



Tender is the Night

Есаян Диана



Содержание

1. Background info
2. Цель работы
3. Ход работы 1
4. Результаты 1
5. Ход работы 2
6. Результаты 2
7. Выводы

Background info

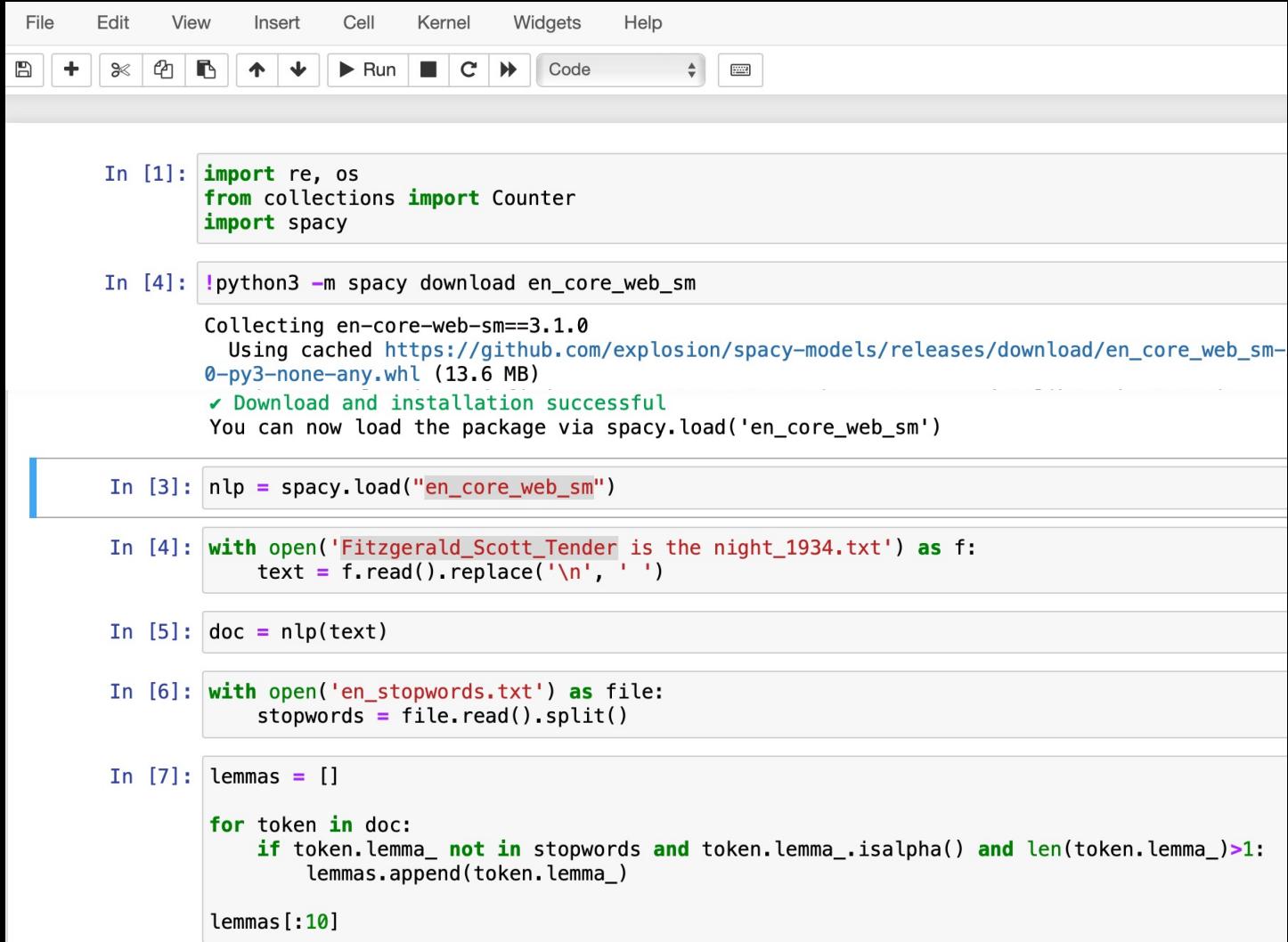
- «Ночь нежна» - последний завершённый роман американского писателя Ф. С. Фицджеральда, который был опубликован в 1934 г.
- Первым персонажем, с которым знакомится читатель, является Розмари. Позже появляются Дайверы и семья Мак-Киско.
- Изначально, читателю может показаться, что Розмари и её любовь к Дику Дайверу будет главной сюжетной линией романа. Однако потом писатель обращает внимание читателя на семью Дайверов, и не остается никаких сомнений в том, что их взаимоотношения являются основой романа.
- Возникает вопрос: кто из героев, в частности Николь или Дик Дайвер, является центральной фигурой романа?

Цель работы

- Определить главного героя романа «Ночь нежна» Ф. С. Фицджеральда
- Охарактеризовать главных героев с помощью Cytoscape



Ход работы



The screenshot shows a Jupyter Notebook interface with the following code cells:

```
In [1]: import re, os
from collections import Counter
import spacy

In [4]: !python3 -m spacy download en_core_web_sm
Collecting en-core-web-sm==3.1.0
  Using cached https://github.com/explosion/spacy-models/releases/download/en_core_web_sm-0-py3-none-any.whl (13.6 MB)
  ✓ Download and installation successful
  You can now load the package via spacy.load('en_core_web_sm')

In [3]: nlp = spacy.load("en_core_web_sm")

In [4]: with open('Fitzgerald_Scott_Tender is the night_1934.txt') as f:
    text = f.read().replace('\n', ' ')

In [5]: doc = nlp(text)

In [6]: with open('en_stopwords.txt') as file:
    stopwords = file.read().split()

In [7]: lemmas = []
for token in doc:
    if token.lemma_ not in stopwords and token.lemma_.isalpha() and len(token.lemma_)>1:
        lemmas.append(token.lemma_)

lemmas[:10]
```

- Сперва я использовала код, над которым мы работали во время урока.
- Так как для данного исследования я использовала английский текст, мне нужно было загрузить модель *en_core_web_sm*. Однако для этого в отличие от русской модели мне пришлось её скачать (*!python3 -m spacy download en_core_web_sm*)

Ход работы (1)

- В процессе лемматизации я решила в список лемм добавить токены с длиной больше единицы, так как большинство самых частотных пар включали слово “I”.

```
In [5]: doc = nlp(text)

In [6]: with open('en_stopwords.txt') as file:
stopwords = file.read().split()

In [7]: lemmas = []

for token in doc:
    if token.lemma_ not in stopwords and token.lemma_.isalpha() and len(token.lemma_)>1:
        lemmas.append(token.lemma_)

lemmas[:10]

Out[7]: ['BOOK',
 'one',
 'pleasant',
 'shore',
 'French',
 'Riviera',
 'half',
 'way',
 'Marseilles',
 'italian']

In [8]: pairs = []

for i in range(len(lemmas) - 1):
    pair = min(lemmas[i], lemmas[i+1]) + ',' + max(lemmas[i], lemmas[i+1])
    pairs.append(pair)

pairs[:10]
```

Ход работы (1)

```
In [9]: counts = Counter(pairs).most_common()
```

```
In [10]: counts[:10]
```

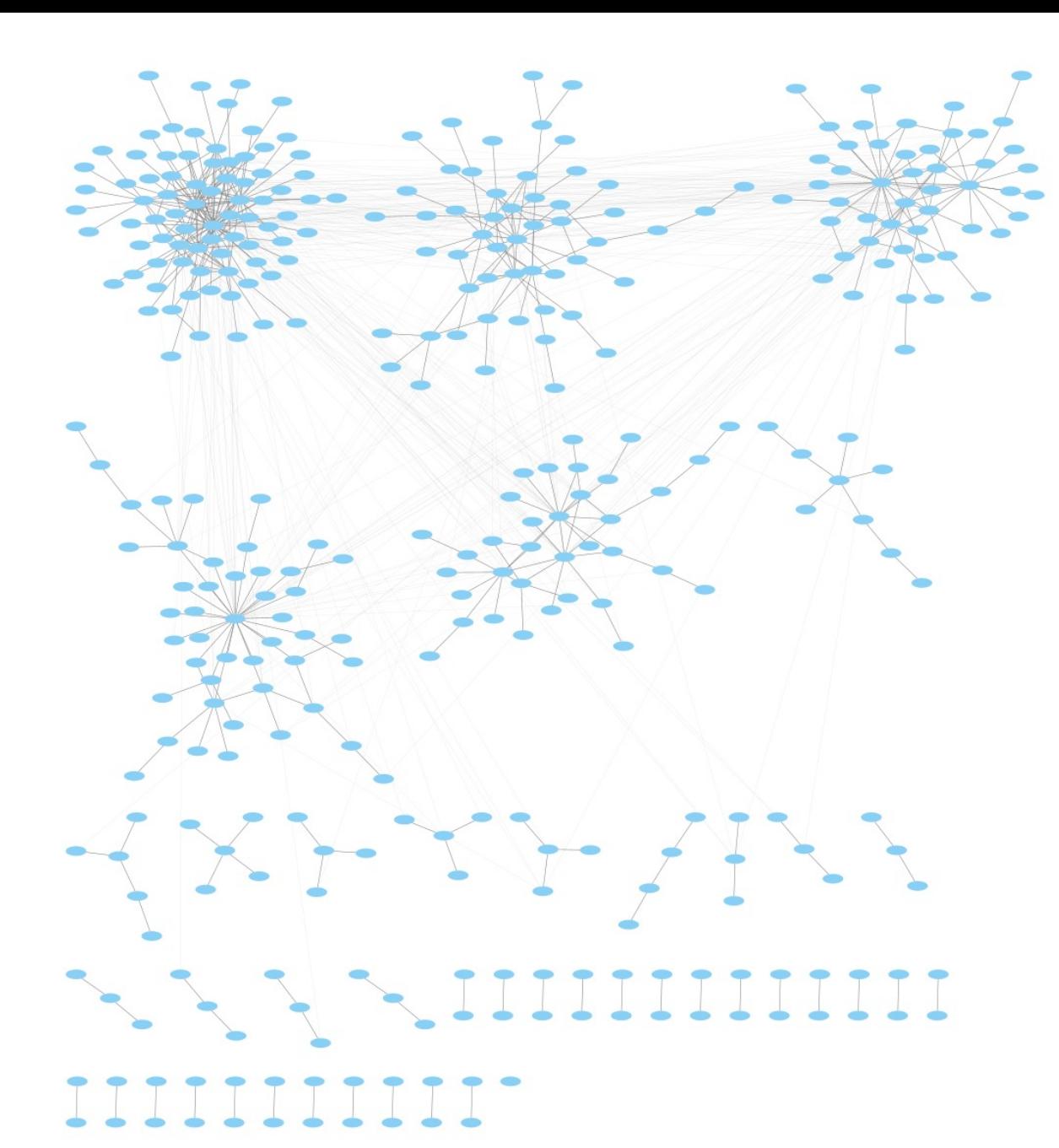
```
Out[10]: [('Dick,say', 79),  
          ('Dick,see', 30),  
          ('Diver,Doctor', 29),  
          ('back,go', 27),  
          ('go,want', 25),  
          ('back,come', 25),  
          ('Nicole,say', 25),  
          ('Doctor,Dohmler', 24),  
          ('Rosemary,say', 23),  
          ('Abe,say', 23)]
```

```
In [11]: csv = 'source,target,weight\n'  
  
for count in counts:  
    if count[1] > 3:  
        line = count[0] + ',' + str(count[1]) + '\n'  
        csv += line  
  
with open ('tender_keywords.csv', 'w', encoding = 'utf-8') as f:  
    f.write(csv)
```

- После я создала csv файл с самыми частотнымиарами слов и назвала его *tender_keywords.csv*
- https://drive.google.com/file/d/18bErVjXDQM8e5WNwQ_6PfW9beWpOnRQ2/view?usp=sharing
- И наконец, загрузила данный файл в Cytoscape для выявления и анализа результатов

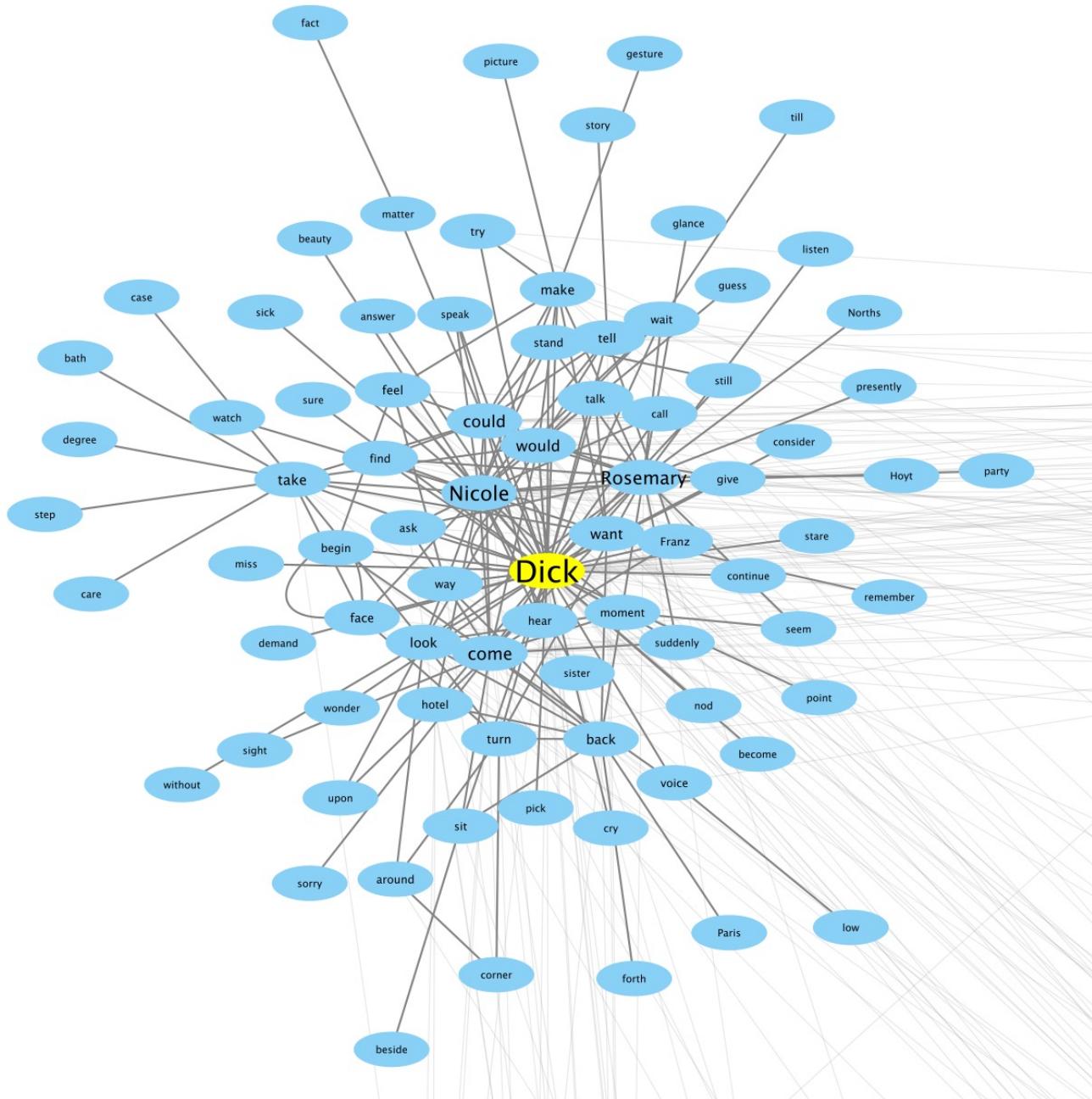
Ход работы (1)

- С помощью clusterMaker Cluster Network --> Community Cluster (GLAY) я разбила граф на кластеры/communities
- Однако для исследования мне понадобиться проанализировать только первый кластер, так как он включает в себя большинство героев романа



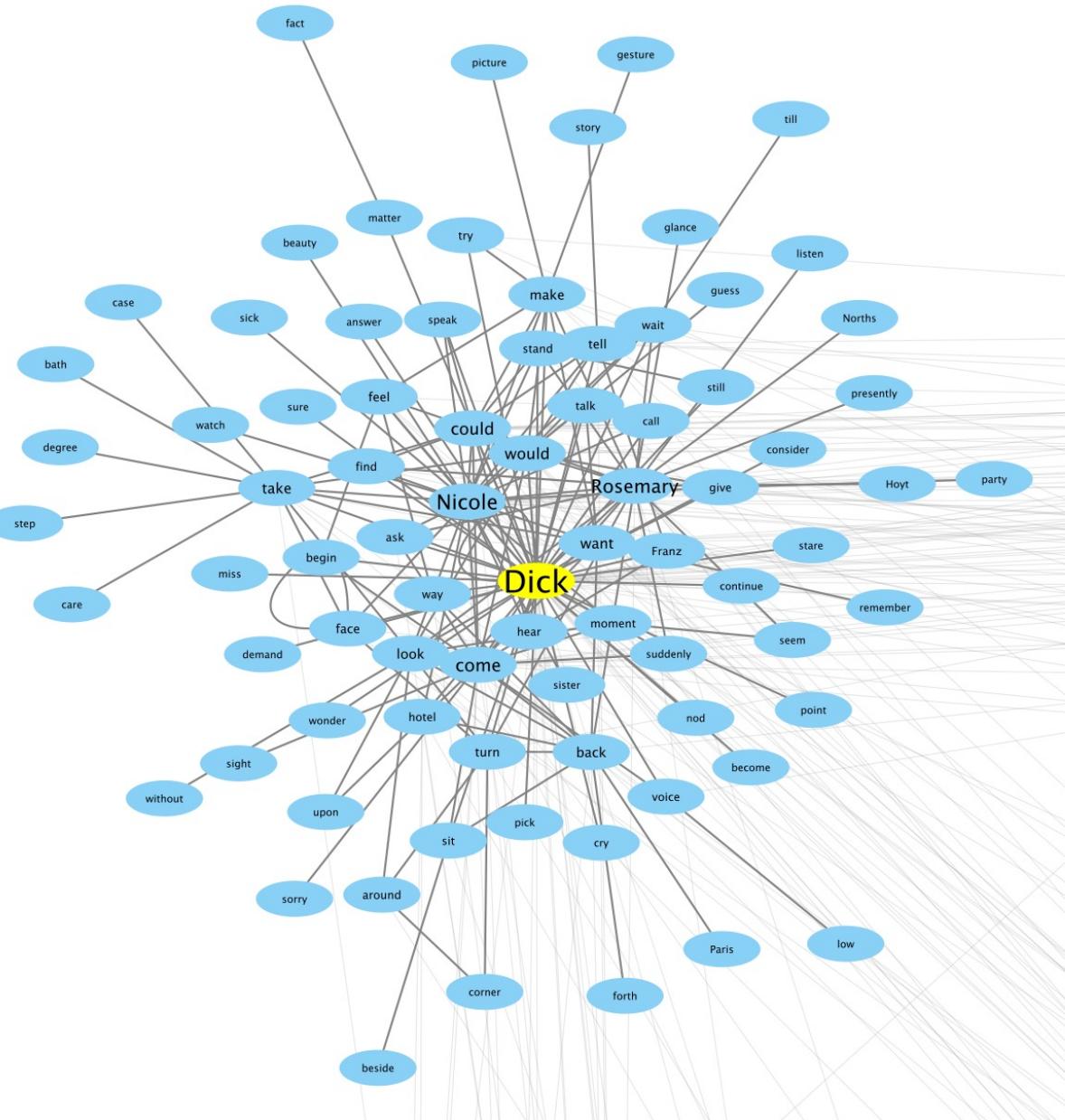
Ход работы (1)

- Во-первых я провела анализ сети с помощью прибора Analyze Network в разделе tools
- Потом в разделе style я выбрала характеристики узлов (nodes) и поменяла форму на эллипс, а размер лейблов в зависимости от их degree (continuous mapping)



Результаты (1)

- Граф выявил три узла с самым высоким значением degree: Dick (самое высокое), Nicole, Rosemary
 - Однако так как все три героя оказались в одном кластере, очень сложно определить связи каждого героя с остальными словами
 - Единственное, что можно, вынести из данной визуализации, это то, что вокруг этих трёх героев происходило большинство событий, так как они окружены большим количеством глаголов.



Ход работы (2)

```
In [12]: couples = []

for sent in doc.sents:
    entities = sent.ents
    if len(entities)>1:
        for i in range(len(entities) - 1):
            for j in range(i, len(entities)):
                if i != j:
                    couple = min(entities[i].lemma_, entities[j].lemma_) + ',' + max(entities[i].lemma_, entities[j].lemma_)
                    couples.append(couple)

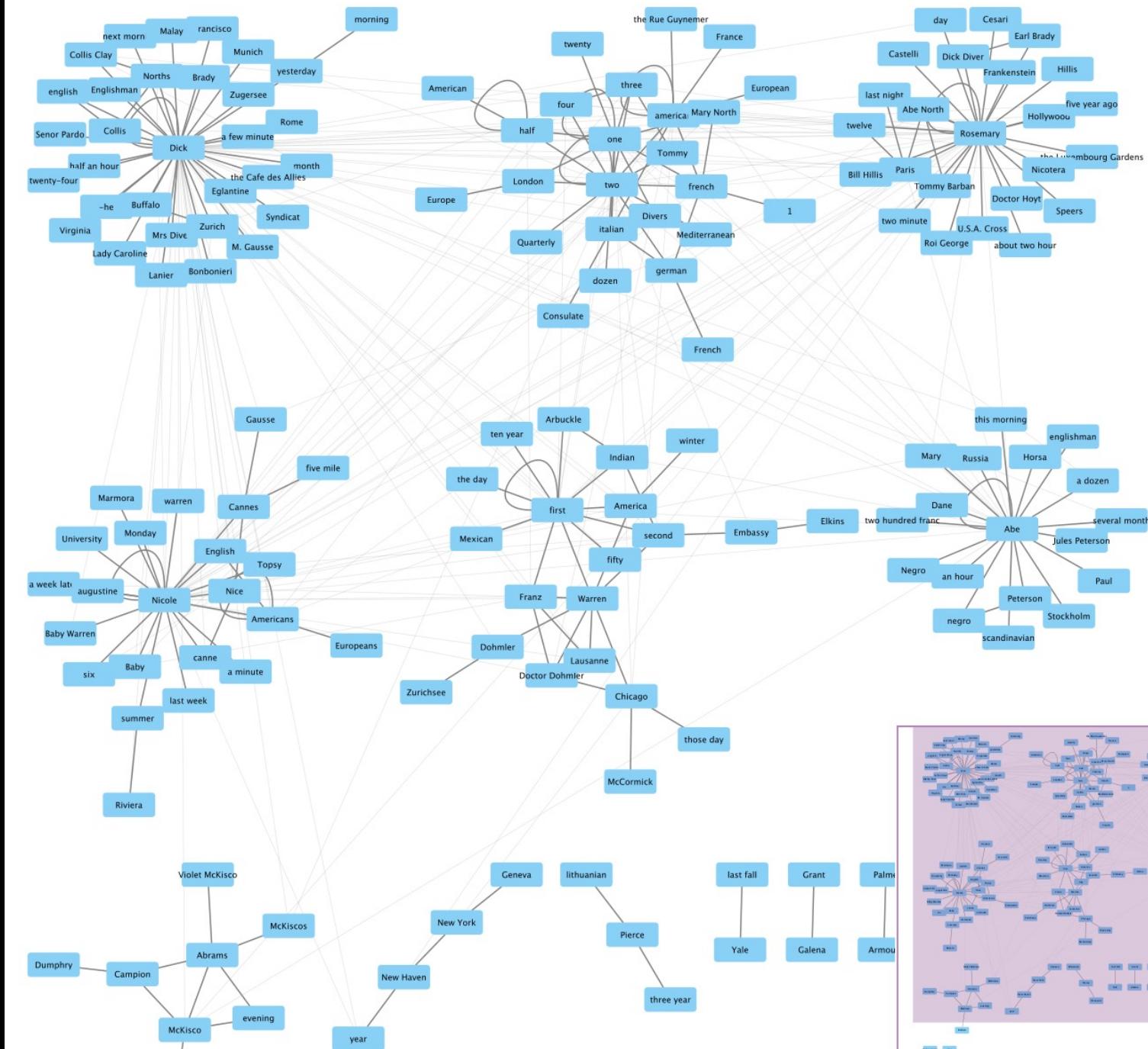
In [13]: couples[:15]
Out[13]: ['one,the French Riviera',
'about half,one',
'Marseilles,one',
'italian,one',
'about half,the French Riviera',
'Marseilles,the French Riviera',
'italian,the French Riviera',
'Marseilles,about half',
'about half,italian',
'Marseilles,italian',
'a decade ago,summer',
'english,summer',
'April,summer',
'a decade ago,english',
'April,a decade ago']

In [14]: counts_couples = Counter(couples).most_common()
```

- Потом я вернулась к коду и и создала список частотных пар именованных сущностей.
- Однако здесь возникли проблемы, которые видны даже в первых 15 пар, изображенных на картинке. Спейси работает нечисто, поэтому такие слова как one, second, summer, about half, a decade ago также были извлечены.

Ход работы (2)

- После того, как я создала файл с самыми частотными парами именованных сущностей и загрузила его в Cytoscape, я разбила граф на кластеры/communities с помощью clusterMaker Cluster Network --> Community Cluster (GLAY)
 - <https://drive.google.com/file/d/1p95uDJ4uXaxqIdE4xv9m3SvOU27OuYsJ/view?usp=sharing>



Результаты (2)

- Граф поделился на кластеры с именами Dick, Rosemary, Nicole и Abe.
 - Мы можем обратить внимание на то, что Nicole, Dick и Rosemary образуют треугольник.
 - Также, мы видим, что Dick взаимодействовал с остальными героями больше всех

Выводы

1. С помощью списков с самыми частотными парами и их визуализации в Cytoscape, мне удалось выявить самого центрального/главного героя романа – Дика Дайвера.
2. Вокруг него происходят все события, поэтому и первый граф показал, что у него самая высокая степень. Также его узел связан с другими узлами – глаголами.
3. Второй граф выявил и то, что Дик имеет больше всех взаимоотношений. Он активно общается со многими героями, в то время как Розмари, не говоря уж о Николь, общаются с довольно ограниченным числом людей.
4. Итак, в то время как читателю сложно определиться с главным героем романа (многие верят, что это Николь, тогда как остальные думают, что это Дик), исследование показало, что Дик однозначно является центральной фигурой романа.