데이터 스토리

대한민국 스타트업 생태계 시각화

[사회 이슈]

작성자 : 한국인사이트연구소 이경현 / 김덕진
E-Mail : khlee@ki.re.kr / socialkim@ki.re.kr
홈페이지 : www.ki.re.kr

스타트업이란 신생 창업기업을 뜻하는 말로 미국 실리콘밸리에서 처음 사용되었고
보통 혁신적인 기술과 아이디어를 보유하고 있지만 자금력이 부족한 경우가 많고,
기술과 인터넷 기반의 회사로 고위험, 고수익, 고성장 가능성을 지니고 있습니다. (출처 : 한경 경제용어사전)
특히, 일반인들의 스타트업에 대한 관심과 긍정적인 인식이 증가 하고 있는데 취업준비생들 설문조사 결과,
스타트업 이직·취업 고려율이 전년조사(23.0%)에 비해 7.5%p 오른 30.5%를 조사되기도 했습니다.(www.venturesquare.net/839574)
스타트업은 시장의 고객들의 불만사항(Pain Point)과 미충족욕구(Unmet Needs)를 기술 뿐 아니라 다양한 비즈니스모델(BM)을 통해 해결해 나가며 혁신해 나가고 있어
기존의 산업분류 체계로 기업들의 특성을 정의하기 어려운 한계점이 존재하기도 합니다.
본 분석에서는 스타트업의 제품과 서비스 소개 비정형 데이터를 활용하여
유사 스타트업 간의 네트워크를 연결한 스타트업 생태계를 시각화하여 산업분류 체계의 한계를 해결할 수 있는 방법에 대해 살펴보고자 합니다.

본 스토리에서는 '디지털 산업혁신 빅데이터 플랫폼'의 스타트업 기업 데이터인 '서비스제품정보'와 '투자유치정보'를 분석하였습니다.
이를 통해 스타트업이 생산 또는 제공하는 제품과 서비스에 대한 소개를 기반으로
기업의 비즈니스와 기술 등을 반영하여 유사한 스타트업의 그룹을 발견할 수 있고,
투자 유치 정보를 결합하여 투자가 활발히 이루어지고 있는 비즈니스와 기술 분야를 살펴 볼 수 있습니다.
또, 스타트업의 제품과 서비스 소개에서 사용되는 단어를 분석하여 유사한 단어를 사용하는 스타트업 간에 네트워크를 연결하고
투자금액에 따라 스타트업의 노드 크기를 다르게 하여 한눈에 스타트업 생태계를 살펴볼 수 있도록 하였습니다.

이를 위해 활용한 데이터 셋은 다음과 같습니다.

구분	원천 데이터셋 링크	비고
서비스제품정보	https://www.bigdata-dx.kr/product/DX062000040001	디지털 산업혁신 빅데이터 플랫폼 제공
투자유치정보	https://www.bigdata-dx.kr/product/DX062000010001	디지털 산업혁신 빅데이터 플랫폼 제공

데이터 내의 서비스 제품정보와 투자유치정보의 전체 필드값은 다음과 같습니다.

구분	필드값
서비스제품정보	제품일련번호, 기업일련번호, 제품국문명, 제품영문명, 한줄소개내용, 제품상세설명, 검색태그내용, 홈페이지url, 앱스토어url, 구글플레이 url
투자유치정보	투자일련번호, 기업일련번호, 투자일자, 투자단계명, 투자금액, 기업가치금액, 투자유치비고, 보도자료발행사명, 보도자료제목, 보도자료 url, 투자기관번호

분석을 위해 먼저 데이터 통합을 진행하였습니다, 서비스제품정보와 투자유치정보 데이터를 기업 일련번호를 기준으로 하나의 데이터셋으로 병합하였고,
한줄 소개내용과 투자일자, 투자금액의 필드 값을 활용하여 분석을 진행하였습니다. 분석을 위해 사용된 데이터 필드값과 데이터는 다음과 같이 구성되어 있습니다.

구분	필드값
기업일련번호	스타트업을 구분하기 위해 기업마다 숫자로 고유의 일련번호를 부여
한줄소개내용	스타트업이 생산 또는 제공하고 있는 제품과 서비스에 대한 한 줄 소개 정보
투자일자	스타트업이 투자를 받은 일자를 연도, 월, 일순으로 8자로 표기
투자금액	스타트업이 투자자로부터 투자받은 금액(단위: 원)

기업일련번호	한줄소개내용	투자일자	투자금액
135105	젊은 연구자들의 지식 커뮤니티	20191117	100000000
100807	::Our mission is to organize information and make it connectable and useful resources	20190130	500000000
149535	깨끗한 물을 제공하고	20210317	520000000
130223	GS RETAIL 통합커머스 플랫폼	20210604	500000
20640	실시간 줌 강좌 컨텐츠와 라이브 커머스 BM 개발	20210602	
124965	블록체인 공증 기능과 보상 시스템 등이 내재된 강력한 콘텐츠 매니지먼트 솔루션	20210604	
136103	최대 12명 까지 다인 화상이 가능한 화상 모임 어플리케이션입니다.	20210128	200000000
149838	자동차 매매 플랫폼	20210526	1200000000
149832	삼진글로벌넷의 다양한 제품들을 만나보세요	20200407	8000000000
112737	하루만에 내 명품을 전국 딜러에게 비교견적 받고 팔기 App 서비스	20210602	2100000000

[병합된 데이터 일부]

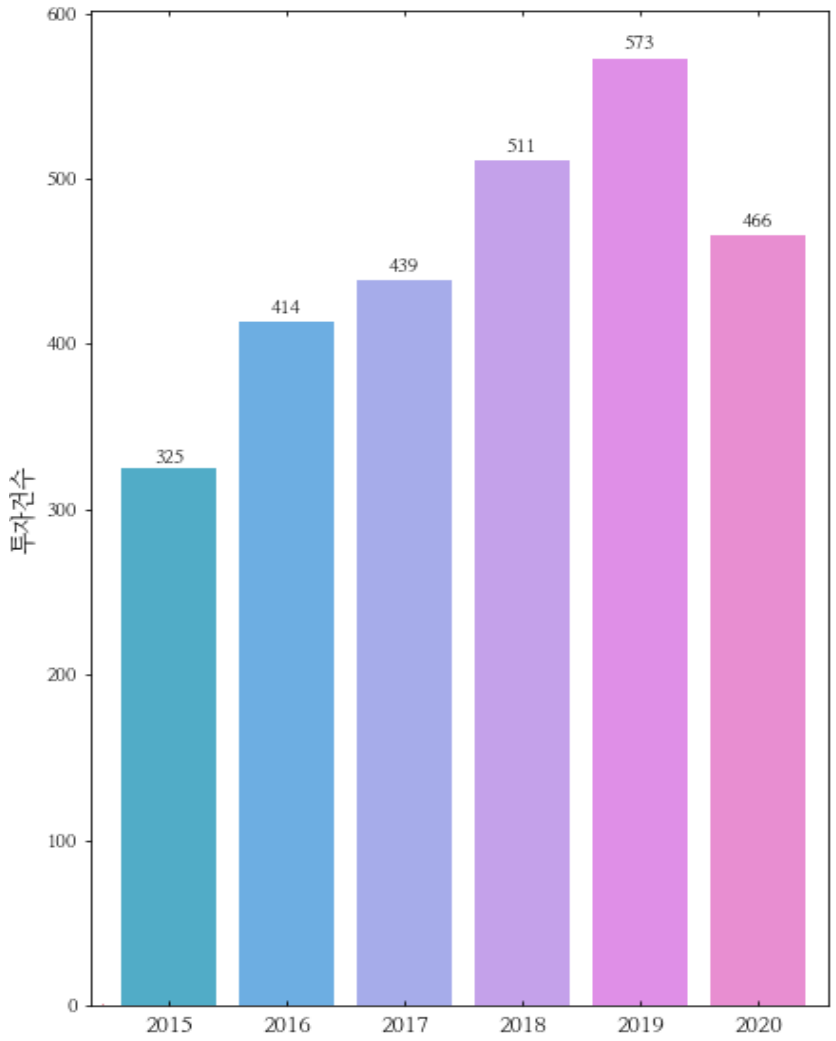
다음 페이지 계속 ↓

[스타트업 투자데이터 기초분석]

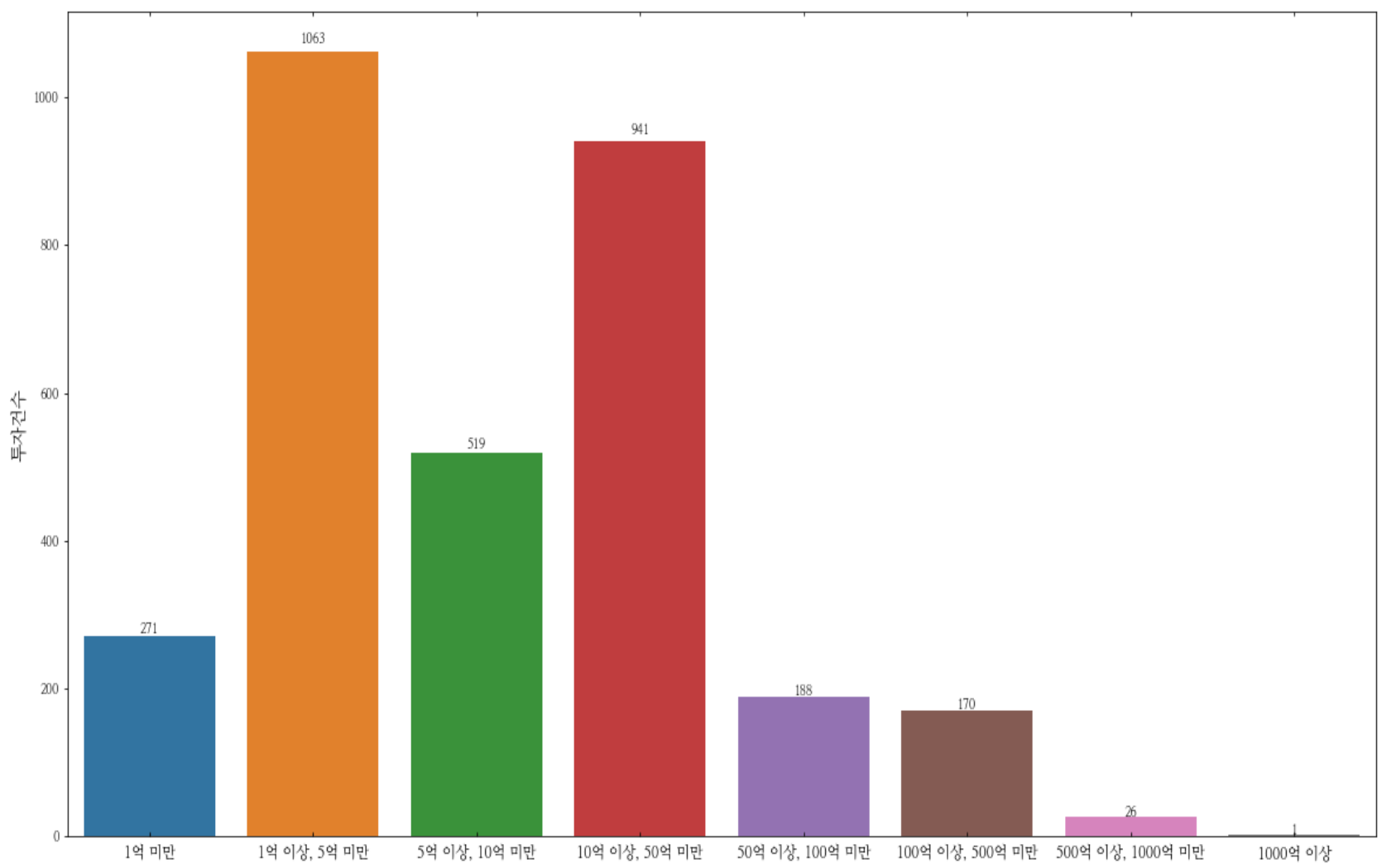
데이터셋을 병합한 후 결측치와 이상치에 대한 가공작업을 진행하였습니다.
한줄 소개내용, 투자일자, 투자금액의 결측치가 있는 데이터는 제거하고 너무 작은 투자 금액과 고액의 투자금액의 이상치를 제거하였습니다.

가공작업을 마친 데이터셋으로 먼저 연도별 투자가 얼마나 활발하게 진행 되었는지 살펴 보았으며,
연도별 투자 건수는 스타트업 투자가 활발하게 이루어 지기 시작한 2015년부터 2020년까지 분석하였습니다.

연도별 투자건수



투자금액 구간별 투자 건수



스타트업 투자는 일반적으로 초기투자금(Seed) – 시리즈(Series A/B/C/D) –기업공개(IPO)/인수합병(M&A) 의 단계를 거치게 됩니다.
초기투자금은 비즈니스의 매우 초기 단계에 집행하는 투자로 보통 친구나 가족의 투자, 엔젤투자자 등으로부터 받게 되고
비즈니스를 확장해 나가면서 벤처 캐피탈로부터 시리즈 투자를 단계적으로 받게 됩니다.
이후 IPO(기업공개)와 M&A를 통해 지분매각(Exit)의 과정을 거치게 됩니다. 각 단계를 거치면서 보통 기업의 가치와 투자금액이 증가하게 됩니다.

연도별 투자 건수를 시각화하면 투자자들의 스타트업에 대한 관심이 얼마나 증가하고 있는지 파악해 볼 수 있습니다.
연도별 투자 건수의 경우 2015년 이후 지속해서 투자 건수가 증가하는 것을 확인할 수 있으며 2020년 투자 건수가 감소하였는데
이는 코로나19 팬데믹 등의 외부적 요인으로 인해 스타트업 투자 시장 역시 얼어붙었기 때문이라고 해석해 볼 수 있습니다.

이어서 투자금액을 금액 구간별로 구분하여 시각화하면 스타트업의 어떤 단계에서 투자가 많이 이루어 지고 있는지 살펴 볼 수 있습니다.
본 분석에서는 2000년부터 2021년까지의 투자금액을 8구간(1억미만, 1억이상~5억미만, 5억이상~10억미만, 10억이상~50억미만, 50억이상~100억미만, 100억이상~500억미만,
500억이상~1,000억미만, 1,000억 이상)으로 구분하여 분석하였습니다.

투자금액 구간별 투자 건수를 살펴보면 1억 이상 50억 미만에 대부분의 투자가 집중된 것을 확인할 수 있었습니다.
이를 스타트업 투자 단계로 살펴보면 Seed에서 시리즈 A 사이의 투자가 대부분 집중된 것을 알 수 있습니다.

[스타트업 생태계 네트워크분석]

스타트업의 특징 중 하나는 다양한 기술과 서비스의 융합, 하드웨어와 소프트웨어의 결합 등을 통해 시장과 고객의 문제를 해결해 나가는 혁신성이라고 할 수 있습니다.
이러한 이유로 기존 산업분류를 통해 스타트업을 정의하고 유사한 기업을 그룹화하는 것은 거의 불가능에 가깝습니다.
산업분류로는 그들의 변화와 문제해결 영역을 정의하기 힘들기 때문이지요.
이러한 문제점을 해결하기 위해 스타트업의 제품과 서비스 소개 데이터의 주요 단어를 통해 유사 기업을 구분하고 이를 네트워크화하여 시각화한다면
기존 산업분류에서 볼 수 없었던 스타트업 생태계를 볼 수 있을 것이라고 생각하여 네트워크 분석을 진행하였습니다.

스타트업 생태계 네트워크 분석은 스타트업이 작성한 자사의 제품과 서비스 소개 텍스트 데이터를 형태소 단위로 분리하여 단어 네트워크를 구성하고
정보 검색과 텍스트 마이닝에서 이용하는 보편적인 가중치인 TF-IDF(Term Frequency – Inverse Document Frequency) 분석을 통해 단어간의 유사관계를 파악하였습니다.
TF-IDF는 어떤 단어가 특정 문서 내에 얼마나 주요하게 출현하는지를 통계적 수치를 이용하여 측정하는 값으로
이를 통해 스타트업의 제품과 서비스 정보 간 유사도를 파악할 수 있으며
이를 기반으로 스타트업 기업간 얼마나 비슷한 제품과 서비스군으로서 묶일 수 있는지 유사도를 정의하는 데 활용하였습니다.

TF-IDF를 구하는 방법을 간단하게 소개하겠습니다. TF-IDF의 값은 TF와 IDF 값을 곱한 값으로 정의되는데요,.

$$tf(t,d) = f(t,d)$$

다음 페이지 계속 ↓

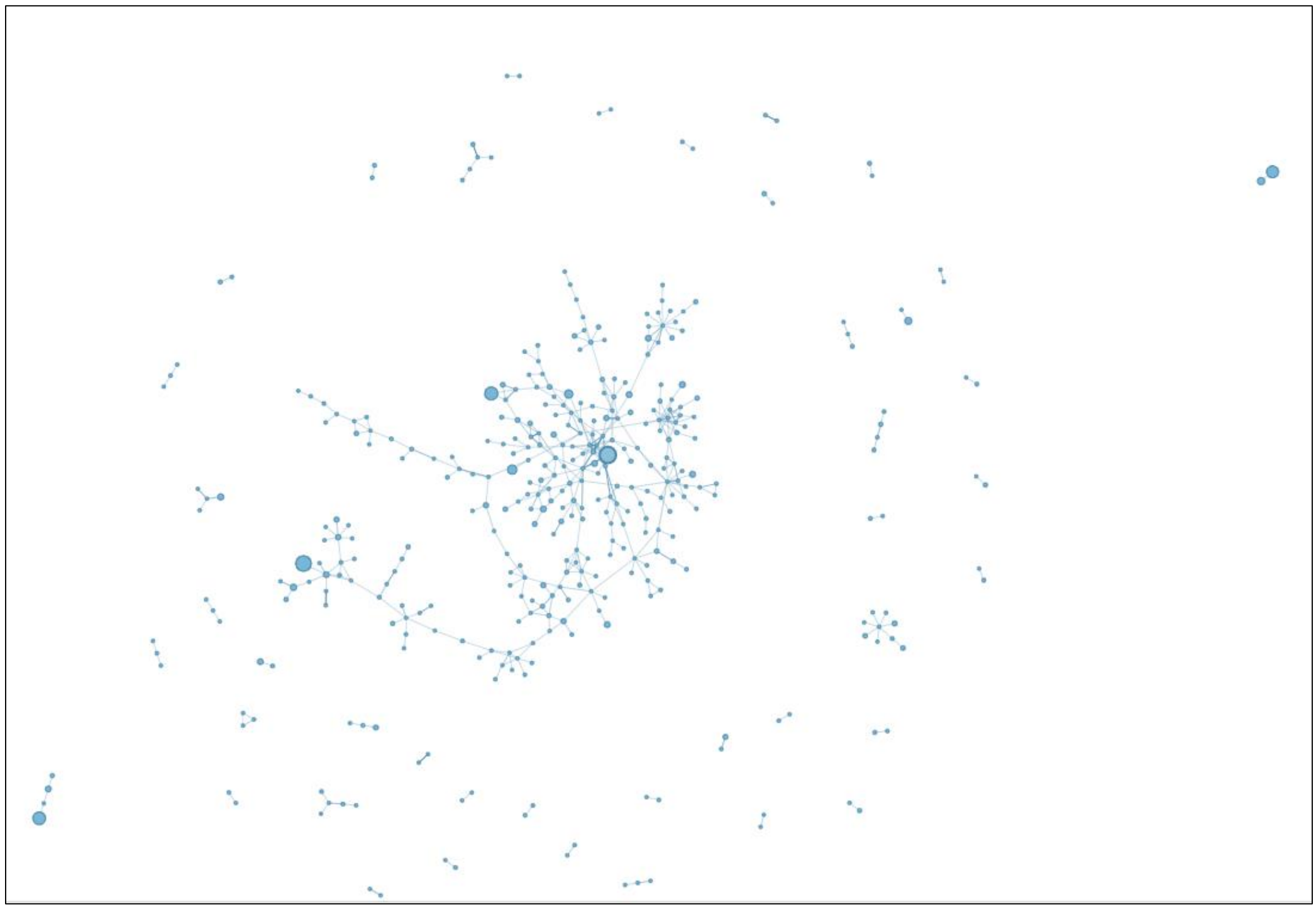
단어 빈도(Term Frequency)는 특정 단어가 문서 내에 얼마나 자주 등장하는지 나타내는 값으로 값이 증가함에 따라 해당 단어의 중요도 또한 증가하게 됩니다.
문서 빈도 (Document Frequency)는 단어가 문서 군 내에서 자주 사용되는 경우 해당 단어가 흔하게 등장하는 것을 의미하는데요,
즉 IDF(Inverse Document Frequency)는 특정 단어 t가 등장한 문서의 수 DF 값의 역수(반비례)로 한 단어가 문서 전체에서 얼마나 공통으로 나타나는지 알 수 있는 수치입니다.

데이터 가공 결과를 이용하여 소셜네트워크 분석에서 각각 분석 단위 사이의 유사도를 측정하기 좋은 방법의 하나인
코사인 유사도(Cosine Similarity)를 통해 시각화하면 아래와 같이 스타트업 생태계 네트워크를 나타낼 수 있습니다.
1) 이는 파이썬을 활용하여 작업하였으며 이때 코사인 유사도를 파이썬으로 코딩시 수식을 0으로 나누는 경우 발생하는 문제가 발생하는데요
이때는 eps를 적용해야 합니다.2)

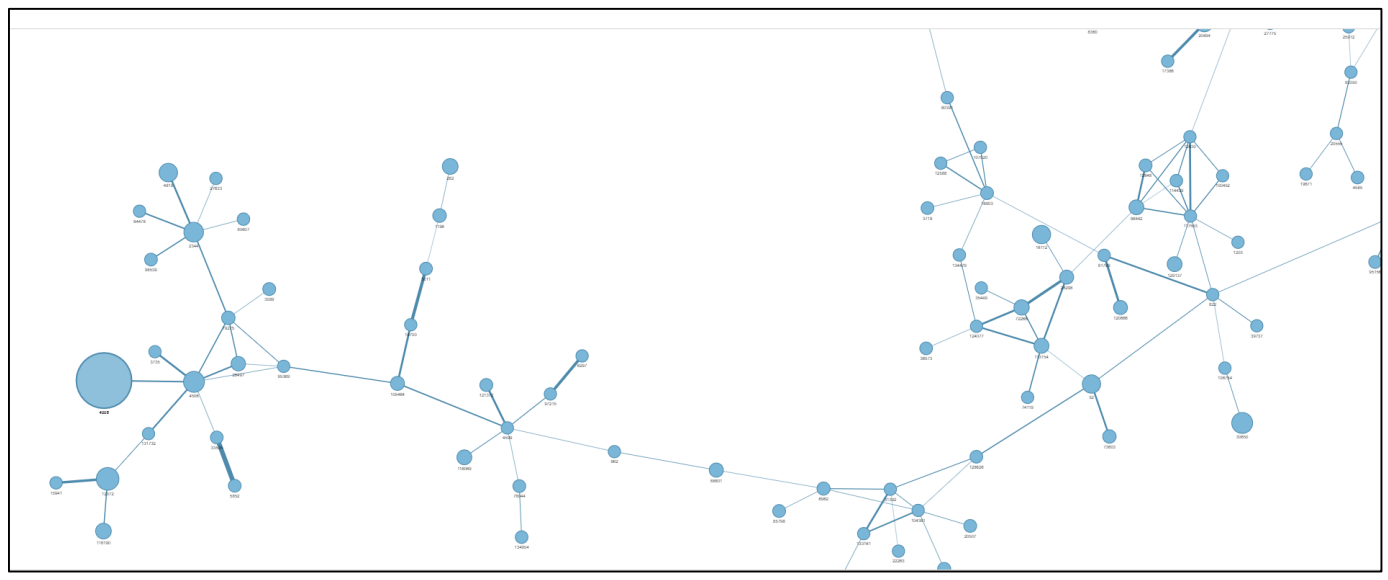
코사인 유사도 코드

```python
import numpy as np
def cos_similarity(x, y, eps=1e-8):
    nx = x / (np.sqrt(np.sum(x ** 2)) + eps)
    ny = y / (np.sqrt(np.sum(y ** 2)) + eps)
    return np.dot(nx, ny)
```

스타트업 생태계 네트워크 맵



(html 파일 별첨)

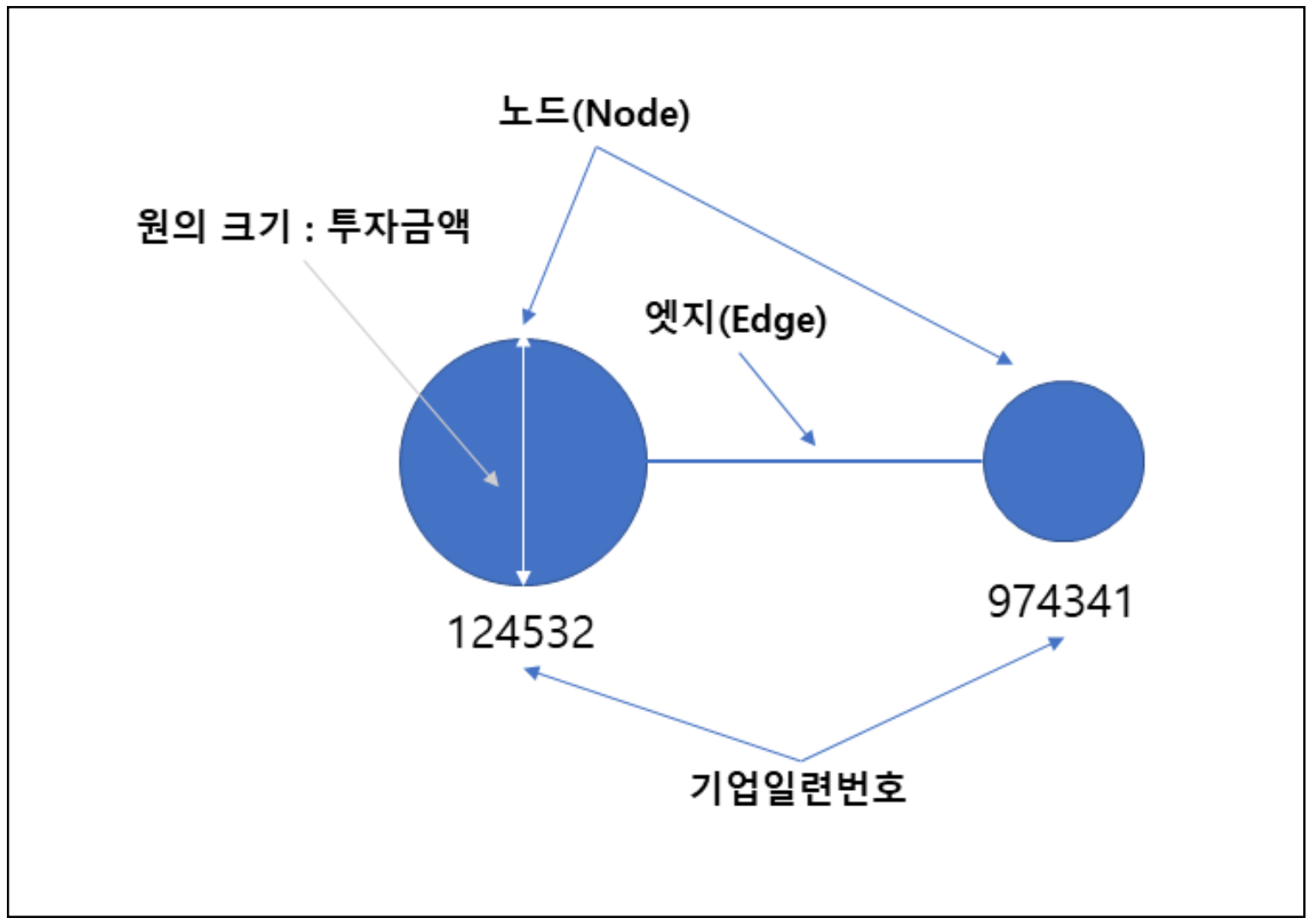


**스타트업 네트워크 맵을 마우스 스크롤을 활용하여
줌인/줌아웃 하여 마우스로 탐색하면 이와 같이
연결된 기업들의 흐름을 자세하게 관찰 할 수 있습니다.**

다음 페이지 계속 ↓

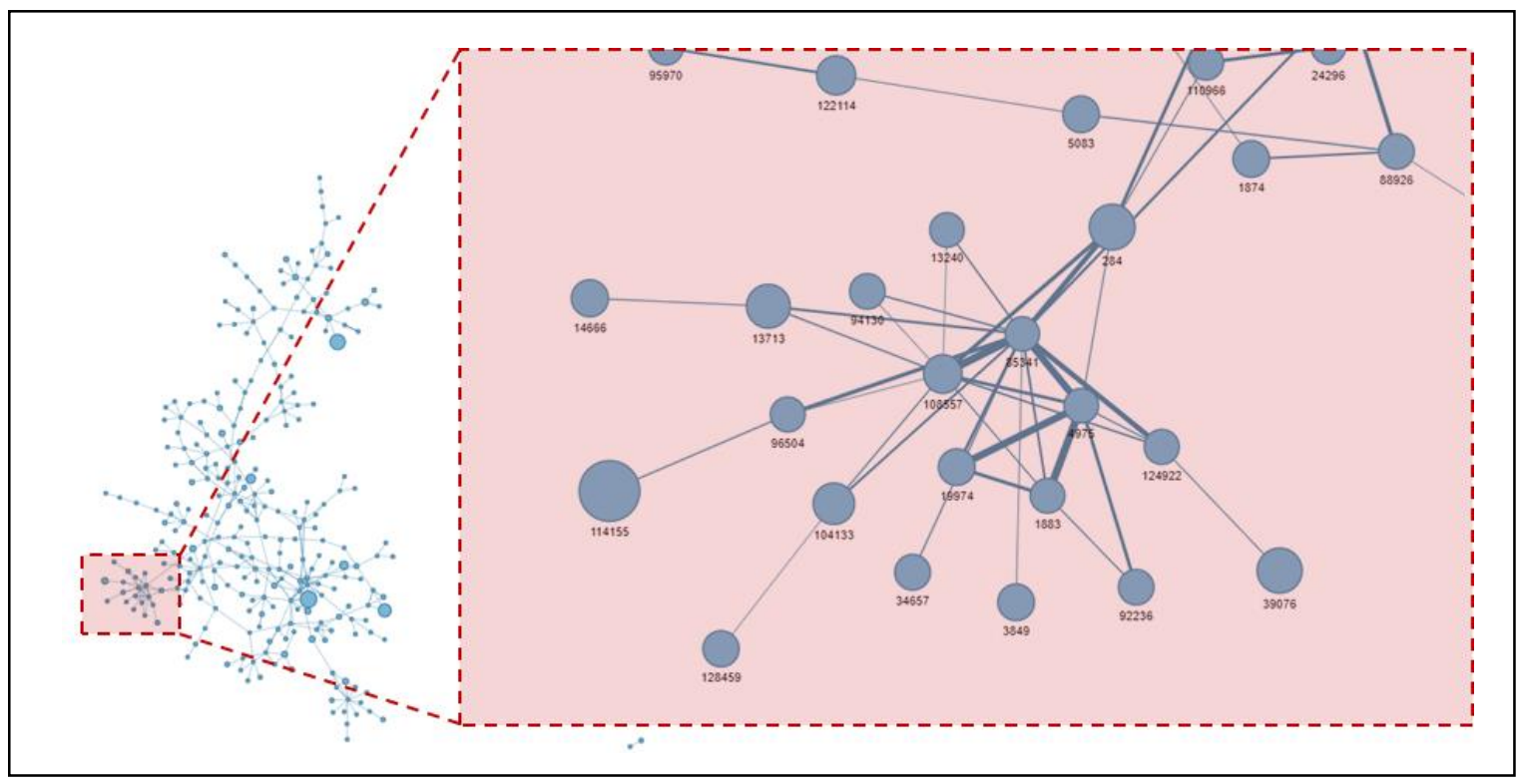
위의 그림이 서비스 제품정보 데이터에 포함되어 있는 10980개의 스타트업을 기업의 특징에 따라 연결한 스타트업 생태계 네트워크의 전체 맵입니다.
이를 마우스 스크롤을 통해 줌인하면 각 기업을 상세히 볼 수 있게 됩니다.
이와 같이 네트워크 시각화를 진행하게 되면 각각의 기업은 동그라미로 표현이 되는데요, 네트워크 시각화 분석에서는 이를 노드(Node)라고 칭합니다.
이러한 노드간를 연결한 선을 엣지(Edge)혹은 링크(Link)라고 부르는데요,
스타트업 생태계 네트워크에서는 엣지가 연결된 기업들은 제품과 서비스에 유사한 키워드를 가지고 있어 유사한 스타트업으로 해석할 수 있습니다.
노드의 크기는 투자금액의 규모를 알 수 있도록 투자금액을 20~300사이의 값으로 Scaling하여 시각화 하였습니다.

[스타트업 네트워크 시각화 개요]



이 분석 결과물을 활용하면 스타트업 생태계 전체 시각화 자료에서 연결되어 있는 그룹들을 통해 비슷한 영역의 업을 하게 되는 기업군을 분석할 수 있습니다.
아래의 이미지 영역은 전체 네트워크에서 일부를 줌인하여 찾아본 내용입니다.
TF-IDF값을 기준으로 그룹화된 기업들의 소개내용을 보게 되면 인공지능 관련 솔루션 및 응용 서비스를 제공하고 있는 유사 스타트업으로 이루어진 그룹임을 알 수 있습니다. 또한,
인공지능 관련 유사 기업 그룹으로부터 인공지능을 활용한 다양한 비즈니스로 확장된 스타트업들을 네트워크 결과와 원 데이터들을 통해 탐색할 수 있는데요,
이와 같이 사람이 특정 단일 기준값으로 구분하기 어려운 스타트업 그룹들을 비정형 텍스트 분석기법과 네트워크 분석기반의 시각화를 통해 탐색할 수 있는 형태로
데이터 분석이 가능함을 볼 수 있습니다.

[인공지능(AI) 키워드의 유사 기업 그룹과 노드 관련 원 데이터 정보]



기업일련번호	제품국문명	제품영문명	한줄소개내용
85341	마인드트립	MindTrip	인공지능(AI) 플랫폼
13240	집현전	ziphz	인공지능 부동산 추천/중개 서비스
4975	Fluenty.ai	Fluenty.ai	대화형 인공지능 플랫폼
108557	예스플리즈 에이아이	YesPlz AI	온라인 상에서 좋아하는 옷을 찾아주는 인공지능을 만듭니다.
94130	루카스	LuCAS-plus	인공지능 흉부CT 분석 진단보조SW
124922	주키퍼	ZuKeeper	인공지능이 만드는 주식 플랫폼
19974	영상합성 솔루션	Video Synthesis Solution	대화형 인공지능 모델 전문 플랫폼
92236	DeepNatural AI	DeepNatural AI	대화형 인공지능을 위한 고품질 코퍼스를 생산합니다.
124922	주키퍼	ZuKeeper	인공지능이 만드는 주식 플랫폼
39076	Fastcall	Fastcall	실시간 주식 투자 정보 제공
13713	루시	Lucy	인공지능 기반 VR 인지 건강 관리 솔루션
96504	네이버스		인공지능 기반 통합 모빌리티 플랫폼
104133	토플뱅크	TOEFL BANK	맞춤형 인공지능 토플 튜터
114155	모빌리티 디지털 광고	Mobility Digit Ad and Urba	모빌리티 기반 Adtech 및 Urban data platform 개발 회사

다음 페이지 계속 ↓

지금까지 데이터 시각화 분석 과정을 통해서, 스타트업 생태계에 대한 특징들에 대해 분석한 내용들을 정리하면 아래와 같습니다.

※ 스타트업 생태계의 특징 :

- 대한민국 성장동력으로 주목받고 있는 스타트업에 대한 지속적인 투자가 증가하고 있다.
- 스타트업은 대부분 초기 기업으로 1억 이상 50억 미만의 투자 규모에 가장 많은 투자가 집중되어 있다.
- 기존 산업분류를 통해 유사 스타트업을 그룹화하고 생태계를 설명하는 데 한계가 존재한다.

결론 : 보다 정확한 스타트업 생태계 이해를 위해서는 데이터 시각화 필요

이처럼 스타트업의 제품과 서비스 소개를 비정형 데이터 분석을 통해 네트워크 시각화를 하는 경우,
스타트업의 다양한 비즈니스 모델 요소에 대한 단어를 반영할 수 있어 스타트업 생태계에 대한 이해가 가능해지게 됩니다.
더 나아가 스타트업과 대기업 및 중견기업의 새로운 비즈니스 발굴 및 오픈 이노베이션 탐색을 위한 시각화 도구로 이용할 수 있고
투자자 또한 투자대상 스타트업을 발굴하는 도구로 활용할 수 있습니다.
이러한 분석 결과는 단편적인 현재의 지점을 보는 것보다
서비스를 통해 시각화 된 네트워크 맵과 이와 연계된 실제 원데이터를 인터랙티브하게 보는 것이 더욱 효율적입니다.
더욱 효과적인 데이터 탐색을 위해 디지털 산업혁신 빅데이터 플랫폼 (https://www.bigdata-dx.kr)에서는
‘스타트업 생태계 지도 시각화’ 서비스를 추후 선보일 예정입니다.

[주석내용]

1) 코사인 유사도 : 하나의 벡터를 다른 벡터에 사영하여 같은 방향에 해당하는 길이를 곱한 값으로 표현 할 수 있음.

$$\overline{A} \cdot \overline{B} = |A||B|\cos\theta$$

$\cos\theta$ 값은 -1 ~ 1의 범위를 가지며, 두 벡터가 같은 방향일 때 최대값인 1을 나타냄. θ 가 0°일 때 1이고 180°일 때 −1의 값을 가짐.

코사인 유사도는 이러한 성질을 이용하여 두 벡터가 얼마나 유사한지를 확인할 수 있는 공식

2) eps는 x와 y의 요소들이 모두 0인 경우 분모가 0이 되는 것을 막기위해 더하는 값으로 eps는 매우 작은 값을 가지고 있어 대부분의 경우 반올림기 때문에 계산결과에 영향을 주지 않음