글로벌 다이렉트 인덱싱
개발 기획서

조문근 | 건국대학교 컴퓨터공학 & 수학
김준우 | 건국대학교 컴퓨터공학
김민주 | 건국대학교 스마트ICT융합공학 & 문화콘텐츠

목차

1. 데이터

사용할 데이터는 크게 4종류이며, 각각의 간략한 개요는 아래와 같습니다.

데이터 종류	출처	국내 데이터	해외 데이터	특이사항
뉴스데이터	미래에셋증권	수집 완료	수집 완료	제공데이터
사업보고서	DART(금융감독원), SEC	수집 완료	개발 중	공시데이터
주가 및 종목 데이터	네이버 금융, yfinance	수집 완료	개발 중	공시데이터
ETF와 지수 데이터	KRX, GlobalX	수집 완료	개발 중	상업적 이용 시 지수 라이센싱 필요

1.1. 23년도 2분기 뉴스데이터 (제공)

1.1.1. 최종데이터

사용할 최종 데이터베이스 테이블 명세는 아래와 같습니다.

테이블명	필드명	데이터타입	필드설명
NEWS	DATE_TIME	DATETIME	뉴스 게시 날짜(연월일)
	TITLE	TEXT	뉴스의 제목
	ITEM_NAME	VARCHAR(20)	연관 종목명
	IMPORTANCE	FLOAT	뉴스의 중요도

1.1.2. 전처리

1) 날짜 형식 수정
2) 동일 날짜 뉴스 최대 중요도 제외 제거
3) 미사용 컬럼 제거

1.2. 23년도 2분기 분기보고서 (공시)

1.2.1. 최종데이터

사용할 최종 데이터베이스 테이블 명세는 아래와 같습니다.

테이블명	필드명	데이터타입	필드설명
BIZ_REPORT	ITEM_NAME	VARCHAR(20)	연관 종목명
	TITLE	VARCHAR(20)	사업의 개요 | 주요 제품 및 서비스
	CONTENT	TEXT	분기보고서의 내용

[1]OpenDartReader
금융감독원 전자공시시스템인 Open DART 서비스 API를 손쉽게 사용할 수 있도록 돕는 오픈소스 라이브러리
https://github.com/FinanceData/OpenDartReader

1.2.2. 데이터 수집 및 전처리

Python Package Index(PyPI)의 OpenDartReader[1] 패키지를 사용하였습니다. 분기보고서 내용 중 1) 사업의 개요 2) 주요 제품 및 서비스 부분만 사용합니다.
해외 사업보고서는 아래 주소에서 크롤링할 예정입니다.
https://www.sec.gov/edgar/searchedgar/companysearch

1.3. 23년도 2분기 주가 및 종목 데이터 (공시)

1.3.1. 최종데이터

사용할 최종 데이터베이스 테이블 명세는 아래와 같습니다.

테이블명	필드명	데이터타입	필드설명
KR_PRICE	DATE	DATETIME	날짜
	TICKER	VARCHAR(20)	종목코드
	CLOSE	FLOAT	종가

테이블명	필드명	데이터타입	필드설명
OS_PRICE	DATE	DATETIME	날짜
	TICKER	VARCHAR(20)	종목코드
	CLOSE	FLOAT	종가

1.3.2. 데이터 수집 및 전처리

국내 주가는 네이버 금융의 수정종가(Adj close)를 크롤링하였습니다.
해외 주가는 PyPI의 FinanceDataReader 패키지를 사용할 예정입니다.

1.4. 23년도 2분기 ETF와 구성종목 데이터

사용자가 기준 지수로 사용하는 TIGER 테마 ETF는 fnguide, KRX 등에서 개발한 테마 지수를 추종하고, Global X 테마 ETF는 Indxx 등의 지수 개발 업체에서 개발한 지수를 추종합니다. 대회 진행 과정에 있어 해당 데이터의 법적 이슈가 존재한다면 4.1의 키워드 기반 지수 유니버스 개발을 진행할 예정입니다.

1.4.1. 최종데이터

테이블명	필드명	데이터타입	필드설명
ETF_PDF	PARENT_ETF_TICKERS	VARCHAR(20)	ETF 종목코드
	TICKERS	VARCHAR(20)	종목코드

1.4.2. 데이터 수집 및 전처리

(국내) 한국거래소의 PDF 데이터를 크롤링 하였습니다.

http://data.krx.co.kr/contents/MDC/MDI/mdiLoader/index.cmd?menuId=MDC0201030105

(해외) Global X의 PDF 데이터를 크롤링할 예정입니다.
https://www.globalxetfs.com/funds/bkch

[2]유동비율
1-(비유동주식수/총발행주식수)*100
총 발행주식수에서 최대주주 및 특수관계인 보유주식, 정부 보유주식, 자사주 등을 제외한 주식수.

[3]시가총액가중(상한적용)
단순시가총액으로 지수를 계산할 경우 상대적으로 시가총액이 작은 기업의 주가흐름이 반영되지 않는 문제가 있습니다. 따라서 시가총액이 큰 기업의 비중에 제한을 두는 방식입니다.

Fnindex, Wiseindex (fnguide)
https://www.wiseindex.com/About/CeilingLogic

iSelect index (NH투자증권)
https://download.nhqv.com/CommFile/0000000002/257/5/20230717125922835.pdf

2. 모델

2.1. 지수 모델링

2.1.1. 동일 가중 모델 (Equal-Weight Index, EWI)

N개의 종목으로 이루어진 유니버스에 대해, 아래와 같은 가중치를 적용합니다.

$$w_i = \frac{1}{N} \text{ where weights } W = \begin{bmatrix} w_1 \\ w_2 \\ \cdots \\ w_N \end{bmatrix}$$

2.1.2. 시가 총액 가중 모델 (Market-Weight Index, MWI)

N개의 종목으로 이루어진 유니버스에 대해, 시가총액 비율 대로 가중치를 적용합니다. 이 때, 시가총액은 기준시가총액을 사용합니다. 기준시가총액 산출 방법은 아래와 같습니다.

$$C = P \cdot S \cdot FFR$$

C: 기준시가총액, P: 종가, FFR: 유동비율(Free Float Rate)[2], S: 지수산정 주식수

2.1.3. 시가 총액 가중 모델 (상한 적용)[3]

N개의 종목으로 이루어진 유니버스에 대해, 유동시가총액 비율 대로 가중치를 적용하되 상한을 적용합니다.
종목 당 상한이 $c(< 1)$ 일 때, 상한을 넘는 종목 $k(< n)$개에 대해 시가총액을 모두 y로 조정합니다.

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_k \end{bmatrix}$$

x_i는 기업i의 시가총액이고, y_i는 시가총액 비율이 c이상인 기업의 조정 후 시가총액입니다. 이를 수식으로 나타내면 아래와 같습니다.

$$kc = \frac{\sum Y}{\sum Y + (\sum^{n} X - \sum^{k} X)}$$

여기서 $y_i = y_j = y \ \forall i,j \leq k$ 이므로 $\sum Y = ky$ 로 치환하고, 이 외 상수들도 묶어서 문자로 치환해 정리합니다.

$$kc = \frac{ky}{ky + A}$$

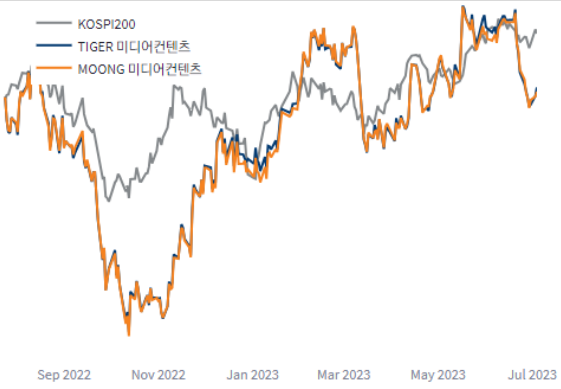
이를 조정된 시가총액 y 에 대해 정리하면 아래와 같습니다.

$$y = \frac{Ac}{1 - kc}$$

해당 모델을 구성 종목들에 적용하면, 기존 ETF와 동일한 가격변화를 보이는 것을 알 수 있습니다.

그림1. TIGER2차전지테마 지수 생성

그림2. TIGER미디어컨텐츠 지수 생성



[4] 영향도 산출
ETF의 구성종목들의 비율(PDF)을 알 수 없을 때는 ETF의 수익률에 대해 각 종목들의 수익률을 대상으로 PCA(주성분분해)를 사용한 후, 계수를 비중으로 간주합니다

[5] langchain
LLM 연결과 프롬프트 엔지니어링, 최종 서빙까지 도와주는 파이썬 라이브러리

[6] Nisan Stiennon 외. 2023. https://openai.com/research/learning-to-summarize-with-human-feedback

2.3. 테마 요약 모델

2.3.1. 종목별 영향도 산출 모델

시점 t에 리밸런싱한 ETF의 구성종목i의 시점 t'에서의 영향도 e_i 는 다음과 같이 산출됩니다.[4]

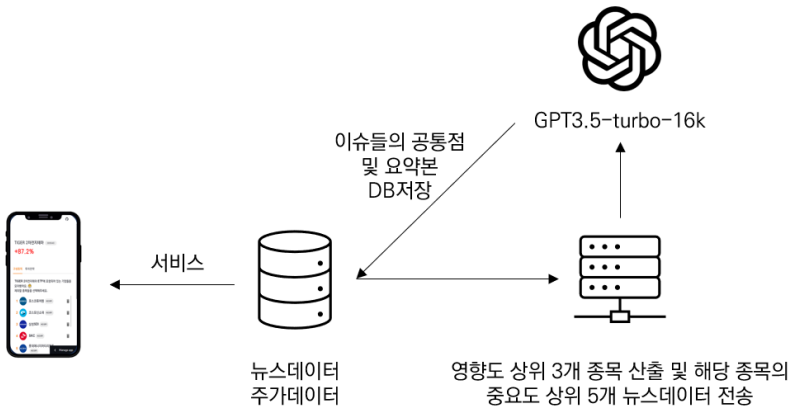
$$e_{i,t'} = r_{i,t'} \cdot w_{i,t}$$

$r_{i,t'}$: 구성종목 i의 시점 t'에서의 수익률
$w_{i,t}$: 구성종목 i의 리밸런싱 시점 t에서의 비중

2.3.2. 사업보고서 및 뉴스 요약 모델

2.3.1에서 추출한 영향도 상위 3개의 주식에 대한 뉴스와 사업보고서를 기반으로 해당 종목들의 공통점대해 요약[5]합니다. Openai의 GPT 3.5 turbo 16k 모델을 활용하며, PyPI의 langchain[6] 패키지를 사용합니다.

그림3. 뉴스 요약 모델 구조도



프롬프트 엔지니어링

현재 뉴스에 대한 공통점 요약까지 진행되었으며, 본선 진출 시 아래와 같은 방식으로 사업계획서 또한 프롬프트에 추가할 계획입니다.

그림4. 프롬프트 엔지니어링

프롬프트 I'm working on an industry analyst in Miraeasset Security.

Please write commonalities of the events of stocks such as large investment, factory

construction, etc. based on the information below.

If you don't know the answer, just say that you don't know, don't try to make up an answer.

Write a Humble, Clear, Factful summary in Korean.

```
Follows are the top news titles related with {stock_names}, which are the best stocks in the second quarter of FY23.

{'TITLE': {'20230411': '[속보]에코프로, 1분기 영업익 1795억…전년比 233% ↑', '20230427': '에코프로머티리얼즈, 상장예비심사 청구서 제출', '20230509': '에코프로 외인 지분율 4년來 최저', '20230609': '에코프로머티리얼즈, ‘1400억 투자’ 포항 RMP2 공장 준공', '20230616': 'FTSE 편입 호재…에코프로 다시 고점 돌파할까'}}

{\'TITLE\': {\'20230425\': \'포스코그룹, 포항에 전구체 공장 추진…배터리 소재 밸류체인 완성 [biz-플러스]\', \'20230426\': "포스코퓨처엠, LG엔솔과 \'30조원\' 규모 양극재 공급계약 체결", \'20230428\': \'포스코퓨처엠, 포항에 첫 하이니켈 양극재 전용공장 착공\', \'20230503\': \'포스코퓨처엠, 포항에 1조7000억원 투자…전구체?음극재 공장 짓는다\', \'20230530\': \'포스코퓨처엠-GM 합작사, 캐나다 정부 지원금 2900억원 확보\'}}

{'TITLE': {'20230409': '신통찮은 美증시에 국내 눈돌리는 서학개미', '20230411': '[속보]에코프로비엠, 1분기 영업익 1073억…전년比 161%↑', '20230414': '에코프로비엠, 5000억 CB발행 추진…주가 영향은?', '20230427': '[마켓인]나신평, 에코프로비엠 신용등급 ‘A-’ 상향…“사업 기반', '20230517': '[2023 1분기 실적] 2차전지가 받쳐줘도 반도체가 무너지자… 코스닥 영업익 40% ‘뚝’'}}
```

GPT3.5 응답 에코프로, 포스코퓨처엠, 에코프로비엠은 2023년 2분기에 가장 좋은 주식으로 평가되었습니다. 이 주식들과 관련된 일반적인 사건들은 큰 투자와 공장 건설 등이 있습니다.

에코프로는 2023년 1분기 영업이익이 전년 대비 233% 증가한 1795억 원을 기록했습니다. 또한, 에코프로머티리얼즈는 상장예비심사 청구서를 제출하였으며, 포항 RMP2 공장을 준공하면서 1400억 원을 투자했습니다. FTSE 편입도 예상되어 에코프로 주식이 다시 고점을 돌파할 수 있을 것으로 전망됩니다. 또한, 에코프로의 외인 지분율은 4년 동안 최저치를 기록했습니다.

포스코퓨처엠은 LG엔솔과 30조 원 규모의 양극재 공급계약을 체결하였습니다. 또한, 포항에 첫 하이니켈 양극재 전용공장을 착공하였으며, 1조 7000억 원을 투자하여 전구체와 음극재 공장을 건설할 예정입니다. 포스코퓨처엠과 GM의 합작사는 캐나다 정부로부터 2900억 원의 지원금을 확보했습니다.

에코프로비엠은 2023년 1분기 영업이익이 전년 대비 161% 증가한 1073억 원을 기록했습니다. 또한, 5000억 원 규모의 CB 발행을 추진하고 있습니다. 에코프로비엠의 신용등급은 A-로 상향 조정되었으며, 코스닥 영업이익은 40% 감소했습니다.

이러한 사건들은 에코프로, 포스코퓨처엠, 에코프로비엠 주식의 성장과 관련이 있으며, 큰 투자와 공장 건설 등이 이루어지고 있다는 것을 알 수 있습니다.

[7]Streamlit
AI,빅데이터 프로젝트에 대해, 웹 지식이 없어도 Python만으로 웹사이트 배포를 도와주는 라이브러리

3. 플로우차트

3.1. 서비스 아키텍처

데이터베이스는 MySQL, 사용언어는 Python이며, 배포는 NAVER CLOUD PLATFORM과 Streamlit[7]을 사용하였습니다.

	사용 언어	배포
웹 서버	python	Streamlit
데이터베이스 서버	MySQL	NAVER CLOUD PLATFORM

3.2. 플로우차트 & 서비스 설명

테마 ETF를 기반으로 나만의 ETF를 만들 수 있는 다이렉트 인덱싱 서비스입니다. 크게 3단계로 진행되며, 각각의 하부 메뉴는 아래와 같습니다.

3.2.1. ETF 선택

다이렉트 인덱싱을 할 종목 유니버스를 선택합니다. 2차전지, 퓨처모빌리티, 메타버스 등이 있습니다. 이는 연관 TIGER ETF의 PDF를 기반으로 합니다.

3.2.2. ETF 편집

종목 유니버스를 기반으로 다이렉트 인덱싱을 진행합니다. 구성종목 편집을 통해 기존 종목 유니버스에서 제외할 종목을 선택할 수 있습니다.

구성종목 편집

유니버스에 속한 종목을 제거할 수 있으며, 각 종목들의 비중을 수동으로 설정할 수 있습니다.

[8] bt
ffn(financial functions for python)
의 백테스팅 확장 라이브러리

[9] 사용자 지정 ETF
예시로 동일가중을 사용하였습니다.

투자전략 편집

▲ 동일가중, 시가총액가중으로 종목들의 비중을 선택할 수 있습니다. 기본값은 시가총액가중입니다. ▲ 1개월, 3개월, 1년의 리밸런싱 주기를 선택할 수 있습니다. 기본값은 3개월입니다. ▲ 나만의 ETF의 이름을 설정할 수 있습니다. 기본값은 “TIGER”입니다.

3.2.3. 모니터링

하부 기능으로는 1) 수익, 2) 손실, 3) 리밸런싱 내역이 있으며 Python Package index(PyPI)의 bt[8] 패키지, 시각화는 Plotly express로 구현합니다.

수익

1) 구성된 포트폴리오(MOONG)의 과거 수익률, 2) 대표지수(KOSPI200) 과거 수익률, 3) 기준지수(TIGER) 과거 수익률을 보여줍니다.

그림5. MOONG 2차전지 동일가중[9] 수익률

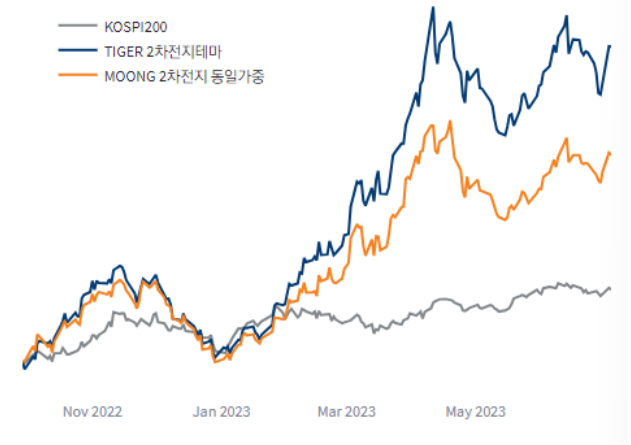


그림6. MOONG 미디어컨텐츠 동일가중 수익률

손실

1) 구성된 포트폴리오(MOONG)의 과거 하루 손실률, 2) 대표지수(KOSPI200) 하루 손실률, 3) 기준지수(TIGER) 하루 손실률을 보여줍니다.

그림7. MOONG 2차전지 동일가중 MDD

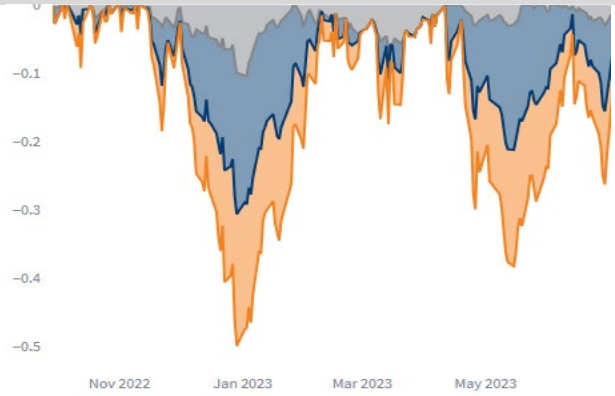
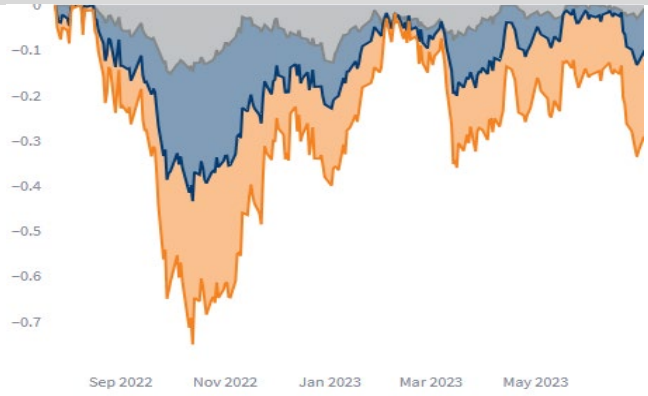


그림8. MOONG 미디어컨텐츠 동일가중 MDD



리밸런싱 내역

구성된 포트폴리오의 리밸런싱 내역을 보여줍니다.

그림9. MOONG 2차전지테마 동일가중 리밸런싱[8]

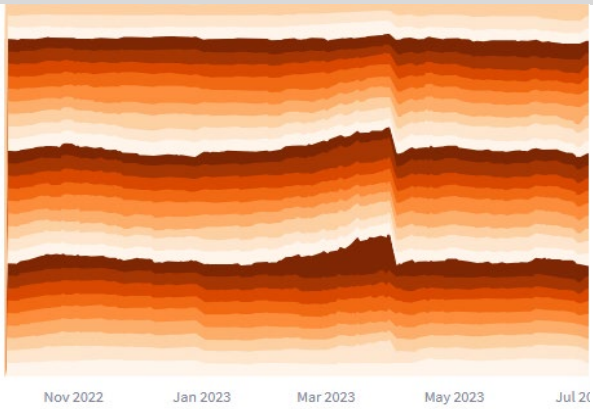
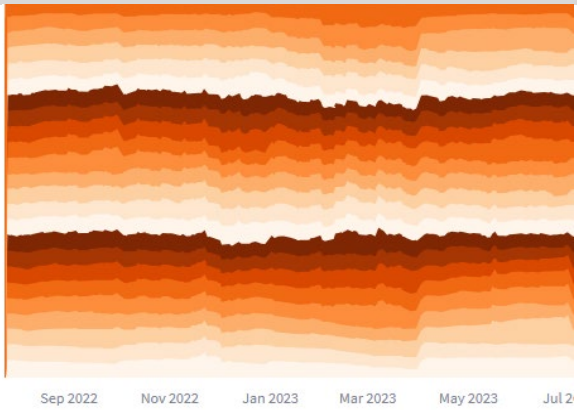


그림10. MOONG 미디어컨텐츠 동일가중 리밸런싱



[9] 키워드 기반 종목 선정

Fnguide FnIndex, WiseIndex

http://file.fnguide.com/fnindex/FnGuide_Metaverse%20Theme_Index_Methodology_Book_미래에셋자산운용_(20210810)_v1.0.pdf

NH투자증권 iSelect지수

https://download.nhqv.com/CommFile/0000000002/257/0/20230118133256870.pdf

4. 기타

4.1. 키워드 기반[9] 지수 구성 종목 산출 방식

유가증권시장 및 코스닥 상장 종목 중 기초 필터링을 통과한 종목들을 유니버스로 하고, 유가증권시장 및 코스닥에 상장된 기업 들에 대해 [테마 대표 단어, ex: 메타버스] 관련 키워드 기반 머신러닝으로 종목별 키워드 유사도 스코어링을 진행하여, 유니버스 포함 종목 중 [테마 대표 단어]가 높은 종목을 선정하여 지수를 구성합니다.

유니버스 선정

종목선정일 기준 유가증권시장 및 코스닥 상장 보통주 중에서, 아래의 조건을 모두 만족하는 종목을 유니버스로 합니다.

- ✓ 관리종목 또는 투자주의환기종목으로 지정되었거나, 상장폐지가 확정된 종목 제외
- ✓ 선박투자회사, 인프라투자회사, REITs, ETF, ETN, SPAC 제외
- ✓ 유동비율 10% 미만 종목 제외
- ✓ 시가총액 1000억원 미만인 종목 제외
- ✓ 20영업일 평균 거래대금 시장 하위 10% 미만 종목 제외

키워드 유사도 산출

종목선정일 직전 월 기준 과거 1년치 사업계획서를 TF-IDF 기법을 통해 벡터화 합니다, 다음으로 핵심 단어 키워드 벡터를 이용해 각 문서 벡터들 간의 코사인 유사도 값을 계산하여 [테마 대표 단어] 관련 키워드와 종목의 유사도를 산출합니다.

- ✓ 키워드 분석 문서 유니버스: 종목선정일 직전월 기준 과거 1년치 사업보고서(분기, 정기 보고서) 中 사업의 내용
- ✓ 기초 키워드 : [테마 대표 단어] 관련 키워드에 각각 고정된 가중치가 부여됨
- ✓ TF-IDF : Information Retrieval을 위해 제안된 Term Weighting의 한 방법이며, 개별 문서에서 자주 등장하는 단어의 영향력은 높이되, 어느 문서에서나 등장하는 단어의 영향력은 줄이는 기법
- ✓ $TF-IDF(w, d) = TF(w, d) * \log(N\ DF(w))$
- ✓ DF : 단어 w 가 등장한 문서의 개수

- ✓ N : 문서 집합에서 문서의 총 개수

문서별 스코어링

- ✓ TF-IDF 기법을 통해 벡터화된 문서 벡터(B) 간의 코사인 유사도 값을 계산
- ✓ $Similarity = cos(\theta) = A \cdot B \, \|A\|\|B\|$

종목별 스코어링

종목별 문서 매칭 기준, 문서별 스코어링 내림차순으로 지수단위 역가중합