다.
    - 3. **압력 측정 시스템**: 공정 유체로 확장 가능한 압력 감지 프로브와 전기적 특성이 공정 유체의 압력에 따라 변하는 압력 센서를 포함하는 시스템입니다. 광물 절연 케이블은 금속 쉬스를 포함하고, 그 쉬스의 끝은 압력 감지 프로브에 부착되며, 금속 쉬스 내에 여러 도체가 전기적으로 절연된 건조 광물에 의해 서로 간격을 두고 확장됩니다.
    - 4. **프레스티지 액체 저장 용기를 위한 표시 어셈블리**: 압력이 가해진 컨테이너에 연결된 밸브 어셈블리와, 유체 압력에 반응하여 종 방향으로 이동하는 피스톤을 포함합니다. 어셈블리는 피스톤의 종 방향 이동에 반응하여 이동하는 표시 멤버를 더 포함하며, 이는 압력 임계값 아래를 나타내는 첫 번째 조건과 임계값 위를 나타내는 두 번째 조건 사이에서 이동할 수 있습니다.
    - 5. **유체로부터 열 에너지 제어 제거 시스템**: 열전도성 금속 도관 내의 유체로부터 열 에너지를 제어적으로 제거하는 방법, 장치 및 시스템을 제공합니다. 이 시스템은 유체의 흐름을 정지시킬 수 있는 가역적인 플러그의 현장 형성을 허용하며, 특히 도관에 열 유발 응력 균열이나 침범을 유발하지 않습니다. 이 장치와 시스템은 도관을 제어적으로 다시 가열하여 열 유발 응력 균열이나 침범 없이 도관을 통한 유체 흐름을 복원할 수 있습니다.

    ```
    위 문서들을 비롯해 총 201개의 문서들에서 뽑아진 토픽이 합쳐지고 계산되서 위 사진에 나온 Topic 0 토픽들이 된 것. (돌릴때마다 조금씩은 달라짐.)
    ```

    .
    .
    .
    Topic 1, 2, 3, 4 ... 268까지 이렇게 뽑힌거임
    TEST 01보다는 좀 더 명확하게 나오는 듯 !?

추가적으로, 데이터 전처리를 하고도 돌려봤는데,

(`spaCy`라이브러리 활용하여 진행)
품사 필터링과 불용어 제거 : 명사와 형용사에 해당하는 토큰만 선택
표제어 추출 : "running" -> "run" 뽑기
소문자 변환

이렇게 하고 돌리니깐 생각보다 잘나오지 않음.
-> BERTopic이 문장과 문장, 문단과 문단을 읽다보니 **'명사','형용사'** 만 남기는 방식은 잘 맞지는 않는듯??

-> 아직 뭐 좀 더 알아봐야되긴함.