プロジェクト
最終発表

目標

- Word2vecとgoogle検索からの連想語から作成されたネットワークの比較を行う。

- 二種類のネットワークと単語ごとの関係性に対し、理解を深める。

ネットワーク（グラフ）作成に使用したライブラリ

- Matplotlib

ネットワークの大きさの指定と、保存、表示に使用

- NetworkX

ネットワークにノード、エッジを追加するために使用。

ここで、出来るだけネットワークが重複しないように、下記を設定

```
pos = nx.spring_layout(G,k=1.5,iterations=10000)
```

引数kは大きくなるほど、ノード間の距離が広がる。

Iterationsは最適化を行う反復回数を指定。多いほど、レイアウトの精度が向上し、計算時間も増加する。（参考資料1）

Google検索 知識抽出

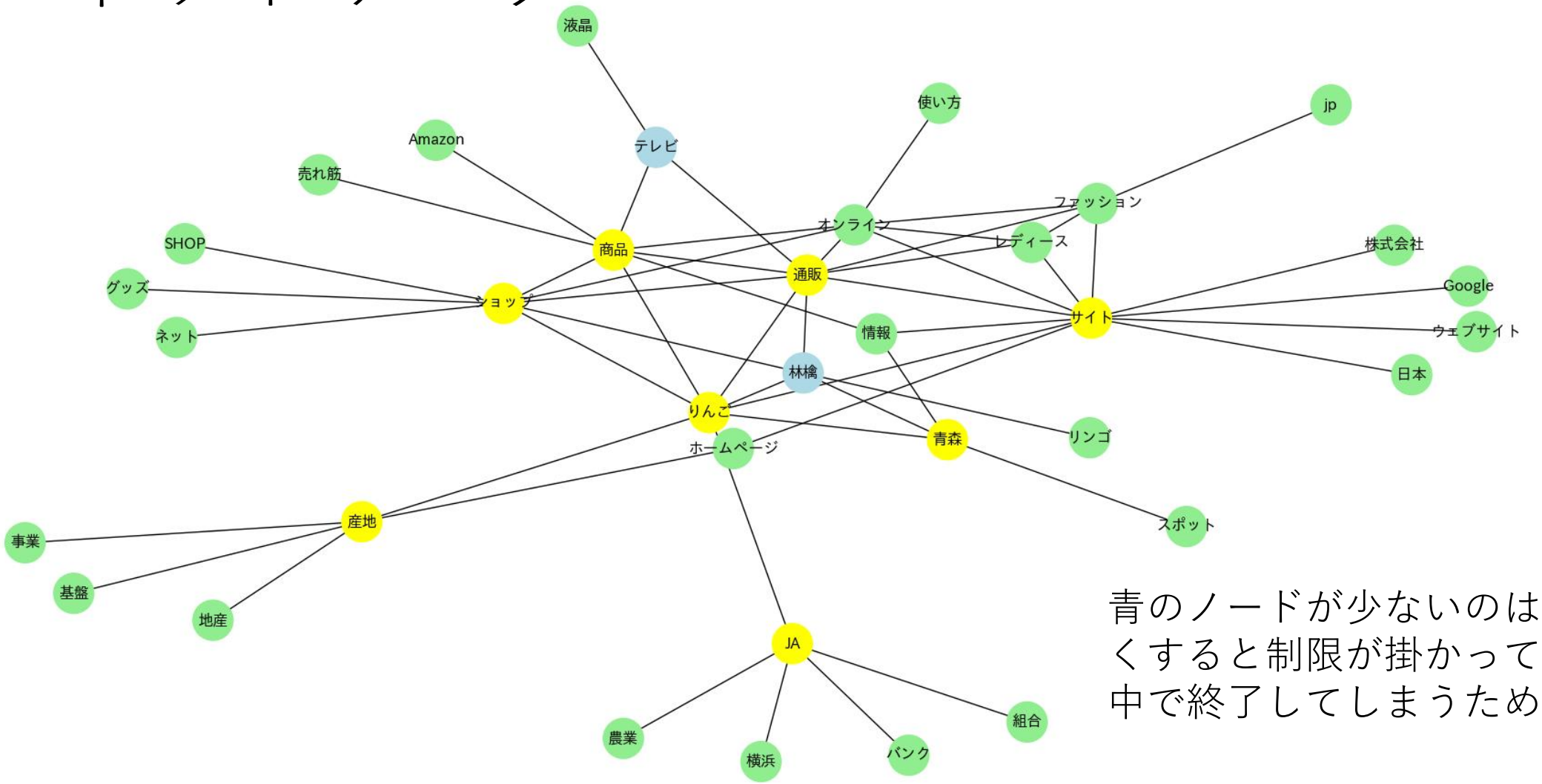
- 検索されるページの数を増やす。
  上位10ページを対象にする
- 対象にするページを増やすと、スクレイピング制限にかかる時間が早まる。そのため、timeモジュールを用いて実行間時間を増やし、制限がかかるまでの時間を極力伸ばす。

```
import time
time.sleep(1) # 1秒停止
```

- リストに含まれる単語の数を増やし、足切りを行うことで関連性が低い単語を減らす。

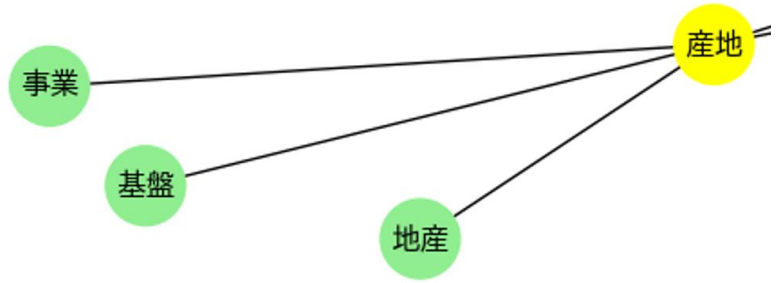
```
c = collections.Counter(saihin)
a = [i[0] for i in c.items() if i[1] >= 5] #5回以上出現のみ許容
```

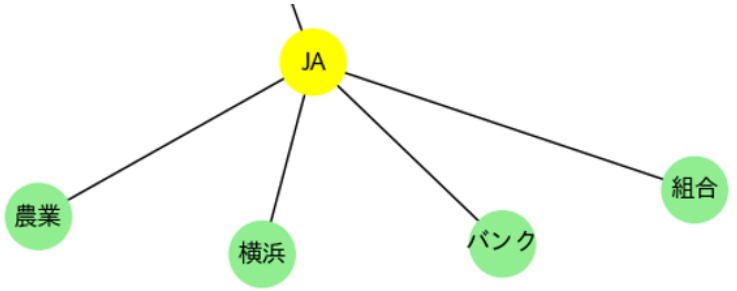
Google検索での関連語句から作成したネットワーク

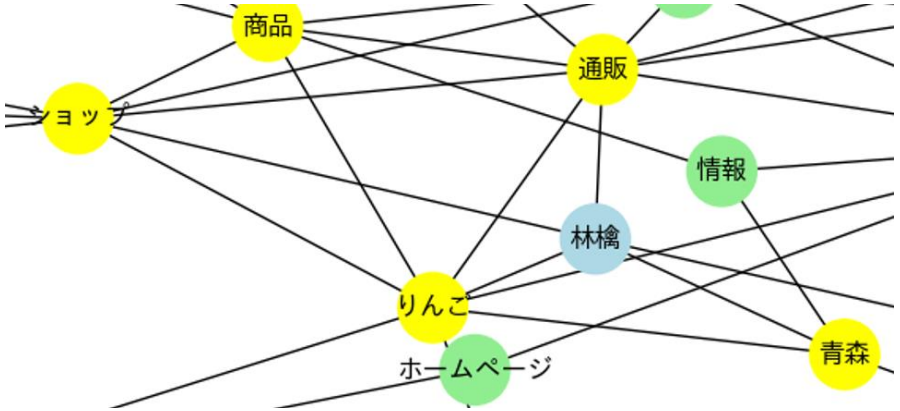


青のノードが少ないのは、検索を深くすると制限が掛かってしまい、途中で終了してしまうため。

Google検索での関連語句から作成したネットワーク







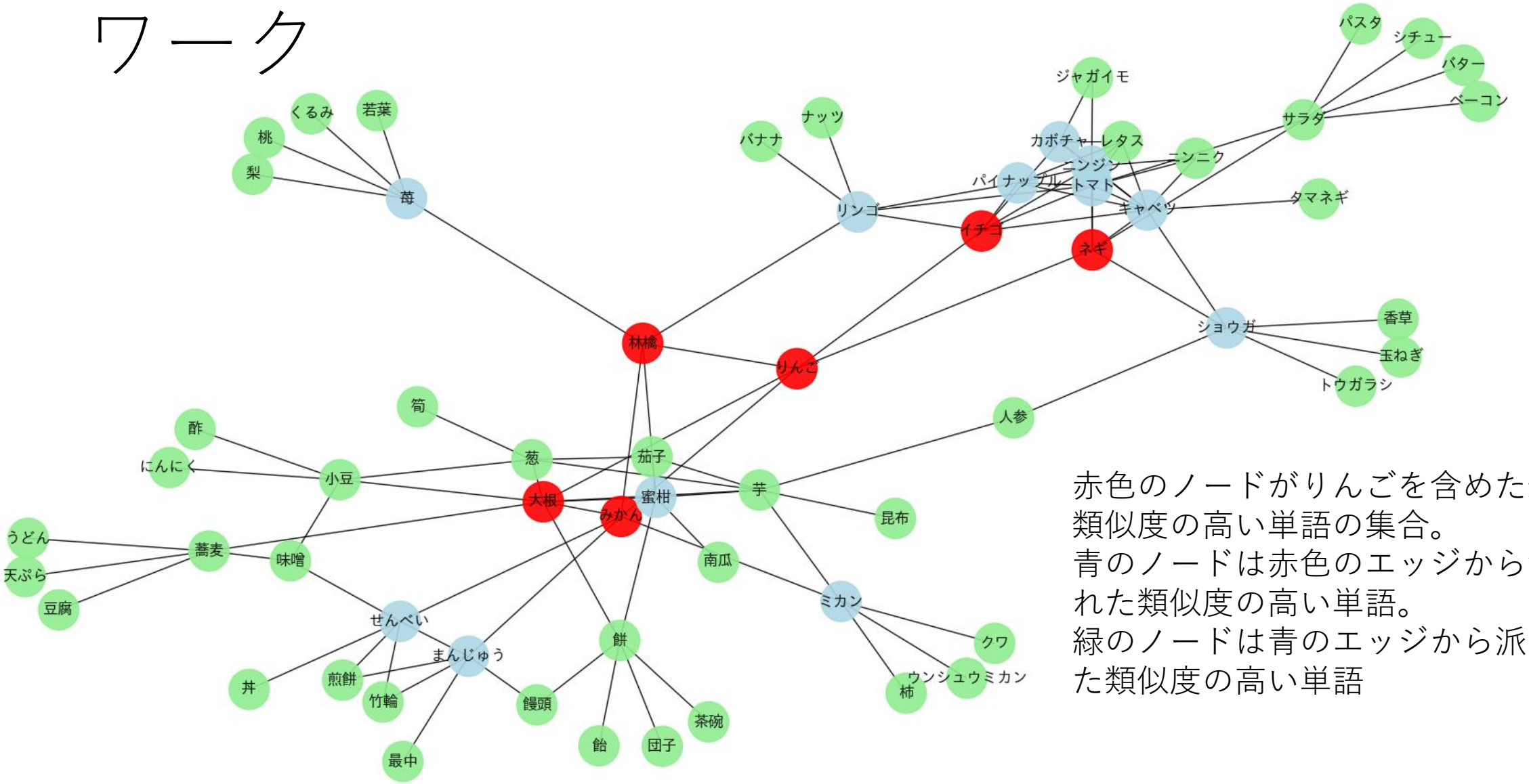
黄色のノードが最初に関連度の高い単語の集合。
青のノードは黄色のエッジから派生された単語の集合。
緑のノードは青のエッジから派生された単語の集合

Word2vec 知識抽出

- 日本語Wikipediaエンティティベクトルのバイナリ版を使用
  txt版より動作が軽く、実行時間が短くなる。

- 主な使用したライブラリ
  genism
KeyedVectorsを用いてバイナリファイルからモデルを作成。
類似度が高い単語を繰り返し調べ、ネットワークを作成する。

Word2vecでの類似度で作成したネットワーク



赤色のノードがりんごを含めた最初に類似度の高い単語の集合。
青のノードは赤色のエッジから派生された類似度の高い単語。
緑のノードは青のエッジから派生された類似度の高い単語

結果

- 人間が連想する単語はgoogle検索寄りだと考える。

- Word2vecではりんごから野菜や果物、食べ物に繋がる

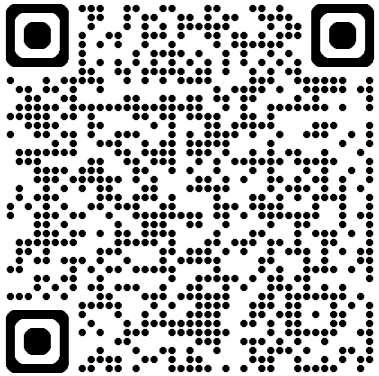
  これは、同じような文面の時にその単語の代わりとなる単語がそのようなものが多いためだと考える。

- 逆に、google検索は人間が実際に検索し、その検索結果でのクリック数により表示場所が変動するため、人間が想定する単語がネットワークに表示されやすくなる。

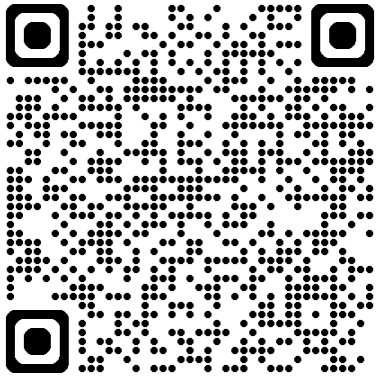
  他にも複数の要因は存在する（参考資料2）

参考資料

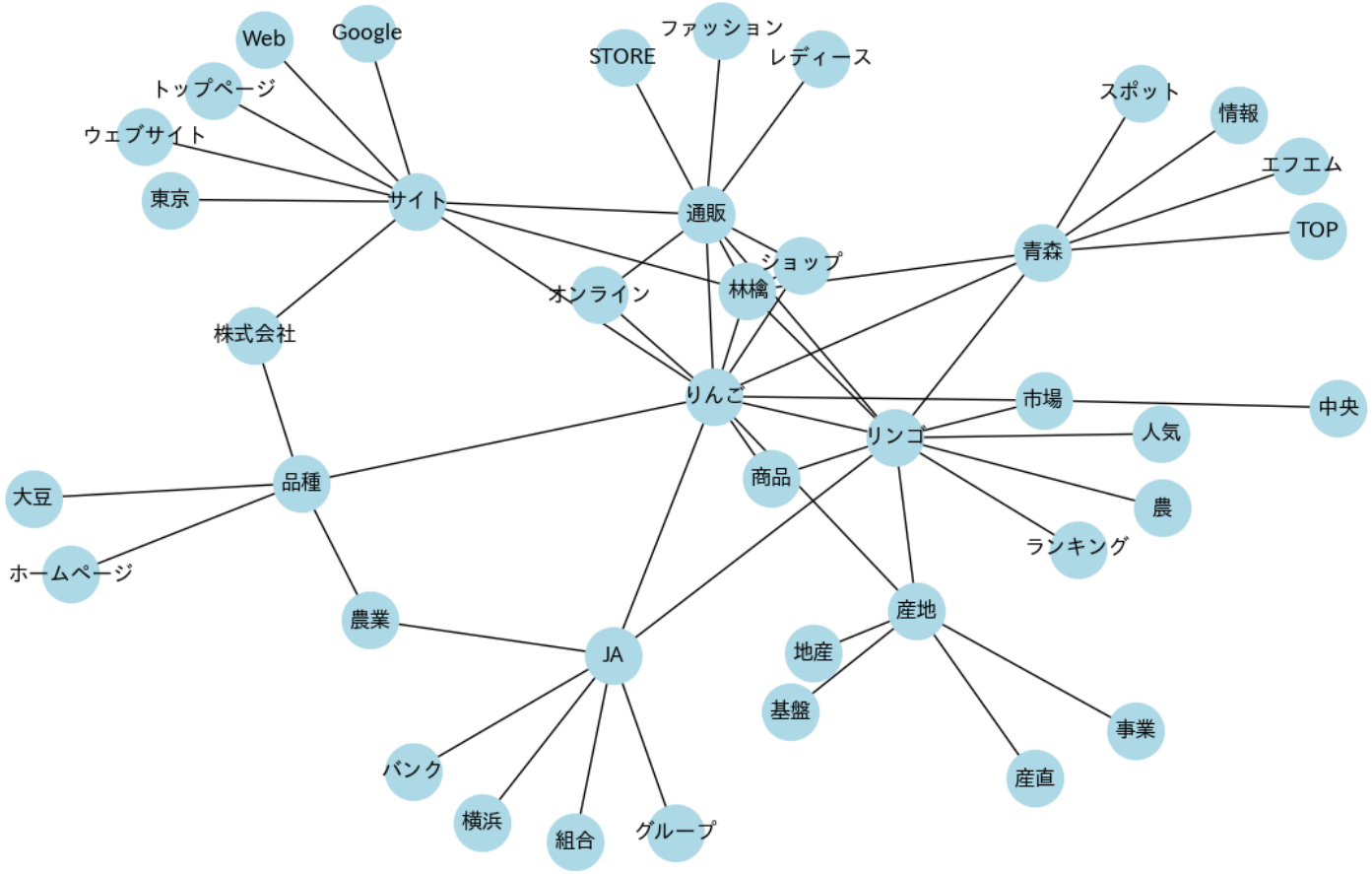
- 1 NetworkX Developers, spring_layout
spring_layout — NetworkX 3.3 documentation

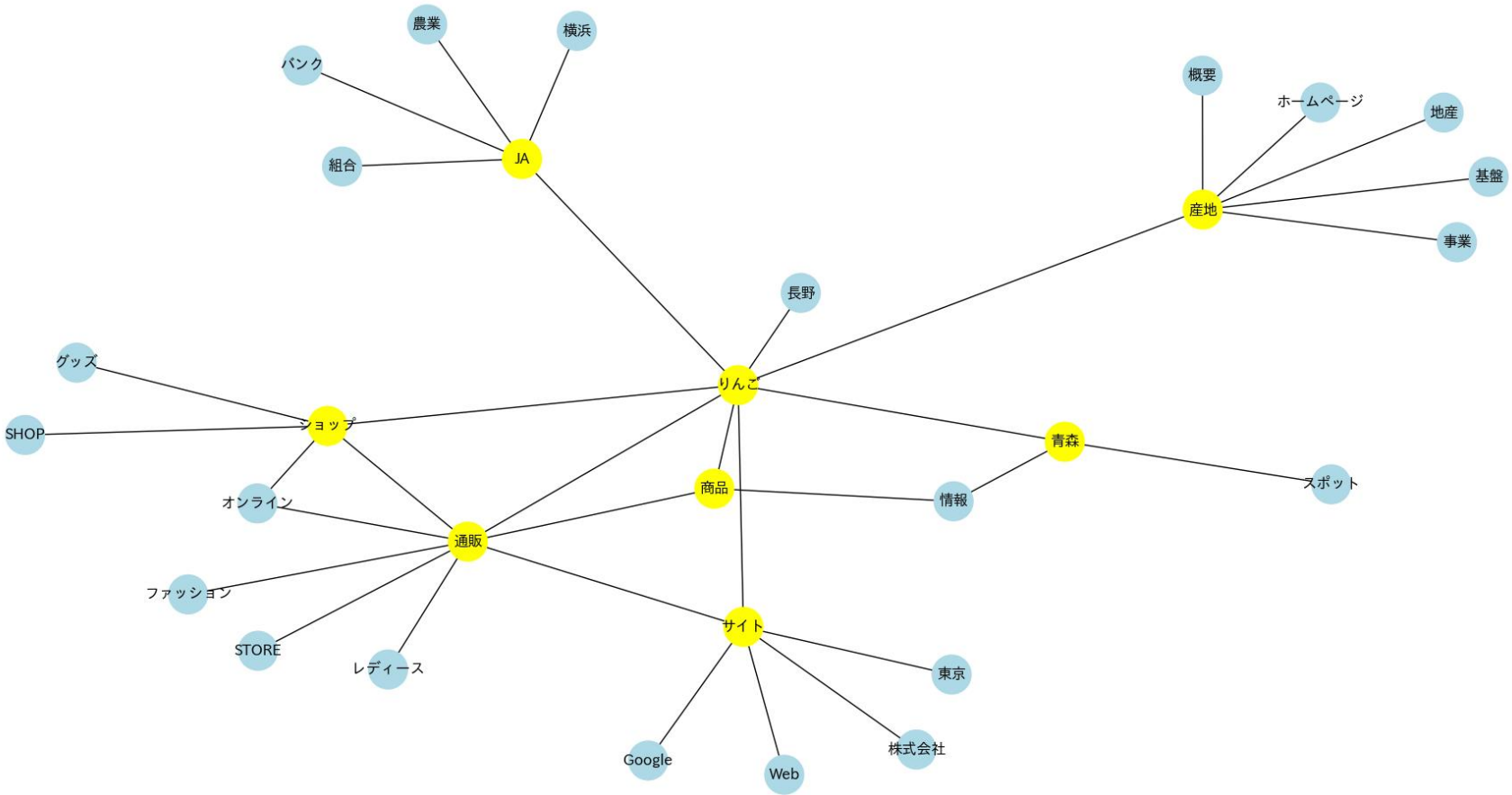


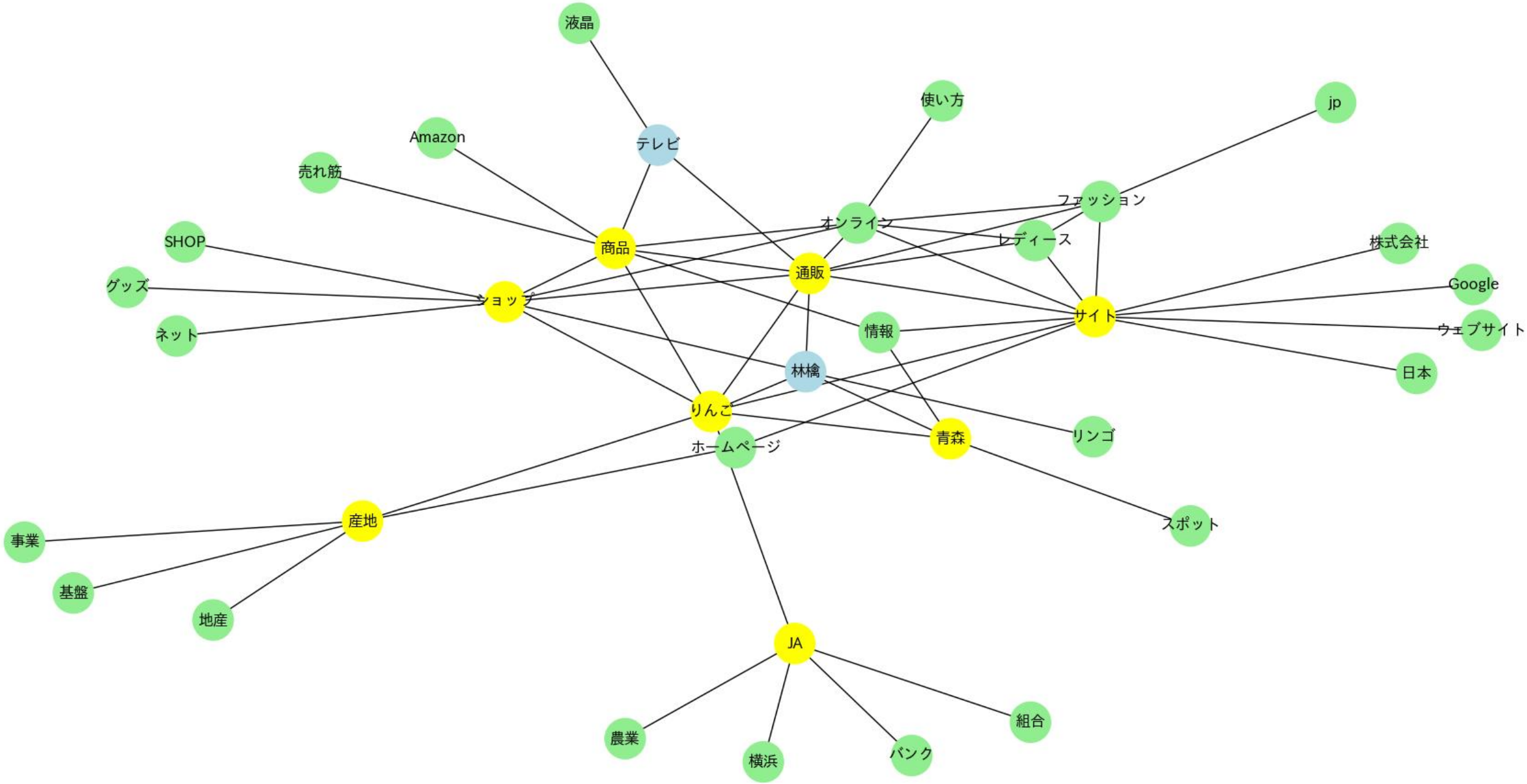
- 2 Google, 結果を自動的に生成する仕組み
ランキング結果 – Google 検索の仕組み



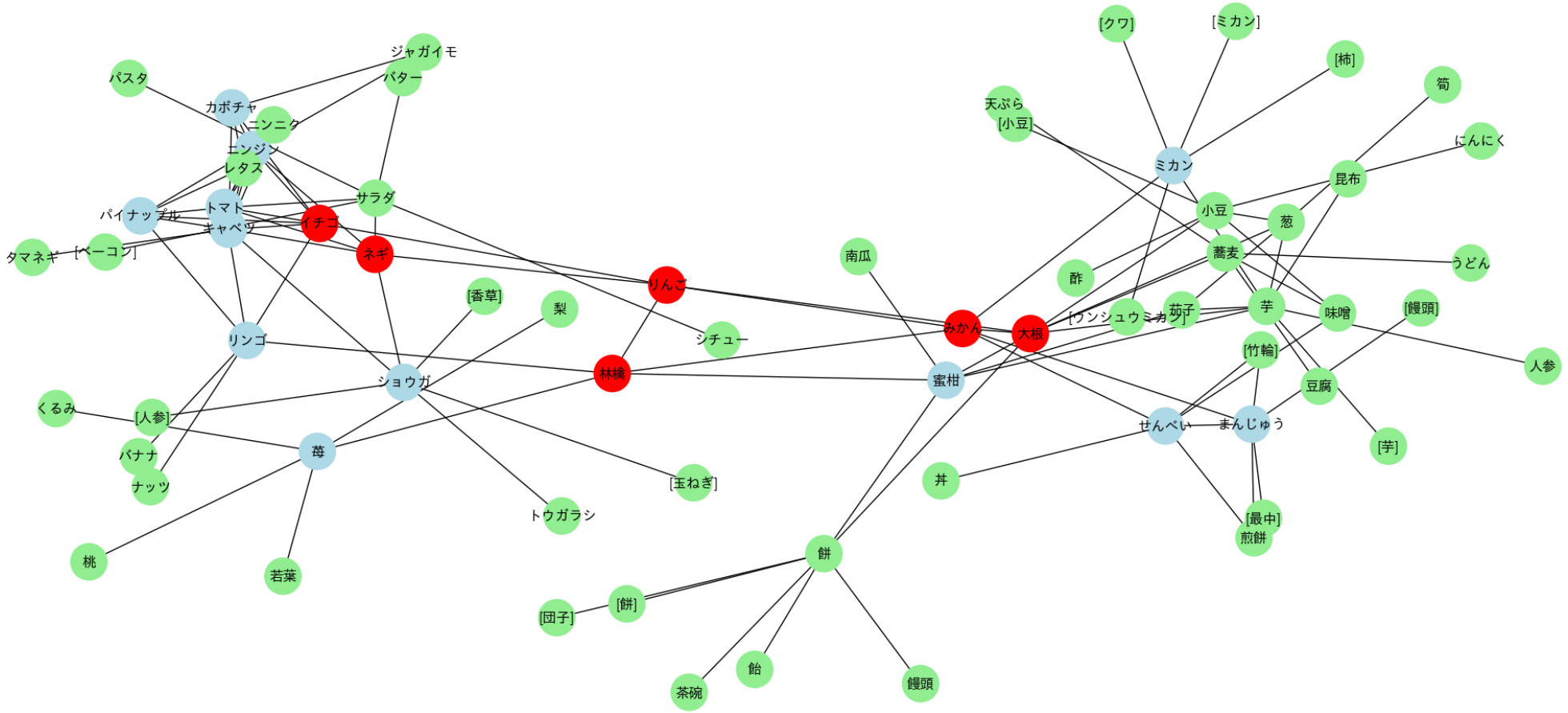
Google検索で単語の関係の表示

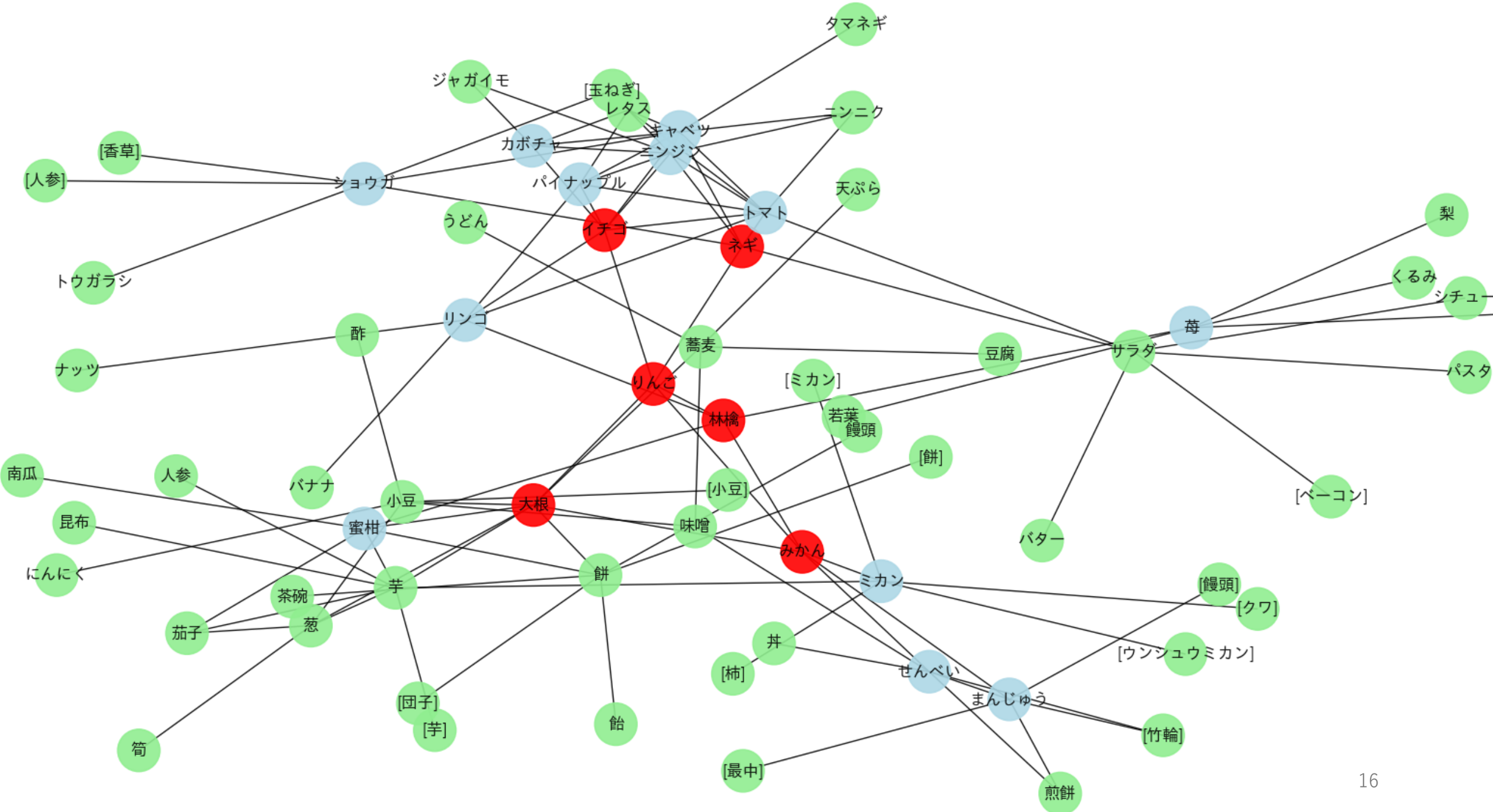


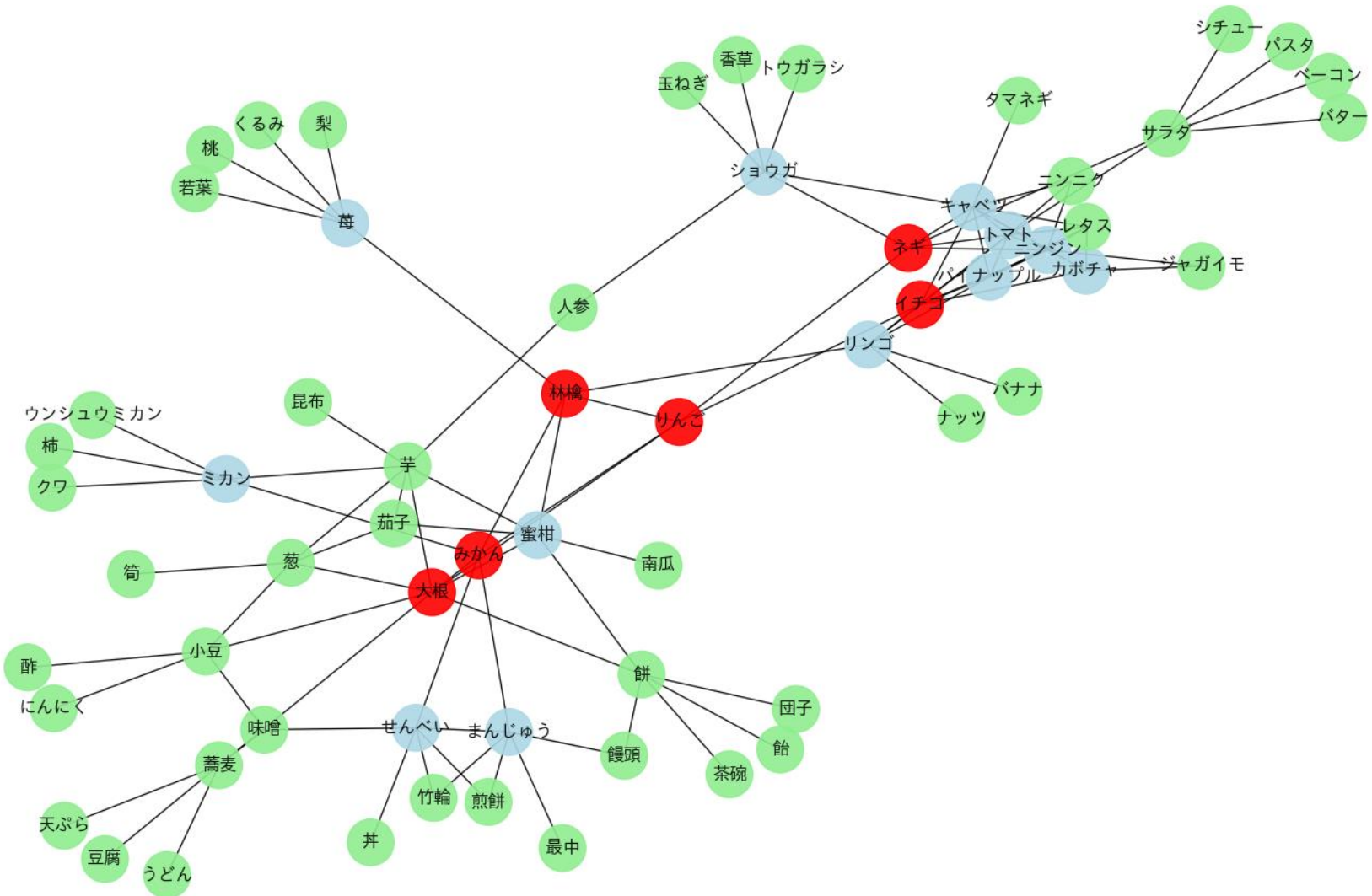

農業
横浜
バンク
JA
組合
概要
ホームページ
地産
基盤
産地
事業
長野
グッズ
りんご
SHOP
ショップ
青森
スポット
商品
情報
オンライン
通販
ファッション
STORE
レディース
サイト
東京
株式会社
Google
Web

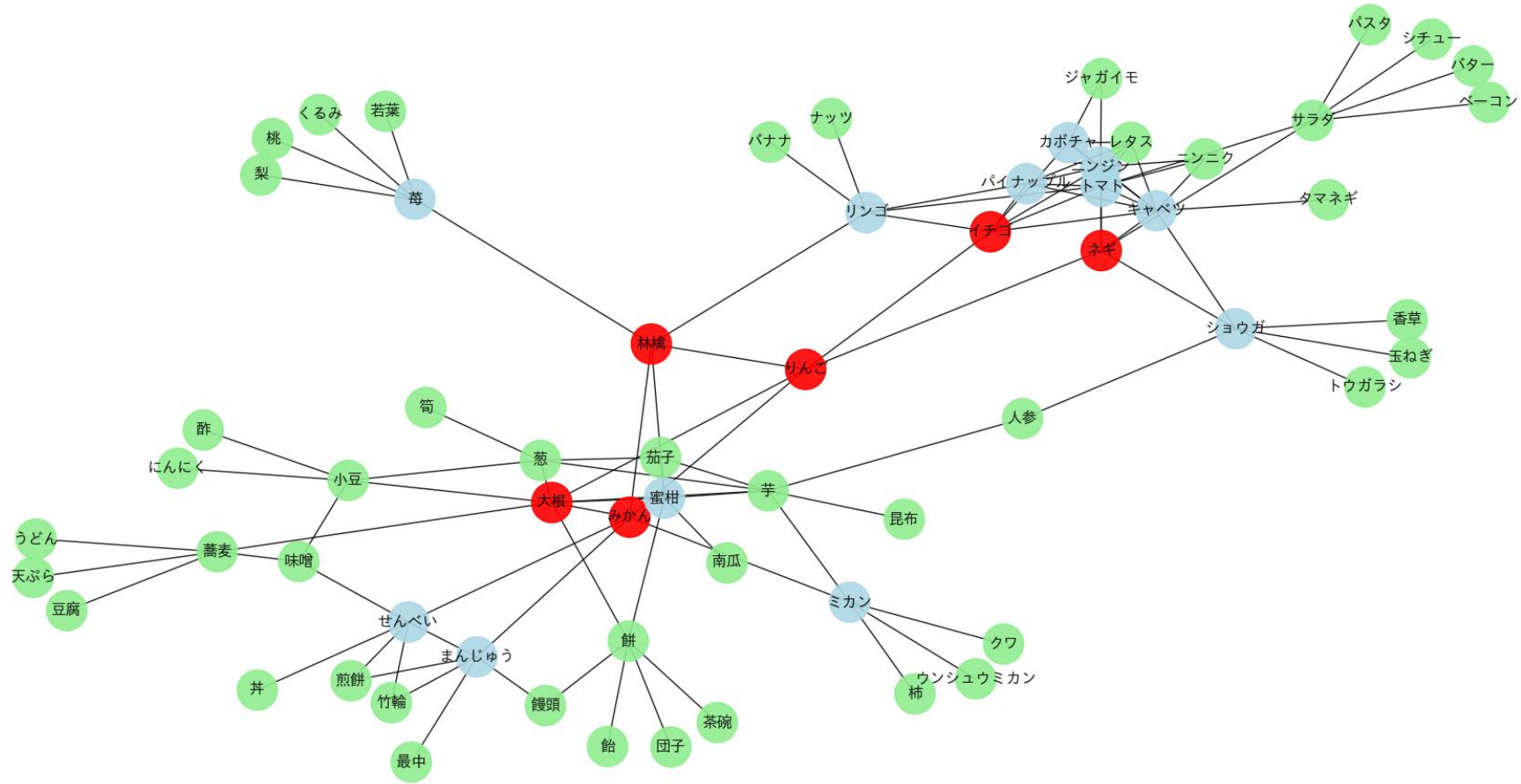

液晶
テレビ
使い方
jp
Amazon
売れ筋
ファッション
オンライン
SHOP
商品
レディース
株式会社
グッズ
通販
Google
ショップ
サイト
ネット
情報
ウェブサイト
林檎
日本
りんご
リンゴ
青森
ホームページ
産地
スポット
事業
基盤
地産
JA
組合
農業
横浜
バンク

word2vec

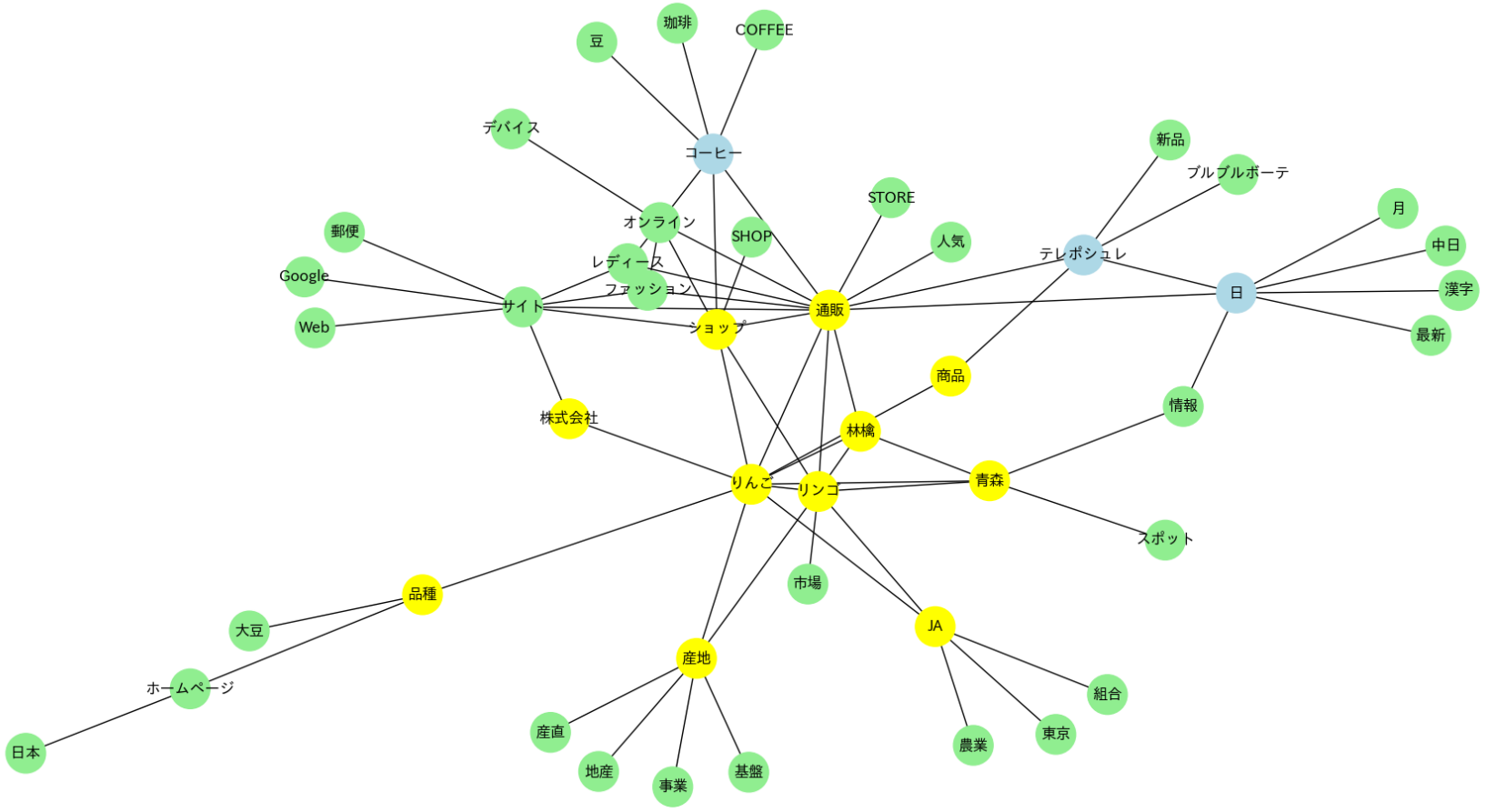

ジャガイモ
パスタ
バター
カボチャ
ニンニク
エンジン
レタス
サラダ
パイナップル
トマト
キャベツ
イチゴ
タマネギ
[ベーコン]
ネギ
りんご
[香草]
梨
シチュー
リンゴ
ショウガ
林檎
くるみ
[人参]
バナナ
ナッツ
苺
[玉ねぎ]
トウガラシ
桃
若葉
南瓜
みかん
大根
蜜柑
[クワ]
[ミカン]
[柿]
筍
天ぷら
[小豆]
ミカン
にんにく
昆布
小豆
葱
蕎麦
うどん
酢
[ワンシュウミカン]
茄子
芋
味噌
[饅頭]
[竹輪]
人参
豆腐
せんべい
まんじゅう
[芋]
丼
[最中]
煎餅
餅
[団子]
[餅]
飴
茶碗
饅頭


タマネギ
ジャガイモ
[玉ねぎ]
レタス
キャベツ
ニンニク
カボチャ
ニンジン
[香草]
[人参]
ショウガ
パイナップル
天ぷら
トマト
うどん
イチゴ
ネギ
トウガラシ
梨
くるみ
シチュー
リンゴ
酢
蕎麦
豆腐
苺
サラダ
ナッツ
りんご
[ミカン]
パスタ
若葉
林檎
饅頭
[餅]
南瓜
人参
バナナ
小豆
[小豆]
大根
[ベーコン]
昆布
蜜柑
味噌
バター
みかん
にんにく
餅
ミカン
[饅頭]
茶碗
芋
[クワ]
茄子
葱
[ウンシュウミカン]
丼
[柿]
せんべい
まんじゅう
[団子]
[竹輪]
[芋]
飴
筍
[最中]
煎餅


シチュー
パスタ
ベーコン
バター
サラダ
香草
トウガラシ
玉ねぎ
タマネギ
ショウガ
ニンニク
キャベツ
レタス
トマト
ニンジン
カボチャ
ジャガイモ
パイナップル
ネギ
イチゴ
リンゴ
バナナ
ナッツ
くるみ
梨
桃
若葉
苺
人参
林檎
りんご
昆布
ウンシュウミカン
柿
クワ
ミカン
芋
茄子
蜜柑
みかん
南瓜
筍
葱
大根
小豆
酢
にんにく
餅
団子
飴
茶碗
味噌
せんべい
まんじゅう
饅頭
蕎麦
竹輪
煎餅
天ぷら
豆腐
うどん
丼
最中


パスタ
シチュー
バター
ベーコン
サラダ
ジャガイモ
カボチャ
レタス
ニンジン
ニンニク
パイナップル
トマト
キャベツ
タマネギ
バナナ
ナッツ
リンゴ
イチゴ
ネギ
くるみ
若葉
桃
梨
苺
林檎
りんご
ショウガ
香草
玉ねぎ
トウガラシ
人参
筍
酢
にんにく
小豆
葱
茄子
芋
大根
蜜柑
みかん
昆布
うどん
天ぷら
豆腐
蕎麦
味噌
南瓜
ミカン
クワ
ウンシュウミカン
柿
せんべい
まんじゅう
餅
井
煎餅
竹輪
饅頭
茶碗
餡
団子
最中

最新 google


豆
珈琲
COFFEE
デバイス
コーヒー
新品
ブルブルボーテ
STORE
オンライン
SHOP
人気
月
中日
郵便
レディース
テレポシュレ
Google
ファッション
日
漢字
サイト
通販
Web
ショップ
最新
商品
株式会社
情報
林檎
りんご
リンゴ
青森
スポット
市場
品種
JA
大豆
産地
ホームページ
組合
産直
農業
東京
日本
地産
事業
基盤