Network Graph Clustering with Instagram Hashtags

2019. 01. 22.

Lee Sungjae
Woo Hyeonwoong

Overview

1. Previous Research : Tag Clustering
2. RawData Preprocessing
3. Plan

1. Previous Research : Tag Clustering

Previous Research

연관 태그의 군집 알고리즘의 설계 및 구현

박병재, 우종우 — 2009, 한국IT서비스학술회

> Delicious / Flickr 웹페이지의 특정 태그를 크롤링하여 군집화 및 시각화
> 유클리디안 유사도 함수를 이용하여 군집 형성 및 평가

연관 태그의 군집화를 위한 클러스터링 기법 비교 연구

한승희 — 2009, 한국문헌정보학회지

> 위의 논문과 유사한 태그 데이터를 이용하여 코사인 유사계수, 피어슨 상관계수로 연관성 분석
> 연관성과 계층적, 비계층적 클러스터링 알고리즘을 조합하여 최적의 모델 구현

Graph Clustering

Satu Elisa Schaeffer — 2007, Computer Science Review

> 기본적인 그래프 이론과 distance 계산을 통한 그래프 Build 방법
> Spectral Clustering 에서의 Cut 을 이용한 그래프 군집화 방법에 대한 설명

Other Papers

Automated Tag Cluster : Improving search and exploration in the tag space

Graph Clustering Based on Structural / Attribute Similarity

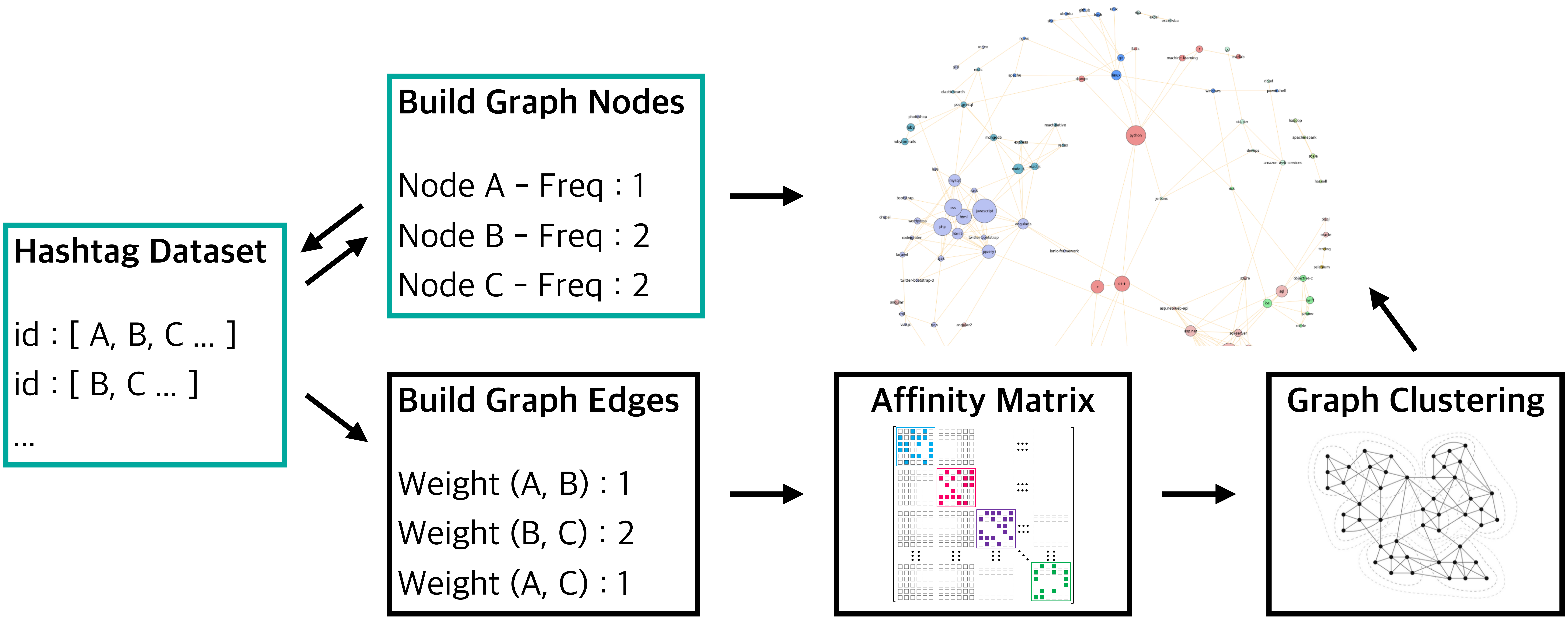
Community Detection in Networks

Semi-Supervised Clustering : A Kernel Approach

Deep Spectral Clustering Learning

Keyword : Tag Clustering / Graph Clustering / Community Detection

Tag Clustering Workflow



Hashtag Dataset

id : [A, B, C …]

id : [B, C …]

…

Build Graph Nodes

Node A – Freq : 1

Node B – Freq : 2

Node C – Freq : 2

Build Graph Edges

Weight (A, B) : 1

Weight (B, C) : 2

Weight (A, C) : 1

Affinity Matrix

Graph Clustering

Tag Clustering Workflow

Hashtag Dataset

id : [A, B, C …]

id : [B, C …]

…

Build Graph Edges

Weight (A, B) : 1

Weight (B, C) : 2

Weight (A, C) : 1

코사인 유사계수
Cosine Coefficient

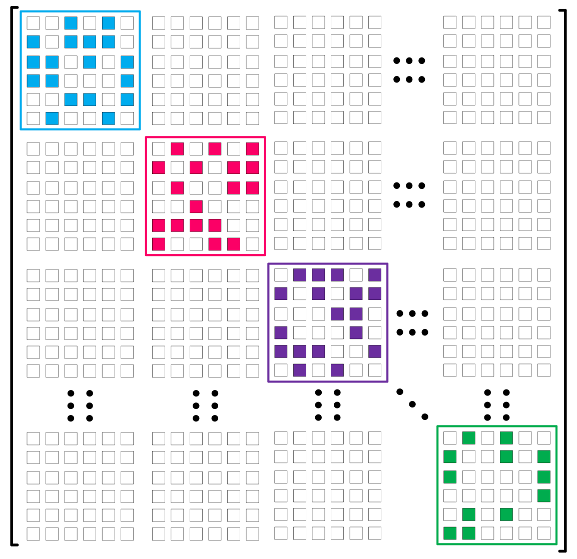
피어슨 상관계수
Pearson Correlation Coefficient

$$\cos(x,y) = \frac{\sum_i (x_i y_i)}{\sqrt{(\sum_i x_i^2)(\sum_i y_i^2)}}$$

$$r(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

$$(\bar{x} = \frac{1}{n}\sum_i x_i, \quad \bar{y} = \frac{1}{n}\sum_i y_i, \quad i = 1 \ldots n)$$

Affinity Matrix

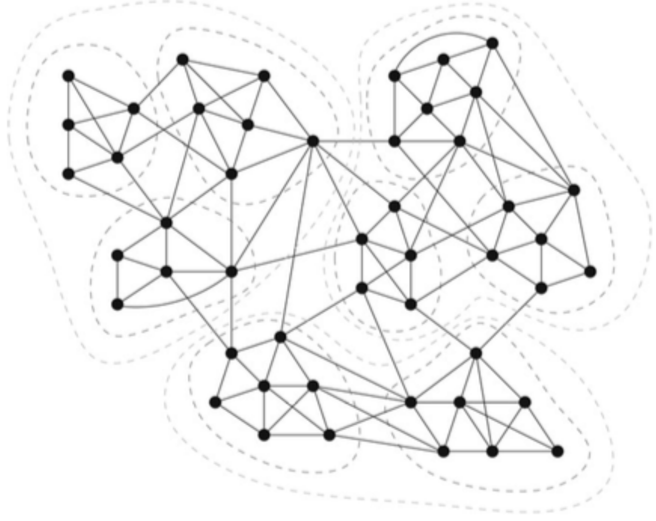


스펙트럼 클러스터링
Spectral Clustering

계층적 클러스터링
완전연결 / 단일연결 / 집단평균 / 워드

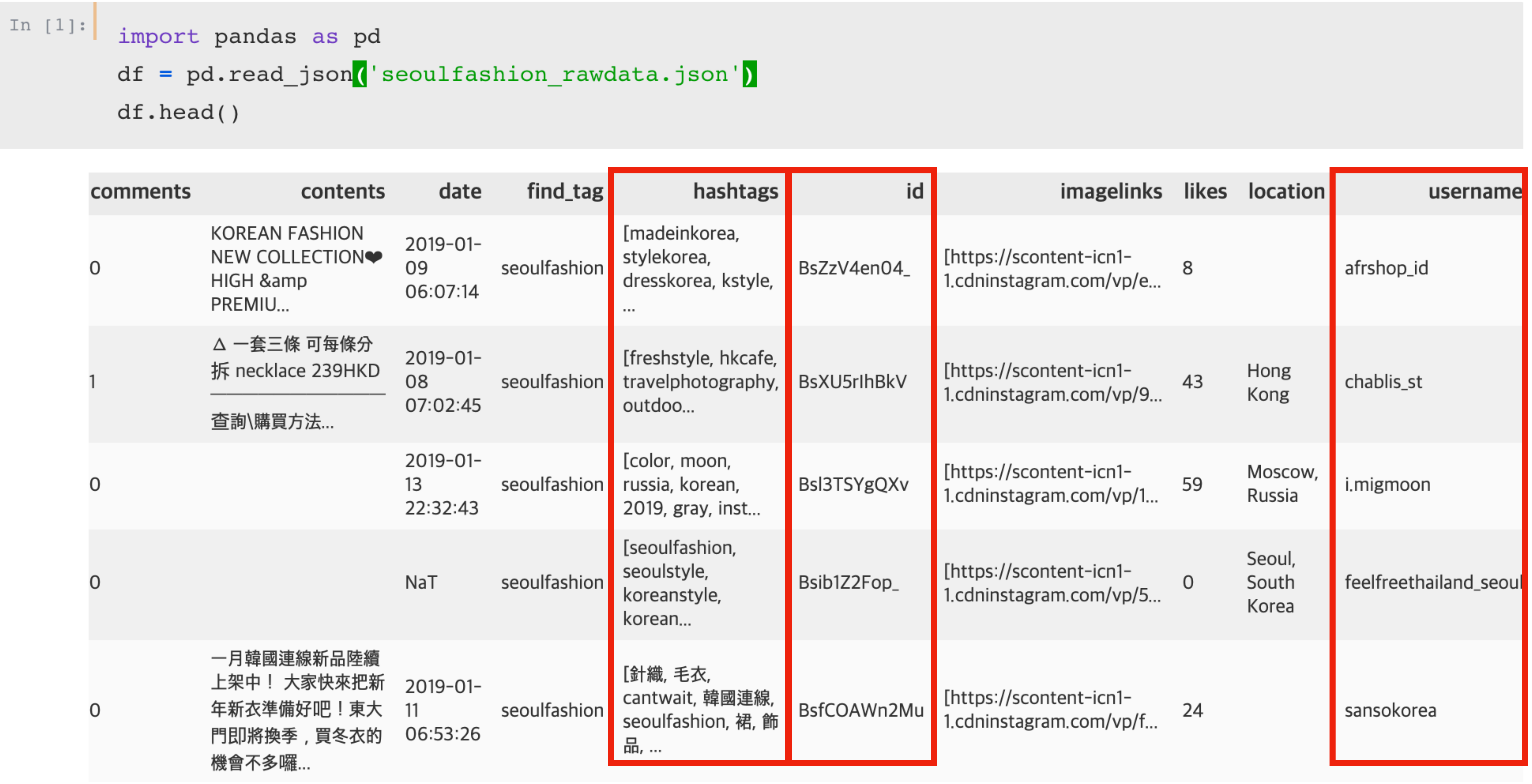
비계층적 클러스터링
K-Means

Graph Clustering



2. RawData Preprocessing

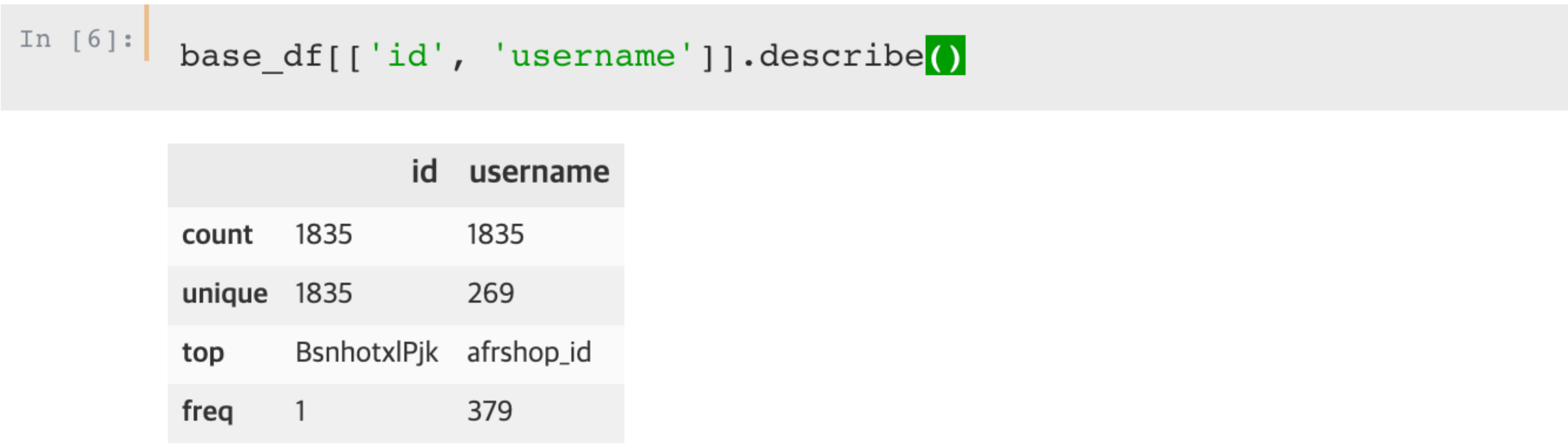
RawData Preprocessing



```
import pandas as pd
df = pd.read_json('seoulfashion_rawdata.json')
df.head()
```

comments	contents	date	find_tag	hashtags	id	imagelinks	likes	location	username
0	KOREAN FASHION NEW COLLECTION❤ HIGH &amp PREMIU...	2019-01-09 06:07:14	seoulfashion	[madeinkorea, stylekorea, dresskorea, kstyle, ...	BsZzV4en04_	[https://scontent-icn1-1.cdninstagram.com/vp/e...	8		afrshop_id
1	△ 一套三條 可每條分拆 necklace 239HKD ———————— 查詢\購買方法...	2019-01-08 07:02:45	seoulfashion	[freshstyle, hkcafe, travelphotography, outdoo...	BsXU5rIhBkV	[https://scontent-icn1-1.cdninstagram.com/vp/9...	43	Hong Kong	chablis_st
0		2019-01-13 22:32:43	seoulfashion	[color, moon, russia, korean, 2019, gray, inst...	Bsl3TSYgQXv	[https://scontent-icn1-1.cdninstagram.com/vp/1...	59	Moscow, Russia	i.migmoon
0		NaT	seoulfashion	[seoulfashion, seoulstyle, koreanstyle, korean...	Bsib1Z2Fop_	[https://scontent-icn1-1.cdninstagram.com/vp/5...	0	Seoul, South Korea	feelfreethailand_seoul
0	一月韓國連線新品陸續上架中！ 大家快來把新年新衣準備好吧！東大門即將換季，買冬衣的機會不多囉...	2019-01-11 06:53:26	seoulfashion	[針織, 毛衣, cantwait, 韓國連線, seoulfashion, 裙, 飾品, ...	BsfCOAWn2Mu	[https://scontent-icn1-1.cdninstagram.com/vp/f...	24		sansokorea

RawData Preprocessing

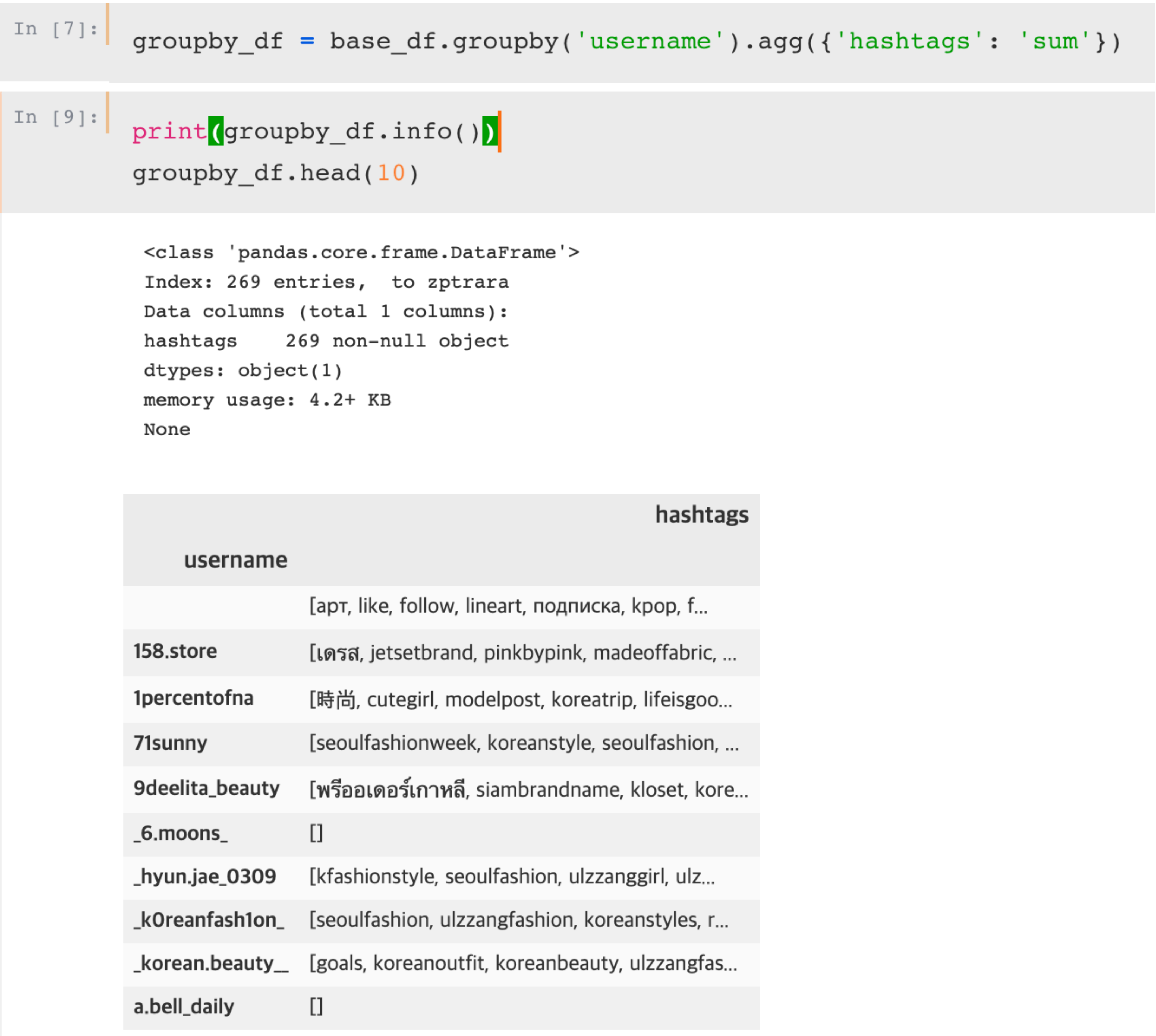


```
In [6]: base_df[['id', 'username']].describe()
```

	id	username
count	1835	1835
unique	1835	269
top	BsnhotxlPjk	afrshop_id
freq	1	379

base_df → groupby_df

username - all hashtag list
+ delete empty list



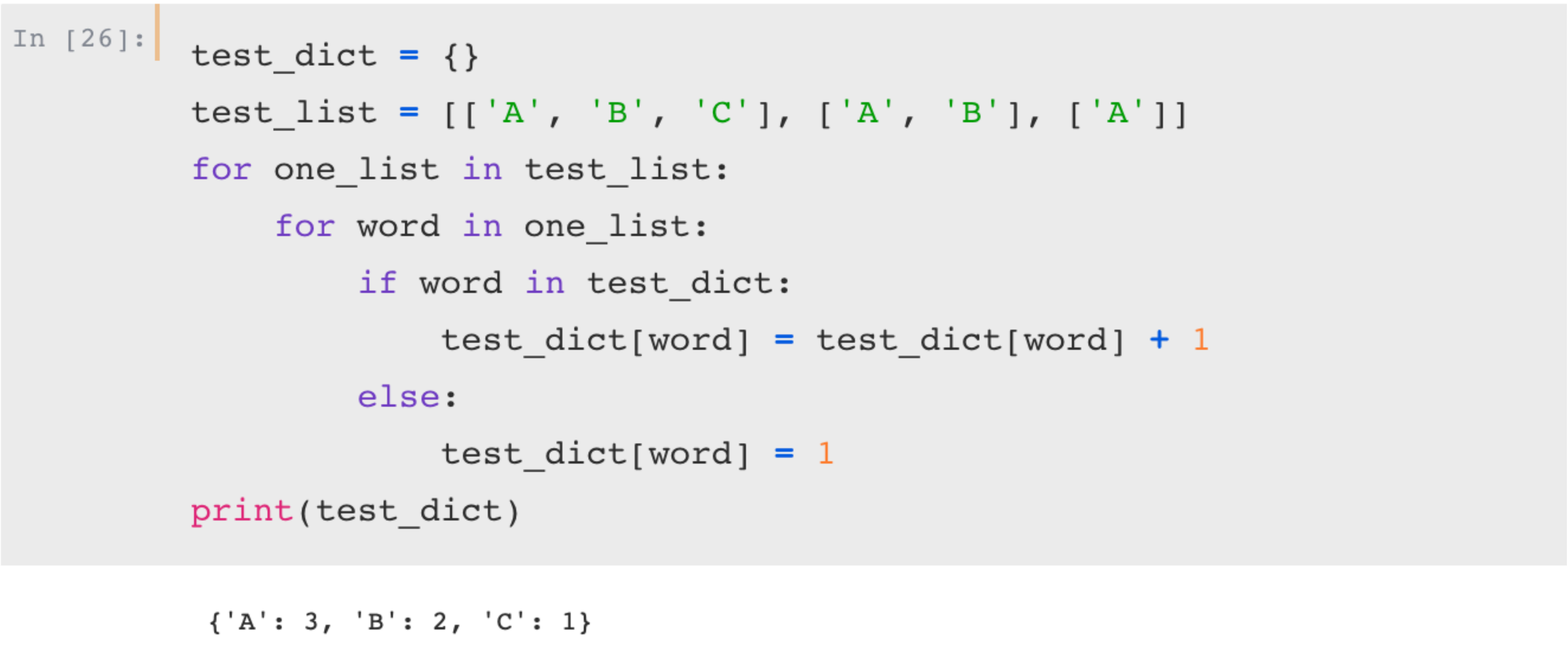
```
In [7]: groupby_df = base_df.groupby('username').agg({'hashtags': 'sum'})
```

```
In [9]: print(groupby_df.info())
        groupby_df.head(10)
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 269 entries,  to zptrara
Data columns (total 1 columns):
hashtags    269 non-null object
dtypes: object(1)
memory usage: 4.2+ KB
None
```

	hashtags
username	
	[арт, like, follow, lineart, подписка, kpop, f...
158.store	[เดรส, jetsetbrand, pinkbypink, madeoffabric, ...
1percentofna	[時尚, cutegirl, modelpost, koreatrip, lifeisgoo...
71sunny	[seoulfashionweek, koreanstyle, seoulfashion, ...
9deelita_beauty	[พรีออเดอร์เกาหลี, siambrandname, kloset, kore...
6.moons	[]
_hyun.jae_0309	[kfashionstyle, seoulfashion, ulzzanggirl, ulz...
k0reanfash1on	[seoulfashion, ulzzangfashion, koreanstyles, r...
_korean.beauty__	[goals, koreanoutfit, koreanbeauty, ulzzangfas...
a.bell_daily	[]

RawData Preprocessing



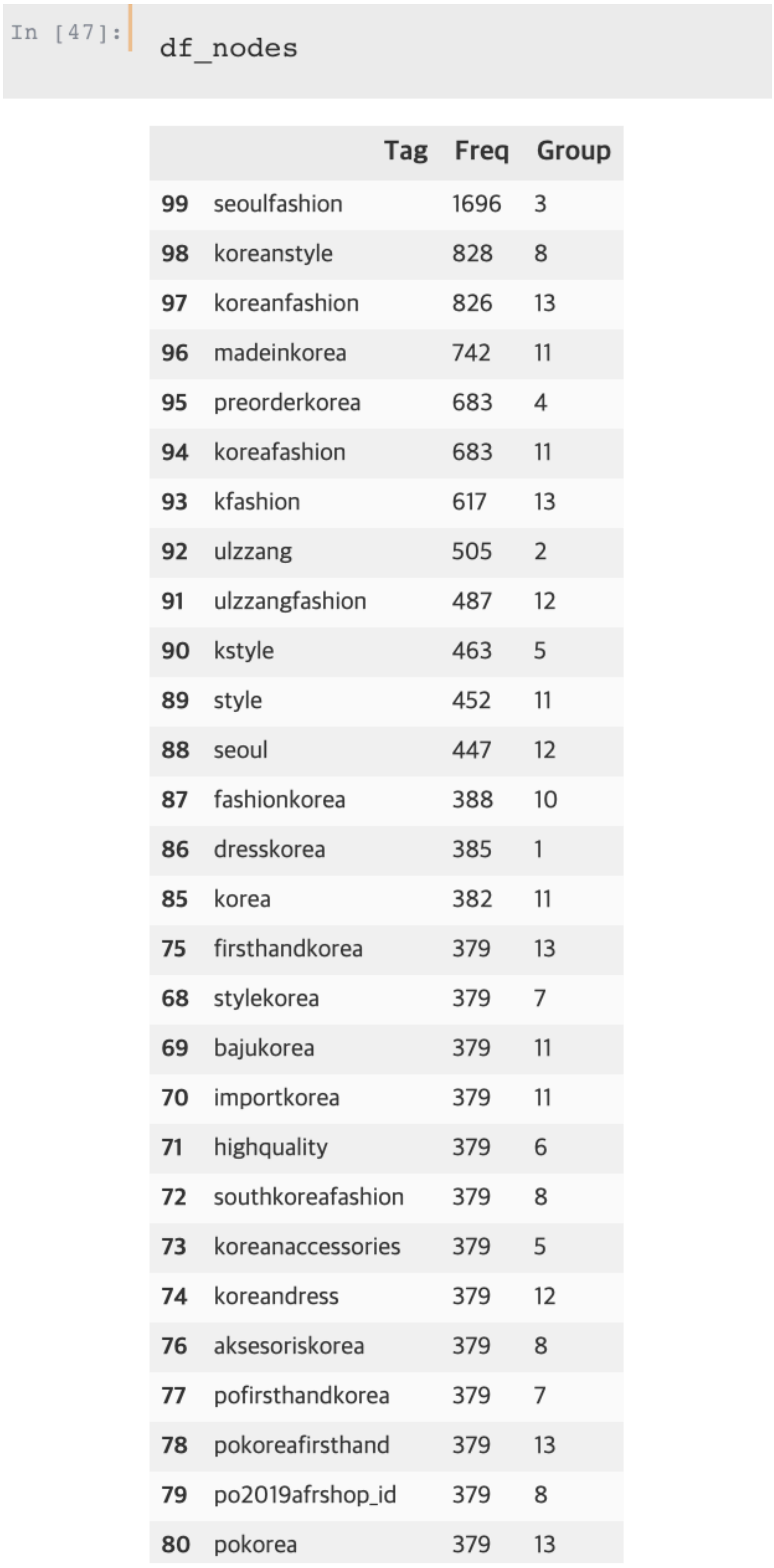
```
In [26]: test_dict = {}
         test_list = [['A', 'B', 'C'], ['A', 'B'], ['A']]
         for one_list in test_list:
             for word in one_list:
                 if word in test_dict:
                     test_dict[word] = test_dict[word] + 1
                 else:
                     test_dict[word] = 1
         print(test_dict)

         {'A': 3, 'B': 2, 'C': 1}
```

groupby_df → df_nodes

Hashtag frequency count

Sort by frequency

Select Top 30 Hashtag



```
In [47]: df_nodes
```

	Tag	Freq	Group
99	seoulfashion	1696	3
98	koreanstyle	828	8
97	koreanfashion	826	13
96	madeinkorea	742	11
95	preorderkorea	683	4
94	koreafashion	683	11
93	kfashion	617	13
92	ulzzang	505	2
91	ulzzangfashion	487	12
90	kstyle	463	5
89	style	452	11
88	seoul	447	12
87	fashionkorea	388	10
86	dresskorea	385	1
85	korea	382	11
75	firsthandkorea	379	13
68	stylekorea	379	7
69	bajukorea	379	11
70	importkorea	379	11
71	highquality	379	6
72	southkoreafashion	379	8
73	koreanaccessories	379	5
74	koreandress	379	12
76	aksesoriskorea	379	8
77	pofirsthandkorea	379	7
78	pokoreafirsthand	379	13
79	po2019afrshop_id	379	8
80	pokorea	379	13

seoulfashion

koreanstyle

koreanfashion

madeinkorea

preorderkorea

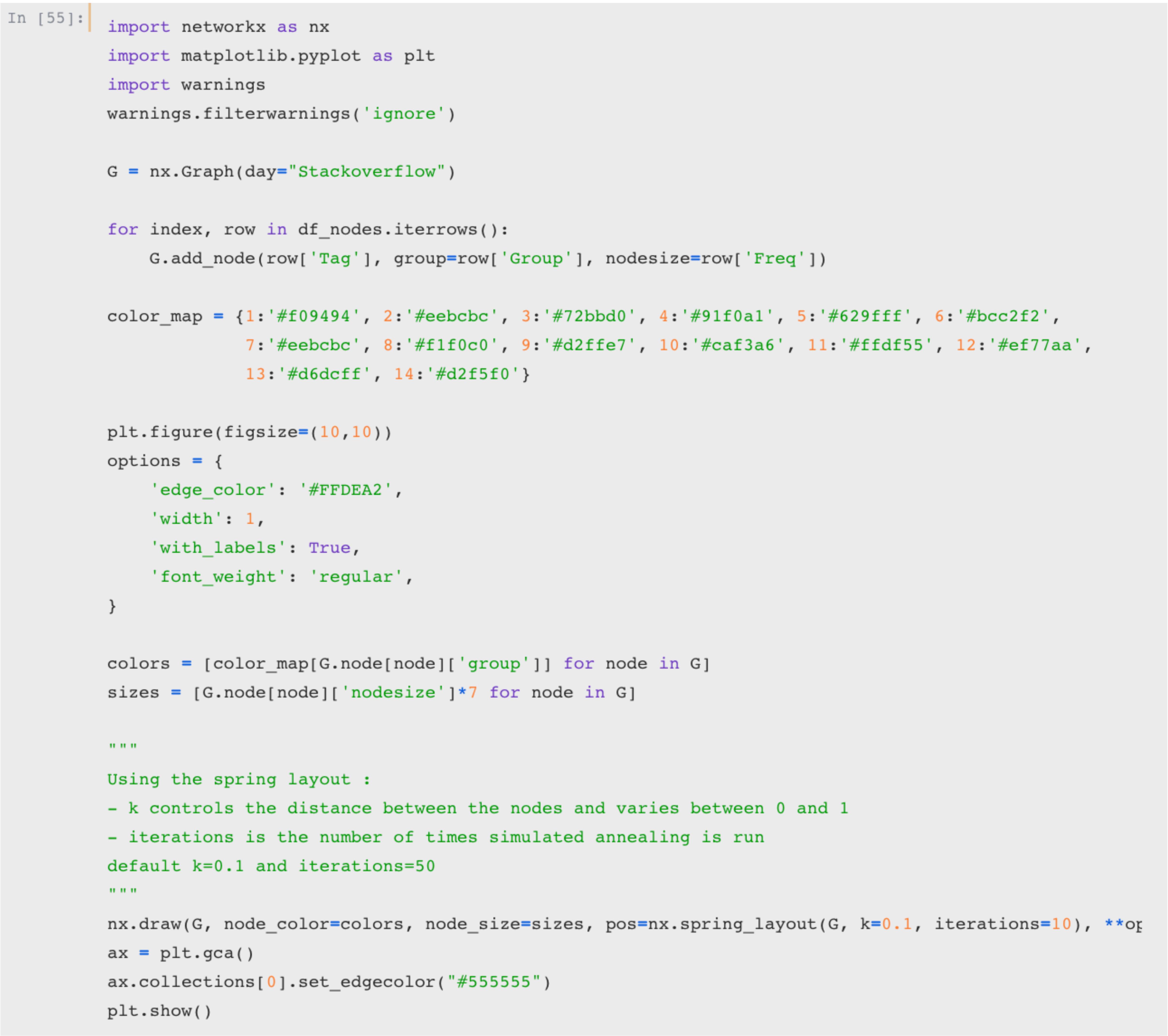
koreafashion

kfashion

ulzzang

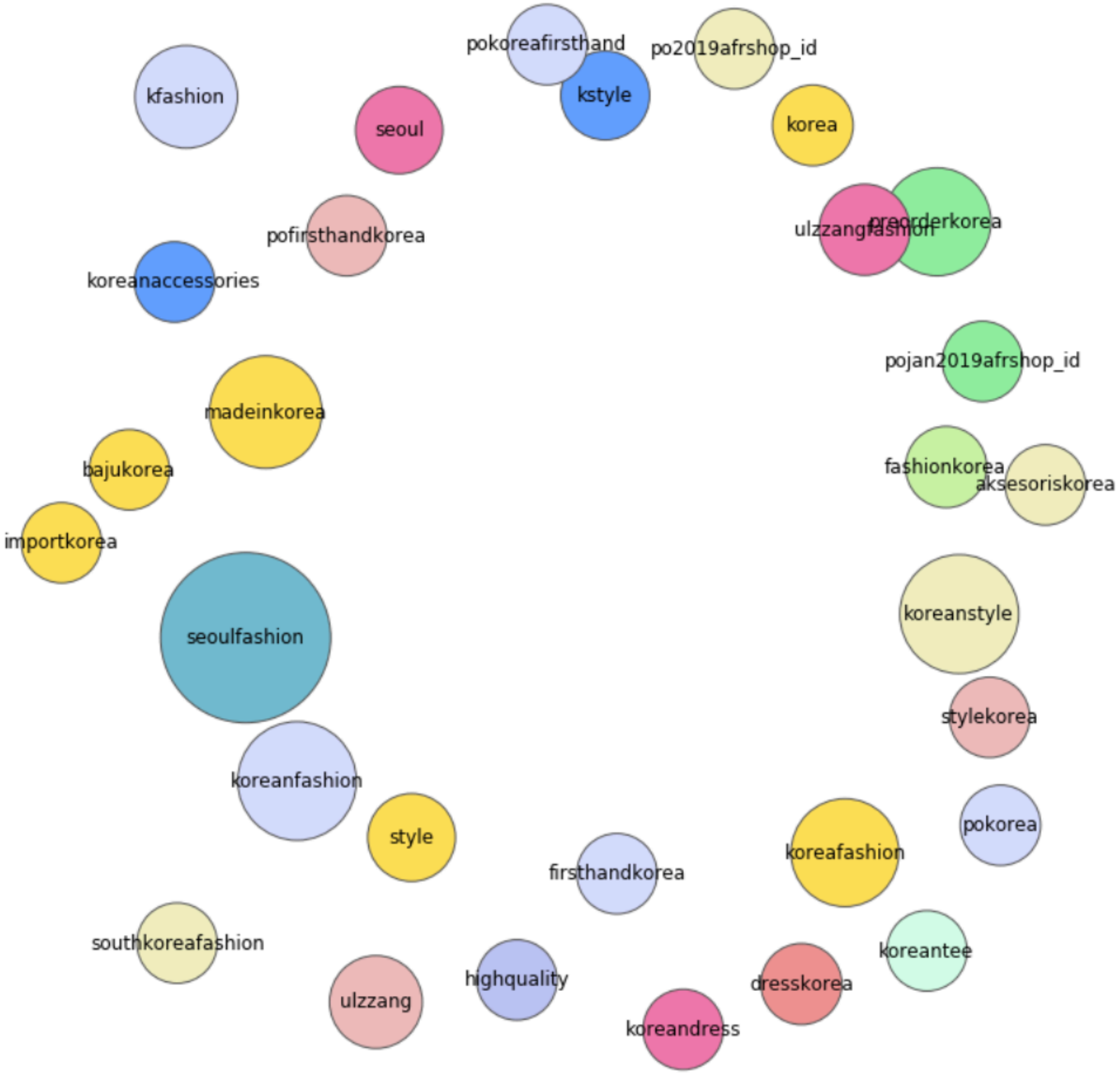
...

Build Node with Top30 Hashtag



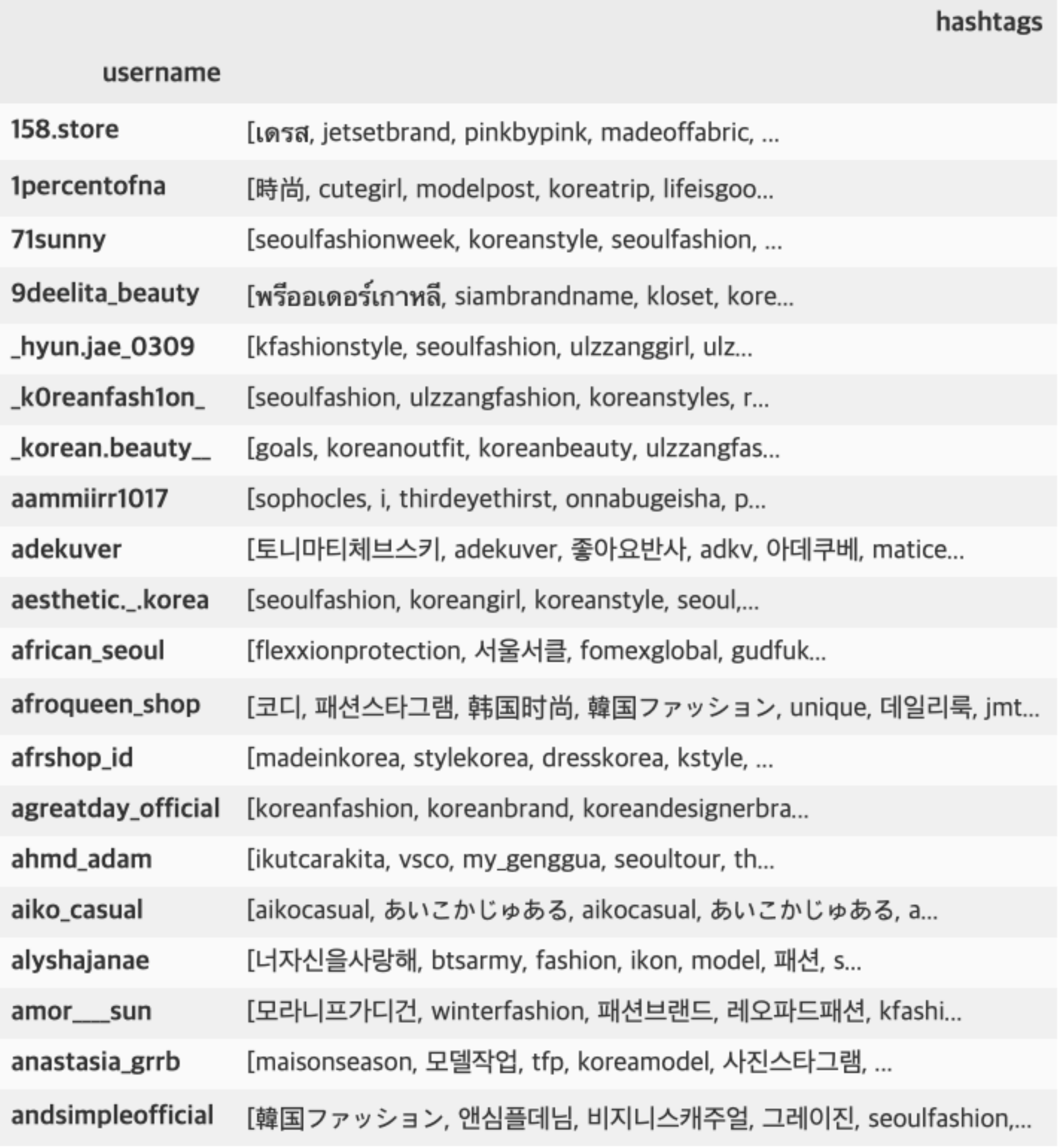
In [55]:

```
import networkx as nx
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')

G = nx.Graph(day="Stackoverflow")

for index, row in df_nodes.iterrows():
    G.add_node(row['Tag'], group=row['Group'], nodesize=row['Freq'])

color_map = {1:'#f09494', 2:'#eebcbc', 3:'#72bbd0', 4:'#91f0a1', 5:'#629fff', 6:'#bcc2f2',
             7:'#eebcbc', 8:'#f1f0c0', 9:'#d2ffe7', 10:'#caf3a6', 11:'#ffdf55', 12:'#ef77aa',
             13:'#d6dcff', 14:'#d2f5f0'}

plt.figure(figsize=(10,10))
options = {
    'edge_color': '#FFDEA2',
    'width': 1,
    'with_labels': True,
    'font_weight': 'regular',
}

colors = [color_map[G.node[node]['group']] for node in G]
sizes = [G.node[node]['nodesize']*7 for node in G]

"""
Using the spring layout :
- k controls the distance between the nodes and varies between 0 and 1
- iterations is the number of times simulated annealing is run
default k=0.1 and iterations=50
"""
nx.draw(G, node_color=colors, node_size=sizes, pos=nx.spring_layout(G, k=0.1, iterations=10), **op
ax = plt.gca()
ax.collections[0].set_edgecolor("#555555")
plt.show()
```



3. Plan

Plan 1. Build Affinity Matrix & Clustering



	hashtags
username	
158.store	[เดรส, jetsetbrand, pinkbypink, madeoffabric, ...
1percentofna	[時尚, cutegirl, modelpost, koreatrip, lifeisgoo...
71sunny	[seoulfashionweek, koreanstyle, seoulfashion, ...
9deelita_beauty	[พรีออเดอร์เกาหลี, siambrandname, kloset, kore...
_hyun.jae_0309	[kfashionstyle, seoulfashion, ulzzanggirl, ulz...
k0reanfash1on	[seoulfashion, ulzzangfashion, koreanstyles, r...
_korean.beauty__	[goals, koreanoutfit, koreanbeauty, ulzzangfas...
aammiirr1017	[sophocles, i, thirdeyethirst, onnabugeisha, p...
adekuver	[토니마티체브스키, adekuver, 좋아요반사, adkv, 아데쿠베, matice...
aesthetic._.korea	[seoulfashion, koreangirl, koreanstyle, seoul,...
african_seoul	[flexxionprotection, 서울서클, fomexglobal, gudfuk...
afroqueen_shop	[코디, 패션스타그램, 韩国时尚, 韓国ファッション, unique, 데일리룩, jmt...
afrshop_id	[madeinkorea, stylekorea, dresskorea, kstyle, ...
agreatday_official	[koreanfashion, koreanbrand, koreandesignerbra...
ahmd_adam	[ikutcarakita, vsco, my_genggua, seoultour, th...
aiko_casual	[aikocasual, あいこかじゅある, aikocasual, あいこかじゅある, a...
alyshajanae	[너자신을사랑해, btsarmy, fashion, ikon, model, 패션, s...
amor___sun	[모라니프가디건, winterfashion, 패션브랜드, 레오파드패션, kfashi...
anastasia_grrb	[maisonseason, 모델작업, tfp, koreamodel, 사진스타그램, ...
andsimpleofficial	[韓国ファッション, 앤심플데님, 비지니스캐주얼, 그레이진, seoulfashion,...

코사인 유사계수
Cosine Coefficient

피어슨 상관계수
Pearson Correlation Coefficient

유클리디언 유사도 함수
Euclidean Similarity

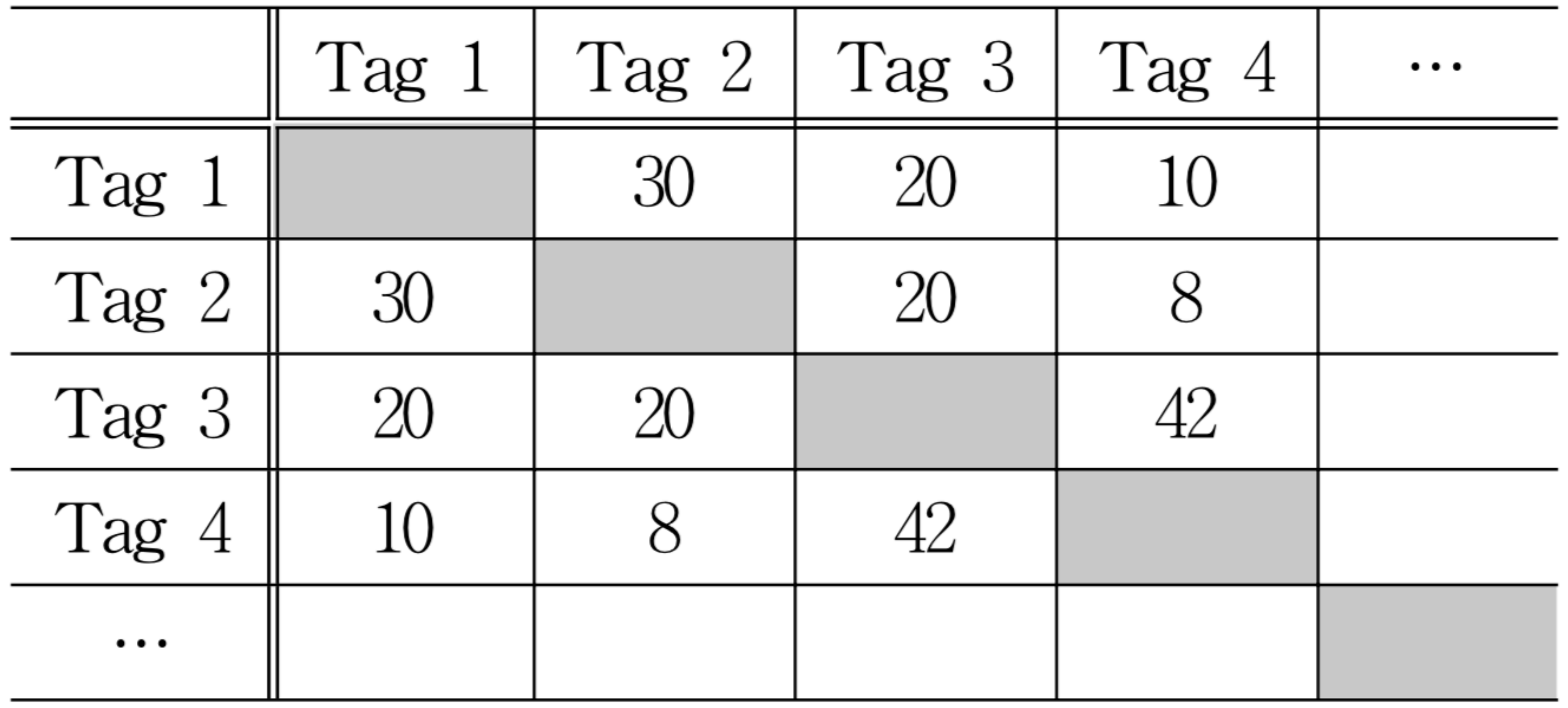
Similarity Calculation

S (코디, 패션스타그램)
S (Fashion, Koreafashion)
...

	Tag 1	Tag 2	Tag 3	Tag 4	⋯
Tag 1		30	20	10	
Tag 2	30		20	8	
Tag 3	20	20		42	
Tag 4	10	8	42		
⋯					

Plan 2. Graph Clustering Algorithms

Affinity Matrix



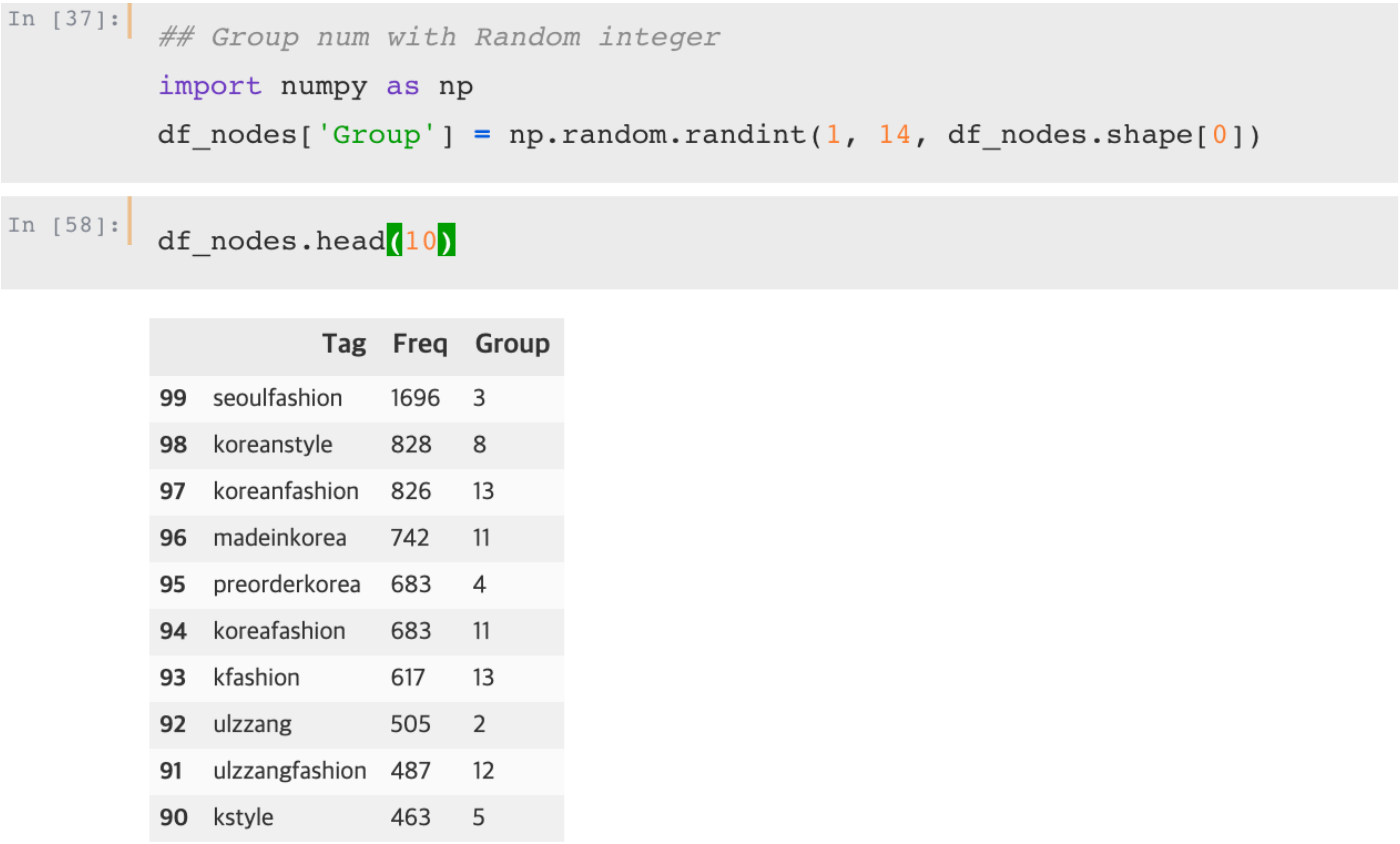
	Tag 1	Tag 2	Tag 3	Tag 4	…
Tag 1		30	20	10	
Tag 2	30		20	8	
Tag 3	20	20		42	
Tag 4	10	8	42		
…					

스펙트럼 클러스터링
Spectral Clustering

계층적 클러스터링
완전연결 / 단일연결 / 집단평균 / 워드

비계층적 클러스터링
K-Means

Grouped Tag Dataframe



```
In [37]: ## Group num with Random integer
         import numpy as np
         df_nodes['Group'] = np.random.randint(1, 14, df_nodes.shape[0])
```

```
In [58]: df_nodes.head(10)
```

	Tag	Freq	Group
99	seoulfashion	1696	3
98	koreanstyle	828	8
97	koreanfashion	826	13
96	madeinkorea	742	11
95	preorderkorea	683	4
94	koreafashion	683	11
93	kfashion	617	13
92	ulzzang	505	2
91	ulzzangfashion	487	12
90	kstyle	463	5

Plan 3. Where to use ?

기존의 연구

> 군집별 태그 추천을 통한 사용자 검색능력 향상
>
> 개인 맞춤형 태그 추천 시스템

새로운 연구 목표 (택1)

> 태그 군집분석을 통한 특정 기간의 트렌드 파악 : 패션 / 여행 / 축제 등
>
> 태그 군집과 사용자 정보를 활용한 광고성 / 악성 사용자 탐지
>
> 실시간 데이터 수집 및 시간별 태그 군집분석 (Stream Data Analysis)

References

스펙트럼 알고리즘 기초 : https://elecs.tistory.com/169

계층적 클러스터링 알고리즘 기초 : https://bab2min.tistory.com/219

LSA / LDA 알고리즘 기초 : https://bab2min.tistory.com/585

단어간 유사도 측정 수학 공식의 기초 :
https://stats.stackexchange.com/questions/289400/quantify-the-similarity-of-bags-of-words