Máster en Business Analytics y Big Data
Edición 2014 / 2015

Asignatura: *Introducción* al análisis de redes sociales

Título del trabajo: *Análisis de la red social de Juego de Tronos*

Autor: *Pablo González Fuente*

1. Índice

1. Introducción

Juego de Tronos es el nombre que se le da a una famosa serie de novelas de fantasía épica de fantasía épica escritas por el novelista y guionista estadounidense George R. R. Martin. El nombre real de la serie es "Canción de hielo y fuego" aunque todo el mundo la llama por el nombre de su primera novela ("Juego de Tronos").

La historia de Canción de hielo y fuego se sitúa en un mundo ficticio medieval. Se caracteriza por su gran número de personajes. Hay tres líneas argumentales en la serie (Wikipedia) que interactúan fuertemente entre sí y que se cuentan de manera simultánea:

- La crónica de la guerra civil dinástica por el control de Poniente entre varias familias nobles.
- La creciente amenaza de los Otros y los salvajes, apenas contenida por un inmenso muro de hielo que protege el norte de Poniente.
- El viaje de Daenerys Targaryen, la hija exiliada del rey que fue asesinado en otra guerra civil hace quince años, quien busca regresar a Poniente a reclamar sus derechos.

La serie está narrada en tercera persona a través de los ojos de varios personajes (no necesariamente los protagonistas) que interactúan con otros muchos cientos de personajes. Es muy característico de las novelas la aparición y desaparición de personajes principales de manera inesperada.

La serie ha tenido una gran acogida a nivel mundial, incluso se ha realizado una serie de televisión, pero debido a la forma de la narración de varias historias simultáneas y tanta variedad de personajes hay muchas personas que se pierden o les resulta difícil seguir el hilo argumental.

Para facilitar la tarea de explicar los personajes y sus historias creo que es muy apropiado construir un grafo con la red social de los personajes. Además, el lector "experto" en la serie también podrá beneficiarse de los análisis extraídos en este trabajo, ya que muchos de los resultados resultan muy curiosos y no tan evidentes.

2. Análisis de la fuente de datos

El primer reto consiste en elegir la fuente de datos correcta. Para elegir correctamente la fuente de datos es necesario que cumpla al menos dos condiciones: Que sea exhaustiva/completa y que la información esté estructurada para facilitar la extracción en las siguientes fases del proceso.

Existen muchas páginas web en internet con información muy detallada sobre la serie de libros y de televisión, alguna de ellas bastante bien estructurada. La primera de las fuentes donde acudí es la Wikipedia…

Wikipedia

Disponen de mucha cantidad de información de los libros pero en formato texto y en forma de resumen. Aparecen solo los personajes principales sin apenas entrar en detalle de las relaciones en la historia ni las relaciones familiares entre los personajes.

Con respecto a la serie de televisión disponen de información más detallada, con un resumen de cada capítulo dividido por escenas y se podría extraer para cada capítulo los personajes que aparecen en cada escena analizando por párrafos.

Un análisis más profundo revela que no hay resumen para todos los capítulos de la serie (faltan bastantes) y al estar redactado para que lo lea un humano los nombres de los personajes no están "normalizados", es decir, se usan pronombres, apodos y todo tipo de referencias que dificultarían el procesado posterior.

IMDB

Tiene información detallada de cada uno de los capítulos de la serie de televisión, además de forma tabulada (por actores) y que sería fácilmente extraíble. La pega es que no está dividido por escenas y dado que la serie cuenta varias historias simultáneas en el tiempo pero muy separadas geográficamente podría llevar a datos erróneos, ya que todos los personajes que aparecen en el mismo capítulo tendrían relación pese a no haberse visto nunca en la historia.

Hielo y Fuego Wiki

Finalmente encontré esta wiki con información muy exhaustiva de todo lo relacionado con el universo de Juego de tronos. En particular tienen un resumen de cada capítulo del libro junto con una lista de los personajes que aparecen en cada uno de ellos.

Ilustración 1: Ejemplo de página de Hielo y Fuego Wiki

Esta organización es perfecta para nuestro propósito ya que la forma en la que está narrada la novela (cada capítulo es desde el punto de vista de un personaje) podemos suponer que los personajes que aparecen en un mismo capítulo han tenido una relación.

3. Extracción de los datos

Para extraer los datos primero es necesario descargar el html[1] de las páginas de cada uno de los capítulos de la web, para ello utilizo el siguiente trozo de código, que reconstruye las urls de cada uno de los capítulos y los guarda en disco:

```
base_url = "http://hieloyfuego.wikia.com/wiki/"
books = [(1,"Juego_de_Tronos"),(2,"Choque_de_Reyes"),(3,"Tormenta_de_Espadas"),
(4,"Festín_de_Cuervos"),(5,"Danza_de_Dragones")]
CAPITULO = "-Cap%C3%ADtulo_"
for (index,book) in books:
    print "Starting book", book
    for i in range(1,100):
        if(glob.glob('./'+book+'/'+str(i)+'.html') == None):
            url = base_url+book+CAPITULO+str(i)
            page = requests.get(url)
            if(page.status_code == 200):
                with codecs.open(book+'/'+str(i)+'.html','w','utf8') as f:
                    f.write(page.text)
                    f.close()
print "Finished"
```

Una vez que tenemos los htmls es hora de empezar con el scrapping. He elegido una de las herramientas de scrapping más utilizadas: BeautifulSoup. El siguiente trozo de código busca el título “Aparecen” y va leyendo los elementos de tipo “<li>” que se va encontrando hasta que encuentra el texto “Mencionados”. Esta aproximación no es la mejor, pero la estructura de la web , aunque lo parece, no es uniforme para todos los capítulos y no es posible utilizar una expresión xpath:

```
for (book_index,book) in books:
    files = sorted(glob.glob('./'+book+'/*.html'))
    print "Starting book", book, '(', len(files), ')'
```

[1] Se puede hacer “online” pero es mejor descargarlo para reducir el tiempo del resto del código y minimizar el tráfico a la web que nos proporciona los datos.

```
    for f in files:
        #print "Chapter", f
        with open(f, 'r') as chapfile:
            soup = BeautifulSoup(chapfile,'html5lib')
            #Buscamos solo los personajes que aparecen
            a = soup.find(id='Aparecen')
            siblings = a.parent.next_siblings
            characters = []
            for sibling in siblings:
                if(type(sibling) is bs4.element.Tag):
                    if(sibling.find(id="Mencionados")):
                        break
                    listitems = sibling.find_all('li')
                    for item in listitems:
                        character_name = unidecode(item.text).strip()
```

Ahora que ya tenemos los nombres de los personajes que aparecen por cada capítulo es hora de construir el grafo.

4. Construcción de la red

Existen muchas librerías en Python para construir una red, para este ejercicio he elegido networkx que permite trabajar de una manera muy cómoda con los nodos y aristas y además incluye muchas funciones para calcular métricas y hacer operaciones sobre los grafos.

La primera parte es decidir qué serán los nodos y qué significan las aristas, como ya hemos explicado en apartados anteriores, nuestro objetivo es ver las relaciones de los personajes en la historia de los libros, así que nuestros nodos serán cada uno de los personajes y cada arista representará que dos personajes han aparecido juntos en un capítulo.

El algoritmo para añadir las aristas a partir de la lista de personajes es sencillo:

1. Me construyo una lista vacía de personajes para cada capítulo, que iré rellenando según vayan apareciendo en la página web
2. Cada vez que aparezca un personaje en la página web, creo un nodo en el grafo (si

está repetido la librería lo ignora)

3. Añado una relación con cada uno de los personajes de la lista del punto 1. Para el primer personaje no crearé aristas pero las crearé según vayan apareciendo el resto de personajes (no importa el orden, es un grafo no dirigido)

4. Por último añado el personaje a la lista y continúo con el siguiente (punto 2) o paso al siguiente capítulo (punto 1) si ya no quedan más personajes.

Como este algoritmo es un poco pesado, al finalizar el proceso guardo el grafo en un fichero gml para poder recuperarlo desde ahí y no tener que repetir este proceso cada vez que quiera hacer un análisis.

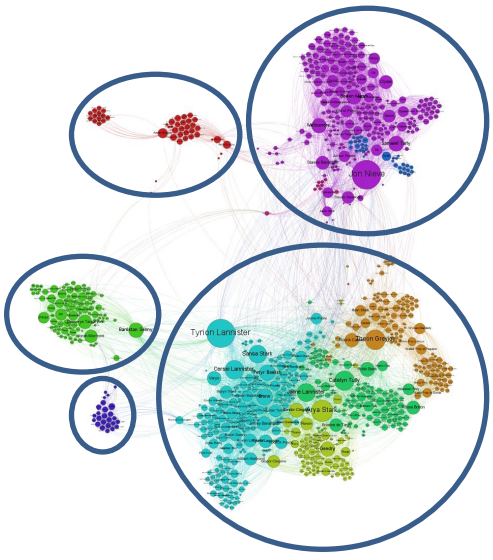
5. Análisis de la red

Con el grafo construido ya podemos empezar a analizar la red social y empezar a extraer información. Utilizaremos dos herramientas: Una más visual, que nos va a ayudar a extraer los dibujos para encaminar nuestros estudios y explicar mejor los resultados y la otra herramienta será la propia librería de grafos, con la que obtendremos las métricas del grafo y de cada uno de sus subconjuntos.

Gephi

Utilizaremos Gephi para obtener una representación del conjunto del grafo y estudiar las comunidades de manera visual.

Utilizando un layout de tipo ForceAtlas 2, usando el Pagerank para el tamaño de los nodos y las comunidades (10) como colores, obtenemos esta representación (imagen ampliada en Anexos):



Aquí nos encontramos con 5 clusters principales, atendiendo a los personajes que los conforman vemos que que relacionan perfectamente con las 5 localizaciones e historias independientes de la serie y sus tramas secundarias:

1. La guerra por el trono de hierro, el grupo más numeroso y formado por los nodos azul claro, amarillo, verde turquesa y marrón
   1.1. En azul turquesa la historia de las intrigas en la capital, sus nodos principales son los personajes que habitan en el palacio real.
   1.2. Los nodos amarillos son los de la historia de la fuga de Arya Stark del palacio real
   1.3. En verde, vemos la parte de la historia de la guerra desde el punto de vista de Catelyn Tully (las batallas que libra Robb Stark y cuando capturan a Jaime Lannister)
   1.4. En marrón, aparecen los nodos de los personajes que se quedan como señores del norte una vez muerto Ned Stark
2. La historia en el Muro y los salvajes del norte son los nodos morados, azules oscuro y rosas
   2.1. Los personajes del Muro aparecen en color morado
   2.2. Los salvajes que atacan el Muro salen en azul
   2.3. Los personajes que interaccionan con Sam y Eli durante su huida del Muro aparecen en rosa
3. La historia de Daenerys, en otro continente, de color verde
4. La historia de los hombres de las islas del hierro (en el norte) son los nodos rojos
5. La historia más reciente del nuevo heredero Targaryen en azul oscuro

Networkx

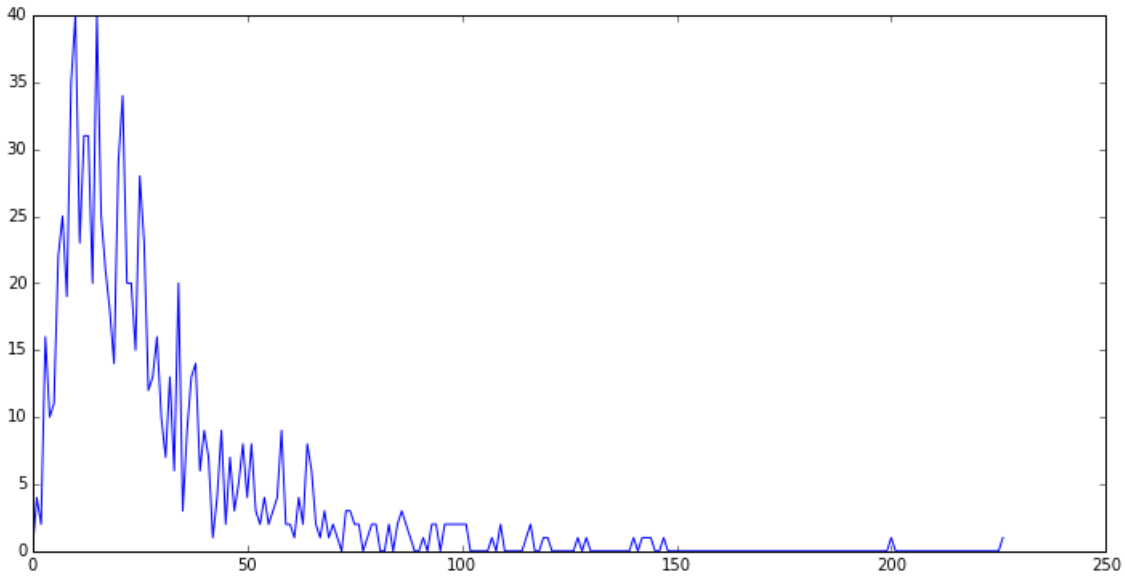
Utilizando las funciones que nos da la librería podemos sacar las estadísticas del grafo:

Nodos	906
Aristas	12918
Componentes conectados	1
Tamaño del component gigante	906
Average clustering	0.827

Observamos que todos los nodos del grafo están conectados y que el tamaño del componente gigante coincide con el número total de nodos.

Un average clustering tan alto nos indica que los nodos están fuertemente organizados en comunidades.

Si estudiamos la distribución de los nodos en función de su grado obtenemos esta gráfica:



Aquí podemos observar que se trata de una función log-normal y que la mayoría de los nodos tienen un grado entre 0 y 50, pero hay casos extraños que tienen más de 200 aristas, si observamos quiénes son coinciden con los protagonistas de la serie:

Name	Degree
Jon Nieve	226
Tyrion Lannister	200
Cersei Lannister	147
Catelyn Tully	144
Arya Stark	143
Sansa Stark	142
Theon Greyjoy	140
Jaime Lannister	129
Bowen Marsh	127
Clydas	120

Sin embargo, los últimos de esta clasificación no podrían considerarse protagonistas (o por lo menos no entre los 10 primeros). Esto puede ser porque hayan interactuado con varios grupos de personajes y no porque sean personajes principales.

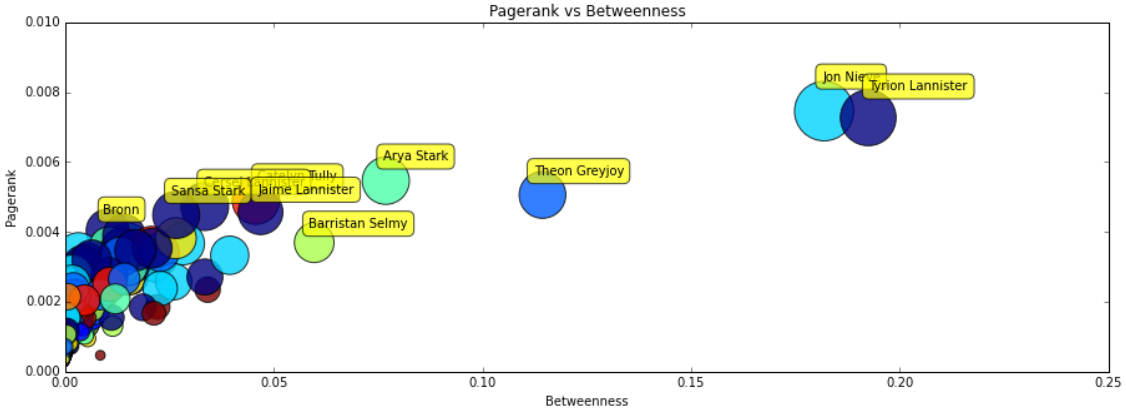
Vamos a probar con otras métricas de centralidad para localizar a los personajes principales.

Name	Betweenness
Tyrion Lannister	0.192455
Jon Nieve	0.181893
Theon Greyjoy	0.114413
Arya Stark	0.076878
Barristan Selmy	0.059718
Jaime Lannister	0.046865
Catelyn Tully	0.045744
Stannis Baratheon	0.039488
Aeron Greyjoy	0.034193
Jeyne Poole	0.033541

Name	Pagerank
Jon Nieve	0.007458
Tyrion Lannister	0.007268
Arya Stark	0.005465
Theon Greyjoy	0.005059
Catelyn Tully	0.004895
Cersei Lannister	0.004728
Jaime Lannister	0.004566
Sansa Stark	0.004487
Bronn	0.004036
Petyr Baelish	0.003895

A primera vista la métrica del Pagerank es la que mejor ha identificado a los personajes principales desde el punto de vista de la historia, pero con el betweenness también sabemos quiénes son los personajes que hacen de nexo entre los distintos grupos. Para salir de dudas pintaremos en una gráfica todas las métricas y los personajes principales serán aquellos que mejor salgan en conjunto.

En este gráfico hemos representado cada nodo en sus coordenadas de Betweenness y Pagerank, el tamaño del nodo es su Grado y el color corresponde a su comunidad.



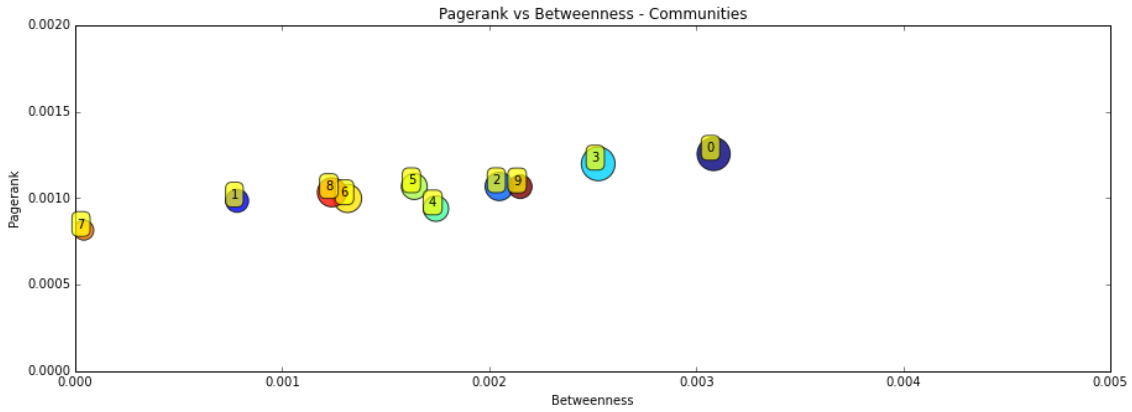
Con esta representación vemos claramente que “Jon Nieve” y “Tyrion Lannister” destacan en todas las métricas, seguidos por “Theon Greyjoy” y en menor medida por “Arya Stark”. Después hay otra serie de personajes importantes (“Catelyn Tully”, “Jamie y Cersei Lannister”, “Sansa Stark” y “Barristan Selmy”)

En la serie aparece repetidamente el concepto de casas nobiliarias y muchas de las historias también se organizan en función de estas familias, sería interesante realizar el

mismo estudio pero con las comunidades obtenidas de la red.

Si calculamos las comunidades y la media de cada una de sus métricas obtenemos esta tabla:

Community	Degree	Betweenness	Pagerank
0	35.776119	0.003084	0.001255
1	17.240000	0.000782	0.000984
2	26.000000	0.002049	0.001067
3	37.156250	0.002527	0.001198
4	21.032967	0.001743	0.000939
5	21.967391	0.001638	0.001067
6	25.818182	0.001316	0.000998
7	12.687500	0.000042	0.000814
8	27.611111	0.001240	0.001033
9	18.266667	0.002150	0.001066



Aunque todas las comunidades parecen tener un Pagerank similar, si atendemos al Betweenness, la comunidad "0" es la que a primera vista parece la más importante en todas las métricas. Si observamos sus miembros es posible que podamos descubrir a qué familia noble de la serie hace referencia:

Name	Degree	Betweenness	Pagerank	Community
Tyrion Lannister	200	0.192455	0.007268	0
Cersei Lannister	147	0.033492	0.004728	0
Jaime Lannister	129	0.046865	0.004566	0
Sansa Stark	142	0.026663	0.004487	0
Bronn	119	0.010229	0.004036	0

Esta comunidad es la que hace referencia a los Lannister y todos los personajes relacionados con ellos.

Por otro lado, hay una comunidad (la 7) que destaca por su bajo Betweenness, podría deberse a que es una comunidad aislada del resto de las otras, vamos a ver una lista de sus personajes principales:

Name	Degree	Betweenness	Pagerank	Community
Ygritte	44	0.000653	0.002147	7
Errok	15	0.000024	0.000935	7
Forunculo	12	0.000000	0.000800	7
Henk el Timon	12	0.000000	0.000800	7
Quort	12	0.000000	0.000800	7

Se trata de los "habitantes más allá del muro" una serie de personajes que no pertenecen a ninguna casa nobiliaria y que solo interactúan con los guardianes del muro (los llamados "Guardia de la Noche") de ahí que tengan muy poco Betweenness, sin embargo su Pagerank es similar al del resto de grupos porque su principal componente "Ygritte" fue novia de uno de los personajes principales ("Jon Nieve").

Por último, otra de las características de esta serie es que el autor tiende a eliminar de la serie a lo que parecen los personajes principales al final de cada libro, también puede ser interesante repetir este análisis pero en el ámbito de cada libro (ver anexo).

Observamos que a medida que avanzan los libros aparecen menos personajes principales y además con menor intensidad ya que los nodos poco a poco se desplazan hacia el origen de coordenadas. Además, hay un libro en el que se ve una distribución muy distinta a la de los otros 4, donde se ve que hay muchos personajes con mucha importancia (Pagerank alto) y con un Betweenness bajo. Si miramos la descripción del libro nos encontramos con este párrafo:

"Debido a la complejidad y extensión que estaba alcanzando los editores decidieron separarla en dos tomos, publicados simultáneamente, pero George R. R. Martin prefirió publicar sólo una parte de lo que inicialmente tenía previsto, siguiendo únicamente los acontecimientos de parte de los personajes principales, y ampliando los detalles de la trama con nuevos capítulos. La segunda parte, simultánea en el tiempo de narración y que incluye el resto de personajes, es Danza de Dragones."

Es decir, este libro es una separación "artificial" de la historia, donde solo se encuentran parte de los personajes principales de la trama y por lo que se ve, no hay personajes que hagan de nexo entre las historias.

6. Conclusiones

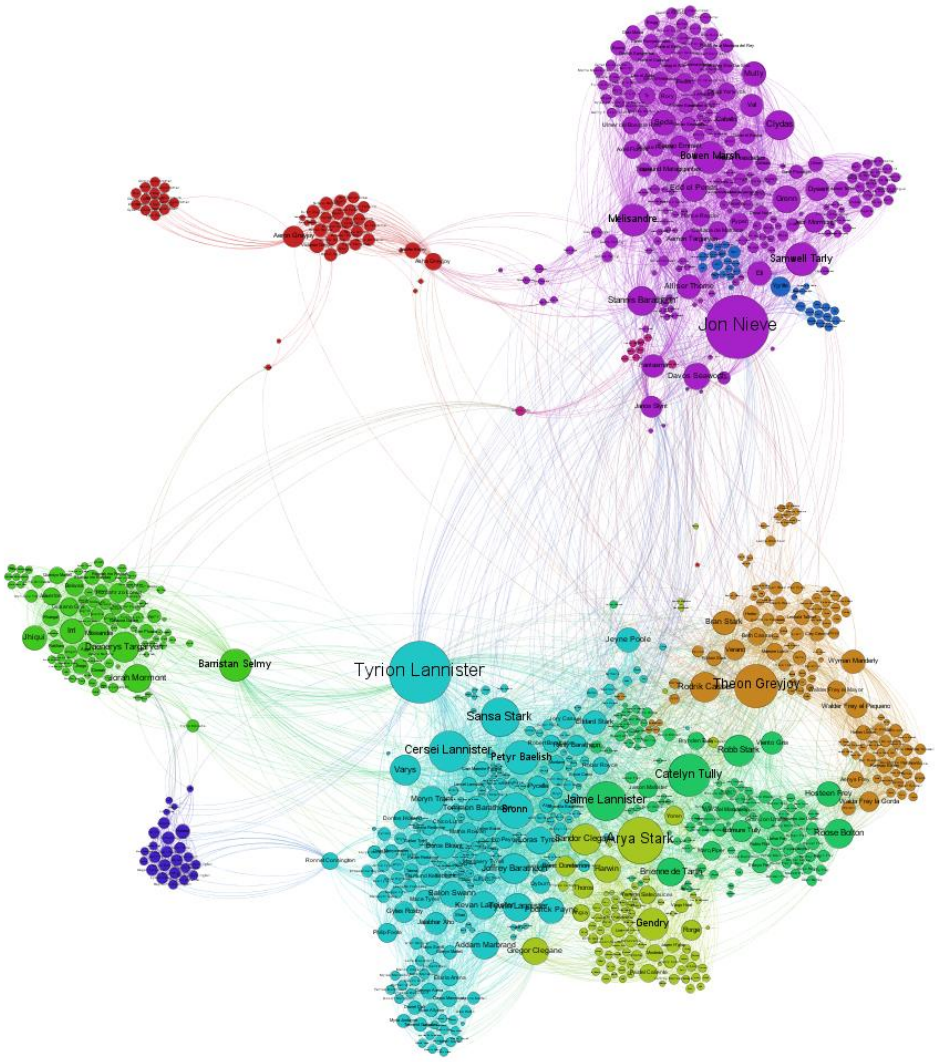
Gracias al estudio de la centralidad y la detección de comunidades hemos sido capaces de explicar muchas de las características de una historia muy compleja y con muchos personajes.

Para la detección de personajes principales ha sido muy útil el estudio de la centralidad de la red, no obstante, nos hemos dado cuenta que una única medida no siempre nos da una visión completa de la realidad y hemos observado que es mejor utilizar varias métricas y su representación en gráficos.

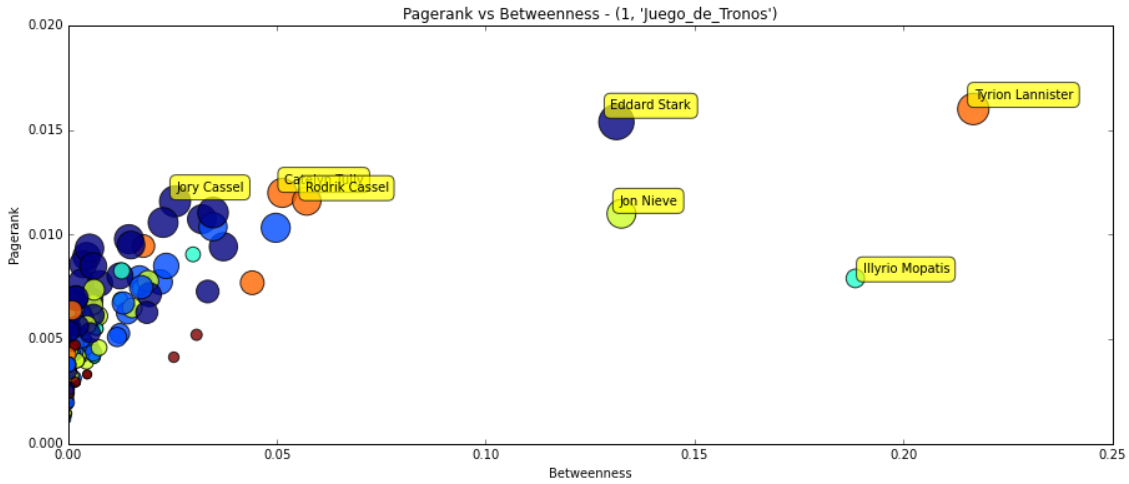
Para la detección de tramas argumentales nos hemos ayudado de la agrupación por comunidades, donde de una manera muy acertada el software ha sido capaz de discriminar las historias principales y alguna de las secundarias. Es posible que si aumentamos la precisión del algoritmo de detección de comunidades obtuviéramos más tramas secundarias en los grupos con menos nodos.

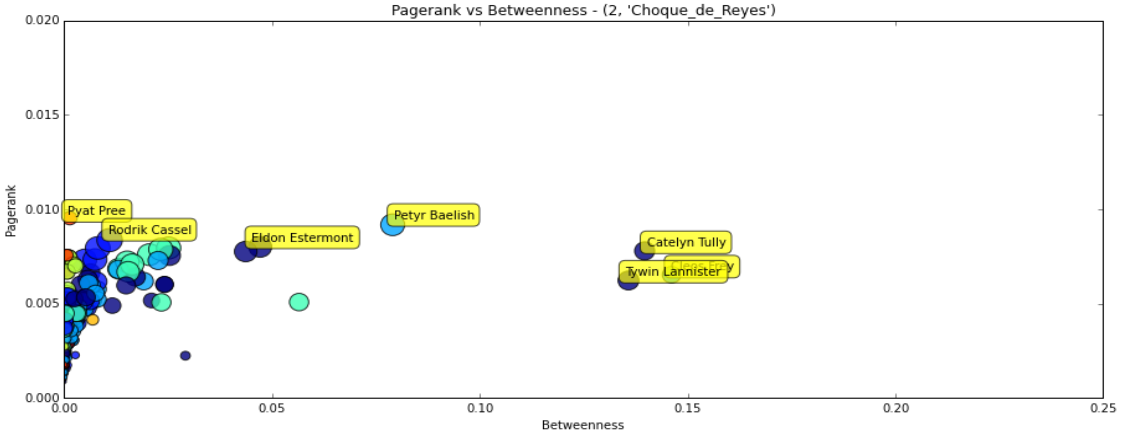
7. Anexos

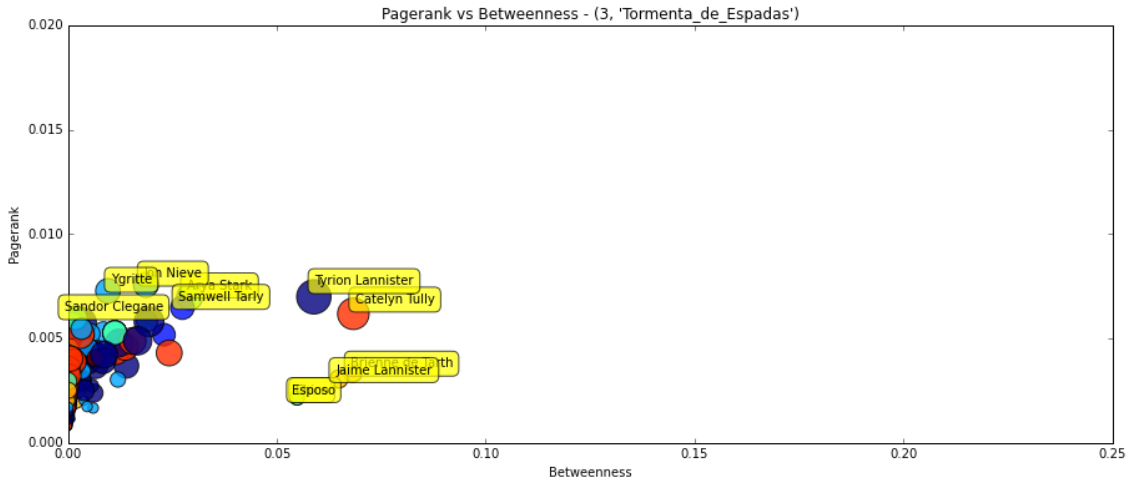
Representación en Gephi

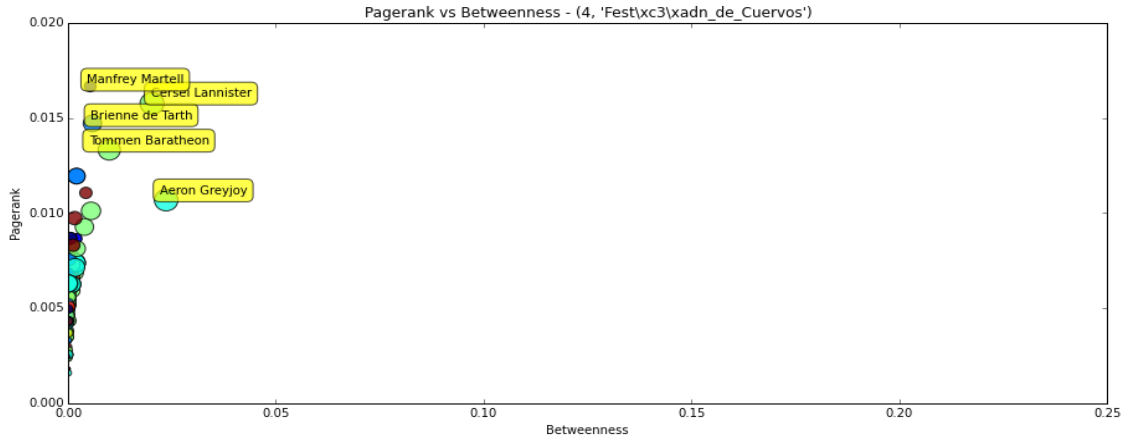


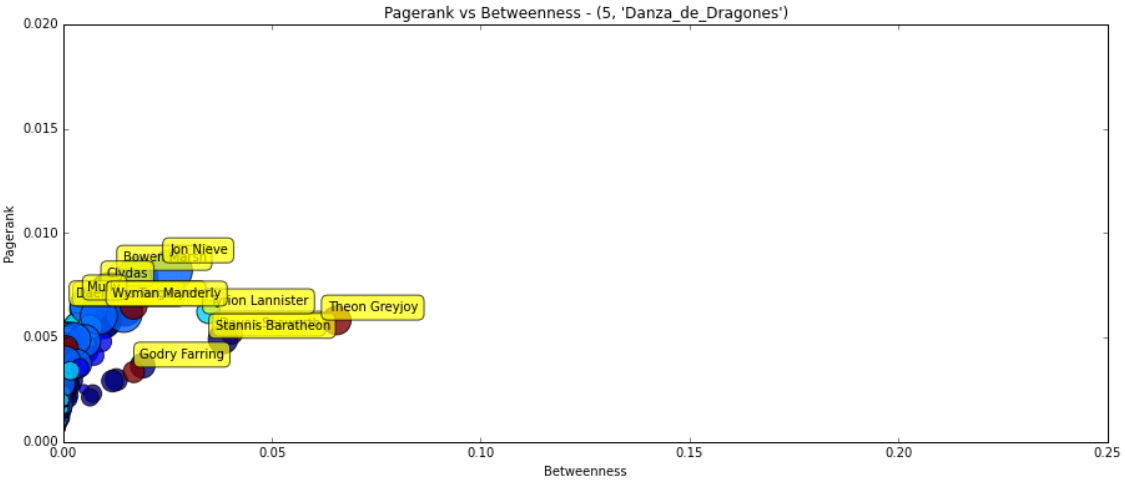
Análisis de personajes por libro











Notebook Ipython

Game of Thrones networkx.ipynb