수리적 빅데이터 개론 기말과제 보고서

2014024007
김민기

1. 명사의 중요도

```
>>> df_weightedDegree.iloc[:20,:]
      NODE_ID   DC
967       New  975
940     Stark  939
566         A  925
945      meet  916
897      Tony  901
53    Avengers  872
1136      York  828
211        Man  804
72       Earth  762
1038      Iron  741
1293    Banner  729
209      Bruce  729
1       America  720
232       Just  711
1175    SHIELD  678
1183   Captain  658
512      Peter  655
1253        US  651
894       Fury  649
615     Howard  633
```

Degree Measure의 상위 20개를 뽑아봤을 때, 'New York','Tony Stark','Avengers','Bruce Banner','Captain America','SHIELD','Peter','Fury','Howard Stark','Iron man' 가 수치가 높은 것을 볼 수 있다. 동시발생행렬로부터 Network를 만들었으므로, Degree가 높다는 것은 그 만큼 상위 Degree를 기록한 명사들이 많이 다른 명사와 함께 자주 등장했다는 것을 알 수 있고, 따라서, 위의 명사들이 Degree Measure로 봤을 때, 중요한 명사라고 할 수 있다.

```
>>> df_weightedBetweeness.iloc[:20,:]
      NODE_ID        BC
945      meet  0.042533
566         A  0.040546
72      Earth  0.036278
967       New  0.034114
53    Avengers  0.027198
940      Stark  0.022324
1136      York  0.020130
897       Tony  0.018423
615     Howard  0.018225
211        Man  0.018066
232       Just  0.017039
1253        US  0.015566
512      Peter  0.014060
1038      Iron  0.012449
1      America  0.012263
1175    SHIELD  0.012129
209      Bruce  0.012077
1293    Banner  0.012077
933     London  0.010602
165        War  0.009620
```

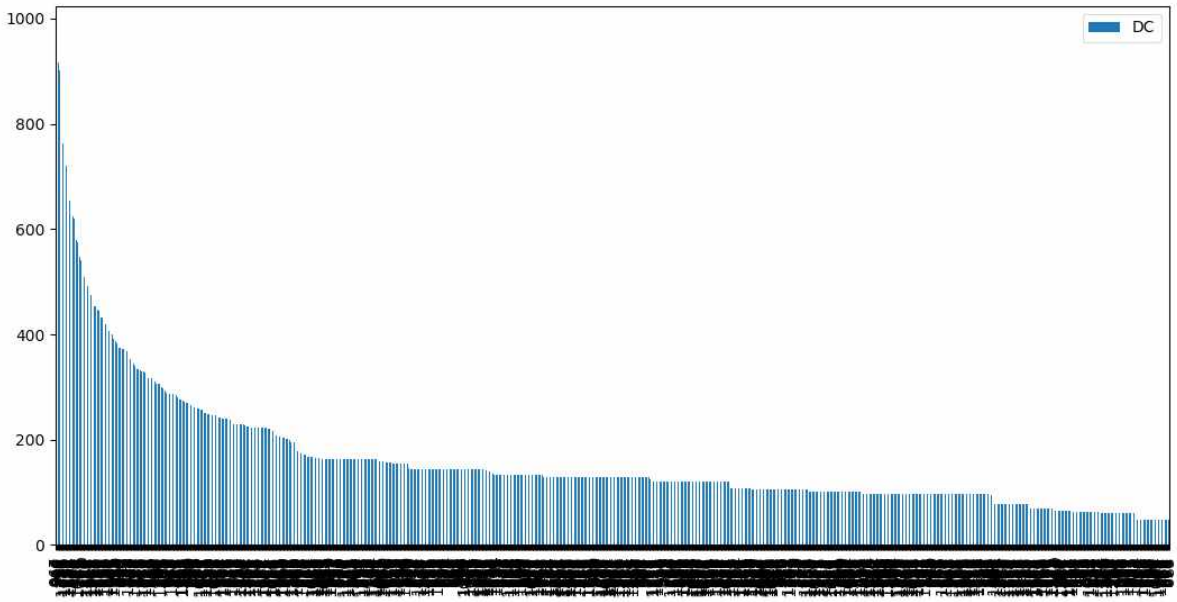
Betweeness Centrality Measure로 봤을 때, 'New York','Tony Stark','Howard Stark','Peter','Captain America','SHIELD','Bruce Banner','London','War','Iron Man'이 수치가 높았는데, 이것은 각각의 명사들이 위의 명사들을 많이 거쳐가야 한다는 뜻이고, 이것은 다른 명사들과 관련이 많다는 것을 뜻한다. 따라서 위의 명사들이 중요한 명사임을 알 수 있다.

```
>>> df_weightedClose.iloc[:20,:]
      NODE_ID        CC
967       New  0.798530
940     Stark  0.781306
566         A  0.774807
945      meet  0.770686
897      Tony  0.763913
53    Avengers  0.751152
1136      York  0.732584
211        Man  0.722838
72       Earth  0.706392
1038      Iron  0.698447
1293    Banner  0.693986
209      Bruce  0.693986
1      America  0.690678
232       Just  0.687401
1175   SHIELD  0.675648
1183   Captain  0.668718
512      Peter  0.667691
1253        US  0.666326
894       Fury  0.665646
615     Howard  0.660253
```

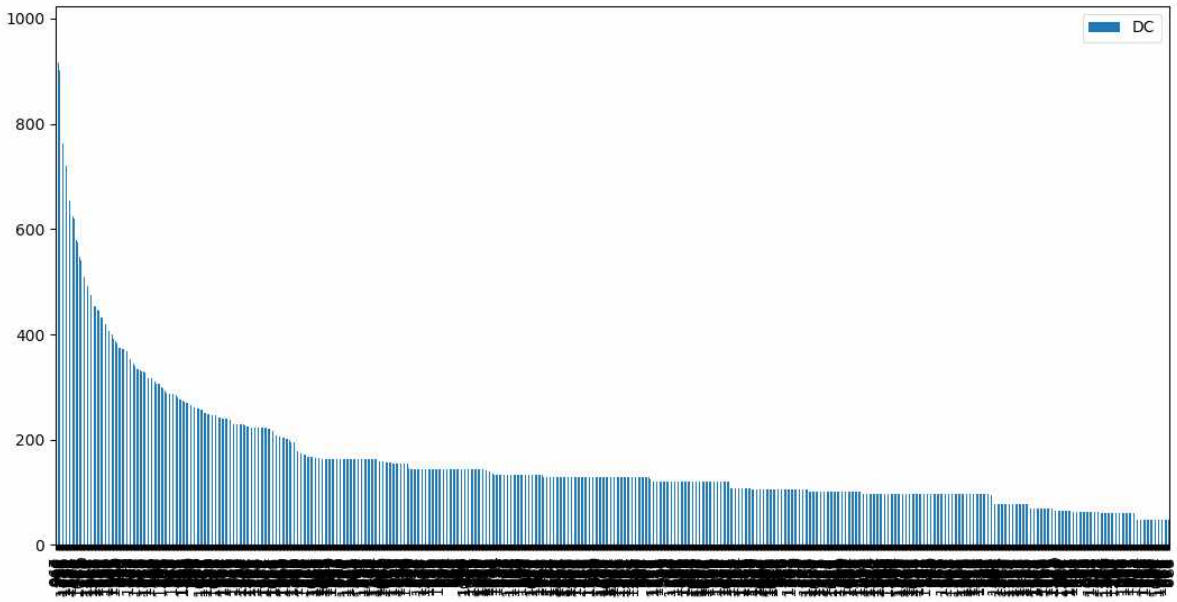
Closeness Centrality Measure로 봤을 때, 'New York','Tony Stark','Howard Stark','Peter','Captain America','SHIELD','Bruce Banner','London','War','Iron Man','Fury'이 수치가 높았는데, 이것은 위의 명사들이 다른 명사들과 거리가 가까운 것을 의미하고, 이것은 BC와 마찬가지로, 다른 명사들과 연관이 높다는 것을 의미한다. 따라서, 위의 명사들이 중요도가 높다고 해석할 수 있다.

DC,BC,CC의 Measure를 종합적으로 생각해 볼 때, 흔히 어벤져스의 주인공이라 생각되는 아이언맨과 캡틴아메리카, 스파이더맨, 퓨리국장이 중요한 명사로 생각되는 것을 알 수 있고, 추가적으로, 어벤져스의 주 활동무대인 뉴욕,런던등의 장소도 중요한 명사로 나타나는 것을 알 수 있다.

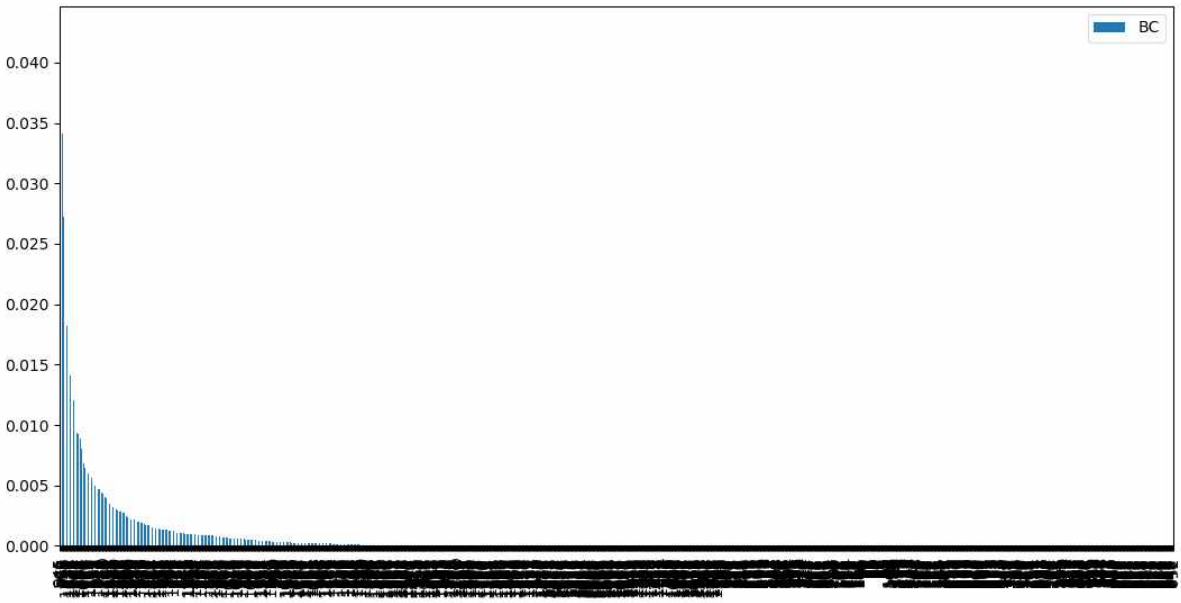
2. 네트워크의 종류



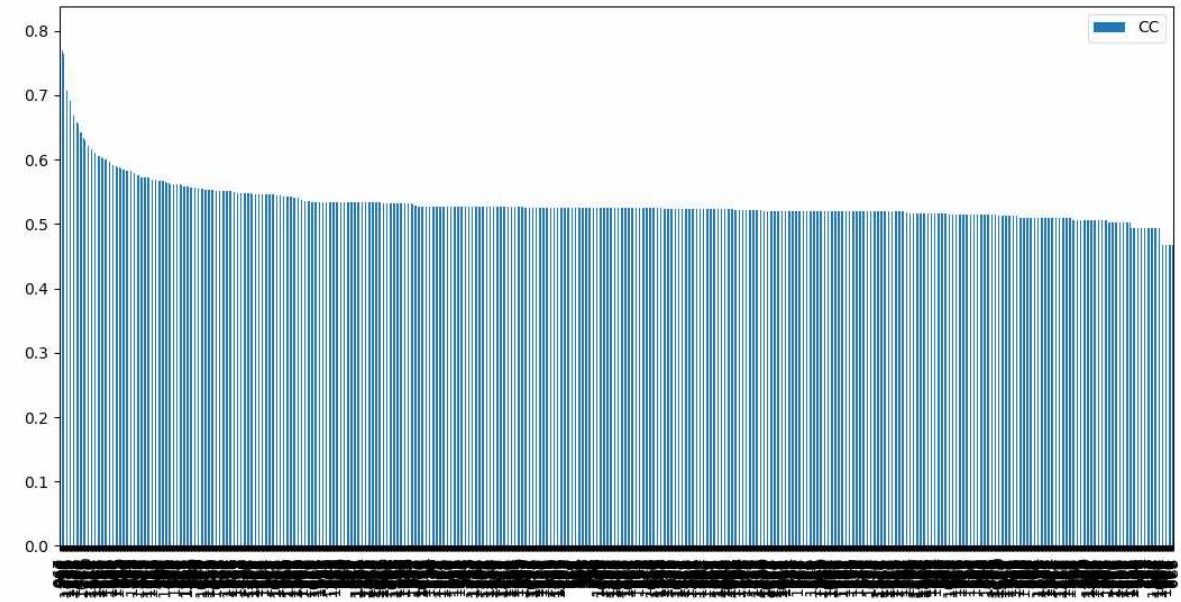
<Weighted Degree>



<Unweighted Degree>



<Weighted Betweeness Centrality>



<Weighted Closeness Centrality>

Weighted Degree는 동시발생행렬을 만들었을 때, 각 원소의 값을 Edge의 weight로 연결한 graph의 Degree이고, Unweighted Degree는 각 원소의 값을 0이아니면 모두 동일한 가중치로 Edge를 연결한 graph의 Degree이다. 결과를 보면, 가중치를 주었을 때와 주지 않았을 때는 결과가 같은 것을 알 수 있고, 위의 그래프를 바탕으로 생각해봤을 때, 가장 유사하게 보이는 네트워크의 종류는 'Scale Free' 네트워크라고 판단된다. DC,BC,CC를 봤을 때, 먼저, CC의 측면으로 보면, 대부분의 명사들끼리의 거리는 가까운 것을 알 수 있다. BC의 측면으로 보면, 몇몇의 명사들을 거쳐야만 다른 명사들로 갈 수 있는 Hub의 역할을 하는 명사가 존재한다는 것을 알 수 있고, DC의 측면으로 보면, 몇몇의 명사들이 다른 명사들과 압도적으로 많이 연결되어 있다는 것을 알 수 있다.

위의 Measure들을 종합해봤을 때, 각각의 명사들끼리의 거리는 가까운 편이지만, 다른 명사로 가기위해 거쳐가야하는 Hub 명사가 존재하는 것을 알 수 있고, 그 명사들의 Degree가 높다는 것을 알 수 있다. 따라서, 위의 네트워크는 'Scale Free' 네트워크라고 생각된다.

3. 명사의 빈도수

```
>>> df_frequency[:30]
Tony        336
Peter       211
Steve       205
Stark       160
Thor        131
Fury        115
Loki        111
Man          90
Hulk         76
Bruce        76
Bucky        75
Thanos       74
Scott        71
Pepper       69
Strange      61
Iron         61
Rocket       60
Rogers       58
Gamora       56
Ross         55
```

각각의 영화의 등장한 단어들을 모두 합한 후, 상위 30개를 뽑아 봤을 때, 대부분 'Tony Stark','Peter','Steve Rogers','Thor','Hulk','Fury','Loki','Bucky','Pepper','Doctor Strange','Gamora','Ross'와 같은 중요 인물들이 자주 등장했음을 알 수 있다.

4. 명사의 중요도와 명사의 빈도수 비교

Degree, Betweeness Centrality, Closeness Centrality로 추출한 명사의 중요도와 명사의 빈도수로 추출한 명사의 중요도를 봤을 때, Measure들로 추출한 명사의 중요도는 어벤져스 시리즈 내에서의 주요인물들과 장소들이 명사중요도 목록 상위에 rank되었다. 반면, 명사의 빈도수로 추출한 명사들은 Measure들로 추출한 명사들보다 등장인물 들이 훨씬 많이 출현하는 것을 알 수 있다.

이 분석을 통해, 대체적으로 많이 등장하는 명사들이 네트워크에서 큰 중요도를 가질 수 있지만, 꼭 그런 것은 아니다 라는 것을 알 수 있고, 텍스트 분석과 같은 것을 진행할 때, 여러 가지 척도를 고려해서 분석해야 해야 텍스트가 주는 정보를 놓치지 않을 수 있음을 알 수 있었다.