El género, la película

Gastón Bujía, Mariano Nicolini, Sasha Smolarchik

Diciembre 2018

gastonbujia@gmail.com, mariano.nicolini.91@gmail.com, sasha95@gmail.com

Resumen

Utilizando una combinación de técnicas de procesamiento natural del lenguaje y de redes complejas, se analizó si es posible determinar el género de un conjunto de películas a través del texto de las mismas, i.e., sus diálogos y narraciones, creando una noción de similitud entre ellas y a partir de esta, confeccionando redes para posteriormente detectar una estructura en comunidades que refleje dichos géneros. Las comunidades evidenciaron un aglomeramiento de películas bajo tópicos similares, y dichos tópicos resultan tener una correlación con algunos géneros particulares dependiendo de cada caso. Además, se caracterizó la homofilia en género de la red y resultó positiva con un alto grado de confianza. Esto significa que LSA tiende a conectar más películas del mismo género que lo esperado por azar (si se preserva la estructura de la red).

1. Introducción

Una pregunta que resulta interesante de formular en cuanto a nuestra capacidad como humanos de clasificar el mundo que nos rodea es qué tanto de lo que entendemos por el *género* de una película emerge del lenguaje, es decir, de los diálogos de las mismas. Se combinarán técnicas de procesamiento natural de lenguaje con técnicas de análisis de redes complejas para intentar responder estos interrogantes.

Con este propósito, utilizaremos la técnica *Latent Semantic Analysis* (LSA), la cual crea representaciones vectoriales o *word embeddings* para un cuerpo de documentos, y a partir de éstos es posible definir una noción de similitud entre ellos. Luego, crearemos una red en donde los nodos representan a las películas y existe un enlace entre dos de ellas cuando los subtítulos que en mismas aparecen sean similares.

Una vez creada la red, se aplicarán algoritmos de detección de comunidades para ver cómo éstas se corresponden con los géneros mediante técnicas cuantitativas y cualitativas.

Por último, se hara una descripción de la homofilia presente en la red construída en la variable *'genres'* de las películas.

2. Armado de la red

En esta sección se describen cuáles fueron los datos utilizados y cuál fue la metodología de armado de las redes que estudiaremos.

2.1. Datos

Los datos utilizados consisten en aproximadamente 8000 subtítulos en inglés de películas de los últimos 5 años, obtenidos de opensubtitle.org. Además, contamos con los datos de los géneros de las películas como fueron etiquetadas en la página de IMDB, así como la información de los directores, guionistas, puntuación y duración.

2.2. Procesamiento de texto

El texto de los subtítulos fue procesado previamente de la siguiente forma:

1. Se eliminaron las denominadas *stopwords*, palabras que no aportan al significado del texto por ser demasiado frecuentes. Ejemplos de estas palabras son los conectores y los pronombres.
2. Cada una de las palabras se llevó a su base eliminando el plural y las conjugaciones entre otras cosas.
3. Se impusieron filtros de frecuencia a los términos, es decir, se estableció una frecuencia máxima y mínima para los mismos con el objetivo de deshacerse de términos que no son significativos para modelar el lenguaje de la película.
4. Se descartaron los nombres de personas, pero no de lugares.

Una vez normalizado el texto, para poder construir la red necesitaremos crear una representación de cada película en torno al lenguaje que utiliza. La técnica elegida con este propósito fue *Latent Semantic Analysis*[1]. Esta técnica es un método para extraer y representar el significado de uso contextual de las palabras o los textos, buscando así tópicos o temas sobre los cuales estas hablan.

La técnica LSA consiste, en primero construir una matriz de ocurrencias W de términos y documentos. En nuestro caso cada fila representa una película y cada columna un término del corpus. El corpus se compuso de todas las palabras que aparecen en todos los subtítulos más todos los bigramas, es decir, términos compuestos por dos palabras consecutivas. Cada entrada de la matriz luego se pesará por la frecuencia del término y la frecuencia inversa de la película, también conocido como el puntaje tf-idf:

$$w_{pt} = tfidf(p,t) = tf(p,t) \times idf(t) = tf_{pt} \times \log \frac{N}{df_t}$$

Cada uno de estos valores representa:

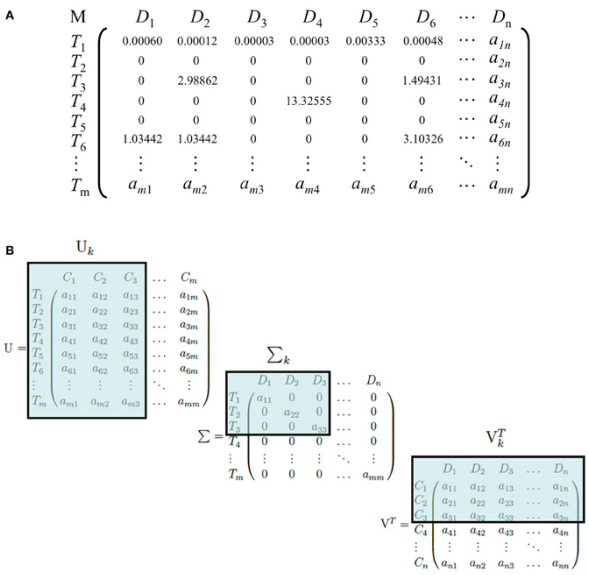
- tf_{pt} es la frecuencia relativa del término t en la película p.
- df_t es la cantidad de películas que contienen al término t.

En nuestro conjunto de datos la matriz W tuvo dimensiones del orden de 8 millones de términos por 8000 películas antes de imponer los filtros y luego fue reducida a 17000 términos gracias al preprocesamiento. Llamaremos T a la cantidad total de términos luego de aplicar los filtros y P a la cantidad de películas.

Una vez construida la matriz, LSA descompone esta matriz con la descomposición SVD truncada, para esto primero calcula la descomposición en las matrices $U \in \mathbb{R}^{T \times T}$, $\Sigma \in \mathbb{R}^{T \times P}$ y $V \in \mathbb{R}^{P \times P}$:

$$W = U \Sigma V^t$$

Y luego trunca la matriz Σ quedándose con los k valores principales más grandes donde $k << T$, donde k es un parámetro a determinar. Finalmente, de esta forma obtenemos que cada película p puede ser representada como un vector fila $\mathbf{w}_p \in \mathbb{R}^k$.



(a)

Figura 1: A: Ejemplo de una matriz luego de tf-idf. En este caso las filas representan términos, las columnas documentos, en nuestro caso sería al revés. B: Las tres matrices de la descomposición SVD. La matriz Σ sólo tiene elementos en la diagonal, ordenados de mayor a menor. El truncamiento consiste en quedarse con los valores principales más grandes (hacer 0 todo lo que no esté en el rectángulo azul que se ve en la figura). Volver a hacer el producto con U y V^t devuelve la reducción dimensional deseada.

2.3. Construcción de la red

Dado que tenemos una representación de las películas en forma vectorial en base a su lenguaje, el criterio que utilizaremos para construir la red es el siguiente:

1. Primero determinaremos k, para esto elegiremos varios valores. En LSA, esperamos que cada una de esas k dimensiones represente los tópicos o temas que uno busca encontrar para los documentos que uno trabaja. Como contamos con 23 géneros, elegiremos a $k = 20, 50, 100$, donde los valores más grandes se explican por la cantidad de combinaciones que aparecen de los géneros.

2. Determinaremos una distancia a utilizar que en nuestro caso será la distancia coseno:

$$d(v, w) = 1 - \frac{v\ddot{w}}{\|v\| \, \|w\|}$$

3. Se construyó la matriz $D \in \mathbb{R}^{P \times P}$ de distancias entre cada par de películas, es decir $d_{ij} = d(\mathbf{w}_{p_i}, \mathbf{w}_{p_j})$.

4. Se decidió un valor de corte c que permite definir un enlace de nuestra red. De esta manera, se define una matriz de adyacencia:

$$A_{ij} = \begin{cases} 1 & d_{ij} \leq c \\ 0 & d_{ij} > c \end{cases}$$

Es decir, las redes no dirigidas que utilizaremos tienen como nodos a las películas y dos películas p_1 y p_2 estarán conectadas sí y solo sí $d(\mathbf{w}_{p_1}, \mathbf{w}_{p_2}) < c$. Los valores que se consideraron de c involucraron el promedio de las distancias entre todas las películas y el promedio de los desvíos:

$$\mu = \frac{1}{P}\sum_{i=1}^{P}\mu_i = \frac{1}{P^2}\sum_{i=1}^{P}\sum_{j=1}^{P} d_{ij}$$

$$\sigma = \frac{1}{P}\sum_{i=1}^{P}\sigma_i = \frac{1}{P}\sum_{i=1}^{P}\sqrt{\frac{1}{P-1}\sum_{j=1}^{P}(d_{ij}-\mu_i)^2}$$

Utilizamos como criterio para elegir este valor qué grado medio y densidad tenía la red resultante, cuantificando la conectividad de la red. Excepto que se indique lo contrario, para los análisis se utilizó una red construida a partir de un LSA de 20 componentes y 20 iteraciones, con un valor de corte $c = \mu - 2{,}5\sigma$.

A continuación se presenta una tabla con algunas características de las redes construidas.

Componentes LSA (k)	Corte (c)	Nodos	Enlaces	Grado medio	Densidad	Clustering medio
20	$\mu - 2{,}5\sigma$	7423	543815	146.52	0.0197	0.513
30	$\mu - 2{,}5\sigma$	7424	604696	162.90	0.0219	0.484
50	$\mu - 2{,}5\sigma$	7424	608931	164.04	0.0221	0.444
100	$\mu - 3{,}5\sigma$	7426	221293	59.60	0.008	0.459
100	$\mu - 3{,}0\sigma$	7427	447028	120.38	0.0162	0.488
100	$\mu - 2{,}0\sigma$	7429	795905	214.27	0.0288	0.332

Tabla 1

3. Estudio de comunidades

El objetivo de este trabajo es estudiar las comunidades que aparecen en las redes recién descriptas, es decir buscar las subredes altamente conectadas que subyacen en la red. Con este fin consideraremos dos técnicas ampliamente utilizadas, una de ellas basada en modularidad, el *algoritmo de Louvain* y la otra basada en flujo, el *algoritmo Infomap*. Ambos algoritmos fueron elegidos por tener objetivos diferentes y ser los que más rápido corren con una complejidad media de $O(Nlog(N))$[2] donde N es la cantidad de nodos de la red.

El *algoritmo de Louvain*[3] tiene como objetivo ir iterativamente aumentando el valor de la modularidad. El mismo tiene dos etapas, en la primera comienza determinando una partición inicial de comunidades a cada nodo, es decir una comunidad por nodo. Luego, para cada nodo i de la red consideramos todos los nodos vecinos j y evaluamos como varía la modularidad resultante de eliminar a i de su comunidad y agregarlo a la comunidad de j. Solo si la ganancia es positiva, agregamos a i a la comunidad que le genere mayor ganancia de modularidad. Si no es posible obtener una ganancia positiva, me quedo en su comunidad original. Este proceso iterativamente para todos los nodos hasta que no se pueda realizar ninguna mejora adicional. Finalizado este paso, la segunda etapa del algoritmo consiste en armar una red donde los nodos son las comunidades encontradas en la primera etapa. En esta red las aristas tendrán pesos asociados a los pesos a la suma de los pesos intercomunidades y aristas que salen y vuelven a entrar a la misma comunidad con la suma de los pesos intracomunidad. Este proceso se repite hasta que no mejora la modularidad.

Por otro lado, el *algoritmo Infomap*[4] busca optimizar una función de calidad sobre las comunidades de la red. En particular, explota la dualidad entre la estructura de flujo de una red y la mínima cantidad de bits requeridos para describir una trayectoria aleatoria. Si se desea codificar de la manera más eficiente la trayectoria de un caminante aleatorio en esta red, el código ideal debería aprovechar el hecho de que el caminante aleatorio tiende a quedar atrapado en las comunidades, permaneciendo allí durante mucho tiempo. Estos *módulos* pueden etiquetarse con algún código de entrada y salida, lo cual va a permitir repetir etiquetas (a modo de ilustración, si hay una comunidad A y una B, se podría tener un nodo 1 en cada comunidad y llamarlo A1 y B1 sin caer en ambigüedades). Este proceso efectúa una compresión de la información, que está asociado a la estructura de la red. Las comunidades que devuelve el algoritmo de infomap son estos módulos. Es interesante que LSA también busca, de otra manera, comprimir

información, proyectando vectores de un espacio de dimensionalidad inmensa a un subespacio de dimensión reducida en donde su proyección es más grande (buscando el mejor trade-off entre la menor pérdida de información en relación a la reducción de dimensionalidad).

Vale destacar que aunque las comunidades que tendremos de referencia dadas por los géneros no son disjuntas, sino que son solapadas, ya que una película puede tener asociado mas de un género, solo utilizaremos estos algoritmos de comunidades disjuntas por cuestiones de complejidad computacional.

4. Resultados

En la figura 2 se ve una visualización de la red, con nodos y enlaces coloreados por pertenencia a comunidad detectada por infomap (red de 20c de LSA).

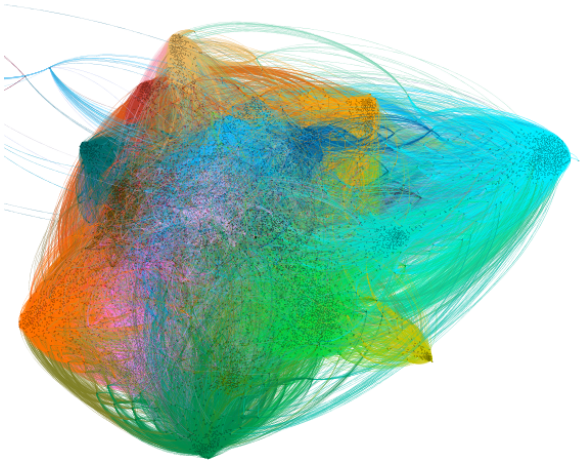


Figura 2: Visualización de la red coloreada por comunidades de infomap.

Con las comunidades ya detectadas en la red, se procedió a calcular la información mutua entre las particiones obtenidas con cada algoritmo y las comunidades de referencia, obtenidas de considerar la información externa que tenemos acerca de los géneros con los que fueron etiquetadas las películas analizadas. La ecuación tradicional de información mutua para comunidades no solapadas,

$$I_E(C1, C2) = \sum_{C1} \sum_{C2} p(C1, C2) log_2\Big(\frac{p(C1, C2)}{p(C1)p(C2)}\Big) \tag{1}$$

fue modificada con el fin de tener en cuenta el solapamiento que existe entre las comunidades de los géneros de referencia, ya que una dada película puede estar etiquetada con más de un género. Al calcular las probabilidades $p(C1)$ y $p(C2)$, y así como también la probabilidad conjunta, $p(C1, C2)$, los nodos pertenecientes a la intersección entre comunidades fueron contados con un peso inversamente proporcional a la cantidad de comunidades en las que participan. En particular, al calcular $p(C1, C2)$, cada nodo fue pesado como el inverso de la cantidad máxima de comunidades en las que participa entre las dos particiones.

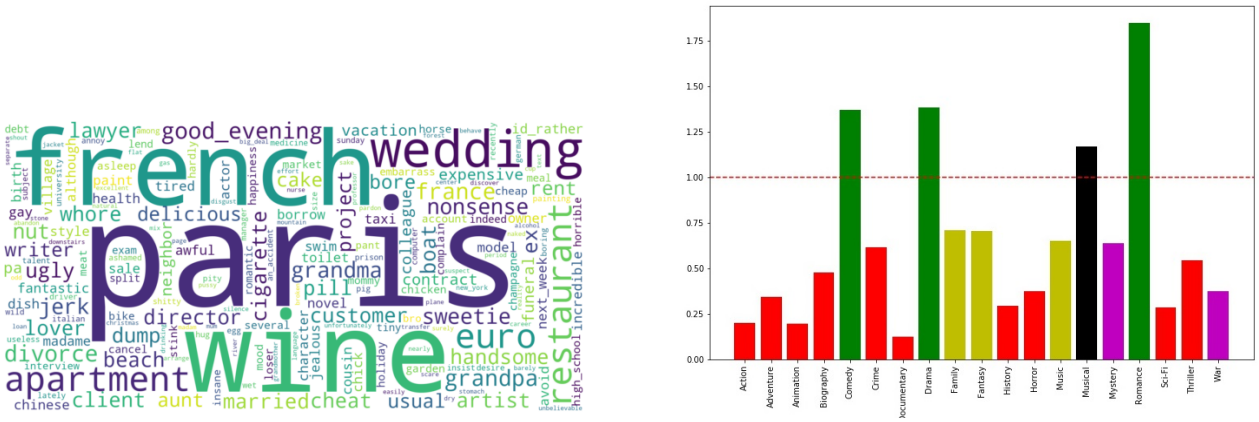
Los cálculos de información mutua para los algoritmos de detección de comunidades de Infomap y Louvain devolvieron valores que se pueden apreciar en la tabla 2, junto con la modularidad.

Infomap	Información Mutua	Modularidad
Red 1 (LSA 20c)	0.122	0.720
Red 2 (LSA 50c)	0.101	0.556
Red 3 (LSA 100c)	0.0934	0.550
Louvain	**Información Mutua**	**Modularidad**
Red 1 (LSA 20c)	0.100	0.711
Red 2 (LSA 50c)	0.072	0.580
Red 3 (LSA 100c)	0.055	0.577

Tabla 2

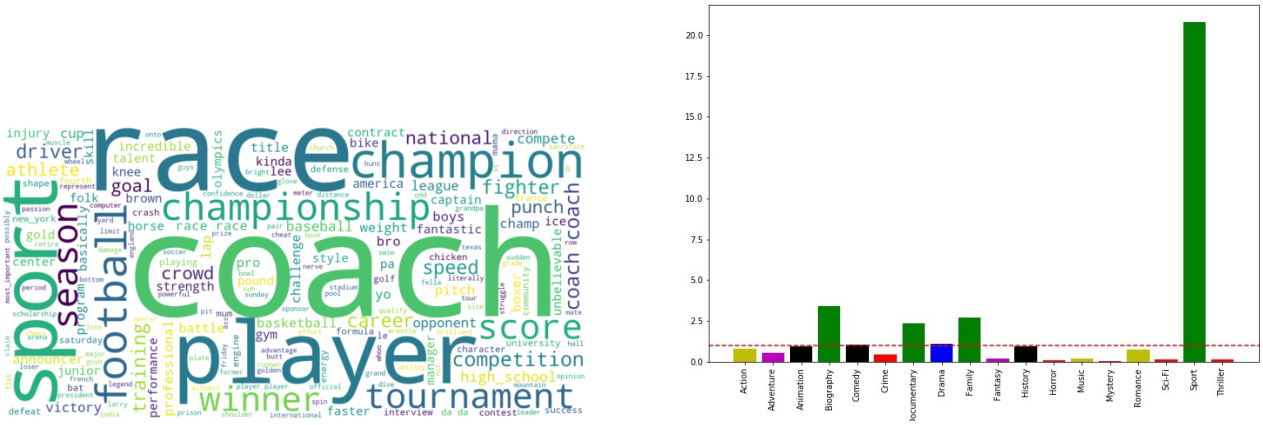
Estos valores relativamente bajos de la información mutua entre las comunidades detectadas por los algoritmos y las formadas por la etiqueta externa de los géneros parecerían marcar que los mismos no emergen de la estructura de comunidades de la red formada a partir de las técnicas de procesamiento natural de lenguaje. Aun así, las redes parecen ser modulares, lo cual podría significar que se está encontrando algún tipo de orden subyacente en su estructura.

Posteriormente se realizaron análisis más profundos de cada comunidad para evidenciar si existe esta identidad subyacente que esté aglomerando a las películas en la red. Por un lado, se confeccionaron gráficos donde se puede observar la distribución de géneros de cada comunidad, y por otro lado se construyeron las nubes de palabras más frecuentes (*WordClouds*) de cada comunidad, los cuales permiten visualizar cuáles son las palabras de los subtítulos que más aparecen en cada una de ellas. Se muestran esquemáticamente los resultados obtenidos de este análisis para algunas comunidades en las figuras 3, 4 y 6.



(a) (b)

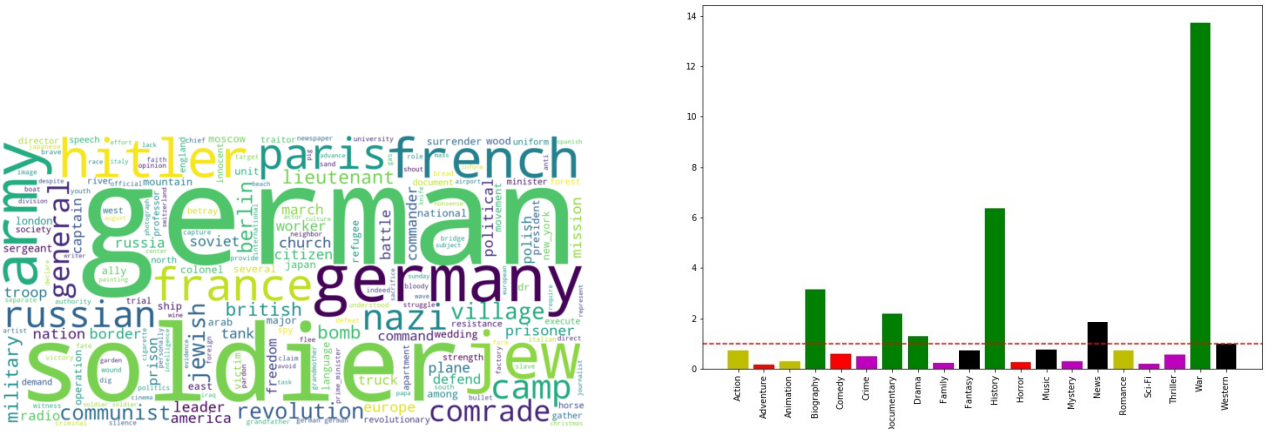
Figura 3: (a) WordCloud de una comunidad detectada por infomap, cuya palabra más frecuente es “paris”. (b) Distribución de géneros para esta comunidad, que contiene muchas más románticas, dramas y comedias que lo esperado.



(a) (b)

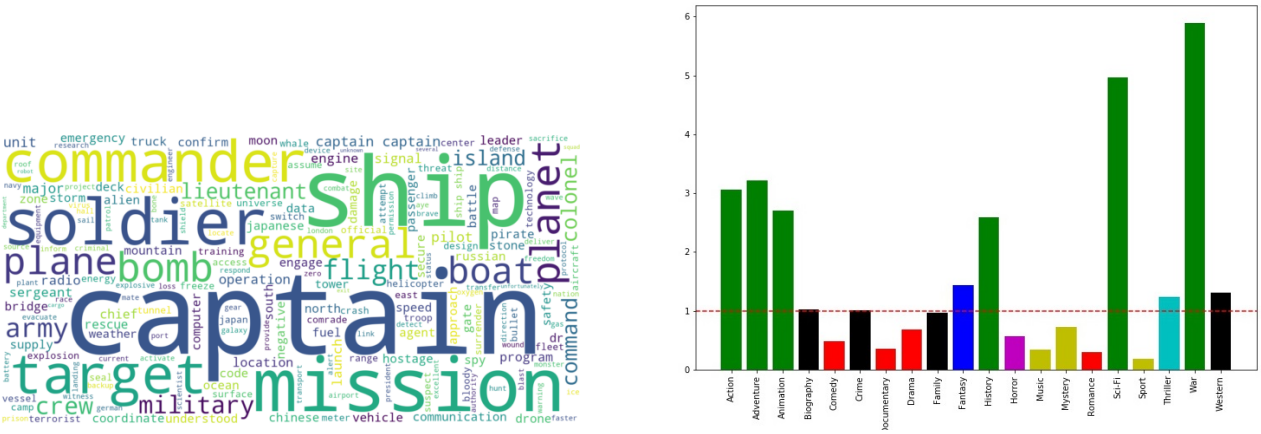
Figura 4: (a) WordCloud de otra comunidad detectada por infomap, cuya palabra más frecuente es “coach”. (b) Distribución de géneros para esta comunidad, que contiene una enorme cantidad de películas de deporte, además de bastantes más biográficas, documentales y familiares de lo esperado.

En los wordclouds, las palabras de tamaño mayor son las que aparecen con más frecuencia. En los gráficos de barra, el eje vertical representa la cantidad de ocurrencias de un género dentro de una comunidad en relación a lo esperado por azar; el valor de 1, indicado por una línea horizontal punteada, indica lo esperado por azar. El color de la barra representa qué tantas más o menos películas hay de las esperadas, medido por la desviación estándar. El orden, con representatividad del género en la comunidad ascendente, es: rojo ($< -3\sigma$), violeta, amarillo, negro (neutral), azul, celeste, verde ($> 3\sigma$).



(a) (b)

Figura 5: (a) WordCloud de una comunidad detectada por infomap, cuya palabra más frecuente es “german”. (b) Distribución de géneros para esta comunidad, que contiene muchas más películas de guerra, historia, biografía, documental y drama que lo esperado.



(a) (b)

Figura 6: (a) WordCloud de una comunidad detectada por infomap, cuya palabra más frecuente es “captain”. (b) Distribución de géneros para esta comunidad, que contiene muchas más películas de guerra, ciencia ficción, aventura, acción, animación e historia que lo esperado. En este caso hay más cantidad de géneros sobre-representados, pero es fácil reconocer como todos ellos suelen tocar la temática sugerida por el WordCloud.

Estas visualizaciones revelaron que las comunidades que emergen de la red están más asociadas a tópicos que a géneros. En el primer ejemplo elegido, la comunidad representada en la figura 3, tiene, entre las palabras más frecuentes de los documentos de cada película, palabras como “paris”, “wine”, “wedding” o “euro”, las cuales nos llevan a pensar que las mismas fueron reunidas bajo una temática de romance francés o europeo. Se puede observar en la distribución de géneros de la comunidad, que hay muchas más películas de romance, drama y comedia que las esperadas completamente por azar. Por lo tanto, estas no fueron agrupadas por un único género, pero, sin embargo, existe un tópico común que representa a la comunidad.

Por otro lado, en la comunidad representada en la figura 4, vemos palabras como “coach”, “player”, “sport” o “race”. Las películas fueron claramente reunidas bajo una temática deportiva. Al observar la distribución de géneros, no resulta sorprendente notar que existe un número de películas etiquetadas bajo el género “Sport” que es órdenes de magnitud mayor al número esperado por azar. En este caso particular, las películas parecen haberse agrupado bajo dicho género, aunque también se observa que el género biográfico y de documentales está sobre-representado, aunque en mucha menor medida. Viendo la extrañeza de este caso en comparación a lo observado en la mayoría de las demás comunidades, las cuales tienen una distribución de géneros en donde la sobre-representación de géneros es más variada, podemos argumentar que los tópicos que se desarrollan en las películas de deportes tienen una tendencia a ser similares entre ellas. Un fenómeno similar puede observarse en una de las comunidades en donde las películas fueron aglomeradas bajo una temática bélica de la segunda guerra mundial, apareciendo entre sus palabras más frecuentes “german”, “soldier”, “jew” y “hitler”, en la figura 5. Al observar la distribución de géneros, vemos también que hay una sobre-representación muy considerable del género “War”, pero resulta evidente que existe una correlación muy grande entre el tópico bajo el cual fueron reunidas estas películas y dicho género. Es decir, ciertas temáticas se correlacionan más con géneros particulares que otras.

Como último ejemplo, en la comunidad representada en la figura 6, encontramos entre las palabras con mayor aparición, “captain”, “ship”, “soldier”, “mission” y “commander”, lo que da a pensar en una temática de aventura y de guerra enfocada alrededor del uso de barcos, aviones y otras vehículos con un cierto prestigio por su uso para cumplir determinadas misiones con fines mayoritariamente bélicos. Al observar la distribución de géneros, vemos que hay muchos de ellos que están sobre-representados, como aventura, acción, animación, ciencia ficción, historia y por supuesto, guerra. Claramente vemos que es una temática abordada por gran cantidad de géneros, aunque en cada caso con diferentes matices. También es interesante que tanto esta comunidad como la anterior se caracterizan por incluir muchas películas de guerra, pero dentro de este género las temáticas pueden ser muy distintas.

En todos los casos puede verse como los módulos de la partición por algoritmos de detección de comunidades (en este caso infomap) están en relación con una partición por temáticas, que se inducen a simple vista por la bolsa de palabras del corpus de texto correspondiente a cada comunidad. Es imposible ahora medir observables

como la información mutua, debido a que no disponemos de una partición externa de referencia de temáticas. Sin embargo, los gráficos de barra muestran que hay géneros sobre y sub representados en cada comunidad, lo cual no significa que LSA+infomap esté detectando clusters por género, sino que posiblemente haya una relación entre temáticas y géneros: las películas de temática 'francia' suelen ser románticas, las que hablan sobre entrenadores, jugadores, carreras y campeones suelen ser de deportes y biográficas. El cuarto caso es ilustrativo: hay palabras que son comunes a varios géneros, tanto por su multiplicidad de significados (ship puede ser una nave marina o espacial) como por su contextualización semántica: muchas películas animadas, aparentemente alejadas de temáticas de guerra o historia, comparten bastante lenguaje en común. Es por este motivo que LSA+infomap probablemente no podría, al menos por sí solo, detectar clusters puramente por género.

Homofilia

Aunque la información mutua entre la partición por género sea baja, el hecho de que los gráficos de barras revelen cierta representación del género nos llevó a estudiar la homofilia de la red. Para medir la misma tuvimos que utilizar un criterio específico. Las películas están etiquetadas, casi siempre, con hasta tres géneros. Sin embargo, hay alrededor de 500 combinaciones de géneros ocurrentes en nuestro conjunto de películas, por lo que los enlaces homofílicos serían siempre muy bajos si se compara la combinación exacta de géneros. El criterio aplicado fue el siguiente: supóngase una película A con un conjunto de géneros G_A, y otra película B con un conjunto de géneros G_B; A y B están enlazadas en la red. Se considera que el enlace entre A y B es homofílico si y sólo si:

$$|G_A \cap G_B| \geq \frac{|G_A \cup G_B|}{2}$$

A partir de la red original, creamos 100 redes, donde se preservó la estructura y se distribuyeron aleatoriamente los géneros. O sea, dados los conjuntos G_i de géneros asociados a los nodos i de la red original, se asignó aleatoriamente cada uno a un nodo $j \neq i$ (podía tocar $j = i$ con una probabilidad muy baja, igual que la de cualquier otro nodo). Se computó la homofilia para cada una de estas redes, y graficamos la distribución de homofilia en la figura 7. La línea vertical roja representa la homofilia de la red original, de $\sim 0{,}56$, muy por arriba de lo esperado por azar (entre 0,34 y 0,35).

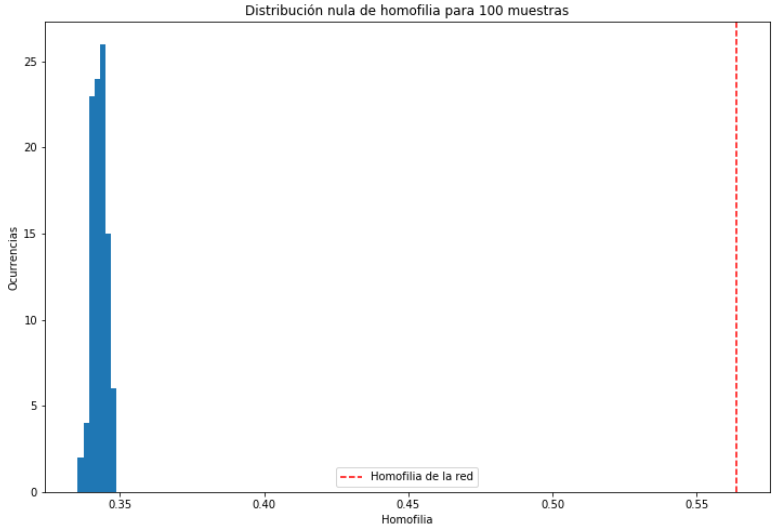


Figura 7

El carácter homofílico de esta red proviene probablemente de la relación entre temáticas y géneros sugerida por

las figuras 3, 4 y 6.

5. Conclusiones

Las redes construidas presentan estructura de comunidades, pero las particiones obtenidas tienen un coeficiente de información mutua bajo con una partición de referencia de género. Esto no significa, sin embargo, que no hubiera relación entre las comunidades y el género de la película: estudiando en detalle cada comunidad, pudimos ver que emergieron temáticas (en los wordclouds) y una distribución no aleatoria de géneros. Además, se caracterizó a la red como homofílica en género (56 % de enlaces entre películas que coincidían en más de la mitad de sus géneros), lo cual significa que LSA tiende a conectar subtítulos de películas del mismo género más de lo que lo haría por azar. A primera vista, comparando las temáticas de cada comunidad con su respectiva distribución de géneros, parece haber una relación entre las temáticas y los géneros más representados por la comunidad, lo cual indicaría que esas temáticas suelen estar vinculadas a esos géneros.

Como continuación a este trabajo, podría buscarse alguna partición externa de las películas por temática y medir la información mutua. Un análisis interesante que puede hacerse es el de la red de palabras. Podría estudiarse la centralidad de las palabras en el corpus de cada comunidad y de toda la red, y ver por ejemplo si son importantes para mantener la conectividad, además de compararlas con algún índice de peso emocional (proveniente de algún estudio psicológico). También podría efectuarse un estudio análogo al de este trabajo, pero haciendo análisis sentimental en el corpus y viendo si hay géneros que se relacionan más con categorías emocionales (por ejemplo, drama y comedia).

También es posible, como se mencionó en la introducción, que la información de género contenida en una película no pueda ser correctamente interpretada sin analizar todos los canales de comunicación disponibles: la imagen y el sonido, en conjunto, pueden utilizarse como medios de transmisión de información sumamente abstractos y de alto contenido emocional.

Referencias

[1] Thomas K Landauer and Susan T Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

[2] Albert-László Barabási et al. *Network science*. Cambridge university press, 2016.

[3] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[4] Martin Rosvall, Daniel Axelsson, and Carl T Bergstrom. The map equation. *The European Physical Journal Special Topics*, 178(1):13–23, 2009.