ETF Recommender

CISC 6352 – Advanced Computational Finance –Project 1

Sheng Yun (syun13), Jingyan Xu (jxu246), Cathy Xueqing Zhang (xzhang347)

CONTENTS

- Holding dataset
  - Set up MongoDB
  - Master List
  - Get holding data from website
- Recommendation
  - Cosine similarly
  - Recommendation
- Discussion

01

Getting Holding Data

MongoDB

Full cloud-based to save data

Easy push and retrieve
No duplicate

Save as dictionary
- ETF ticker is index

ETF list

246 ETFS with AUM bigger than $2bn

BeautifulSoup

iShare.com

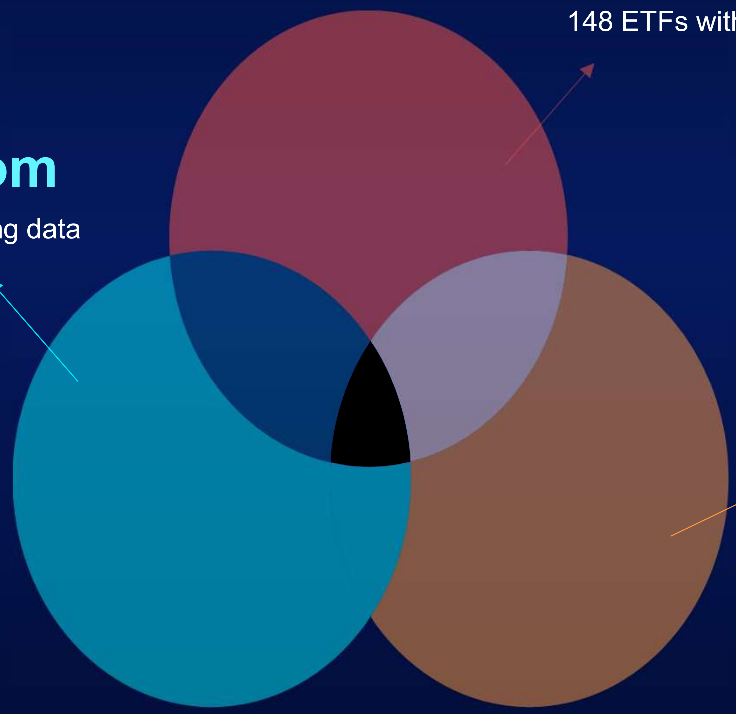
84 ETFs with holding data

stockAnalysis.com

148 ETFs with holding data

Invesco.com

14 ETFs with holding data



iShares ETF Holdings

- 84 ETFs with holding data
- Construct ‘get_ishare_holding’ function to obtain iShares holdings information
  - ‘tick_to_url’ use to obtain url for etf
  - Only focus on equity
  - Clear out stocks for which tickers letters aren’t less than or equal to 4 and contains other characters (ie. numbers, special characters)
  - Re-adjust weights to 1.0 base
- Push to MongoDB


```
{'MTDR': 0.0030659798903487764,
 'IRDM': 0.0029794908519799723,
 'CROX': 0.0029749461621869167,
 'SAIA': 0.00292437858935519,
 'INSP': 0.0028882813684318427,
 'RBC': 0.002849040030471767,
 'EME': 0.0027983942690318615,
 'HALO': 0.0027330797931671455,
 'SWAV': 0.0027023930327407684,
 'TXRH': 0.0026805727101899416,
 'CHX': 0.002667255791457733,
 'MUR': 0.0026541312627728574,
 'ADC': 0.002618845430226225,
 'STAG': 0.0026062440367633027,
 'CMC': 0.0025948735257213186,
 'LNW': 0.002474852360423985,
 'CHRD': 0.0024517157107091163,
 'KRTX': 0.0024036548839709445,
 'KNSL': 0.002382732511009686,
 'SSB': 0.0023264133572238705,
 'NOVT': 0.0023064756156947366,
 'MEDP': 0.002293424408251681,
 'AQUA': 0.0022747215193014135,
 'SIGI': 0.0022307283463217425,
 'MUSA': 0.002229441297605799,
 'CELH': 0.0022226752309369296,
 'EXLS': 0.0022219446254473393,
 'GTLS': 0.0022203103983969949,
 'UFPI': 0.002189239949389255,
 'AIT': 0.0021679840826119936,
 'FLR': 0.0021533296808118442,
 'ATKR': 0.0021224951663822724,
 'SLAB': 0.002114370257980451,
```


Snippet of output
get_iShares_holding(‘IWM’)

Invesco ETF Holdings

- 14 ETFs with holding data
- Easier one
- Direct download from URL
- Construct ‘get_invesco_holding’ function to obtain invesco holdings information
  - Focus on Equity only (Common stock)
  - Clear out stocks' special tickers
  - Adjust weights accordingly
- Push to MongoDB


```
{'MTDR': 0.0030659798903487764,
 'IRDM': 0.0029794908519799723,
 'CROX': 0.0029749461621869167,
 'SAIA': 0.002924378589355519,
 'INSP': 0.0028882813684318427,
 'RBC': 0.002849040030471767,
 'EME': 0.0027983942690318615,
 'HALO': 0.0027330797931671455,
 'SWAV': 0.0027023930327407684,
 'TXRH': 0.0026805727101899416,
 'CHX': 0.002667255791457733,
 'MUR': 0.0026541312627728574,
 'ADC': 0.002618845430226225,
 'STAG': 0.0026062440367633027,
 'CMC': 0.0025948735257213186,
 'LNW': 0.002474852360423985,
 'CHRD': 0.0024517157107091163,
 'KRTX': 0.0024036548839709445,
 'KNSL': 0.002382732511009686,
 'SSB': 0.0023264133572238705,
 'NOVT': 0.0023064756156947366,
 'MEDP': 0.002293424408251681,
 'AQUA': 0.0022747215193014135,
 'SIGI': 0.0022307283463217425,
 'MUSA': 0.002229441297605799,
 'CELH': 0.0022226752309369296,
 'EXLS': 0.0022219446254473393,
 'GTLS': 0.002203103983969949,
```


Snippet of output
get_Invesco_Holdings(‘PHDG’)

StockAnalysis ETF Holdings

- 148 ETFs on stockanalysis.com
- Most difficult one, request rest runtime for full dataset
- Construct ‘get_stockAnalysis_holding’ function to obtain stockAnalysis holdings information
  - Assume all equity, if ticker is letters and less than or equal to 4
  - Adjust Weight
- Push to MongoBD every time
- Manually push for ‘NOBL’:
  - “n/a” as ticker and “n/a” as weights
  - Clean accordingly


```
{'AAPL': 0.07704402515723272,
 'MSFT': 0.06531204644412193,
 'AMZN': 0.032051282051282055,
 'GOOGL': 0.02092404450895017,
 'BRKB': 0.019714562167392355,
 'GOOG': 0.01886792452830189,
 'NVDA': 0.017900338655055636,
 'TSLA': 0.016811804547653606,
 'XOM': 0.016811804547653606,
 'UNH': 0.016086115142718918,
 'JNJ': 0.015602322206095793,
 'JPM': 0.01451378809869376,
 'V': 0.013425253991291729,
 'META': 0.012094823415578134,
 'PG': 0.01173197871311079,
 'HD': 0.011490082244799226,
 'CVX': 0.011248185776487665,
 'MA': 0.011248185776487665,
 'LLY': 0.009554910498306726,
 'MRK': 0.009433962264150945,
 'ABBV': 0.009192065795839382,
 'BAC': 0.008708272859216257,
 'PFE': 0.008708272859216257,
 'AVGO': 0.008466376390904693,
 'KO': 0.008345428156748911,
 'PEP': 0.008224479922593132,
 'TMO': 0.007982583454281569,
```


Snippet of output
get_stockAnalysis_holding('SPY')

Final Dataset

- Retrieve full data from MongoDB
- Convert to dataFrame for follow recommendation system analysis

	_id	ticker	Holdings
0	63d741139b6177006e4ee4e8	IVV	{'AAPL': 0.06406294725137868, 'MSFT': 0.054308...
1	63d741159b6177006e4ee4e9	IEFA	{'NESN': 0.024309578051267204, 'ASML': 0.02037...
2	63d741169b6177006e4ee4ea	IEMG	{'RELIANCE': 0.02691150682230978, 'VALE': 0.02...
3	63d741179b6177006e4ee4eb	IJR	{'ADC': 0.006969187214429515, 'UFPI': 0.005923...
4	63d741189b6177006e4ee4ec	IJH	{'FICO': 0.007450394880033458, 'RS': 0.0060110...
...	...	...	...
241	63d74b67f9a12b008a203509	PTLC	{'AAPL': 0.07736077481840194, 'MSFT': 0.065496...
242	63d74b67f9a12b008a20350a	COPX	{'ANTOL': 0.0601939806019398, 'HK': 0.00449955...
243	63d74b68f9a12b008a20350b	FXN	{'DINO': 0.04439112177564487, 'PDCE': 0.044391...
244	63d74b69f9a12b008a20350c	DEM	{'VALESA': 0.07611990569215224, 'TW': 0.001010...
245	63d74e05f9a12b008a20350d	NOBL	{'CAT': 0.019340615292113436, 'BEN': 0.0188395...

246 rows × 3 columns

ticker	IVV	IEFA	IEMG	IJR	IJH	IWF	IWM	IWD	EFA	ITOT
AAPL	0.064063	NaN	NaN	NaN	NaN	0.118347	NaN	NaN	NaN	0.053962
MSFT	0.054308	NaN	NaN	NaN	NaN	0.100193	NaN	NaN	NaN	0.045755
AMZN	0.02664	NaN	NaN	NaN	NaN	0.049034	NaN	NaN	NaN	0.022452
GOOGL	0.017425	NaN	NaN	NaN	NaN	0.02806	NaN	0.004048	NaN	0.014682
BRKB	0.016353	NaN	NaN	NaN	NaN	NaN	NaN	0.029058	NaN	0.013773
...	...	...	...	...	...	...	...	...	...	...
FIBRAMQMX	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
BCHRBK	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
MERPM	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
FROTOEIS	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
BANPURBK	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

8080 rows × 242 columns

02

Recommendation System

Cosine similarity function

- **similarity** (*df*: DataFrame, *input*: str, *cutoff*: numeric) → dictionary
  - Calculate Cosine Similarity to measures of similarity --- scipy.spatial.distance.cosine()
  - *cutoff* score must be within (0,1)
  - *input* ticker's type **must** be a 1-D vector.
  - Return a **sorted** dictionary --- **descending** order.
    - the ETFs ticker as the key
    - the Cosine similarity score as the value.
- **recommend**(*df*: DataFrame, *input*:str, *cutoff*: numeric, *num*: integer) → dictionary
  - Calling similarity function to get the score
  - Print the recommended ETFs based on the similarity.
    - If the *input* is invalid, will catch the keyerror and print out the message.
    - If the *num* is invalid(<0), default *cutoff* value is 0.8, it will print out the most similar ETF, if the *num* exceeds the return length, it will print the complete candidates

Recommendation Result

```
# Tests:

print(recommend(final_result, 'QQQ',0.8,20))

ticker QQQM, similarity 0.9999999824575585
ticker QYLD, similarity 0.9585477491146559
ticker ONEQ, similarity 0.9402668297849629
ticker SCHG, similarity 0.925952187876004
ticker IWF, similarity 0.9220408830298149
ticker IWY, similarity 0.9216983533797071
ticker VUG, similarity 0.9182151554581917
ticker MGK, similarity 0.9179983369027024
ticker VONG, similarity 0.9090901072228659
ticker IYW, similarity 0.8824953465631118
ticker ESGV, similarity 0.8815702587479439
ticker VONE, similarity 0.8765800230067662
ticker SCHK, similarity 0.8764315286710781
ticker SCHX, similarity 0.8764106424878356
ticker IGM, similarity 0.8720673285871995
ticker PBUS, similarity 0.8682894744358113
ticker OEF, similarity 0.8660374439801083
ticker SPLG, similarity 0.8653587280779103
ticker SCHB, similarity 0.8652650012902074
ticker SPTM, similarity 0.8651547407825559
{'QQQM': 0.9999999824575585, 'QYLD': 0.9585477491146559, 'ONEQ': 0.9402668297849629, 'SCHG': 0.925952187876004, 'IWF': 0.9220408830298149, 'IWY': 0.9216983533797071, 'VUG': 0.9182151554581917,
```

- The cosine distance we got between QQQ and XLK is 0.828605616857576

03

Discussion

Recommendations Analysis

- **Using our recommender system --- get list of similarity ETFs.**
- **When we making final recommendation, we also need to consider:**
  - **Lower-cost ETF**
    - Measured by MER
      - i.e. For QQQ – expense ratio is 0.2%
      - IVV 0.863923 +0.03%
        SPLG 0.865359 +0.03%
        SCHB 0.865265 +0.03%
        SPTM 0.865155 +0.03%
  - **ETF with the lowest volatility**

	etfs	similarity	expense_ratio
0	QQQM	0.9999999824575585	+0.15%
1	QYLD	0.9585477491146559	+0.60%
2	ONEQ	0.9402668297849629	+0.21%
3	SCHG	0.925952187876004	+0.04%
4	IWF	0.9220408830298149	+0.18%
5	IWY	0.9216983533797071	+0.20%
6	VUG	0.9182151554581917	+0.04%
7	MGK	0.9179983369027024	+0.07%
8	VONG	0.9090901072228659	+0.08%
9	IYW	0.8824953465631118	+0.39%
10	ESGV	0.8815702587479439	+0.09%
11	VONE	0.8765800230067662	+0.08%
12	SCHK	0.8764315286710781	+0.05%
13	SCHX	0.8764106424878356	+0.03%
14	IGM	0.8720673285871995	+0.40%
15	PBUS	0.8682894744358113	+0.04%
16	OEF	0.8660374439801083	+0.20%
17	SPLG	0.8653587280779103	+0.03%
18	SCHB	0.8652650012902074	+0.03%
19	SPTM	0.8651547407825559	+0.03%

Thank you

Sheng Yun (syun13), Jingyan Xu (jxu246), Cathy Xueqing Zhang (xzhang347)

1/30/2023