DEPARTAMENTO
DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Análisis cuantitativo de sesgos culturales en películas de Hollywood

December 30, 2017

Valeria Tiffenberg
valetiff@gmail.com

Directores

Edgar J. Altszyler Lemcovich
edgaralts@gmail.com

Ramiro H. Gálvez
ramirogalvez@gmail.com

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - (Pabellón I/Planta Baja)
Intendente Güiraldes 2610 - C1428EGA
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (++54 +11) 4576-3300
http://www.exactas.uba.ar

Resumen

Las películas y series de Hollywood son consumidas masivamente en todo el mundo, y a través de ellas nos vemos expuestos a ciertas normas socio-culturales. En este proyecto, buscamos generar métodos para extraer los sesgos culturales existentes en las películas, focalizándonos en analizar en qué contexto social y cultural aparecen determinados actores.

Para cuantificar la presencia de estereotipos en los grupos a analizar, aplicaremos técnicas de minería de texto. Las técnicas utilizadas se basan en la identificación y comparación de los contextos en los que se menciona a los distintos agentes a estudiar, a partir de un gran corpus de subtítulos de películas desde 1930 a la actualidad.

Este sistema de análisis automático nos permite evidenciar y monitorear los estereotipos que consumimos a través de las películas y su evolución en el tiempo. De esta manera podemos responder preguntas de interés cultural, tales como qué rol se le otorga a la mujer y cómo evoluciona el mismo, y cómo va variando la visión que incorporamos sobre distintos países y religiones con los años, y así reflexionar sobre cómo estos modelos que consumimos podrían condicionar los roles asignados socialmente y viceversa.

Abstract

Hollywood movies and series are globally consumed every day and through them we are exposed to a certain set of sociocultural norms. In this thesis, we intend to find methodologies that allow us to find the cultural biases present in movies, by focusing on analyzing what are the contexts in which certain people are shown.

In order to quantify the presence of stereotypes in each group, we'll apply methods from text mining. Those methods used are based on identifying and comparing the contexts in which the relevant actors are mentioned inside a corpus of movie subtitles ranging from 1930 to today.

This automatic analysis will allow us to present and track those stereotypes we consume through movies, and how they've evolved through time. In this manner, we'll be able to answer culturally relevant questions such as what is the role of women in movies and what changes it has seen, and what is the light in which people who practice certain religions or come from different countries are shown in this media. With these answers we'll be able to reflect on how these models affect societal roles and vice versa.

Agradecimientos

A los directores de esta tesis: Edgar y Ramiro, por su eterno entusiasmo, sus ideas y su tiempo.

A la Universidad de Buenos Aires y el Departamento de Computación, por una educación de calidad.

Al equipo de OpenSubtitles que nos dio acceso a su base.

A mi familia entera, por las reuniones, la compañía y el cariño de estos años.

A todos los amigos que hicieron que la facu se sintiera como una casa, a Gabi por todos los TPs, y a Leo, Polín y Fran por bancarme siempre.

A Javi, por guiarme a esta carrera, y a Pablo, por guiarme a esta tesis.

A mis viejos, por alentarme y escucharme siempre.

A Brian por el cariño, el aguante, y la comida que hizo falta para que esta tesis se hiciera.

Índice general

Capítulo 1

Introducción

> It's the movies that have really been running things in America ever since they were invented. They show you what to do, how to do it, when to do it, how to feel about it, and how to look how you feel about it.
>
> *Andy Warhol*

Las expresiones culturales humanas reflejan ideologías, costumbres y realidades cotidianas de una cierta sociedad. La realidad en la que vive un artista embebe su arte, y a su vez, el arte que consumimos nos forma como seres humanos pertenecientes a un contexto social, y cuanto más masivo sea ese arte, más amplia será la llegada de las ideas que transmite al tejido social que lo consume.

En este trabajo, el foco de estudio será la detección de ideología transmitida a través del cine. Para comenzar la investigación, vamos a partir de resultados documentados en otra bibliografía, tales que evidencian preconceptos y estereotipos en diversas fuentes, como literatura, medios o redes sociales. Aplicando tres metodologías del ámbito computacional del procesamiento de lenguaje y minería de texto buscaremos evidencia de si estos métodos sirven también para detectar estos preconceptos en el cine.

El dataset que da origen a este trabajo corresponde a todos los subtítulos en inglés de la base de datos de OpenSubtitles[1] desde sus primeras películas hasta principios del 2015, con lo cual el análisis a realizar será sobre un corpus compuesto por el texto de los diálogos de cada película, más el agregado de la temporalidad: cuándo se dice cada frase y a qué distancia está de las circundantes.

1.1 Motivación y objetivos

La cultura que produce el cine norteamericano es consumida por la sociedad estadounidense y por buena parte del mundo, y aquello que vemos en el cine afecta la forma en la que pensamos y nuestro accionar (Charlesworth y Glantz, 2005; Linz *et al.*, 1984). Por eso, nos enfocaremos en averiguar si es posible cuantificar tendencias en el cine con métodos normalmente aplicados a texto, en búsqueda de preconceptos, estereotipos o simplificaciones frente a ciertos grupos de gente.

Como inspiración a este trabajo, el consumo personal de películas estadounidenses nos permite saber que existen narrativas que se repiten en numerosas películas y pasan a ser consideradas como parte misma del formato del cine. A partir de esta experiencia elegimos los tópicos en los que se

[1] https://www.opensubtitles.org/

enfoca este trabajo, pero la investigación de estereotipos a través de conceptos asociados en el cine podría pasar por múltiples focos diferentes, y ser tan amplia o específica como se desee. Los métodos que vamos a testear a continuación tendrán cada uno sus aplicaciones y casos ideales de uso, y podrían ser utilizados sobre prácticamente cualquier tema para hallar nuevos resultados.

A partir de estas motivaciones, el objetivo del trabajo será investigar métodos habituales de procesamiento de lenguaje natural para tratar de replicar resultados existentes en otra bibliografía y observar si el análisis de gran cantidad de datos aplica también a diálogos en cine en el mismo grado que aplica a corpus de noticias o de literatura.

Lo primero en lo que queremos hacer foco es en el rol que se le adjudica a las mujeres en el cine. Con la ampliación del mercado laboral a la mujer a partir de la Segunda Guerra Mundial, y luego, la década de los 60 (Donnelly *et al.*, 2016), y el crecimiento del movimiento feminista en los últimos años (Cochrane, 2013), vamos a investigar si la entrada de la mujer en ámbitos tradicionalmente masculinos también se refleja en la imagen cinematográfica.

Por otro lado, vamos a enfocar la visión norteamericana sobre culturas ajenas, en particular la imagen estadounidense del comunismo en Rusia y en la Unión Soviética, y su continuación a través de la mafia; y la población perteneciente al medio oriente, especialmente su cercanía con el concepto de "enemigo" y "terrorista". Nos interesa profundizar cómo el ojo cultural estadounidense traduce las imágenes y costumbres de otros países, buscando si hay evidencia de que los personajes siempre siguen estereotipos comunes, o si hay variabilidad.

Para estos casos, nos resulta interesante evaluar la participación de estas poblaciones en el cine alrededor de eventos históricos que impactaron en todo el mundo: la disolución de la Unión Soviética en 1991, que representa el final de la Guerra Fría entre Rusia y Estados Unidos; y el ataque a las Torres Gemelas en 2001, con el comienzo de la guerra contra el terrorismo ("War on Terror"). El caso del terrorismo y los países involucrados con el ataque a las Torres Gemelas es particularmente relevante en la actualidad por el aumento de la islamofobia a nivel mundial (Lichtblau, 2015).

1.2 Trabajos previos

1.2.1 Culturonomics

Michel *et al.* (2011) acuñaron el término "culturonomics" en 2010 para referirse al estudio de tendencias culturales a lo largo del tiempo a través de corpus de grandes cantidades de datos. En el artículo al cual nos referimos se menciona que este estudio puede ser utilizado en todo tipo de ciencias: lingüística, histórica, social, etc. Muestran así la aplicabilidad del estudio de **n-grams** (secuencia consecutiva de n palabras) en múltiples campos y los resultados obtenidos. Para esa investigación se utiliza un corpus de libros de más de 5 millones de ejemplares, pero incluso como parte de la presentación de esta nueva disciplina, los autores exponen que estas estrategias deberán ser aplicadas a todo tipo de corpus, incluyendo medios de comunicación, manuscritos, mapas, y "un sinnúmero de otras creaciones humanas". En este trabajo vamos a aplicar "culturonomics" a un corpus de cine a partir de los años 30 hasta la actualidad.

En el artículo fundacional de "culturonomics" (Michel *et al.*, 2011) la medida que se utiliza para analizar los campos elegidos es la **frecuencia**, a la que Michel *et al* definen como la cantidad de apariciones de un cierto n-gram en un año (con n entre 1 y 5) dividido por la cantidad total de palabras de ese año (la fórmula que utilizaremos para medir esta frecuencia se encuentra en la sección 1.4.1).

1.2.2 La noción de información mutua

Previo a la invención del término "culturonomics", se realizaron numerosos acercamientos a la idea de analizar datos legibles por computadora con el fin de realizar análisis lingüísticos a mayor escala.

Hasta ese momento, este tipo de observación se hacía de forma manual a través de cuestionarios a algunos cientos de sujetos, método que es costoso y limitado en la cantidad de personas a entrevistar.

Uno de los primeros acercamientos fue el método de búsqueda de asociaciones de palabras llamado "association ratio" (índice de asociación) basado en un concepto de teoría de la información: "mutual information" (información mutua). Éste índice compara las probabilidades de ocurrencia conjunta de dos palabras con las probabilidades independientes de cada una de aparecer en el corpus (Church y Hanks, 1990). Se utiliza para **ventanas** fijas de palabras: centrándose en la palabra a analizar, una cantidad fija de términos previos y posteriores con los que coocurre. El tamaño de la ventana es una variable más: las ventanas chicas (3 o 4 palabras en total) permiten encontrar expresiones comunes, mientras que las ventanas con mayor cantidad de palabras arrojan luz sobre relaciones semánticas.

1.2.3 Medición de distancias en espacios vectoriales

A partir de la noción de cercanía semántica se originaron múltiples métodos de **word embeddings**, cuya idea es modelar palabras a través de insertarlas en un espacio vectorial de manera que aquellas que sean semánticamente similares tenderán a quedar próximas entre sí en este espacio. Estas estrategias son el método más usado para calcular proximidad semántica entre palabras, párrafos o documentos (Jurafsky y Martin, 2014).

El comienzo más habitual para generar word embeddings es a través de la creación de **matrices palabra-contexto** (o palabra-palabra). Éstas registran para cada palabra todas las otras palabras que aparecen en su contexto (se entiende como contexto a una ventana de tamaño fijo), es decir, la cantidad de co-ocurrencias de esas dos palabras en el corpus. Existen también las matrices palabra-documento, donde se genera un vector para cada documento con una cantidad de apariciones para cada palabra, lo que permite la comparación de textos a mayor escala, cosa que ayuda a la investigación semántica de cada documento, pero no de cada palabra, y por lo tanto no es el foco de este trabajo.

Las matrices palabra-contexto son cuadradas y cada dimensión es del tamaño del vocabulario completo para el documento o corpus analizado, pero son a su vez, matrices esparzas con pocos elementos distintos a 0, lo que facilita su manejo utilizando estructuras de lectura rápida para ciertas operaciones de suma o búsqueda, y que ocupan menos memoria que manipular la matriz entera. Jurafsky y Martin (2014) proponen cómo utilizar el índice de asociación de Church, llamado actualmente **Pointwise Mutual Information** (PMI), a partir de estas matrices.

Uno de los principales métodos dentro de los conocidos como word embeddings es un modelo denominado Latent Semantic Analysis (LSA) (Landauer y Dumais, 1997; Landauer *et al.*, 1998) que genera un espacio vectorial como el que se podría pensar a través de los vectores de coocurrencias ya mencionados, pero de una cantidad de dimensiones mucho menor. Existen múltiples beneficios de la reducción de dimensionalidad: la extracción de significados latentes, la reducción de ruido, el aumento de densidad (Turney y Pantel, 2010).

La generación del espacio vectorial de LSA parte del mismo lugar que los cálculos previos: las matrices esparzas de coocurrencia entre palabras y contextos. Pero luego se realizan dos etapas nuevas: a cada celda de la matriz se le aplica una transformación según la importancia de la palabra en la frase, y se aplica Singular Value Decomposition (SVD) para convertir los vectores esparzos en vectores densos que condensan la información en menor cantidad de dimensiones.

Otro beneficio de manejar este espacio vectorial reducido, es la inclusión de mayores órdenes de coocurrencia. Utilizando PMI sobre dos palabras no coocurrentes éstas devuelven una similaridad de menos infinito, pero midiendo distancias entre vectores pueden establecerse relaciones de cercanía semántica entre palabras que pueden no ser utilizadas juntas, pero sí tienen contextos similares (Jurafsky y Martin, 2014).

Múltiples autores han usado este método con éxito para mostrar análisis sobre distintos corpus. Lenton *et al.* (2009) utilizaron LSA sobre un corpus de lecturas estándar para estudiantes estadou-

nidenses desde el comienzo de su escolarización hasta el primer año universitario, para mostrar que en estas lecturas los conceptos de hombre y mujer contienen estereotipos con respecto a los roles y atributos "estándar". Sagi *et al.* (2013) utilizan LSA sobre múltiples ediciones de un diario, para mostrar cómo se contextualizan ("framing") las discusiones sobre "terror" y su relación al terrorismo; y las transcripciones de discursos de senadores estadounidenses para iluminar el framing que se realiza del debate sobre el aborto y evaluar cómo eso puede pesar sobre la opinión pública.

Mikolov *et al.* (2013) propusieron un segundo acercamiento posible a word embeddings a través de redes neuronales. A este conjunto de técnicas se las denominó **word2vec**, y han demostrado ser mejores que LSA para los análisis de grandes corpus de textos (Altszyler *et al.*, 2016).

Los modelos de word2vec están inspirados en modelos de lenguaje neuronal, esto implica entrenarlos para que predigan palabras similares. La tarea de predicción no es el objetivo último, pero a partir de estos modelos se vio que la representación de las palabras sirve también para evaluar similaridad semántica. (Jurafsky y Martin, 2014)

La idea de cómo aprende una red dónde ubicar cada palabra es a partir de comparaciones. Utilizando el algoritmo skip-gram con muestreo negativo, en cada ventana de palabras se toma aquellas que forman el contexto como cercanas y luego se toma la misma cantidad de palabras al azar como lejanas. De esta forma, se organizan los embeddings de una cierta palabra para que esté cerca de sus palabras contexto, y lejos de las palabras al azar. Para cada palabra se guardan dos vectores: uno en una matriz W que conserva aquellos que representan cada palabra y uno en una matriz C que conserva los contextos de todas las palabras. (Jurafsky y Martin, 2014)

1.2.4 Cambios del lenguaje en el tiempo

Dentro del foco de este trabajo se encuentra aportar evidencia a que estos análisis computacionales de corpus extendidos en el tiempo permiten detectar cambios de tendencia a lo largo de los años. Kulkarni *et al.* (2015) utilizan frecuencias relativas, "part-of-speech tagging" y word embeddings para detectar exitosamente cambios sintácticos y semánticos a través del tiempo. Tomando una palabra cuyo significado está documentado que cambió a lo largo del tiempo logran detectar cuáles eran los contextos donde se usaba previamente, cuáles son los nuevos contextos, y por lo tanto, su nuevo significado, y en qué período se produjo la rotación de una definición a la otra.

En segundo lugar, Hamilton *et al.* (2016) utilizan metodologías similares a las mencionadas en esta sección (Positive Pointwise Mutual Information, SVD, Skip-gram with negative sampling, éste último es uno de los métodos dentro de word2vec) para detectar cambios ya detectados en otra bibliografía, y mostraron la eficiencia relativa de cada método para esta tarea. Diuk *et al.* (2012) utilizaron métodos similares y mostraron la eficacia de LSA por encima de la frecuencias de aparición para analizar la evolución del concepto de "introspección" en la literatura, además de mostrar la factibilidad de obtener métricas sólidas sobre conceptos abstractos.

1.3 Descripción del corpus

La base de datos utilizada para este trabajo fue provista por un administrador de OpenSubtitles y se compone de todos los subtítulos en inglés (de originales en inglés y en idioma extranjero) desde lo más antiguo que se conserva en el sitio hasta entrado el año 2015.

El corpus entero tiene cantidades variables de originales (películas,series y juegos) para cada año, aumentando en cantidad de películas fuertemente pasando el año 2000. El material más antiguo del corpus es aquel que sobrevivió al paso del tiempo, y posiblemente no es una representación fiel de todo el cine de esa época. Aún así, para este estudio este material que se conserva es pertinente, ya que si se mantuvo a lo largo del tiempo es gracias a su impacto en el público que lo vio y continúa viendo, y por lo tanto, fue influyente en su época.

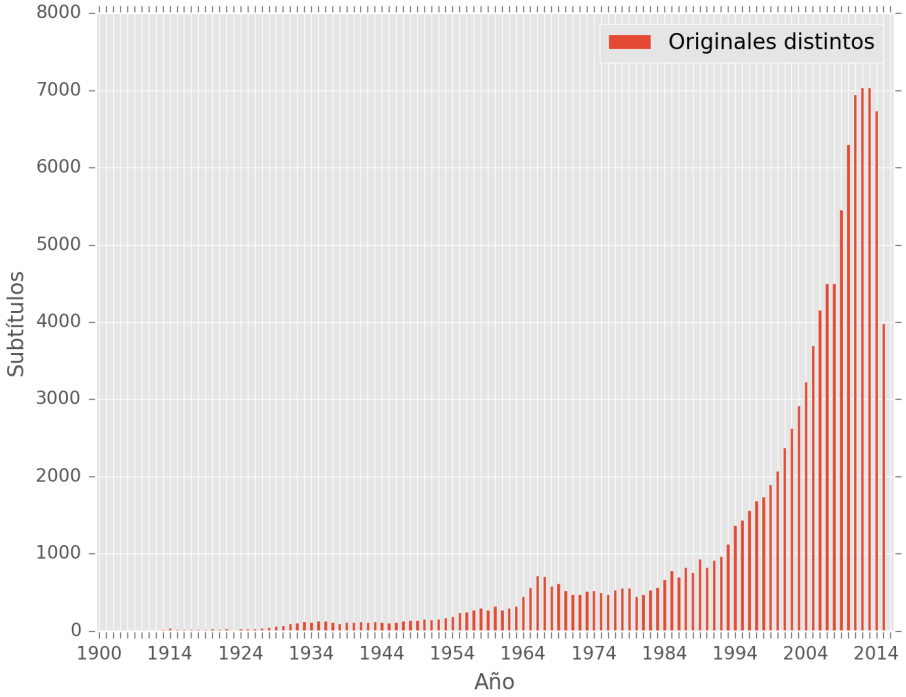


Figura 1.1: Corpus sin filtrar tomando un único subtítulo por original

Todos los subtítulos están en formato .srt, y en inglés, así que no hay necesidad de filtrarlos por estos campos. Cada subtítulo tiene dos números identificatorios: el número único que permite reconocer ese archivo exactamente, y el número que identifica a la película, y que puede repetirse cuando hay varios subtítulos para la misma película. En el diagrama anterior, se está agrupando por número de película, por lo que en la figura 1.1 no está mostrando el total de subtítulos, si no el total de material original diferente.

Para eliminar los repetidos se utilizó como criterio la reputación del usuario que lo subió. El nombre de usuario no estaba entre las columnas pero sí su categoría como "uploader" dentro del sitio. Las categorías tienen un orden de prestigio: primero "super admin", luego "administrator", "subtranslator", "platinum member", "vip plus member", "vip member", "gold member", "trusted", "silver member", "bronze member", "sub leecher", y por último, usuario normal[2]. Para elegir los subtítulos se eligió el del rango más alto posible, o en última instancia, el único que hubiera.

Hasta este punto la catalogación de los subtítulos fue estrictamente extraída de la base de datos de OpenSubtitles: los años que figuran ahí se toman como aquel en el que salió la película, y todo el resto de los datos hacen referencia a los datos del subtítulo como tal. Pero para poder analizar en detalle las películas con las que se iba a trabajar se hizo una descarga de metadata de OMDb Api[3], el sitio web que reúne una base de datos de cine y televisión, es de acceso público, e incluye el elenco y equipo completo de la película, origen, calificaciones y clasificaciones de cada película. El identificador único de OMDb para cada subtítulo era uno de los datos dentro de la base original, por lo que toda la descarga de datos extra se hizo sin necesidad de buscar, si no que cada archivo está vinculado directamente con sus datos.

[2] http://forum.opensubtitles.org/viewtopic.php?t=1991
[3] http://www.omdbapi.com/

Los campos de metadata utilizados dentro del análisis fueron: tipo, idioma y país de origen. Como se ve en la figura 1.2 el tipo nos permitió identificar que, además de películas, la base de datos de OpenSubtitles también contiene series y juegos. Pero los subtítulos responden a 4 categorías: películas, capítulos, series y juegos. No queda clara en el corpus la diferencia entre "series" y "episode", pero por la escasez de subtítulos bajo "series" es posible que se trate de material de un sólo capítulo exclusivo para televisión, o que hayan sido capítulos de series mal catalogados. En cualquier caso, nosotros vamos a trabajar sólo con los subtítulos catalogados como película.

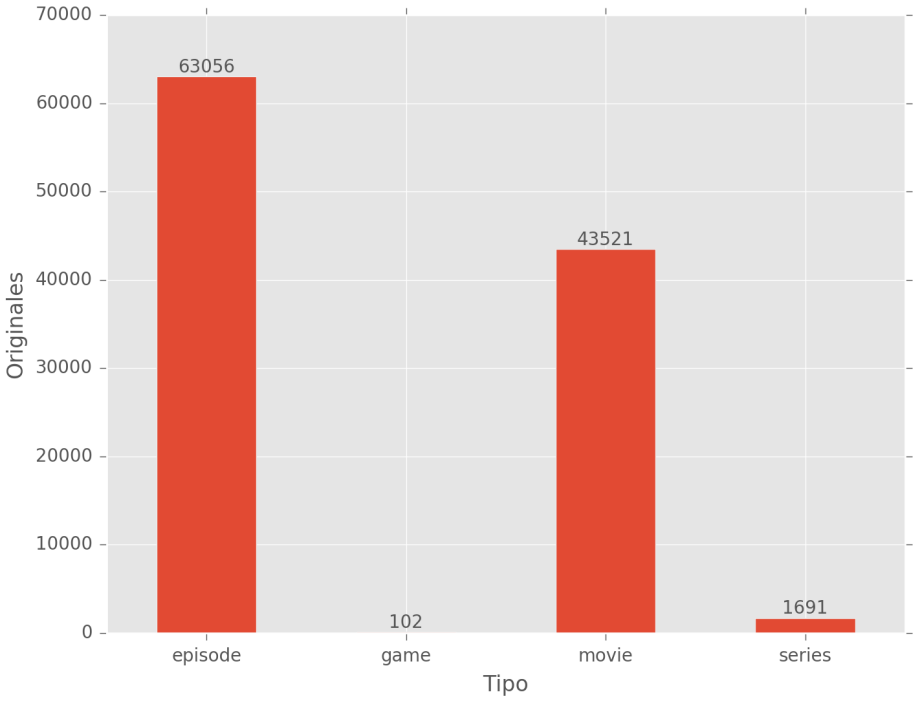


Figura 1.2: Tipos de todos los subtítulos del corpus. Para este trabajo nos interesa únicamente "movie"

Una vez hecho este filtro quedan 43.521 subtítulos de películas diferentes, pero dado que el foco de este trabajo es el cine estadounidense, se buscó el origen e idioma original de estas películas según figuraba en OMDb. OMDb registra el origen de una película como una lista de países que lo produjeron o coprodujeron. Para estudiar el cine estadounidense elegimos quedarnos únicamente con las películas que tuvieran "USA" en esa lista de países. Eso nos deja con 16.970 películas.

El idioma original de la película también se representa en OMDb como una lista de todos aquellos que tienen alguna línea de diálogo. Quedándonos únicamente las películas que tienen inglés entre sus idiomas, el número no se reduce, todas las 16.970 producidas por Estados Unidos tienen inglés como uno de sus idiomas.

Habiendo hecho todos los filtros necesarios, la distribución de películas por año queda como en la figura 1.3. La cantidad de películas es casi inexistente antes de 1932, mejora sustancialmente a partir del año 2000, y sigue subiendo hasta el 2014. El año 2015 fue el de extracción de la base y no está completo, por lo que el último analizado será 2014.

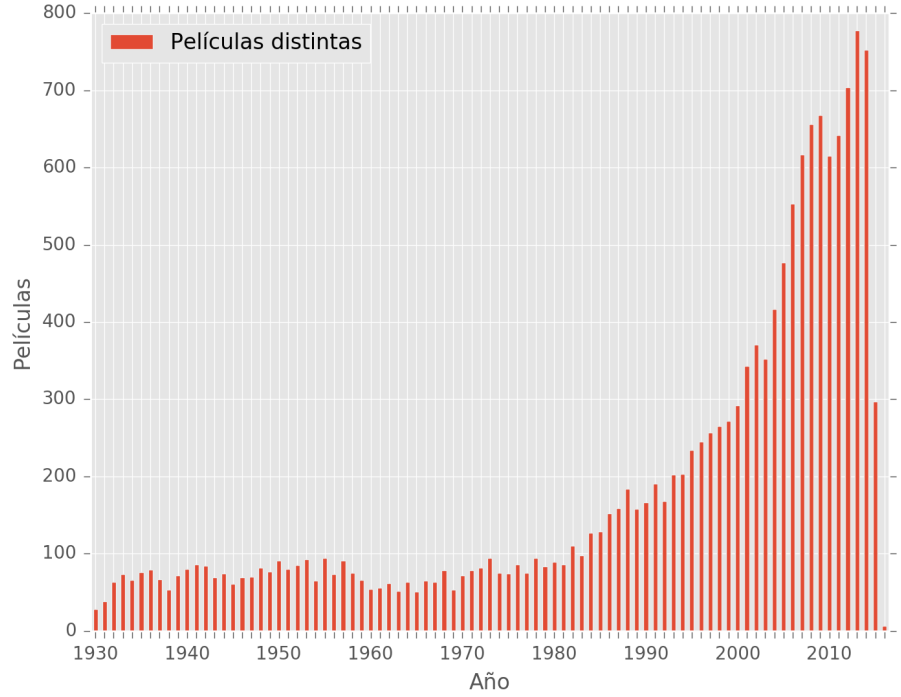


Figura 1.3: Total de subtítulos a analizar por año una vez filtrado por país productor, sólo películas, un subtítulo por cada película y que incluye inglés como idioma original

Una vez obtenido un índice de todos los subtítulos queda por clarificar qué contiene cada archivo de subtítulo. Dado que todos los archivos en la base están en formato .srt, cada subtítulo contiene "frames", y cada frame tiene la siguiente estructura:

```
71
00:09:01,307 --> 00:09:03,025
I'll call you back, thanks.
```

Donde 71 es el número de orden del frame. Los "timestamps" son el tiempo de comienzo donde el frame debe mostrarse, y el de fin en el debe dejar de verse. Por último se encuentra el texto de ese frame.

1.4 Metodología

Los métodos elegidos para analizar este corpus son algunos de los que se utilizan habitualmente para realizar análisis semántico en otros tipos de texto: periodístico, literario, científico, archivo de redes sociales, etc. Pero, hasta donde hemos visto, no han sido utilizados en el contexto del cine.

En esta sección explicaremos el uso de cada uno y el procesamiento necesario del corpus para obtener las mediciones deseadas. Cada instrumento aporta un ángulo distinto sobre las problemáticas a investigar y se complementan entre sí de diversas formas: primero, la frecuencia de palabras será una medida inicial del cine de cada época que aporta una visión sobre el uso y desuso del vocabulario. Segundo, Positive Pointwise Mutual Information y Word2vec serán dos medidas diferentes que permiten apreciar distancias semánticas entre palabras. Por último, la Robustez de asociación

funciona como medida complementaria a PPMI y muestra cuán fuerte es la asociación detectada por esta medida.

1.4.1 Procesamiento de datos

Las métricas que serán detalladas en esta sección se generan a partir de dos estructuras de datos: un índice de frecuencias, y matrices anuales de coocurrencia palabra-palabra. Para generar ambos tipos de estructuras de datos se debe recorrer el corpus completo analizando cada subtítulo y agregando todas las palabras de se contexto (aquellas que se encuentran cercanas en el texto) a la estructura correspondiente.

Es necesario obtener las palabras de cada subtítulo "limpias" de puntuación o contracciones junto con los timestamps para el cálculo de la ventana. Para esto se utiliza pysrt, una biblioteca que "parsea" este formato y divide el texto plano en una lista de frames, donde cada uno tiene un tiempo de comienzo, un tiempo de fin, su duración y el texto.

Cada vez que se analiza un subtítulo se ejecuta la misma serie de pasos a través de un tokenizador:

1. Se limpia el texto:
   (a) Se eliminan los tags HTML del texto, a través de una expresión regular
   (b) Se eliminan las descripciones de acción, sacando todo el texto entre paréntesis y entre corchetes
   (c) Se eliminan los nombres de los personajes que hablan (en subtítulos comúnmente estos están en mayúsculas al principio de una línea y antes de dos puntos)
   (d) Se eliminan los guiones al comienzo de un diálogo entre dos personas en el mismo frame
   (e) Se reemplaza "&" por "and"
   (f) Se eliminan los espacios extras
2. Se tokeniza (divide el texto en palabras, eliminando la puntuación y separando palabras con apóstrofes) el resultado del paso anterior utilizando TextBlob[4], lo que resulta en una lista de palabras
3. Se eliminan las "stopwords" (palabras más comunes del lenguaje, que no aportan a la caracterización semántica) del subtítulo. Para elegir cuáles son, utilizamos la lista presente en NLTK[5], pero conservamos los pronombres de ambos géneros ("she", "he", "her", "him", "hers", "his").

Una vez realizado este proceso tendremos un arreglo de palabras, en orden de aparición, sin stopwords. Esta serie de pasos se realiza con el texto completo de una película o con los textos dentro de una sola ventana (cantidad limitada de palabras antes y después de la analizada), en los casos donde se arman las matrices, dado que necesitan considerar los timestamps.

En el índice de frecuencias se guarda, por año, toda palabra del vocabulario, cada una asociada a un número que representa la cantidad de apariciones. Luego, partiendo del arreglo con todas las palabras, si la palabra ya existe, se suma uno a su cuenta, si no, se agrega con un 1 asociado.

Las matrices palabra-palabra son anuales y específicas al tamaño de ventana: para generarlas se recorren todas las películas de cada año y en cada una se recorre el archivo centrando cada iteración en un frame. La cantidad de palabras que entran en esa ventana es variable: para el caso donde el frame analizado empieza en el segundo t y termina en u, y con una ventana de 10 segundos, entrarán las palabras de todos los frames que terminen pasado $t - 10$ y que empiecen antes de $u + 10$ (figura 1.4). Esto puede generar ventanas de 1 frame, si los siguientes y anteriores están muy alejados, o

[4] https://textblob.readthedocs.io/en/dev/
[5] https://pythonspot.com/en/nltk-stop-words/

de 3 o 4 frames; y cada uno de esos frames tiene una cantidad variable de palabras, comúnmente entre una y diez.

Para cada palabra delimitada por la ventana, se registran todas las demás como coocurrencias tantas veces como aparezcan. En la figura 1.4, para la palabra "buen" se registrarán "hola" y "día" como coocurrencias, pero también todas las palabras en los subtítulos $s1$, $s2$ y $s4$.

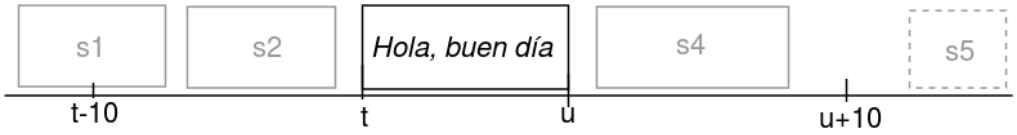


Figura 1.4: Ejemplo de ventana de 10 segundos. La iteración está centrada en el subtítulo en negro.

Para cada matriz anual tenemos una estructura auxiliar secundaria que mapea cada palabra con su índice correspondiente en la matriz. Por simplicidad, utilizamos el mismo índice para cada palabra: tanto cuando es la estudiada (fila de la matriz), como cuando funciona como contexto (columna de la matriz). Por este motivo, la matriz es simétrica y por lo tanto, obtener la cantidad de palabras que tienen a i en su contexto es equivalente matemáticamente a obtener la cantidad de palabras que aparecen en el contexto de i.

Los gráficos en los capítulos siguientes se realizan a partir de los índices y matrices acá descriptos. Pero para simplificar la visibilidad de tendencias todos, excepto aclaración, tienen un "smoothing"[6] de 3 años. Esto quiere decir que, para cada serie, al resultado de cada año se le suman los 3 años anteriores y los 3 años siguientes y el total se divide por 7. El efecto de este cálculo es que la curva se suaviza tomando en cuenta las tendencias de años circundantes.

1.4.2 Frecuencia

El primer método de acercamiento al corpus será el análisis de frecuencias como lo hemos visto en Michel *et al.* (2011) y Kulkarni *et al.* (2015). Se define como la cantidad de apariciones de una palabra dividido el total de palabras de ese año:

$$Frecuencia(x) = \frac{f(x)}{n}$$

con f la cantidad de apariciones de la palabra x en el año, y n el total de palabras de ese mismo año.

Tomamos como primer ejemplo de uso a gran escala el caso de Google Ngram, utilizado en Michel *et al.* (2011) como base de análisis, y también accesible vía web[7]. La medición de frecuencias permite verificar el momento en el cual ciertos términos claves comenzaron a aparecer en la literatura, en el caso de Google Ngrams, o en el cine, en nuestro caso. Es una herramienta útil para ver el efecto de ciertos hitos sociales en el cine, y ver si generan vocabulario nuevo o el crecimiento de algunas palabras, particularmente porque aquello que va a estar más presente en esta base de datos es el cine más "mainstream", aquel distribuido por compañías grandes en circuitos de cine comercial, y no tanto cine independiente, que no suele tener el mismo alcance masivo.

[6] https://books.google.com/ngrams
[7] https://books.google.com/ngrams

Para generar esta medición de forma eficiente armamos el índice, ya mencionado en la sección anterior, con todo el vocabulario de cada año: si una palabra aparece en un subtítulo entonces, en el año de esa película, existe una entrada para esa palabra que tiene la cantidad total de apariciones. Separadamente, tenemos un índice con la cantidad total de palabras por cada año para generar la fórmula vista más arriba.

Como ayuda para visualizar, no necesariamente una palabra, sino un concepto más genérico, en los gráficos se utiliza la suma de más de una palabra sobre el total. De esta forma entendemos todos los pronombres femeninos, por ejemplo, como un sólo concepto de "referencia a mujer", y hacemos el cálculo como si fuese una sola palabra:

$$Frecuencia(mujer) = \frac{f(she) + f(her) + f(hers) + f(herself)}{n}$$

1.4.3 Positive Pointwise Mutual Information

La medida llamada Pointwise Mutual Information (PMI) fue utilizada por Church y Hanks (1990) y la denominaron "índice de asociación". La fórmula que ellos adaptaron de teoría de la información y que utilizamos en este trabajo (en la misma forma en que Jurafsky y Martin (2014) la utilizan para asociaciones en matrices palabra-palabra) es la siguiente:

$$I(x, y) = log_2(\frac{P(x, y)}{P(x)P(y)})$$

En las propias palabras de Church y Hanks, para dos palabras x e y:

> [...] mutual information compares the probability of observing x and y together (the joint probability) with the probabilities of observing x and y independently (chance). If there is a genuine association between x and y, then the joint probability $P(x, y)$ will be much larger than chance $P(x)P(y)$, and consequently $I(x, y) >> 0$. If there is no interesting relationship between x and y, then $P(x, y) \approx P(x)P(y)$, and thus, $I(x, y) \approx 0$. If x and y are in complementary distribution, then $P(x, y)$ will be much less than $P(x)P(y)$, forcing $I(x, y) << 0$.

Luego continúan con cómo se aplica esto a un corpus de palabras y la relación entre ellas, que es el mismo significado que tendrá en este trabajo.

Las probabilidades de cada palabra (o conjunto de palabras) independiente por año se estiman con:

$$P(x) = \frac{f(x)}{n}$$

Donde f es la cantidad de apariciones en el año correspondiente y n es la cantidad de palabras totales en ese año.

La probabilidad conjunta de dos palabras se calcula como:

$$P(x, y) = \frac{f_w(x, y)}{n}$$

Donde f_w es la cantidad de apariciones en un año específico de ambas palabras en una ventana de tamaño w. Para Church y Hanks (1990), y luego también para Jurafsky y Martin (2014), ese tamaño de ventana es un número fijo de palabras antes y después de una palabra. En este trabajo la ventana w va a ser medida en segundos antes y después del subtítulo que se está analizando.

Como lo dice el nombre dado por Church y Hanks (1990), el PMI es un índice de asociación, permite observar qué palabras están vinculadas a alguna en particular a partir de los contextos en los que se utilizan. Entendemos que si una palabra x se usa muy habitualmente en conjunto con otra palabra entonces los significados de estas dos palabras están asociados de alguna forma

(Church y Hanks, 1990; Hamilton *et al.*, 2016). Esta forma puede ser que refieren al mismo tema, que son opuestas, que una es un sustantivo y la otra es algún calificativo muy habitual para él, etc. Esta será la primera medida que nos va a permitir observar cuáles son los usos habituales de los conceptos que elegimos estudiar.

El PMI es sensible al tamaño del vocabulario estudiado, y puede resultar en asociaciones que se entienden como fuertes por ser un número comparativamente alto (frente a otras asociaciones) como consecuencia de muy pocas apariciones conjuntas en las palabras estudiadas. Además, para casos en los que el vocabulario es escaso los valores negativos del PMI son poco confiables (Jurafsky y Martin, 2014). Por esta razón, utilizaremos el **Positive PMI** (PPMI) lo que implica que si un PMI devuelve un valor negativo será reemplazado por cero, y resuelve el problema de cómo visualizar los resultados de $-\infty$ obtenidos cuando no hay ninguna aparición conjunta (Jurafsky y Martin, 2014).

En este trabajo, utilizamos la fórmula dada por Jurafsky y Martin (2014) para matrices palabra-contexto, donde se busca el PPMI entre la palabra en la posición i de la matriz y la palabra contexto en la columna j:

$$PPMI_{ij} = max(log_2 \frac{P_{ij}}{P_{i*}P_{*j}}, 0)$$

La probabilidad de aparición conjunta (P_{ij}) es la cantidad de apariciones conjuntas sobre la suma de todos los contextos del año:

$$P_{ij} = \frac{f_{ij}}{\sum_{i=1}^{W} \sum_{j=1}^{C} f_{ij}}$$

Y por último, las probabilidades independientes (P_{i*} y P_{*j} son la suma de todos los contextos donde se encuentra esa palabra, sobre la suma de todos los contextos del año:

$$P_{i*} = \frac{\sum_{j=1}^{C} f_{ij}}{\sum_{i=1}^{W} \sum_{j=1}^{C} f_{ij}}$$

$$P_{*j} = \frac{\sum_{i=1}^{W} f_{ij}}{\sum_{i=1}^{W} \sum_{j=1}^{C} f_{ij}}$$

En la figura 1.5 vemos cómo se obtienen todos los valores descriptos a partir de las matrices palabra-contexto.

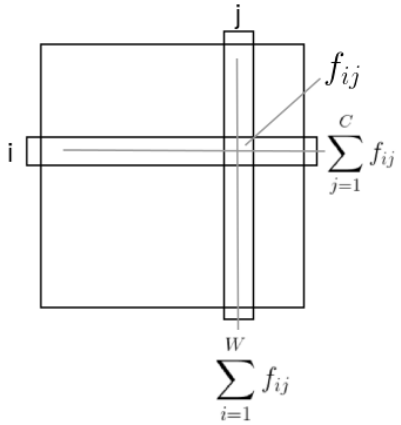


Figura 1.5: Esquema de obtención de valores a partir de matrices palabra-palabra

1.4.4 Similaridad word2vec

Word2vec es un modelo que utiliza redes neuronales para generar un espacio de dimensionalidad reducida donde cada palabra tiene asignado un vector, y estos vectores están cercanos en el espacio siempre y cuando las palabras que representan tengan contextos similares (Mikolov *et al.*, 2013; Kulkarni *et al.*, 2015; Hamilton *et al.*, 2016). Para generar este espacio se debe entrenar el modelo utilizando un corpus, preferentemente de millones de palabras; a mayor información, mejor será el modelo. La descripción de cómo se realiza este entrenamiento se encuentra en la sección 1.2.2

En este trabajo se utilizará la representación vectorial de Mikolov *et al.* (2013) con Word2vec entrenado en el corpus de GoogleNews del año 2014 que tiene 3000 millones de palabras, y cuyos vectores por palabra están reducidos a 300 dimensiones. Este corpus se utiliza habitualmente en investigaciones académicas por tener gran cantidad de palabras, y porque los vectores son directamente descargables sin necesidad de realizar el entrenamiento, lo cual es muy costoso con un corpus tan grande (van Erp y Vossen, 2016; Ouyang *et al.*, 2015).

Para armar nuestra métrica buscaremos todos los contextos de las palabras que queremos testear (las llamaremos palabras de test), luego buscaremos el vector que las representa en word2vec y compararemos este vector con el vector de la palabra objetivo (aquella con la cuál queremos averiguar la similaridad). Una vez que tengamos la similaridad entre todos los contextos y la palabra objetivo, utilizaremos como métrica qué proporción de palabras en el contexto está por arriba de un cierto límite o threshold de cercanía.

El método exacto de cómo se obtiene la similaridad entre ambos conceptos está ilustrado con el pseudo código de la figura 1.6.

El objetivo de esta medida será detectar cuánto de la idea de la palabra objetivo es cuantificable en los contextos de las palabras de test. Y para eso el método es resaltar aquellas palabras que son similares a la idea buscada (Diuk *et al.*, 2012): dentro de los contextos de todas las películas de un año hay mucha variabilidad y, por lo tanto, esta medida busca eliminar el ruido que generan tantas palabras rescatando la proporción que están cerca del concepto buscado.

Con respecto al límite numérico en sí, utilizamos los datos de wordsim353[8] (un dataset que tiene pares de palabras junto con una similaridad promedio a partir de respuestas humanas) para evaluar las similaridades de word2vec. Primero decidimos que nos interesaba quedarnos con un límite que nos acercara a 5.5 (sobre 10) de cercanía humana y dejamos afuera aquellas que no entraran. Luego comparamos las similaridades humanas con las dadas por word2vec para las mismas palabras, aún con el corte a partir de 5.5, el dataset seguía teniendo pares de palabras con similaridad en word2vec muy baja. Elegimos quedarnos con el 80 % más cercano en word2vec de las palabras en ese conjunto, y eso puso el threshold en 0.193171. Esto quiere decir que cuando utilizamos esta medida, tomamos la proporción de palabras tales que la similaridad coseno con el objetivo es mayor o igual a 0.193171.

[8] http://alfonseca.org/eng/research/wordsim353.html

```
vec_objetivo = vector_word2vec("palabra objetivo")
contextos = contextos_para("palabra de test")
// contextos = ["contexto1", "contexto2", ...]
vectores_contexto = [normalizar(vector_word2vec(palabra))
        for palabra in contexto]
similaridades_parciales = [similaridad_coseno(vec_objetivo, vector)
        for vector in vectores_contexto]
similaridades_parciales = [simil if simil >= 0.193171
        for simil in similaridades_parciales]
similaridad_final = longitud(similaridades_parciales) / longitud(contextos)
```

Figura 1.6: Pseudo código para obtener la Similaridad word2vec entre una palabra test y una objetivo

1.4.5 Relación de word2vec con PPMI

Como el PPMI, la similaridad word2vec permite juzgar qué palabras coocurren con el concepto estudiado. El beneficio de este método frente a PPMI pasa por la posibilidad de word2vec de capturar conceptos similares aún cuando no haya coocurrencia en el corpus. Por ejemplo, buscando la cercanía entre tecnología y algo más, si entre los contextos aparece la palabra computadora, ésta va a estar muy cerca del objetivo, y por lo tanto sumar una palabra a la proporción de la similaridad word2vec.

La principal desventaja de este método pasa porque la semántica de las palabras en word2vec depende del entrenamiento que se le haya dado, por lo que el significado de los vectores está atado a los contextos en los que se usa la palabra en el año 2014, y en el contexto de GoogleNews, que tiene sesgos y criterios externos a nuestro corpus.

1.4.6 Robustez de asociación

Esta medida la pensamos como complemento al PPMI. Ya hemos visto que éste puede ser muy sensible a la cantidad de apariciones del vocabulario buscado en el corpus y planteamos la Robustez de asociación para definir si las asociaciones son consistentemente fuertes o si está pesando en exceso la escasez de una palabra en particular.

La robustez de asociación la definimos según el Test Exacto de Fisher (Fisher, 1922) entre las apariciones de la palabra objetivo dentro y fuera del contexto de la palabra a analizar. El test de Fisher para cada año será como sigue: asumiendo que la probabilidad de que una palabra específica esté entre el contexto de otra es similar a la probabilidad de aparición en el corpus en general, ¿cuán probable son la cantidad de apariciones conjuntas efectivamente medidas en el corpus?

Ejemplificamos utilizando el diagrama de la figura 1.7: por año, el Test de Fisher responde a la pregunta ¿qué probabilidad hay de que existan k apariciones conjuntas dado que existen n palabras en los contextos de la palabra de test, Q ocurrencias de la palabra objetivo en los contextos de todo un año, y el total de palabras que aparecen en todos los contextos del corpus para ese año es de N?

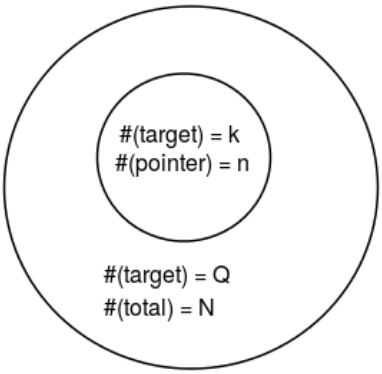


Figura 1.7: Ejemplo conceptual de qué valores se utilizan para calcular la Robustez de asociación

La significancia estadística es medible para cada año (el resultado del test es un p-valor, cuánto más bajo sea este más significativo es el resultado para ese año), pero no es lo que vamos a observar, lo que interesa en este contexto es si la tendencia de la medida a lo largo de los años es consistente. Para eso observamos $1 - p - valor$ a lo largo de los años, si tiene una tendencia consistente por encima de 0.95, entonces juzgaremos que ese resultado es muy significativo.

Cuando observemos que la Robustez de Asociación varía entre márgenes muy amplios, o se encuentra consistentemente alrededor del 0.5, entonces el resultado será poco significativo.

Capítulo 2

Caso de estudio: Rol de la mujer

En la última década se han hecho muchos análisis sobre la representación de la mujer como secundaria al hombre. Ha habido estudios con respecto a las asociaciones que conllevan los pronombres femeninos y la palabra "mujer" en los medios (Sendén *et al.*, 2015), en lecturas genéricas escolares (Lenton *et al.*, 2009), en el cine (Ramakrishna *et al.*, 2015), en el imaginario colectivo (Cejka y Eagly, 1999), en la literatura (Twenge *et al.*, 2012), y probablemente en muchos otros ámbitos. A continuación vamos a presentar análisis sobre la frecuencia de la palabra "mujer", "hombre" y de los pronombres de ambos géneros, las características de personalidad estereotípicamente femeninas y masculinas, cuáles son las profesiones asociadas a cada género, y la visión cinematográfica de profesiones específicas cuya distribución de género cambió en el mercado laboral estadounidense a lo largo de los años.

2.1 Cantidad de referencias a mujeres en el cine

2.1.1 Introducción

En toda la bibliografía analizada se visualiza mayor cantidad de menciones a pronombres masculinos que femeninos, con un margen muy amplio, yendo desde 2 (Twenge *et al.*, 2012) hasta 9 (Sendén *et al.*, 2015) veces más pronombres referentes a hombres que a mujeres.

En este trabajo vamos a investigar esto comparando las menciones a pronombres masculinos y femeninos, y las menciones a "hombre" y "mujer" dentro del corpus, y examinando la frecuencia de estos términos por año.

Twenge *et al.* (2012) analizan la proporción de pronombres masculinos a femeninos en literatura estadounidense a lo largo de los años, y muestran cómo existen variaciones según el período y el status de la mujer en cada época. En su análisis Twenge *et al.* (2012) encontraron 3 períodos históricos diferentes: hasta 1945 una relación relativamente estable de 3.5 pronombres masculinos por cada femenino; a partir de 1945 y hasta 1967, en época de posguerra, se evidencia un giro hacia los personajes masculinos: hasta 4.5 por cada femenino. Luego, a partir de 1968, detectaron un descenso sostenido hasta la actualidad donde se estabiliza alrededor de 2 pronombres masculinos por cada uno femenino.

2.1.2 Hipótesis

Nuestras hipótesis son:

- Siempre se utilizan más pronombres masculinos que femeninos

- Llegando a la actualidad la diferencia entre ambos se reduce, pero se mantiene la mayoría masculina (a partir de los 70)
- La mayor diferencia entre ambos será alrededor de la época de posguerra (entre 1945 y 1970)

Para todas las comparaciones vamos a usar, en primer lugar, los pronombres de cada género ("she", "her", "hers", "herself", y "he", "him", "his", "himself") (los utilizados en Twenge *et al.* (2012)) dado que son de las palabras más habituales del corpus, y por lo tanto, tienen contextos de gran cantidad de palabras lo que los hace muy ricos para analizar.

2.1.3 Resultados y discusión

Tomamos la frecuencia por año de todos los pronombres por género, y la frecuencia de "hombre" y "mujer" separadamente. En la figura 2.1 se ve que el número de pronombres masculinos supera consistentemente a los pronombres femeninos, con una tendencia a reducirse, mientras que los femeninos están relativamente estables. Los usos de "hombre" y "mujer" se mantienen estables a lo largo de los años, con "hombre" superando a "mujer" consistentemente en frecuencia de aparición. En principio, esta primera visualización aporta evidencia a la hipótesis de que los pronombres masculinos superan consistentemente a los femeninos en frecuencia, y se nota una ligera tendencia a la reducción de la diferencia con el avance del tiempo, notablemente más ligera que en la bibliografía vista.

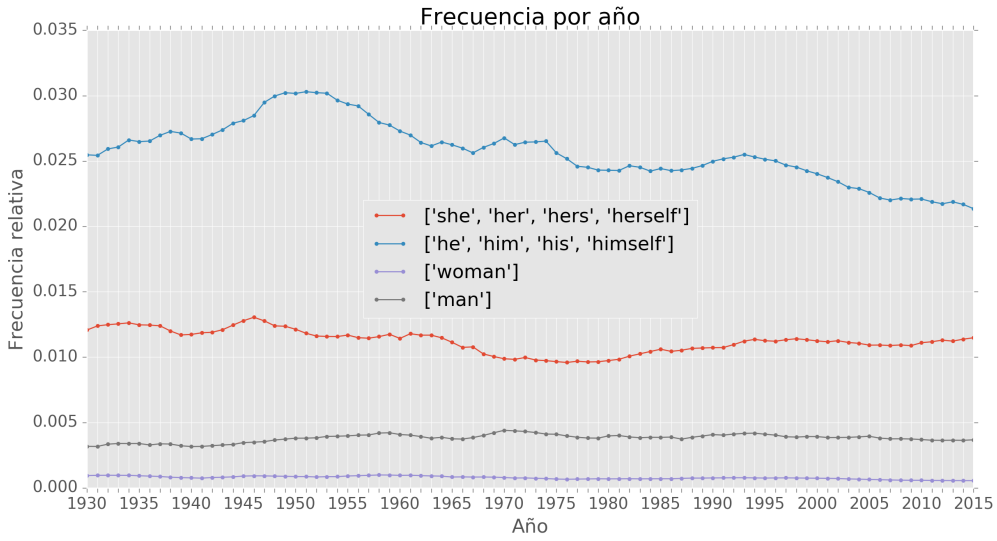


Figura 2.1: Frecuencia relativa de los pronombres de cada género y hombre y mujer (smoothing 3)

Para comparar fácilmente si hubo algún cambio a través de los años en la proporción de pronombres masculinos contra femeninos, la figura 2.2 muestra proporción entre ambos. Los números no son tan extremos como los que se veían en Sendén *et al.* (2015), pero sí están casi completamente sobre 2. Esto quiere decir que por cada referencia a un personaje femenino, hay entre 2 y 2.8 a un personaje masculino. Con esta métrica no podemos extraer el número de personajes en sí, o su importancia, sólo que, al dialogar, los personajes suelen hacer referencia a hombres el doble que a mujeres. Este resultado coincide con la investigación recientemente publicada por Google[1] que

[1] https://www.google.com/intl/en/about/main/gender-equality-films/

extrajo que personajes hombres aparecen en pantalla el doble de tiempo que personajes mujeres y tienen el doble de diálogo.

Con respecto a los períodos resaltados, las etapas no quedan tan claramente evidenciadas como hemos visto en la investigación de Twenge *et al.* (2012), lo que es parcialmente consistente con la hipótesis planteada. No existe un pico en época de posguerra, como habíamos previsto, pero sí se nota el comienzo de un descenso más pronunciado a partir de los 70.

Por otro lado, dado que los pronombres no pasan nunca de 3 masculinos por cada 1 femenino, ni la escalada de la diferencia, ni el descenso son tan pronunciados como lo vimos en la bibliografía estudiada.

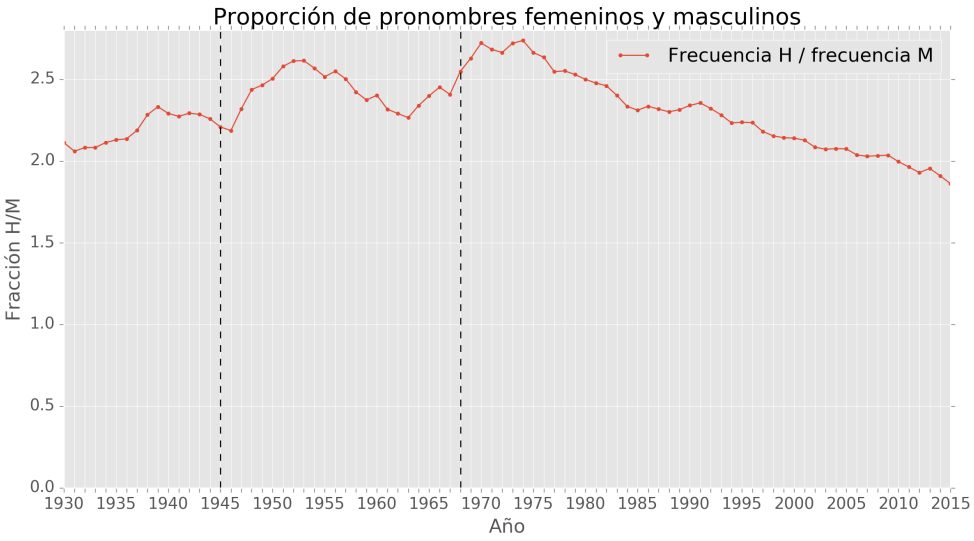


Figura 2.2: Fracción pronombres masculinos / femeninos (smoothing 3)

2.2 Personalidad femenina

2.2.1 Introducción

Como segunda dimensión, en aquellos estudios donde se profundiza sobre la representación estereotípica de las mujeres se suele hacer referencia, por un lado, a las cualidades personales que se les asignan, y por otro, a los roles laborales o profesiones que tienen o se espera que tengan (Lenton *et al.*, 2009; Ramakrishna *et al.*, 2015; Cejka y Eagly, 1999; Sendén *et al.*, 2015). Por comparación al caso estudiado, en la mayoría de estos estudios entra en juego el estereotipo masculino, que tiene un rol igualmente importante en el estudio de representación de género.

Para tratar estereotipos de atributos personales vamos a utilizar el vocabulario proveniente del Inventario de Roles de Género de Bem (Bem, 1979) (BSRI). Dejamos afuera los adjetivos "masculine" y "feminine", y aquellos conceptos expresados en muchas palabras que no pudimos concentrar en un único sinónimo, y reemplazamos los que sí pudimos por una única palabra. Este listado cuenta con palabras asignadas como femeninas, como masculinas y como neutrales, y se encuentra completo en el apéndice A.1. En líneas generales, las palabras identificadas como femeninas tienen que ver con ser afectuosa, maleable y sensible, mientras que las masculinas se relacionan con ser fuerte, decidido y competitivo. Una muestra de las neutrales es "feliz", "confiable", "convencional"

e “ineficiente”.

En Lenton *et al.* (2009) utilizan un vocabulario similar y encuentran asociaciones entre pronombres y atributos femeninos, y relaciones bajas entre pronombres femeninos y atributos masculinos. Los pronombres masculinos resultan en los mismos resultados respectivos, pero los atributos neutros Lenton *et al* los encuentran levemente más asociadas a los pronombres masculinos que femeninos.

2.2.2 Hipótesis

En función de lo estudiado en otra bibliografía vamos a investigar si:

- Los atributos marcados como femeninos tienen una asociación alta con los pronombres femeninos, y el equivalente para los masculinos
- La asociación entre los pronombres masculinos y los atributos femeninos es baja, y lo mismo para el par inverso
- Los atributos neutrales están levemente asociados a pronombres masculinos
- Y por último, viendo los resultados de Twenge *et al.* (2012), si las asociaciones estereotípicas se vuelvan más leves en películas más actuales, y las inversas crecen

2.2.3 Resultados y discusión

Utilizando el PPMI como primera medida de acercamiento (figura 2.3) vemos que la asociación femenina con los atributos estereotípicos es consistente a lo largo de los años. El número que vemos en el eje y es bajo (particularmente más bajo que en capítulos siguientes), pero hemos visto que esta medida es sensible a la cantidad de palabras, y este vocabulario es un porcentaje importante del corpus.

Es perceptible que llegando a la actualidad pasa a ver una asociación levemente más alta entre los pronombres masculinos y estos atributos femeninos, lo que podría indicar una evolución en la variabilidad dentro de los personajes masculinos.

A su vez, podemos ver que se reproducen parcialmente las etapas que analizamos en la sección anterior dadas por el trabajo de Twenge *et al.* (2012): en la época posterior a la Segunda Guerra Mundial, que coincide con una disminución en pronombres femeninos en relación a masculinos, se evidencia un aumento de asociaciones estereotipadas.

Cuando trabajamos con PPMI (o luego, con word2vec) hablamos de los contextos en los que se utilizan estas palabras. Como ya dijimos, estos contextos son aquellas palabras que entran dentro de una ventana de tiempo: en los gráficos dentro de este informe usamos siempre 10 segundos. Pero esta medida es arbitraria a partir de la observación de algunos subtítulos y gráficos preliminares. Con el objetivo de verificar que ese tamaño era apropiado para mostrar el tipo de relaciones que buscábamos, hicimos pruebas con ventanas de 3, 5, 10 y 20 segundos, y los gráficos para el PPMI de los atributos femeninos y los pronombres de ambos géneros se encuentran en el apéndice B para los cuatro tamaños de ventana.

En líneas generales las ventanas más chicas tienden a mostrar expresiones y frases hechas, mientras que las ventanas más grandes permiten percibir similaridades semánticas a mayor escala (Church y Hanks, 1990; Jurafsky y Martin, 2014). Como la cantidad de frames en una ventana de 10 segundos es variable, y el texto de cada frame también, la cantidad de palabras en una ventana varía en cada caso.

En la figura 2.4 se ve cuán fuerte es el vínculo entre los atributos masculinos y los pronombres de cada género. La asociación estereotípica es un poco más leve que la femenina y los cambios son más pronunciados. Llegando a la actualidad la asociación es bastante más baja, pero sorprendentemente la asociación femenina con este vocabulario también es más baja, llegando a casi cero. Ese resultado contradice nuestra hipótesis de que los las asociaciones iban a tender a igualarse entre pronombres

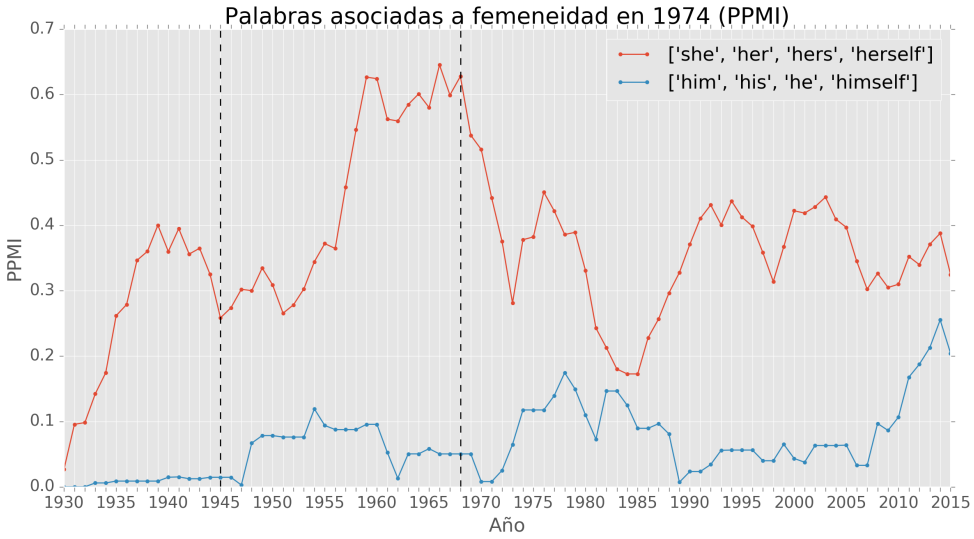


Figura 2.3: PPMI del vocabulario proveniente del BSRI referido a femineidad en contexto de pronombres femeninos y masculinos (smoothing 3)

masculinos y femeninos. Pero combinando este resultado con el anterior, donde los atributos femeninos y los pronombres masculinos tienen mayor asociación, se podría interpretar una tendencia hacia "suavizar" a todos los personajes, tanto femeninos como masculinos. Es decir, más personajes descriptos como afectuosos o sensibles y menos como agresivos o líderes, independientemente de su género.

En las figuras 2.5a y 2.5b se visualiza la Robustez de asociación para los atributos masculinos con smoothing de 3 años y sin ningún smoothing. Se visualiza que, en promedio, la asociación de los pronombres masculinos con los atributos identificados como masculinos se encuentra en el extremo de "más apariciones conjuntas que lo probable por azar", mientras que los pronombres femeninos se encuentran en "menos apariciones conjuntas que lo probable por azar", dado que los pronombres femeninos están mayormente debajo del 0.5 y los masculinos por encima.

Pero se incluye la figura 2.5b para visualizar que este caso del test tiene mucha variabilidad y, excepto en los últimos años donde el alto volumen de películas tiende a marcar una tendencia más firme, es dependiente de las películas que tiene el corpus para ese año. Este caso sirve de ejemplo del uso que tiene esta herramienta: si bien la asociación parecía muy fuerte con los datos de PPMI, este método nos ayuda a juzgar cuán robusta es esa asociación, y lo que vemos es que la asociación existe desde 1930, pero sólo es muy robusta a partir del año 2000 en adelante.

Con respecto a los atributos neutrales, en Lenton *et al.* (2009) habíamos visto que la asociación entre éstos y los pronombres masculinos era más pronunciada que con pronombres femeninos. En este corpus no encontramos apoyo a esta hipótesis: en la figura 2.6 vemos una asociación fuerte con los pronombres femeninos. Lenton concluye de su investigación que el concepto de hombre es más reducido que el de mujer, y propone que eso sea debido a esfuerzos por ampliar la representación, más allá de los motivos, el hecho de que los pronombres femeninos tengan asociaciones fuertes con mayor variedad de atributos puede ser parte del mismo fenómeno.

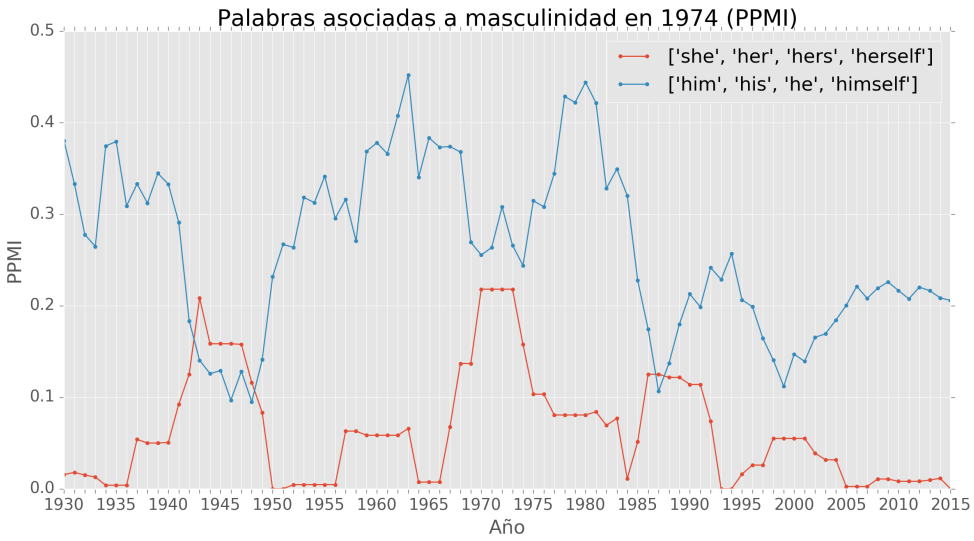


Figura 2.4: PPMI del vocabulario proveniente del BSRI referido a masculinidad en contexto de pronombres femeninos y masculinos (smoothing 3)

2.3 Profesiones y roles estereotípicos

2.3.1 Introducción

Con respecto a las profesiones, los estereotipos se alinean bastante con los atributos. Esto se ve en Cejka y Eagly (1999) ya que en las profesiones juzgadas como "femeninas" se valoran más los atributos habitualmente asignados a mujeres, y en las profesiones "masculinas" aquellos socialmente asignados a hombres. Si se considera que una profesión es más femenina, se asume que ciertos atributos son necesarios para ejercerla, con lo cual se persiste el estereotipo anterior.

Las profesiones o roles elegidos para esta sección son aquellos utilizados en Lenton *et al.* (2009) para evaluar el grado de estereotipos en lectura estándar de escolarización. Ejemplos de roles femeninos son "model", "housekeeper" y "nanny", masculinos: "architect", "carpenter", y "sheriff", neutros: "cashier", "servant", y "doctor". La lista completa se encuentra en el apéndice A.2. En líneas generales, en la lista hay profesiones que tienen mayoría estadística del género asociado, y hay otros roles que tienen que ver con los atributos ya mencionados, como "caregiver" del lado de mujer. En general son todas profesiones muy arraigadas con un cierto tipo de personalidad: profesiones científicas asociadas a lo analítico son asignadas como masculinas, igual que aquellas con un costado de "agresividad" o físico (como deportista o soldado); y profesiones relacionadas con el cuidado asociadas a los atributos de "sensibilidad" designados como femeninos.

2.3.2 Hipótesis

A partir de lo visto en Lenton *et al.* (2009) y en el caso anterior, creemos que:

- Los roles asignados a masculinidad tendrán mayor asociación con pronombres masculinos, y lo mismo para pronombres femeninos
- Los roles neutros tendrán asociaciones de nivel similar entre ambos pronombres

2.3.3 Resultados y discusión

Comenzando con los roles femeninos, la figura 2.7 muestra una asociación entre pronombres femeninos y roles femeninos más fuerte que la vista para atributos en la sección anterior.

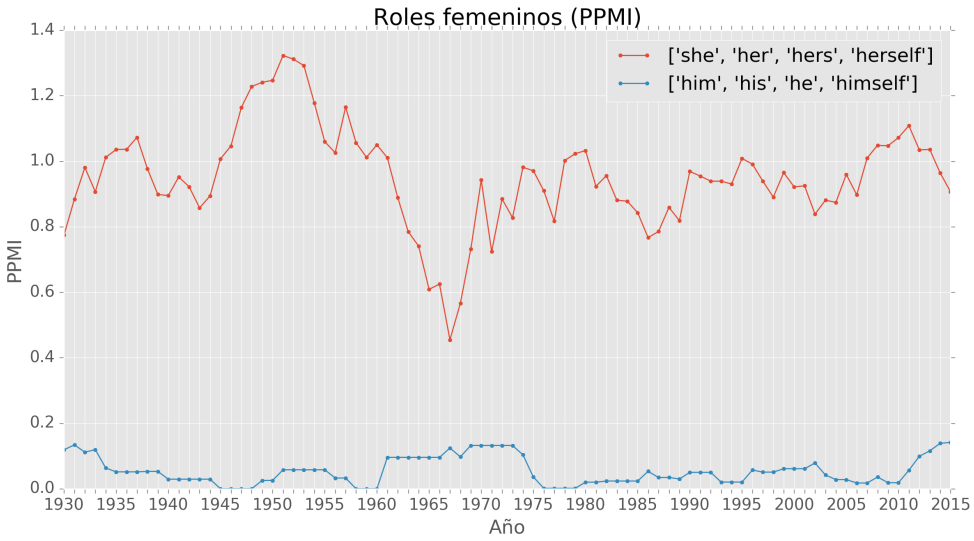


Figura 2.7: PPMI de roles utilizado por Lenton *et al.* (2009) definidos como femeninos en contexto de pronombres masculinos y femeninos (smoothing 3)

Se puede ver en esta figura que si bien hay altos y bajos, el nivel de asociación de la mujer con los roles estereotípicos no disminuye con los años. Es decir, los roles tradicionales a los que se asocia a la mujer siguen estando presentes en el cine, y si bien esto favorece a la existencia de un estereotipo, lo que más pesa es que no hay crecimiento significativo en la asociación de los pronombres masculinos con estos roles. No sería llamativo que estos roles sigan estando presentes en el cine, si no que los únicos personajes asociados a ellos sean mujeres.

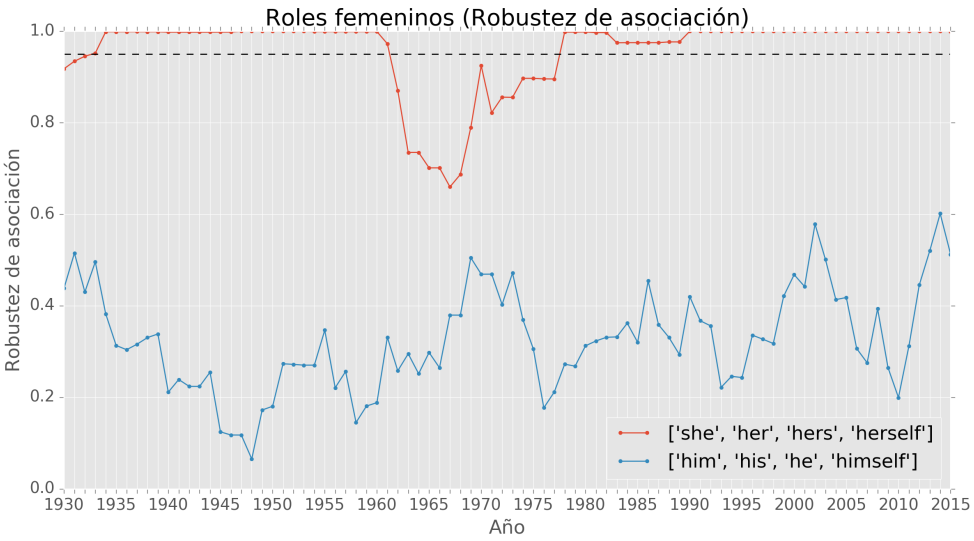


Figura 2.8: Robustez de asociación para PPMI de roles femeninos y pronombres de ambos géneros (figura 2.7) (smoothing 3)

Es interesante observar que hay una disminución en la asociación en un rango de años similar al que vimos en la sección anterior, que coincidía con un aumento de asociación con los atributos femeninos. No parece haber una causa inmediata visible, una forma de dilucidar por qué, sería separar las profesiones y roles y mirar los detalles de cada una. Es posible que si la representación de la mujer estuviese más limitada, se viese menos personajes en roles profesionales, por estereotípicos que fueran, y por eso decae esta asociación.

Al observar la robustez de esta asociación en la figura 2.8 vemos que la asociación entre roles femeninos y pronombres femeninos es muy fuerte. Por otro lado, la escasa asociación vista con los pronombres masculinos es poco robusta. Es decir, no es muy significativa en su asociación ni en la inversa.

En segundo lugar, observamos el PPMI entre los roles masculinos y los pronombres de ambos géneros en la figura 2.9. El grado de asociación es bastante fuerte (aunque menor al de los roles femeninos versus los pronombres femeninos), y se mantiene a lo largo de los años, incluso aumentando hacia la actualidad. Y no existe asociación alguna entre pronombres femeninos y roles masculinos, lo que apoya las hipótesis planteadas (cada rol asociado a su género estereotípico).

También hemos visto en resultados anteriores que actualmente los pronombres masculinos parecen tener mayor asociación con atributos o roles entendidos como femeninos, pero según estos resultados, las mujeres no han podido ganar tanto terreno en roles asociados a la masculinidad. Esto parecería contradecir en el cine aquello que dice en Lenton *et al.* (2009) respecto del idioma inglés estadounidense, y que ya mencionamos anteriormente, respecto al concepto de masculinidad como más limitado que el de femineidad.

Observando la figura 2.10 vemos que la asociación entre pronombres masculinos y roles masculinos es robusta a lo largo de los años. Pero también vemos algo que mirando el PPMI no se puede notar: la asociación entre pronombres femeninos y roles masculinos es robusta en su negatividad. Es improbable que las apariciones conjuntas de estos pronombres y los roles sean tanto menores que lo esperado por azar, y tan consistentemente, sin ningún motivo; es decir que se trata de una asociación inversa. Este tipo de asociación no lo habíamos visto con ninguno de los ejemplos anteriores y es evidencia de que hay pocas menciones a mujeres en contexto de roles tradicionalmente

masculinos. Vamos a continuar esta línea de búsqueda en la próxima sección.

Por último, investigamos los roles considerados neutrales en relación a ambos géneros. En la figura 2.11 se observa que en el cine los roles presentados parecen ser neutrales. Hay asociación de éstos con ambos roles y a grados similares, aunque con más variabilidad en la asociación a los pronombres femeninos. Aunque como se ve en la figura 2.12 el PPMI para estos roles es menos robusto en ambos géneros que los anteriores, hasta los últimos años donde la cantidad de material lo hace más robusto. A partir de 1982 para el caso masculino y de 1995 para el femenino las medidas son muy robustas. Este resultado aporta evidencia a la hipótesis de que los roles neutrales están asociados a los pronombres de ambos géneros.

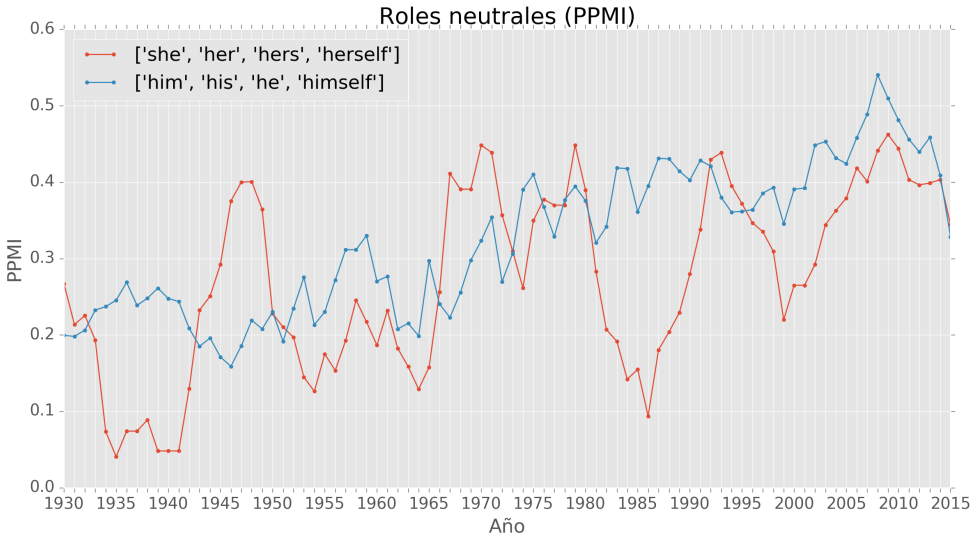


Figura 2.11: PPMI de roles utilizado por Lenton *et al.* (2009) definidos como neutrales en contexto de pronombres femeninos y masculinos (smoothing 3)

En el caso de los hombres, se ve que en esos años el promedio es ascendente, puede que se hable más de las profesiones de los personajes. En el caso femenino alrededor del año '99 hay un descenso para el que no tenemos una explicación inmediata. La tendencia general también es ascendente, pero como previo a los 80 y 90 la asociación no es muy robusta es difícil decir si las fluctuaciones son debido a cambios en el contexto o a que hay pocas palabras relacionadas con estos roles.

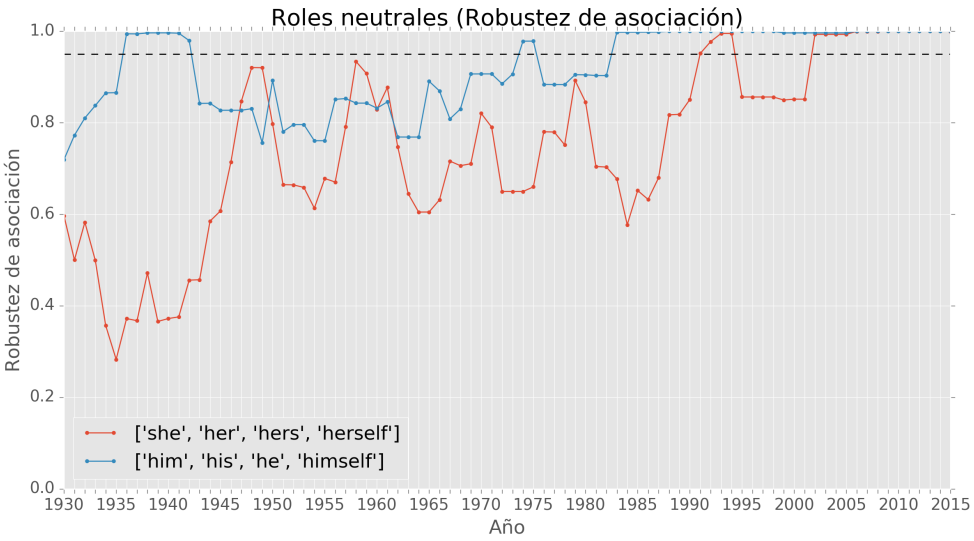


Figura 2.12: Robustez de asociación para PPMI de roles neutrales y pronombres de ambos géneros (figura 2.10) (smoothing 3)

En líneas generales, los roles y profesiones elegidos reflejan los estereotipos de género de forma muy marcada. A duras penas se ven referencias a hombres en roles entendidos como femeninos, y en el caso inverso son casi inexistentes. Como veremos en la próxima sección esto no refleja el estado actual del mercado laboral estadounidense, y es evidencia de que las producciones en Hollywood insisten en mantener los modelos y formas que se ven en el cine hace 50 años o más.

2.4 Evolución de profesiones con cambios en su representación real

2.4.1 Introducción

Para continuar con la idea de la sección anterior buscamos algunas profesiones que hayan tenido cambios significativos a lo largo de los años y otras que hayan mantenido sus proporciones hombre-mujer. Buscamos así enfocarnos un poco más a fondo que utilizando una lista con múltiples profesiones y roles distintos, y dónde sabemos cuál es la situación real hoy. A partir de información del Departamento de Trabajo[2] y del Foro Económico Mundial[3] elegimos una profesión tradicionalmente masculina que se haya mantenido mayoritariamente masculina: "engineer", una profesión tradicionalmente masculina que ahora tiene más mujeres que hombres: "accountant", una profesión tradicionalmente femenina que se mantuvo: "nurse", una profesión tradicionalmente femenina que ahora tiene mayoría masculina: "chef" y una que pasó de tener algo de mayoría masculina a tener prácticamente 50 % de cada género: "realtor" (real estate sales). Notar que ninguna de estas palabras se ven afectadas por el género como podría suceder en español al decir "ingeniero", o en inglés con "waitress", que de por sí ya hacen referencia al género de la persona.

[2]https://www.bls.gov/cps/cps_aa1995_1999.htm
[3]https://www.weforum.org/agenda/2016/03/a-visual-history-of-gender-and-employment

2.4.2 Hipótesis

Dados los resultados de las secciones anteriores, creemos que:

- Las profesiones que se mantienen mayoritariamente de un género se verán más asociadas con ese, y poco o nada con el otro (engineer, nurse)
- Las profesiones que han tenido cambio en su composición hayan aumentado en su asociación con el género opuesto, pero considerando los resultados anteriores, no en la misma medida que la realidad (accountant, chef)
- La profesión con 50 % de cada género esté asociada a los dos grupos de pronombres (realtor)

2.4.3 Resultados y discusión

Los resultados de PPMI para engineer (figura 2.13) muestran que hay pocas ventanas en conjunto con los pronombres femeninos. Las formas de meseta que se observan hasta el año 1970 se deben a un año en el que (posiblemente por una única película) hay un PPMI un poco más alto y el smoothing lo vuelve un arco de 7 años. En cambio, las apariciones conjuntas entre engineer y los pronombres masculinos están divididas en mayor cantidad de años y de películas. La asociación crece y decrece a lo largo de los años, pero es consistentemente mayor a la femenina, y no hay aumentos en los últimos años.

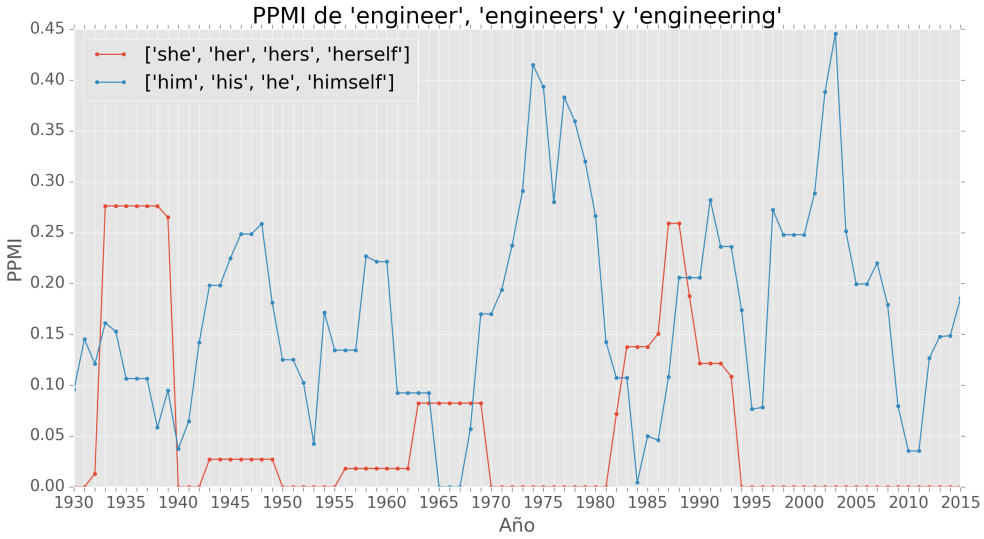


Figura 2.13: PPMI de los contextos de ingeniería y pronombres de ambos géneros (smoothing 3)

Cuando utilizamos word2vec para enfocarnos en profesiones específicas tenemos los problemas que ya mencionamos con respecto a esta medida: a pesar de que los contextos utilizados son con palabras de películas de ese año, los vectores resultantes y el vector contra el que estamos comparando tienen el significado actual de la palabra engineer, pero por otro lado puede ayudar a detectar asociaciones que no podemos ver en PPMI por la escasez de apariciones de este vocabulario tan específico.

Se puede tener una idea aproximada de la semántica de la palabra engineer para word2vec, para saber si es lo esperado y, por lo tanto, la comparación con los contextos tiene sentido, observando

cuáles son las palabras más cercanas en el espacio. Lo haremos para este caso a través de un sitio de prueba[4] por simplicidad.

Dentro de las 20 palabras más cercanas a "engineering" hay múltiples ejemplos relacionados con la aviación que se desvían de lo que estamos buscando, si miramos "engineer" hay referencias a arquitectura y construcción, pero en menor cantidad, y si miramos "engineers" casi no hay, únicamente una referencia a geología y dos relacionadas con la construcción. De esta forma parecería que la semántica más cercana a lo que estamos buscando la encontramos en "engineers".

Entonces, en la figura 2.14 vemos que los contextos de los pronombres masculinos en promedio están consistentemente más cerca de "engineers" que los contextos de los pronombres femeninos. Por lo tanto, ya sea porque el significado actual en los medios de ingeniero está semánticamente más cerca de asociaciones masculinas o porque en el cine cuando se habla de ingeniería se hace en contexto de personajes hombres, se ve a los pronombres masculinos consistentemente más cerca de ingeniería que los femeninos.

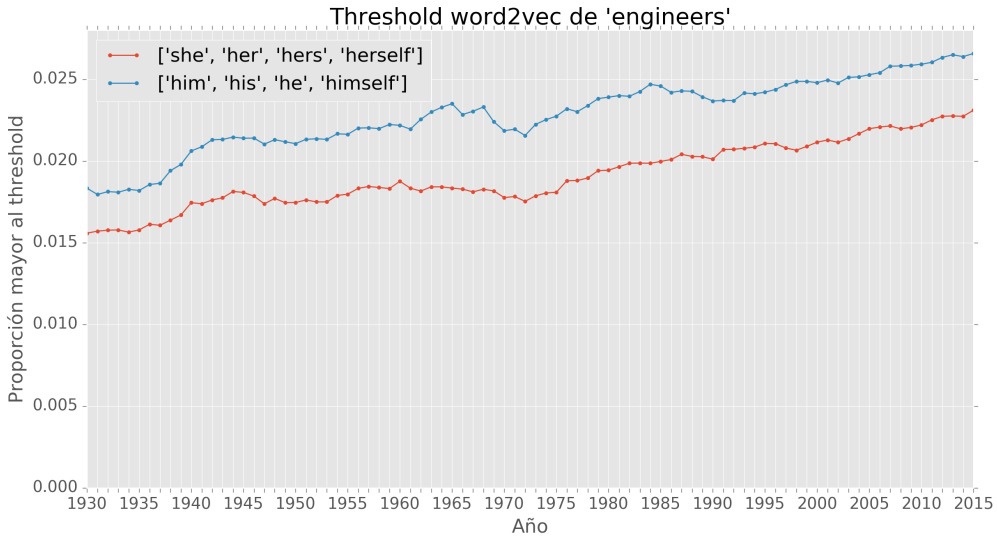


Figura 2.14: Similitud threshold word2vec del vector de ingeniería y los contextos de pronombres de ambos géneros (smoothing 3)

Pasando a accountant, esta es una profesión que era mayoritariamente masculina hasta los años 80 y que en la actualidad tiene una leve mayoría femenina. En la figura 2.15 se puede ver que la asociación a los pronombres masculinos no es tanto más marcada que a los pronombres femeninos como veíamos con la ingeniería, pero que la mayor parte del tiempo es más fuerte la representación masculina que femenina. La tendencia no es muy estable, con lo que es difícil decir si el aumento de asociación femenino y el descenso masculino es consistente o casual, pero es claro que ambos géneros son mencionados cuando se habla de contabilidad.

Este resultado se repite al mirar el resultado de word2vec para la profesión en la figura 2.16, los contextos de ambos géneros están más cerca que en el caso observado de ingeniería, pero consistentemente más cerca de pronombres masculinos que femeninos.

[4]http://bionlp-www.utu.fi/wv_demo/

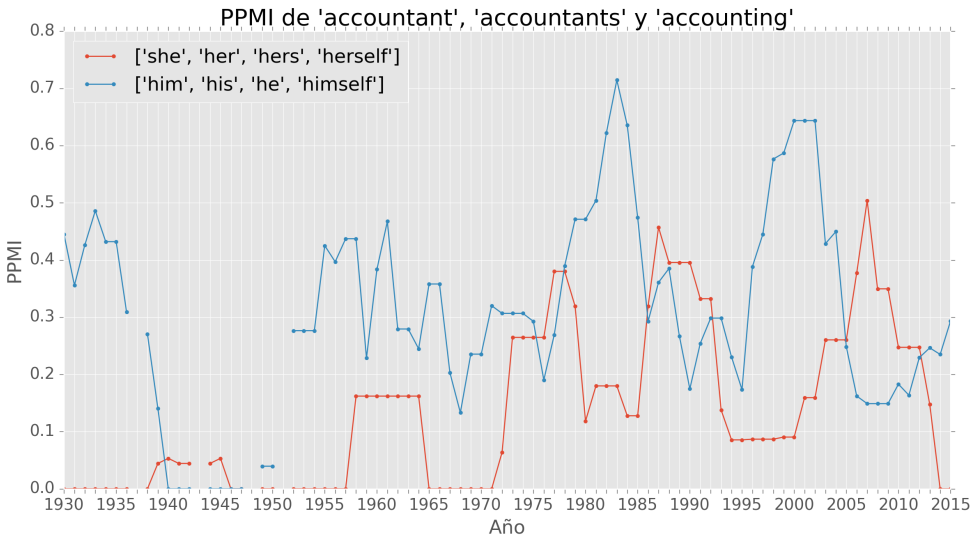


Figura 2.15: PPMI de los contextos de contabilidad y pronombres de ambos géneros (smoothing 3)

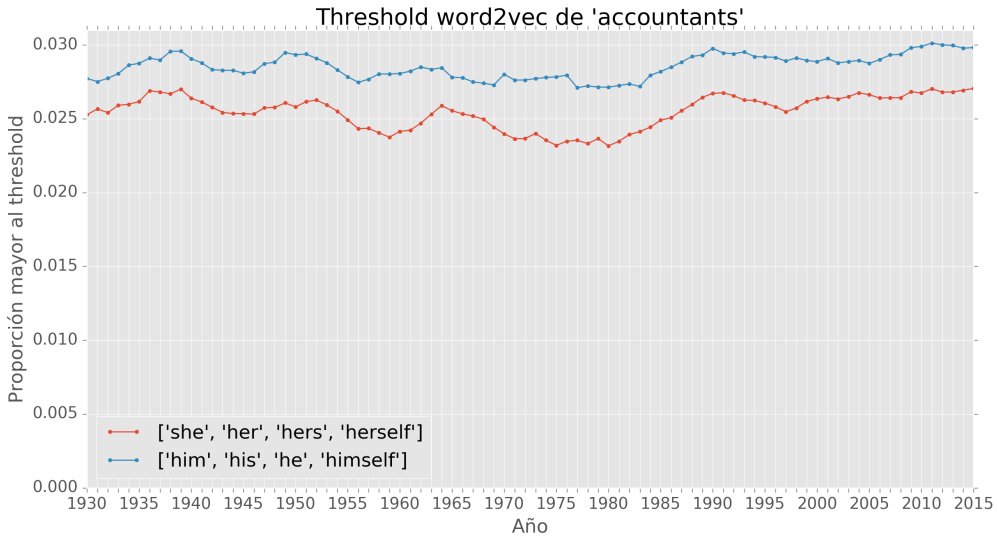


Figura 2.16: Similitud threshold word2vec del vector de contabilidad y los contextos de los pronombres de ambos géneros (smoothing 3)

Con el caso de chef sucede algo distinto en la bibliografía original: los censos solían no tener mucho detalle por profesión, pero con los años fueron agregando roles dentro de cada una. En un principio todo lo que fuera cocina entraba dentro de la categoría "cook", que luego se fue abriendo en "chef", "cook", "kitchen staff", etc. Dado que "cook" en inglés es también el verbo "cocinar" y un apellido, los contextos de la palabra "cook" van a ser tanto los de la profesión como en cualquier otra situación hogareña, como contextos no relacionados con lo laboral. Empezaremos investigando

chef, para evitar los contextos mixtos.

En la figura 2.17 observamos que el PPMI de chef varía entre ambos géneros a lo largo de los años. En la última década parece haber sido mayoritariamente masculino, pero la tendencia no es clara dada la variabilidad hasta ese momento. Esta cercanía entre los géneros y su grado de asociación con chef se mantiene en la figura 2.18 que muestra la cercanía word2vec: parece ser levemente más cercana al pronombre femenino, pero las distancias están muy próximas.

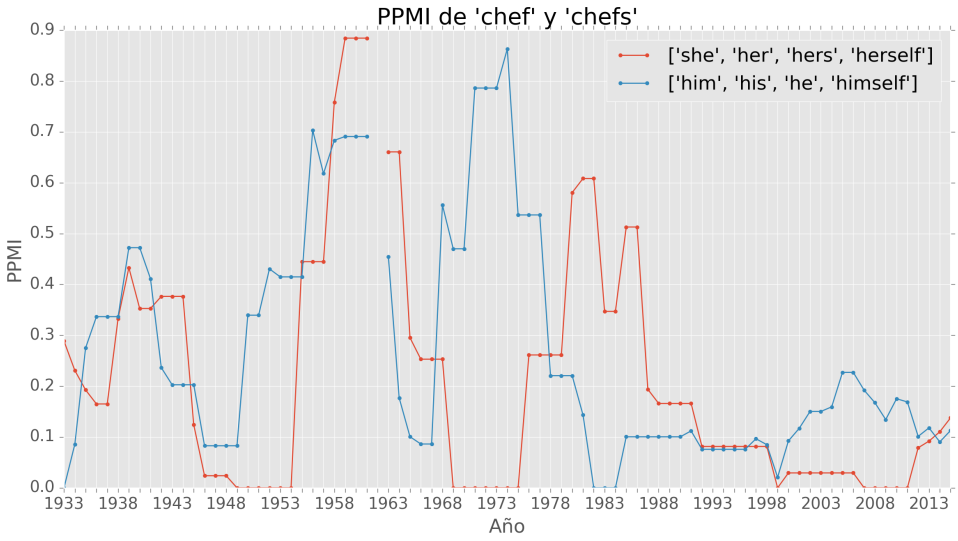


Figura 2.17: PPMI de chef y pronombres de ambos géneros (smoothing 3)

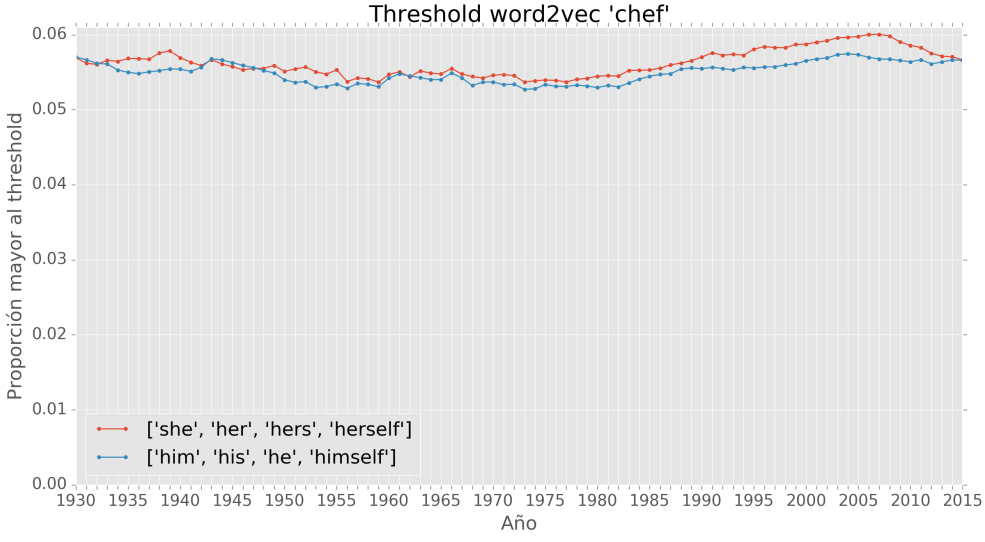


Figura 2.18: Similitud threshold word2vec del vector de chef y los contextos de los pronombres de ambos géneros (smoothing 3)

Por otro lado, si en lugar de chef observamos el comportamiento de cook (figuras 2.19 y 2.20) la asociación se inclina notablemente hacia los pronombres femeninos en ambas medidas. El PPMI da una asociación más consistente entre pronombres femeninos y cook, y los pronombres masculinos no tienen asociación y no hay cambios en la última década. Es posible que las referencias a la cocina sean mayoritariamente hogareñas por encima de laborales, y la asociación que estamos detectando es el cuidado del hogar asignado a las mujeres, más que la cocina profesional. Además, no hay motivo para pensar que el apellido esté más asociado a un pronombre que a otro.

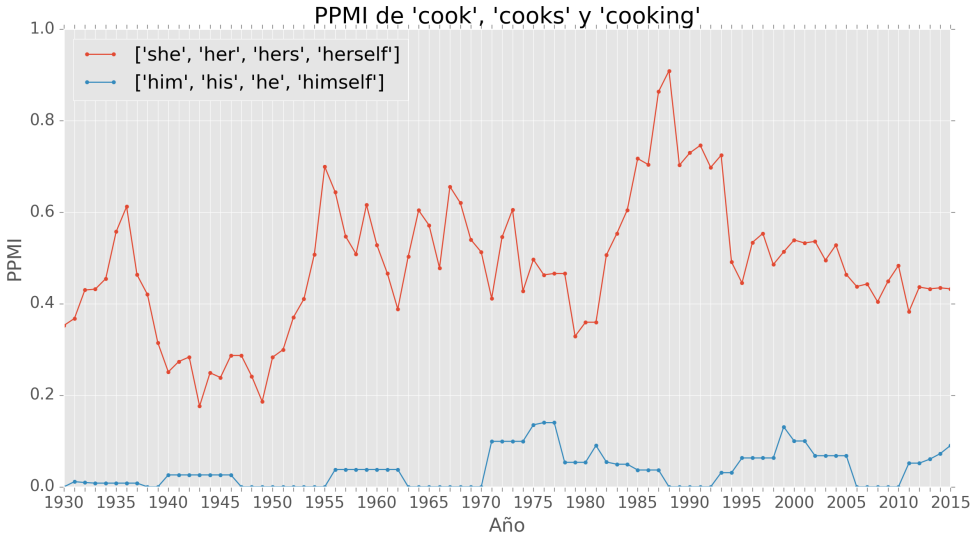


Figura 2.19: PPMI de cocinero y pronombres de ambos géneros (smoothing 3)

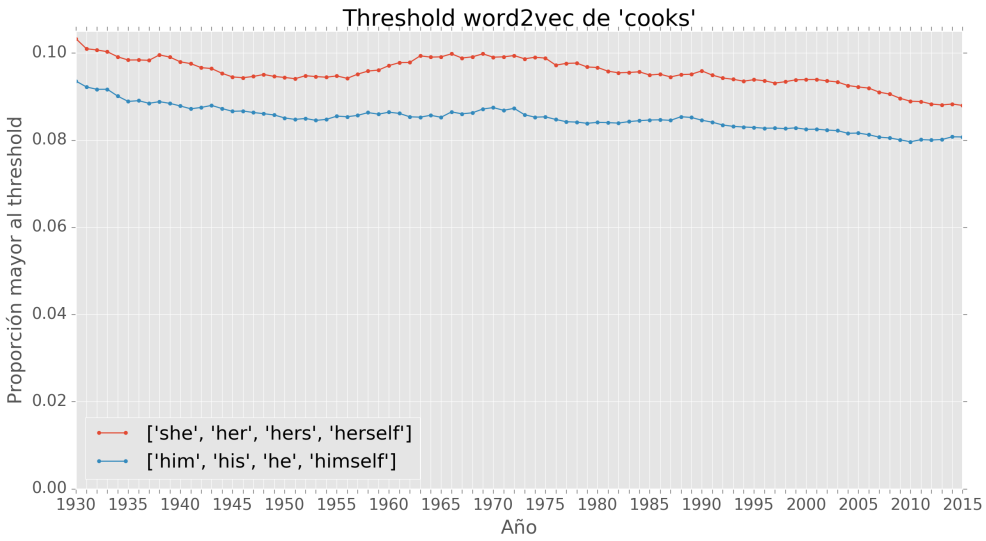


Figura 2.20: Similitud threshold word2vec del vector de cocinero y los contextos de los pronombres de ambos géneros (smoothing 3)

Pasamos a la profesión mayoritariamente femenina y sin cambio en la actualidad, enfermería. En la figura 2.21 vemos el PPMI en relación con cada pronombre, y la asociación de los femeninos es mucho mayor que la de los masculinos. A la vez, no se notan grandes cambios a través del tiempo, la profesión está consistentemente asociada con las mujeres.

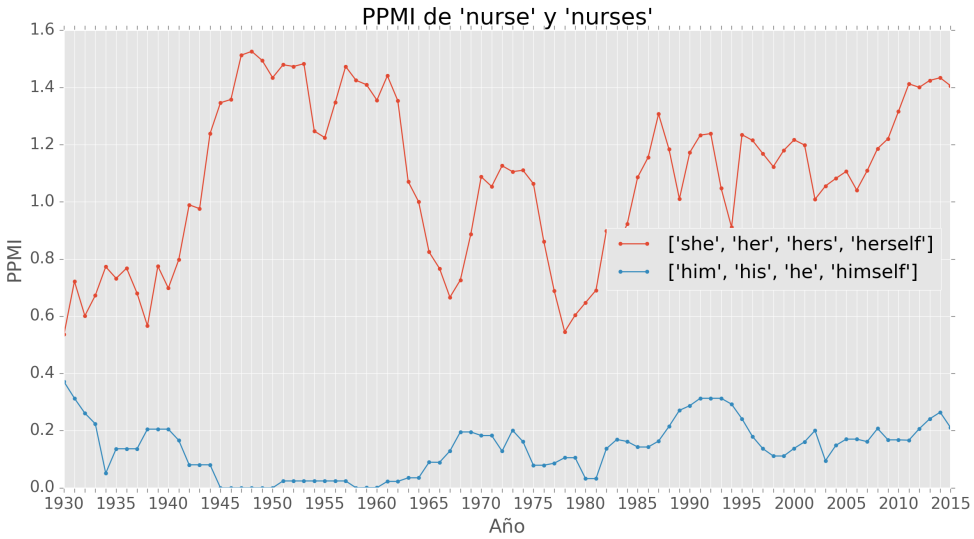


Figura 2.21: PPMI de enfermería y pronombres de ambos géneros (smoothing 3)

Lo mismo sucede cuando observamos word2vec para esta relación, la asociación entre los contextos de los pronombres y la profesión de enfermería es consistentemente más alta en el caso de

los pronombres femeninos.

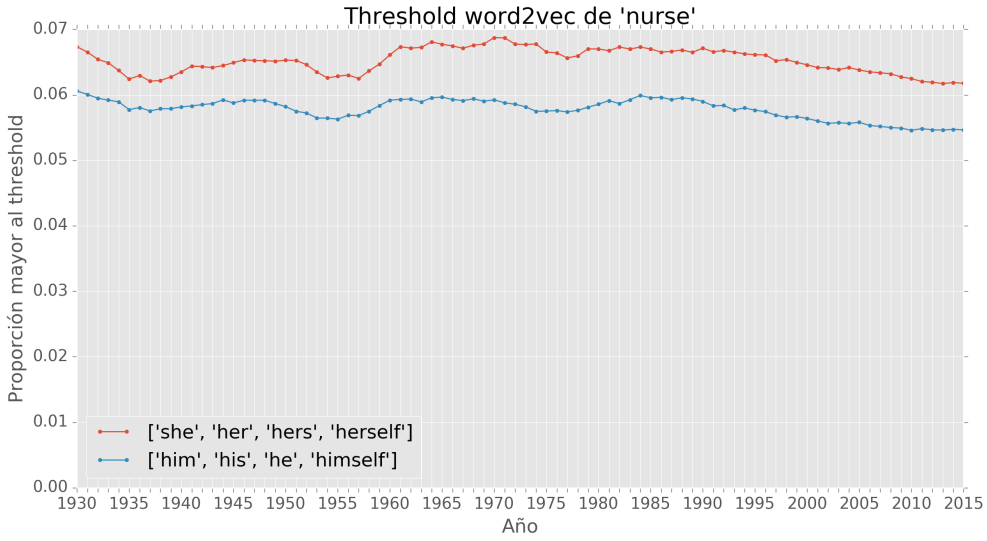


Figura 2.22: Similitud threshold word2vec de enfermería y pronombres de ambos géneros (smoothing 3)

Por último analizamos realtor, una profesión que tuvo mayoría masculina, pero actualmente está casi pareja en su representación de género. Las menciones de realtor comienzan a ser consistentes a partir de 1987, antes que eso, no entra en el vocabulario habitual, y por lo tanto no hay análisis de PPMI. Luego de eso, la asociación con ambos pronombres es baja, y ninguno de los dos predomina por sobre el otro. Parece ser que la terminología de realtor viene asociada con la concepción moderna de la ocupación.

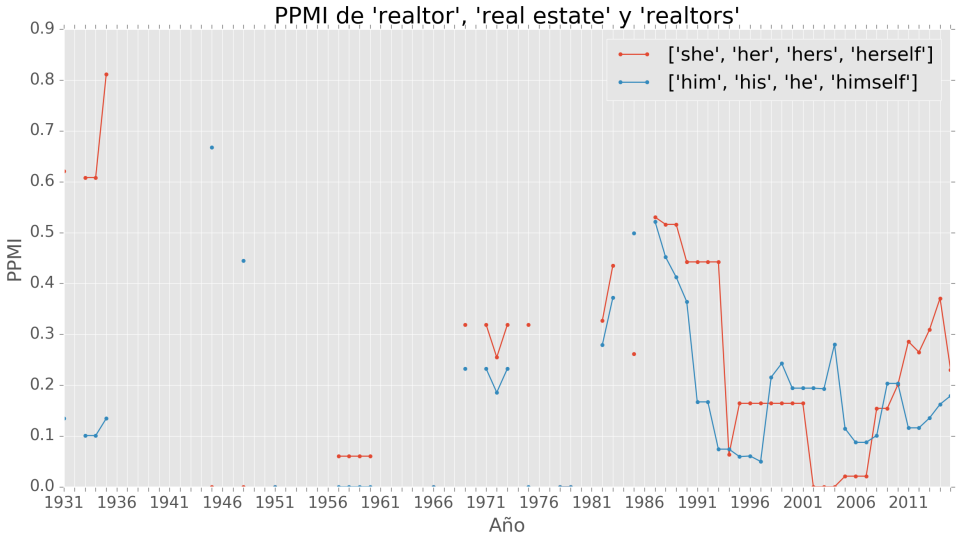


Figura 2.23: PPMI de inmobiliaria y pronombres de ambos géneros (smoothing 3)

En el caso de asociaciones entre word2vec la asociación con los pronombres femeninos es más fuerte, pero ambas son bastante parecidas, y bastante bajas comparando con otras asociaciones vistas en esta sección. El corpus de noticias donde está entrenado este word2vec parece identificar más cercanamente a realtor de contextos femeninos para que la asociación sea tan consistente.

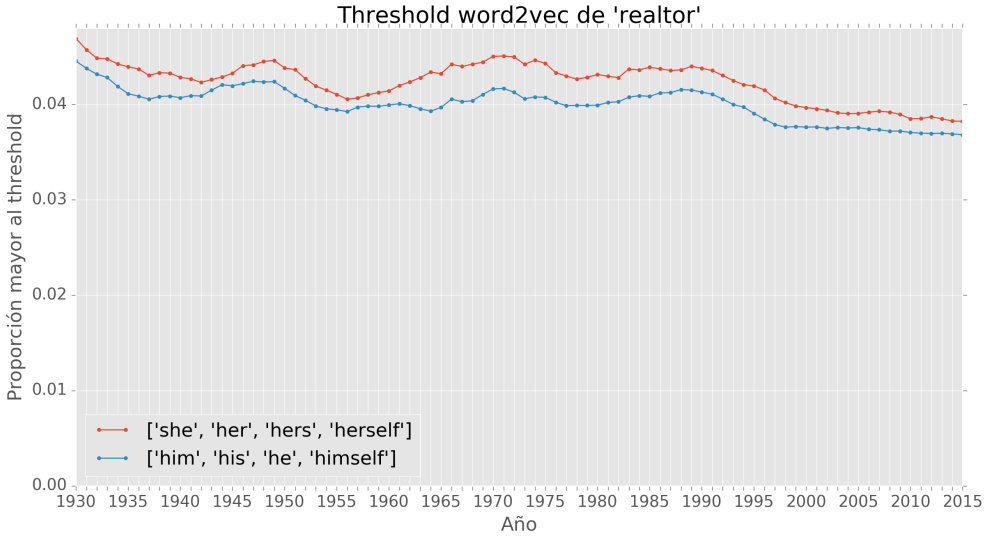


Figura 2.24: Similitud threshold word2vec del vector de inmobiliaria y los contextos de los pronombres de ambos géneros (smoothing 3)

Finalmente, habiendo observado todos los casos de profesiones que hayan variado y se hayan mantenido estables en el tiempo, vemos que el cine parece no estar necesariamente tan estático

en su inclusión de la fuerza laboral como pensamos. Aquellas profesiones que han pasado de tener representación mayoritariamente masculina a mayoritariamente femenina o muy pareja se ven al menos parcialmente reflejadas en el corpus analizado. Y aquellas donde el mercado laboral se encuentra muy dividido, siguen manteniendo su proporción, lo que no es sorprendente.

2.5 Conclusiones

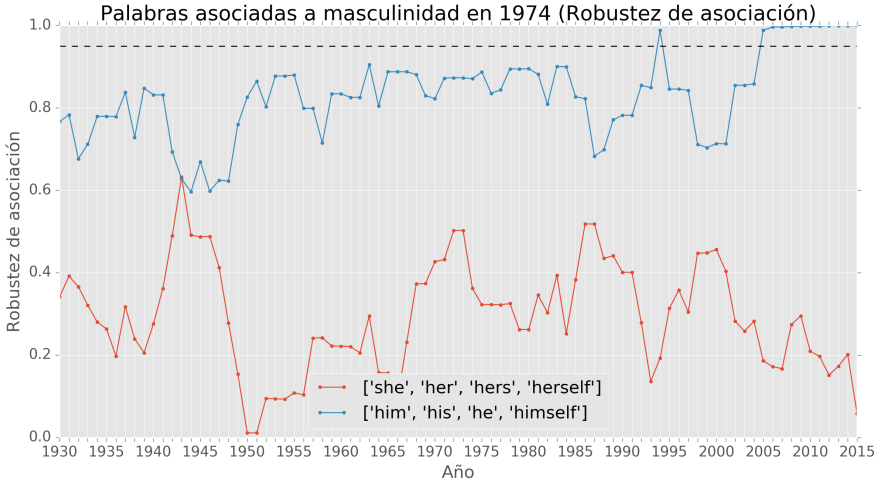
A lo largo de este capítulo pudimos reproducir múltiples resultados: en primer lugar, la cantidad de pronombres masculinos por cada femenino demostró estar en o por arriba del doble. Pudimos observar que la mayor disparidad entre ambos géneros de pronombres sucede alrededor de los 70 y desciende desde entonces.

En segundo lugar, vimos que el vocabulario asignado como atributos femeninos, masculinos y neutrales tiene asociaciones fuertes con los pronombres respectivos, y asociaciones leves con el pronombre opuesto. Aunque un poco más fuerte la asociación masculina con los atributos femeninos que el contrario, especialmente en la actualidad. Pero no encontramos evidencia de que las palabras neutrales estuvieran más asociadas a lo masculino, como en trabajos previos.

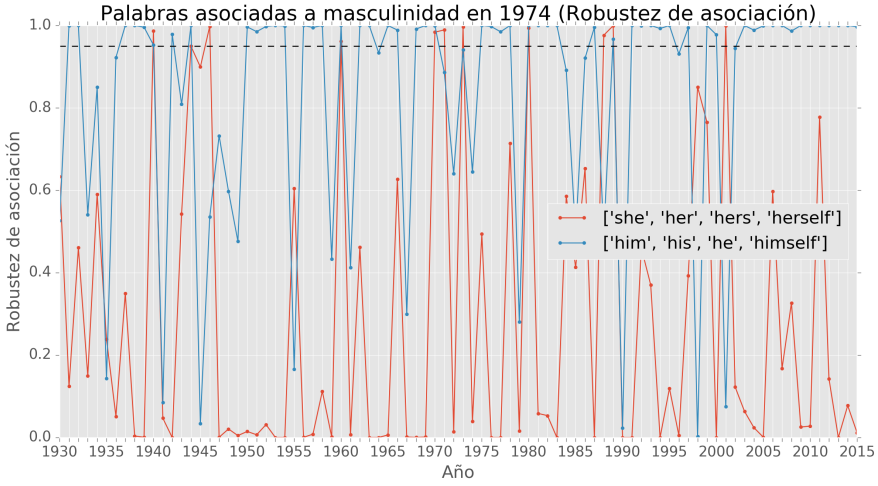
Aún más, las ocupaciones y trabajos marcados como correspondientes a cada género también tienen asociaciones fuertes con éste, incluso más fuertes que los atributos vistos. En ambos casos, el género opuesto tiene una asociación muy leve o nula, y encontramos apoyo para la hipótesis de que los roles neutrales, en promedio, tienen asociaciones similar con los pronombres de ambos géneros. Finalmente nos centramos en profesiones específicas y examinamos un poco más a fondo cada una. Los resultados comparados con las estadísticas reales de cada profesión fueron un poco más realistas que los resultados combinados: aquellas profesiones con mayoría numérica de un género, mantienen una asociación fuerte con él, pero aquellas que han cambiado con el tiempo mostraron algo de esa modificación, con predominio de un género pero más cercana, o relaciones iguales con ambos.

En lo que se refiere a los métodos, entre todos los casos de estudio hemos visto que con un vocabulario tan amplio como son los pronombres por género, el PPMI es una herramienta que permite captar casi todas las asociaciones buscadas.

Por otro lado, al utilizar los vectores word2vec de varias profesiones específicas y compararlos con los contextos de los pronombres los resultados permitieron ver relaciones muy claras entre ambos, aunque las tendencias en el tiempo no fuesen tan claras como con PPMI.



(a) Con 3 años de smoothing



(b) Sin modificar los datos por año

Figura 2.5: Robustez de asociación para PPMI de atributos masculinos y pronombres de ambos géneros (figura 2.4)

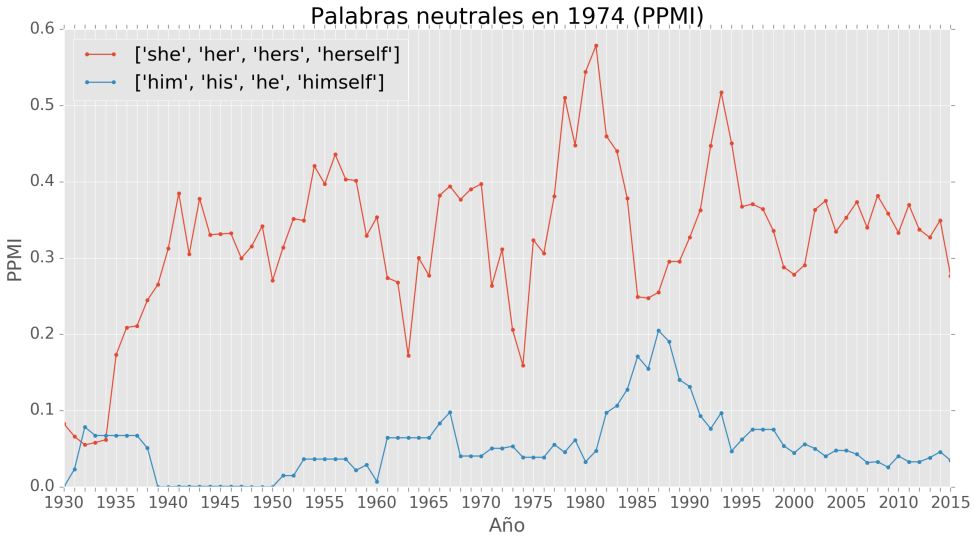


Figura 2.6: PPMI del vocabulario proveniente del BSRI marcado como neutral en contexto de pronombres masculinos y femeninos (smoothing 3)

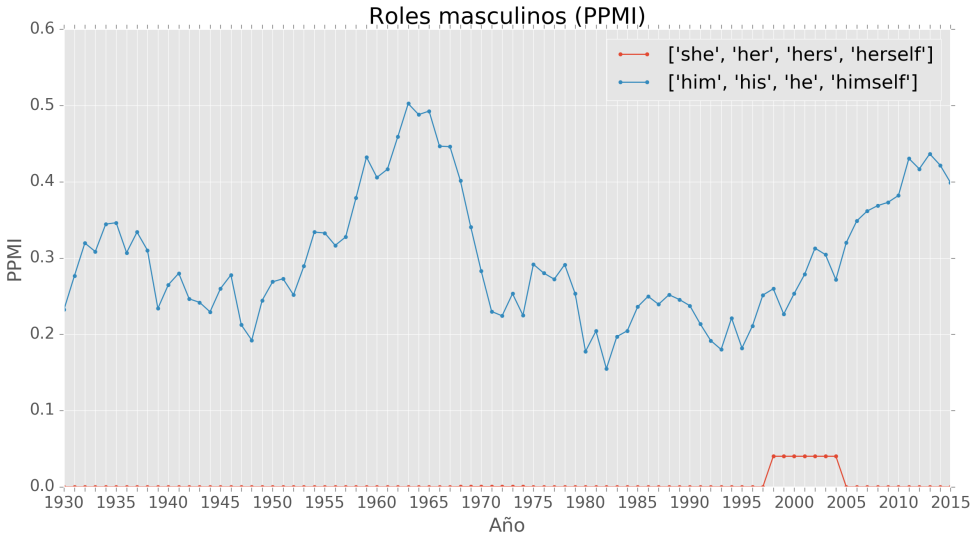


Figura 2.9: PPMI de roles utilizado por Lenton *et al.* (2009) definidos como masculinos en contexto de pronombres femeninos y masculinos (smoothing 3)

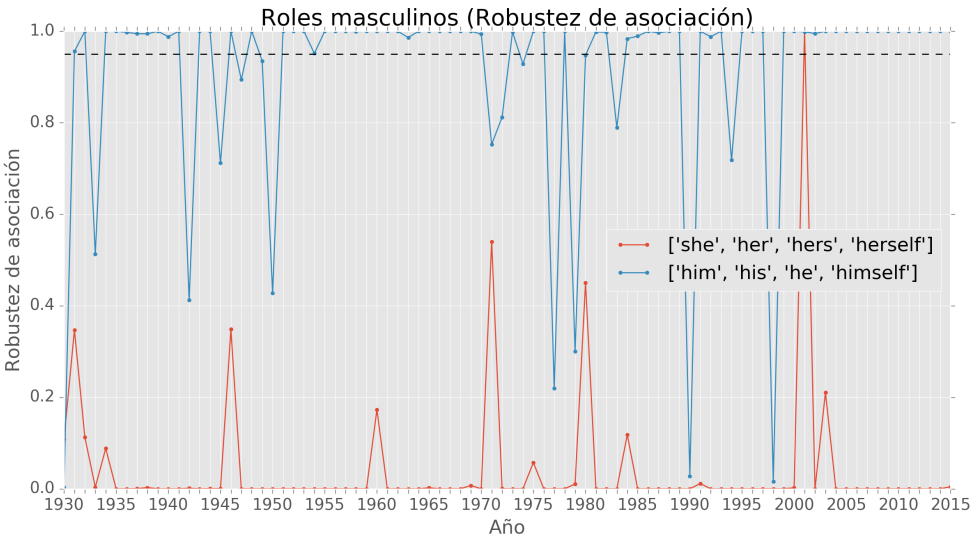


Figura 2.10: Robustez de asociación para PPMI de roles masculinos y pronombres de ambos géneros (figura 2.9) (smoothing 3)

Capítulo 3

Caso de estudio: Terrorismo

El acercamiento al terrorismo como caso de estudio viene motivado por la importancia del ataque a las Torres Gemelas para la cultura estadounidense[1] y el reciente crecimiento de la Islamofobia en el mundo en los últimos años (Lichtblau, 2015).

Como hipótesis inicial presentamos que el terrorismo como enemigo y la lucha en contra del gobierno estadounidense, crecen a partir del 2001, y que está ampliamente relacionado con la religión islámica. Postulamos que existe un estereotipo en el cine estadounidense de terrorista árabe y/o musulmán, y que es la representación mayoritaria de los países árabes.

3.1 Nacionalidades del terrorismo

3.1.1 Introducción

El primer acercamiento que queremos estudiar del estereotipo del terrorista es la asociación a una nacionalidad/etnia en particular. A partir de los resultados de Riegler (2010) seleccionamos árabe y palestino como parte del estereotipo pre-2001. Este artículo estudia casos particulares de representación de terroristas en películas a lo largo de 3 décadas. A través de la revisión de varias películas por década busca sacar conclusiones sobre la imagen del terrorista "estándar" en el cine durante esos años.

Al respecto de la década de los '70, Riegler (2010) dice que la inspiración de Hollywood venía de eventos internacionales, principalmente las acciones de grupos palestinos, dado que no había habido atentados en suelo estadounidense. A partir de ahí, uno de los estereotipos encontrados de terrorista fue el árabe que secuestra un avión.

Por otro lado, las guerras contra el terrorismo a las que se enfrentó Estados Unidos luego del atentado del 2001 fueron en Afganistán, y luego en Irak[2].

Riegler (2010) menciona que los estudios de Hollywood evadieron tratar con el terrorismo inmediatamente después de la caída de las torres por ser un tema muy arriesgado. Recién en 2005 salieron las primeras películas de consumo masivo en las cuáles se trata con ataques terroristas de origen religioso y político.

3.1.2 Hipótesis

A partir de lo estudiado en Riegler (2010) vamos a testear:

- Presencia y asociación con palestinos y árabes durante las décadas de 70 y 80

[1] https://www.globalpolicy.org/war-on-terrorism.html
[2] https://www.globalpolicy.org/war-on-terrorism.html

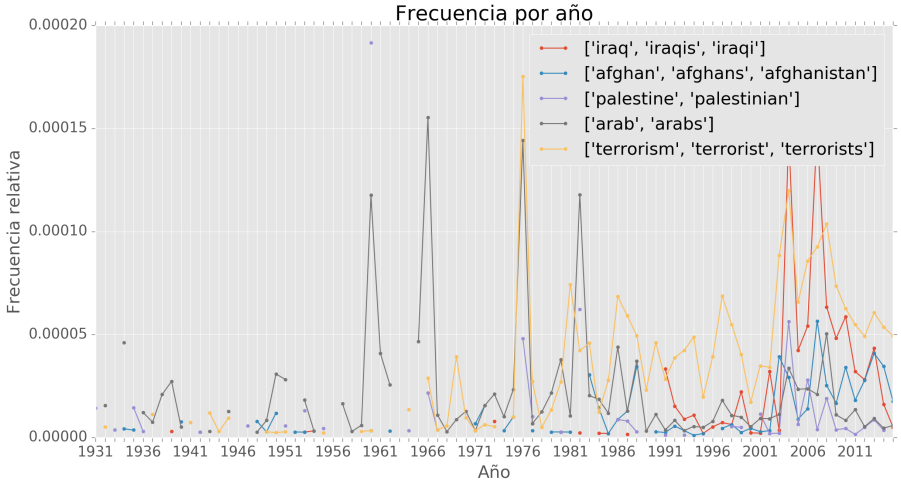
- Luego de 2001 esperamos ver un crecimiento en asociación con Afganistán, y a partir de 2003, con Irak
- No esperamos ver relación entre terrorismo e Italia

Incluiremos Italia en las búsquedas por país como control para comparar con un país que no debiera tener lazos al terrorismo en el cine de Estados Unidos.

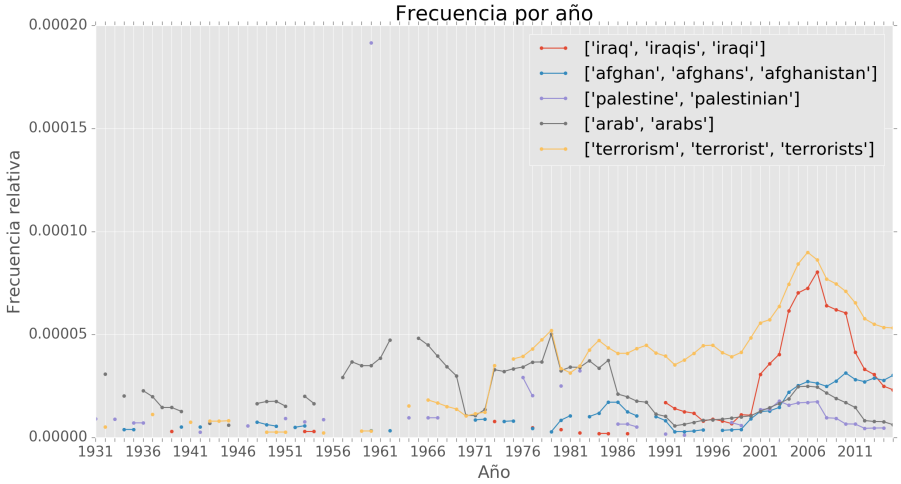
3.1.3 Resultados y discusión

Como panorama introductorio observaremos la frecuencia de los términos que vamos a utilizar en esta sección: la palabra terrorismo en sí y los países involucrados en los conflictos de la Guerra contra el Terror ("War on Terror"). Mostramos tanto la figura sin modificaciones (3.1a) como con smoothing de 3 años (figura 3.1b) para observar la tendencia más fácilmente pero tener visibilidad sobre el momento donde comienza, o no, el efecto del atentado en 2001 sobre las menciones.

A partir de la figura 3.1 vemos que las menciones a Palestina son escasas tanto en los años 70 como después, pero hay múltiples menciones a árabes alrededor de esa época. Por otro lado, aportando a la hipótesis de que Afganistán e Irak sólo entraron en escena a partir del atentado de 2001, ambas curvas crecen a partir del 2003. En 2001 y 2002 hay pocas menciones a terrorismo (esto se ve mejor en la figura 3.1a, dado que en la otra, el smoothing hace que la subida se vea justo en el 2001), haciendo eco de lo que expone Riegler (2010) cuando dice que el impacto de la caída de las Torres Gemelas fue tal que los grandes estudios no se animaron a acercarse al tema hasta varios años después.



(a) Smoothing 0



(b) Smoothing 3

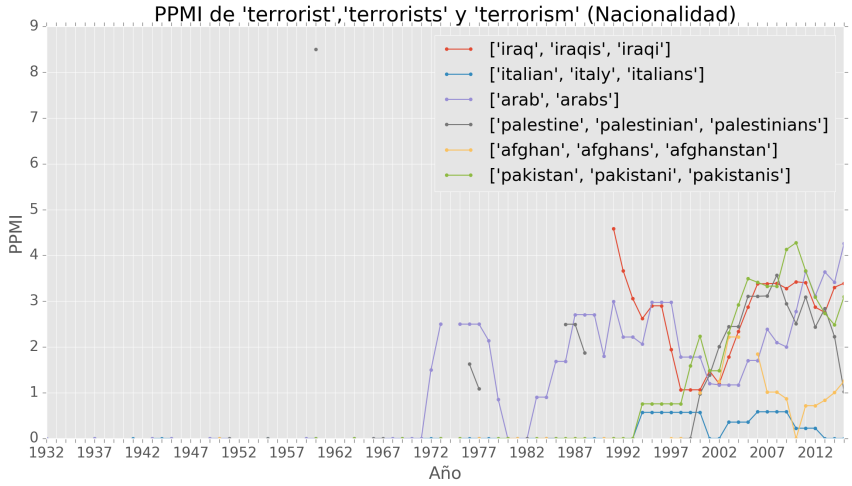
Figura 3.1: Frecuencia relativa de uso de terrorism y los países que vamos a investigar

Observando la figura 3.2a, se ve que las primeras menciones entre terrorismo y cualquiera de las nacionalidades se dan entre 1971 y 1972. En particular, en el año 1970 fue el primer involucramiento de los Estados Unidos en el conflicto "Black September" en Jordania[3].

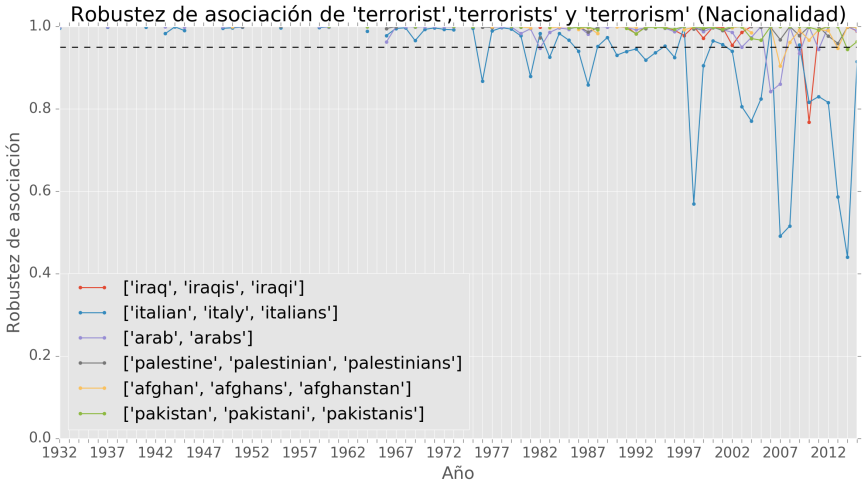
Es también visible que hasta el 2001, las relaciones entre terrorismo y nacionalidades eran casi exclusivamente árabes, palestinos e iraquíes (la historia entre Irak y EEUU es complicada a nivel de ser tanto aliado como objetivo de invasión en este período), pero a partir de ahí se nota lo que también se menciona en Riegler (2010) respecto a la cercanía del personaje estereotípico terrorista con la realidad. Dice en la bibliografía que antes del 2001 era casi una caricatura, acercándose un poco más a hechos reales hacia los años 90. Post 2001 se ve en la figura 3.2a un comienzo de asociación entre nacionalidades involucradas en los conflictos en los que Estados Unidos intervino: Afghanistán como país de mayor importancia siendo allí donde se encontraba Bin Laden y Al-

[3]http://adst.org/2015/07/jordans-black-september-1970/#.WhI7LxNSzUo

Qaeda, Pakistán fue aliado en la búsqueda estadounidense de Bin Laden. Eventualmente Iraq entró en escena con la guerra de 2003.



(a) PPMI de terrorismo para los contextos de los países a investigar (smoothing 3)



(b) Robustez de asociación

Figura 3.2: Asociación entre terrorismo y las nacionalidades investigadas observando PPMI y robustez de esa asociación (smoothing 0)

Teniendo en cuenta que las menciones a terrorismo dentro del corpus son limitadas, observamos la robustez de asociación en la figura 3.2b. El número escaso de menciones en conjunto con las nacionalidades podría haber elevado el PPMI, pero observamos que los resultados obtenidos están lejos de ser probables sin motivo. Excepto el país de control, Italia, que muestran mayor sensibilidad cosa que está dentro de lo esperado, todas las asociaciones son robustas.

Además de la palabra terrorismo en sí, vamos a investigar si temáticas relacionadas al terrorismo se visualizan en los contextos de las nacionalidades elegidas. Es decir, si se habla de terrorismo sustancialmente cuando se menciona a cualquiera de estas nacionalidades.

En la figura 3.3 se visualizan los contextos en los que se habla explícitamente de esos países o

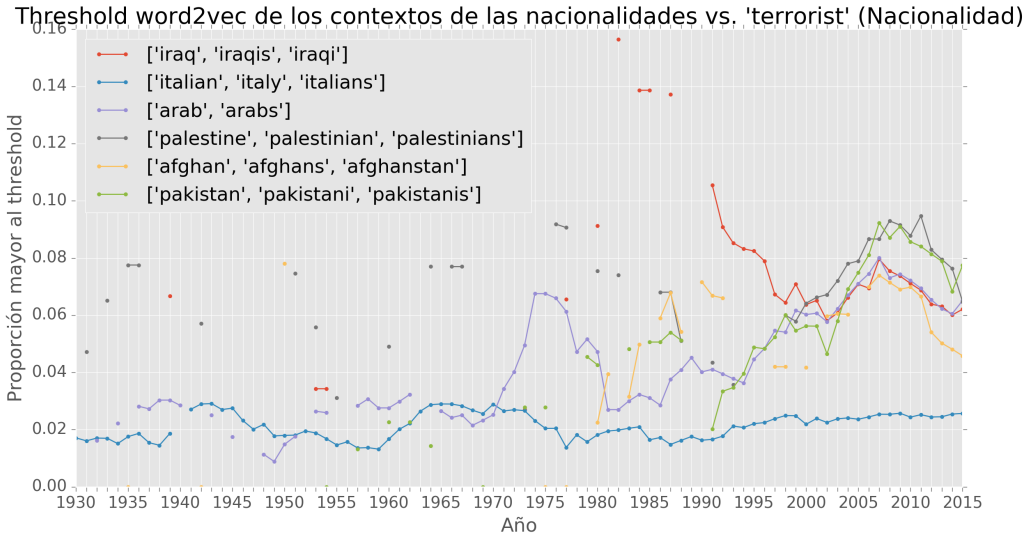


Figura 3.3: Similitud threshold word2vec de los contextos de cada nacionalidad con el vector word2vec de terrorismo (smoothing 3)

de gente de esa nacionalidad. Aquellos años donde no hay datos, es porque no hay menciones en ninguna película. En el caso de Italia no se nota ninguna asociación, pero vemos que un porcentaje mucho mayor del vocabulario contexto del resto de las nacionalidades es cercano al concepto de terrorismo. No podemos decir con certeza que el contexto sea exactamente ese, porque además del vocabulario directamente relacionado, el espacio vectorial cercano a terrorismo incluye: guerra, piratería, cybercrimen, crimen, etc, pero sí encontramos evidencia de que cuando se habla de estas nacionalidades, gran cantidad de veces se hace referencia a situaciones violentas, mientras que cuando se habla de Italia existe ciertas menciones, pero no con esa asiduidad.

Entre los tres últimos resultados se puede observar que Palestina, Afghanistán e Iraq básicamente no tienen menciones excepto en aquellos años donde su relación con terrorismo es alta. Esto aporta evidencia a la hipótesis de que los personajes de estas nacionalidades se encuentran fundamentalmente en películas que involucran terrorismo, y no en otras situaciones.

3.2 Relación con la religión

3.2.1 Introducción

En segundo lugar, enfocamos el análisis en la religión. Hoy en día están en alza los crímenes contra personas musulmanas (Madi, 2017). En los medios estadounidenses la representación de los musulmanes es escasa y generalmente negativa (Team, 2015), por lo que vamos a tratar de visualizar si parte de ese sesgo se da también en el cine.

Durante el período de la Guerra Fría con Rusia, el islam no era enemigo de Estados Unidos porque era anticomunista (Silva, 2017), pero el atentado de septiembre de 2001 generó en su momento una ola de crímenes contra individuos cuyo aspecto fuera estereotípicamente musulmán, lo que llevó a agresiones contra musulmanes, Sikh y personas de países nacidas en Medio Oriente más allá de su religión[4].

[4]https://www.globalpolicy.org/war-on-terrorism.html

En Silva (2017) se describe cómo los medios plantean la violencia del enemigo de Estados Unidos como "terrorista", mientras que la violencia del ejército aliado es necesaria y se matiza la forma en la que se la describe, como por ejemplo "estrategia" de un bando contra "ataque" del otro. Riegler (2010) elabora sobre lo mismo diciendo que a partir de la década de los 80 entra el concepto de "mal menor" como la violencia que ejercen los héroes para desafiar la amenaza terrorista. Para darle un marco lingüístico más claro a esta búsqueda, vamos a utilizar el vocabulario ejemplificado en Steuter y Wills (2009). Este artículo analiza los medios canadienses bajo la premisa de que deberían ser más imparciales que los estadounidenses en su representación del conflicto contra el terrorismo, y termina concluyendo que ambos países utilizan los mismos recursos: asocian al enemigo perteneciente a los países árabes con insectos, roedores y virus a través de analogías con cuevas en lugar de prisiones y describiendo sus movimientos con los verbos correspondientes a acciones animales. De este modo generan una lejanía con las personas que protagonizan esas noticias, y el texto se llena de imágenes que conllevan asociaciones negativas.

Los ejemplos fueron elegidos entre el vocabulario que menciona el artículo como para cubrir varios casos similares: en principio "rat" porque hay varios títulos que hacen referencia directa o indirectamente a estos roedores (indirectamente es a través de verbos asociados como "scurry"), "cage" para referirse al lugar donde se lleva a los presos, "nest" referenciando a lugar de reunión o escondite, y "trap" como sustantivo o verbo en referencia al acto de arrestar. Para comparar, usamos "arrest", "prison", y "hideout".

3.2.2 Hipótesis

Debido a la búsqueda del estereotipo de terrorista como musulmán, vamos a investigar a esta religión con las palabras: "islam", "muslim" y "muslims", y utilizaremos las equivalente para judaísmo y cristianismo como comparación. Vamos a investigar si:

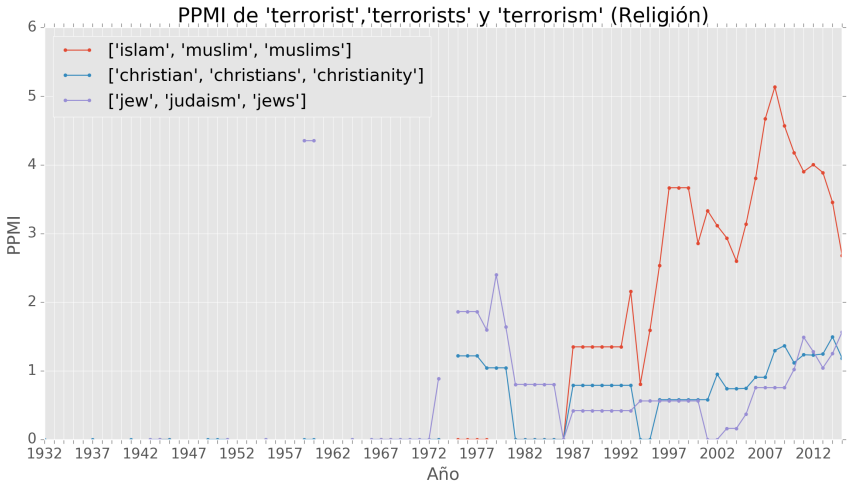
- La religión musulmana tiene mayor relación con terrorismo a partir del año 2001
- Las religiones de control no muestran grandes asociaciones en ningún momento histórico
- Tanto el terrorismo como el islam tienen asociaciones más fuertes a vocabulario relacionado con roedores que con los equivalentes neutros.

3.2.3 Resultados y discusión

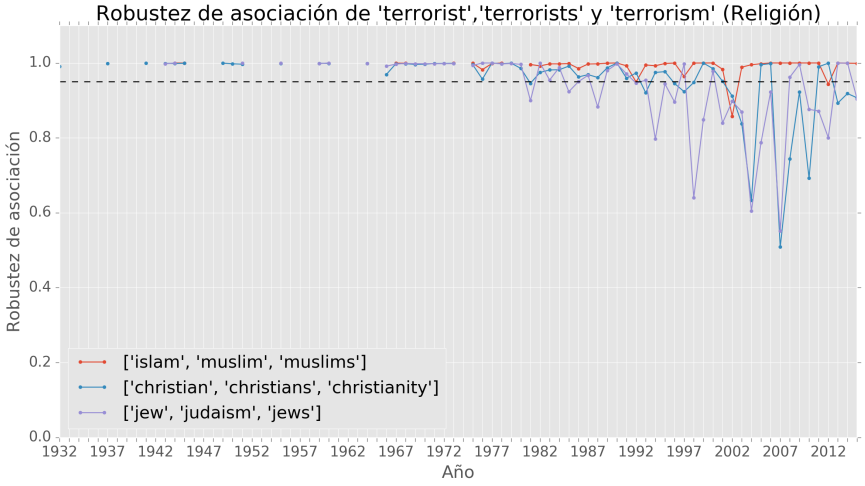
Inicialmente vemos en la figura 3.4a que el crecimiento del acercamiento entre islam y terrorismo se da entre mitad y fin de los 90, cosa que coincide con el final de la Guerra Fría, pero previo a lo que esperábamos. Aparentemente el islam aparecía en el cine en relación al terrorismo previo al 2001. Aunque sí se puede visualizar un aumento en la asociación luego del atentado.

Las relaciones de las religiones de control son mayores a lo esperado, es posible que la religión en general está más conectada semánticamente con el concepto del terrorismo que lo que habíamos pensado en las hipótesis y a partir de la literatura.

En la figura 3.4b vemos la Robustez de asociación para el análisis visto, y si bien el islam es el que tiene la asociación más fuerte, las otras religiones también se ven robustas.



(a) PPMI de terrorismo y religiones (smoothing 3)



(b) Robustez de asociación para la figura 3.4a (smoothing 0)

Figura 3.4: Relación entre terrorismo y religiones judía, cristiana y musulmana

El análisis de los contextos promediados de terrorismo y los vectores para cada religión en word2vec muestra una relación más alta con el islam a partir de los años 80, pero el judaísmo parece más relacionado en época de Guerra Fría. En todo caso, las diferencias son sutiles entre las tres tendencias, y se debe tener en cuenta que el corpus de word2vec está entrenado con noticias del año 2014, y es posible que el concepto de islam como se entiende actualmente en los medios esté cercano al terrorismo Steuter y Wills (2009), y no como se entendía en su momento. Por ejemplo, la cercanía entre los vectores de "threat" y "jew" en de 0.066, mientras que entre "threat" y "muslim" es 0.11[5], sólo con hablar de temas cercanos al terrorismo (como crimen), el corpus ya tiene una distancia más corta relacionada con la historia reciente y la cobertura mediática.

[5]http://bionlp-www.utu.fi/wv_demo/

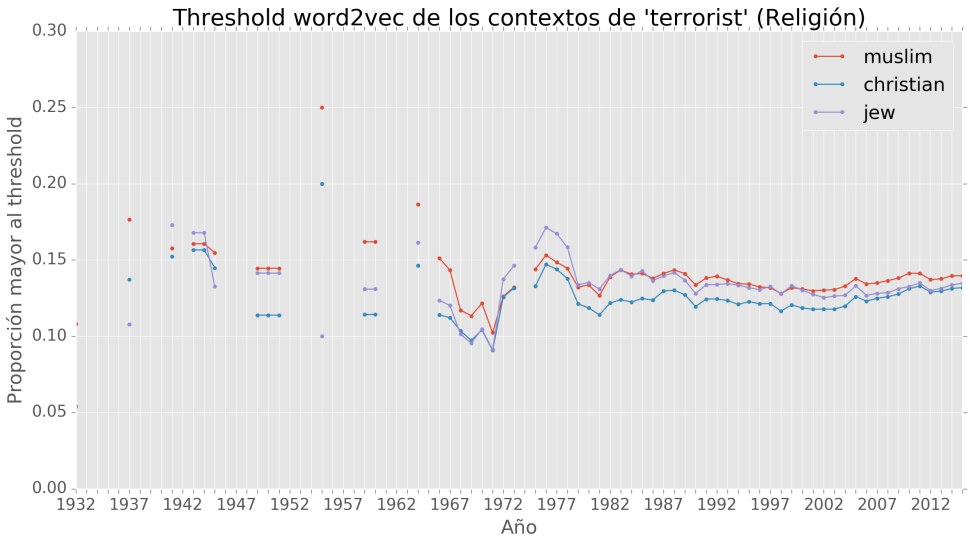
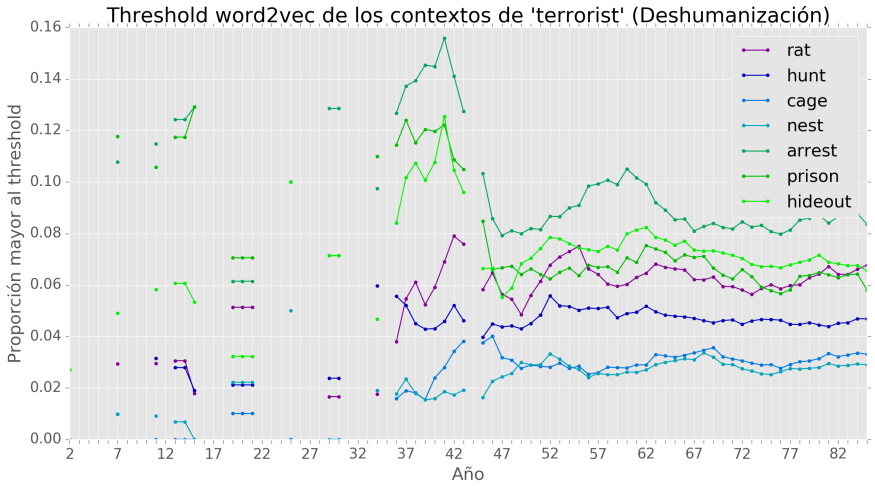


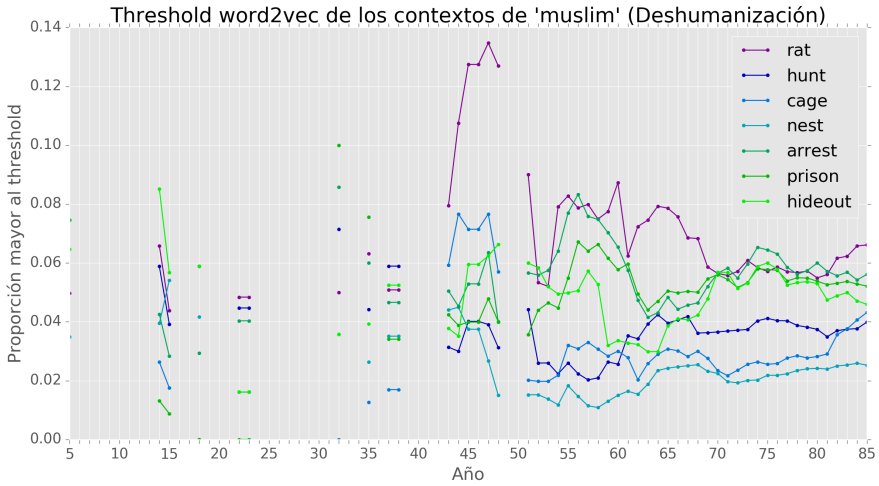
Figura 3.5: Similitud threshold word2vec de los contextos de terrorist con los vectores word2vec de cada religión (smoothing 3)

Por último queremos investigar lo que se expone en Steuter y Wills (2009) respecto del vocabulario deshumanizador que se aplica en los medios a los practicantes del islam. Con este objetivo medimos el vocabulario de ejemplo en su similitud word2vec contra terrorismo y contra el islam.

Los resultados a través de PPMI no permiten observar ninguna asociación dado que el corpus tiene muy pocos usos conjuntos para terrorista o musulmán y el vocabulario elegido (figuras en el apéndice C), por lo que utilizamos los contextos de terrorista para ser comparados con el vocabulario elegido utilizando word2vec.



(a) Similitud threshold word2vec con vectores de lenguaje deshumanizador y los contextos de terrorista (smoothing 3)



(b) Similitud threshold word2vec con vectores de lenguaje deshumanizador y los contextos de musulmán (smoothing 3)

Figura 3.6: Relaciones entre vocabulario deshumanizador por un lado, y terrorismo e islam por el otro

Los resultados de esta búsqueda se encuentran en las figuras 3.6a y 3.6b. Del lado de terrorismo, en la figura 3.6a, los contextos muestran una asociación más fuerte con el vocabulario más humano: "arrest", "hideout", "prison", y algo de cercanía, pero leve con la semántica de "rat". En el caso de "musulmán" (figura 3.6b), también se ven más asociaciones al vocabulario más policial que al deshumanizado (aunque todas son leves), pero además hay una asociación mayor a éstas con "rat". La asociación es significativa, pero sin entrar en los casos específicos de uso es difícil asegurar que sea porque el vocabulario es despectivo, o por otra cercanía.

3.3 Modus operandi

3.3.1 Introducción

Por último, decidimos buscar una asociación que no suele cuestionarse en los medios estadounidenses: la del terrorista con un tipo de atentado particular. Todos los estereotipos que estuvimos analizando hasta ahora vienen ligados a acciones de extranjeros motivados por religión o política, o de psicópatas sin motivaciones racionales. Pero en Estados Unidos existe otro tipo de atentado, que habitualmente se denomina "mass shooting"[6]. Algunos ejemplos son la masacre de Columbine[7], o el tiroteo en un cine de Aurora[8]. Los motivos de estos hechos pueden ser desde personales, como bullying, hasta políticos, como xenofobia, homofobia, o sexismo. Los medios no suelen presentar estos ataques como "terrorismo" sino como "hechos independientes", cometidos por alguien con algún problema de salud mental (Metzl y MacLeish, 2015). Como posible análisis a qué se considera un hecho de terrorismo, agregamos una búsqueda respecto del medio que se utilice para asesinar: la bomba como asociada a los ataques religiosos, incluyendo la inmolación; el arma de fuego, en su genérico "gun", más representativa del crimen habitual en suburbios y ciudades estadounidenses, y el incendio, que si bien recoge varios casos en la historia, no es habitual actualmente.

3.3.2 Hipótesis

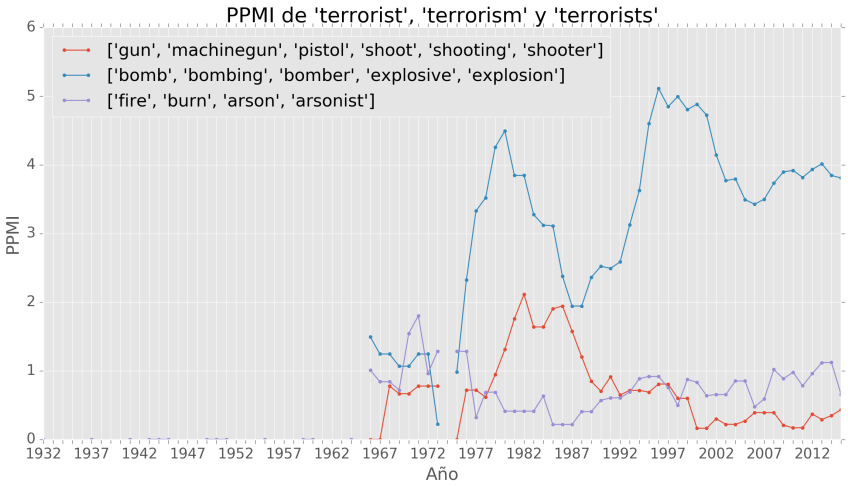
Dada la literatura, vamos a investigar si:

- Existe una asociación fuerte entre terrorismo y bombas
- No existe asociación, o es leve, entre incendios y armas de fuego, y terrorismo

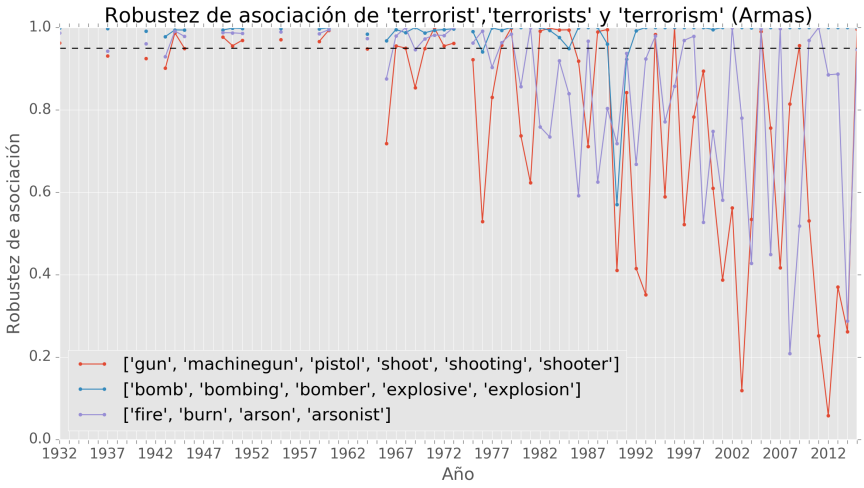
3.3.3 Resultados y discusión

La figura 3.7a aporta evidencia a la hipótesis de que el estereotipo de terrorista en el cine estadounidense viene fuertemente asociado a la idea de la bomba como arma. Nuevamente se observa que las menciones a terrorismo previas a 1965 son muy escasas, y en este caso, no hay apariciones conjuntas con el vocabulario. Utilizando la figura 3.7b observamos que la asociación entre el concepto de bomba y terrorismo es muy fuerte, mientras que las asociaciones que se observan en la figura 3.7a con las otras dos clases de armas, no se ven robustas, especialmente a partir de los años 90.

[6] https://www.nytimes.com/interactive/2017/10/02/opinion/editorials/mass-shootings-congress.html
[7] http://edition.cnn.com/2013/09/18/us/columbine-high-school-shootings-fast-facts/index.html
[8] http://kdvr.com/2017/07/20/remembering-the-12-lives-lost-in-the-aurora-theater-shooting/



(a) PPMI de los contextos de armas y terrorismo (smoothing 3)



(b) Robustez de asociación para PPMI sobre armas y terrorismo (figura 3.7a) (smoothing 0)

Figura 3.7: Relaciones entre terrorismo y tipos de armas: bombas, armas de fuego, incendios

3.4 Conclusiones

En esta sección vimos que el concepto de terrorismo está ligado a distintos países según la época: antes de la década del 2000 lo vimos asociado al genérico de árabe, pero luego del 2001 surgen asociaciones a los países involucrados en la guerra contra el terror: Afganistán e Iraq.

Observamos también que el islam está fuertemente asociado al terrorismo, mientras que otras religiones tienen menos contextos en común, y observamos una asociación entre musulmán y vocabulario despectivo, lo cual tiende a sugerir un estereotipo de musulmán como personaje enemigo. Por último, buscamos asociaciones entre terrorismo y distintos tipos de armas y encontramos que las bombas están fuertemente asociadas, mientras que las armas de fuego no se relacionan con atentados.

El vocabulario estudiado en esta sección tiene menor cantidad de apariciones que lo visto en

la sección anterior. Para estos casos, empezamos a ver la principal limitación del PPMI: si el vocabulario buscado es muy escaso puede no haber apariciones conjuntas de palabras, aún cuando estén relacionadas.

La investigación que más sufrió de este problema fue aquella sobre lenguaje deshumanizador, donde las apariciones de cada palabra específica del lenguaje no eran suficientes para ver ninguna tendencia. Y es allí donde word2vec suple la necesidad de mostrar relaciones entre palabras aún cuando no coocurren.

Por otro lado, creemos que en este caso en particular el entrenamiento de word2vec en un corpus de noticias reciente está pesando en su definición de musulmán, que viene cargada con las noticias de atentados en los últimos años, y que vimos en la bibliografía, suele ser la única representación de esta población (Team, 2015).

Capítulo 4

Caso de estudio: Imagen de Rusia

Rusia sobresale en la historia estadounidense como un país que representa al enemigo, principalmente partiendo de la Guerra Fría, y que luego aparece muy comúnmente en el cine en el formato de némesis del héroe. Para ver si la representación rusa como enemigo es cuantificable, o si los personajes rusos tienen más de una faceta que aquella de enemigo caricaturizado, elegimos medir asociaciones a estereotipos comunes: como comunista y mafioso, y a tradiciones culturales rusas, para saber si existe una representación histórica más profunda.

4.1 Asociaciones estereotípicas

4.1.1 Introducción

De todos los países que interactúan con Estados Unidos, Rusia parece ser una constante en el cine como representación del enemigo. Aún en el cine de los últimos años, donde la acción puede estar sucediendo en la actualidad, en el pasado o en el futuro, los rusos siguen haciendo apariciones como oposición al personaje principal estadounidense.

En el análisis de Fedorov (2013b), se ve un desplazamiento del estereotipo ruso desde los 80 hasta la actualidad: en los años 80 se expresa como "agente de violencia", fuerte, pero malévolo, y principalmente revolucionario comunista. En aquellas películas sobre Rusia no hay componentes románticos, ni familiares.

También dice que avanzados los años 90 aparecen estereotipos nuevos: por un lado, las "novias por encargo" venidas de Rusia para maridos del occidente, la representación de Rusia como país es de empobrecimiento extremo y retraso tecnológico, y muy alejada de cómo se veían realmente las ciudades rusas en ese momento histórico (Fedorov, 2013a).

Por último, hacia el 2000 comienzan a aparecer los rusos como gangsters y mafiosos, abandonando la figura del comunismo como peligroso, se traslada a una figura también relacionada con la violencia.

4.1.2 Hipótesis

Utilizando los términos "russia", "russian" y "russians" para el concepto de Rusia en PPMI (para word2vec usamos "russia" a secas). Queremos testear que:

- Las asociaciones mencionadas existen con respecto a Rusia
  - Mafioso
  - Comunista

  - Novia
  - Empobrecido

- que varían a lo largo del tiempo; en particular:
  - que la asociación comunista es la principal hasta los años 90
  - que a partir de los 90 se pueden ver las asociaciones a novia y a pobreza
  - que a partir del 2000 se puede ver la asociación a la mafia
- Y que son más fuertes con respecto a Rusia que con otros países

4.1.3 Resultados y discusión

Como primer acercamiento vemos, en la figura 4.1, que las referencias a Rusia se mantienen a lo largo del tiempo: post Guerra Fría, las referencias anuales vuelven a la misma cantidad que antes de que sucediera. Las referencias a comunismo, por otro lado, nacen a partir de los 50, y tienen mayor prevalencia entre los 60 y los 90, para decaer los años siguientes. Las referencias a mafia son bastante constantes, con una frecuencia levemente más alta antes de los 40, probablemente a razón de la escasez de películas de esos años (y la existencia de una o dos películas centradas en el tema).

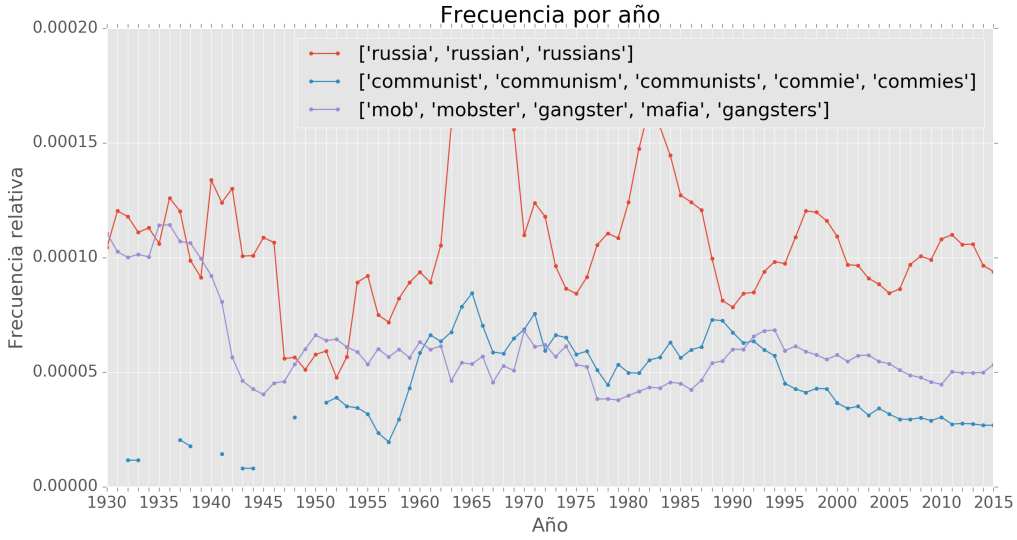


Figura 4.1: Frecuencia de los principales estereotipos mencionados por (Fedorov, 2013b) y Rusia (smoothing 3)

Concentrándonos en las asociaciones buscadas, la figura 4.2 muestra el PPMI de Rusia contra los cuatro conceptos buscados. En primer lugar, vemos que la relación con el comunismo comienza a principio de los 60 (al mismo tiempo que vimos que las menciones de la palabra aumentan), y no disminuye nunca, ni siquiera en la actualidad, a pesar de la disminución en la frecuencia de la palabra comunismo. El cine estadounidense parece continuar hablando de comunismo ruso al día de hoy.

Luego, vemos que la asociación entre Rusia y la mafia entra en juego a partir de los 90 y es muy fuerte, tanto como la de comunismo. No parece reemplazar la asociación comunista, como plantea Fedorov, si no que ambos modelos continúan sucediendo hasta la actualidad.

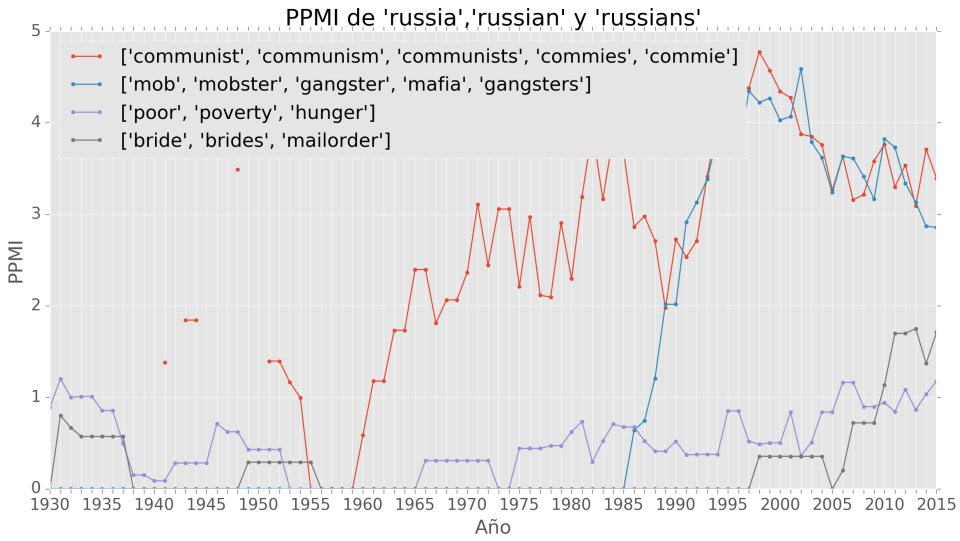


Figura 4.2: PPMI entre los contextos de Rusia y de las asociaciones encontradas en (Fedorov, 2013b,a) (smoothing 3)

En particular la curva de comunismo está relacionada con la historia de la Guerra Fría: ésta comenzó en los años siguientes al fin de la Segunda Guerra Mundial, esto sería 1946 o 1947, contando que la producción de películas lleva tiempo, este es más o menos el momento donde comunismo y Rusia empiezan a estar vinculados, alrededor de 1952. Luego, la tendencia general en la asociación es de crecimiento, pero este crecimiento aumenta aún más con el fin de la Guerra Fría (alrededor de 1989, junto con la caída del Muro de Berlín). Se ve una nueva subida en la asociación a partir del fin del conflicto a principios de los 90.

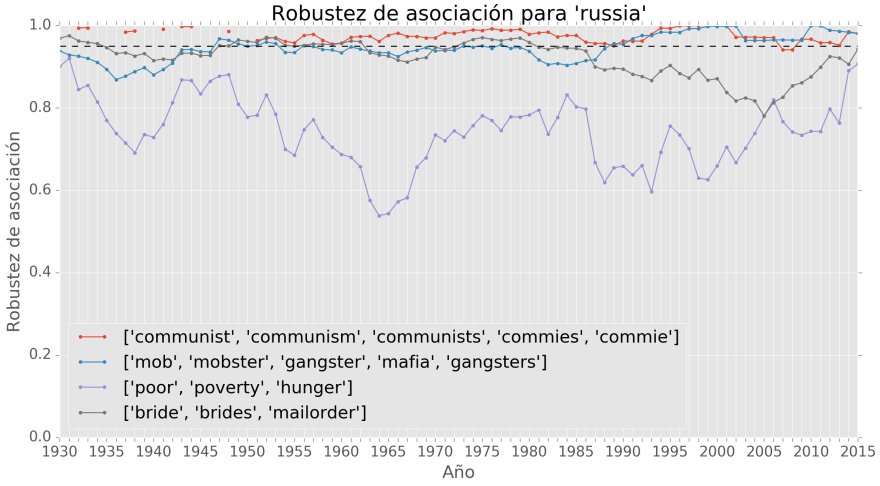


Figura 4.3: Robustez de asociación del PPMI entre Rusia y las asociaciones a estudiar (figura 4.2) (smoothing 3)

Las otras dos asociaciones de las que hablamos no se detectan tan claramente: la asociación de pobreza con Rusia es muy baja y es constante a lo largo del tiempo, no es el resultado del que habla Fedorov en su análisis, vemos a través de la Robustez de asociación (figura 4.3) que la asociación tampoco es muy sólida. Por último las referencias a "novias por encargo" de las que se habla en la bibliografía no acarrean una asociación muy fuerte, pero sí es perceptible un aumento en la cantidad de apariciones conjuntas en los últimos años, posterior a la hipótesis planteada.

Por último, tomamos las dos asociaciones que encontramos como más fuertes en el caso anterior e investigamos la relación comparada entre Rusia y otros países. Los elegidos para este experimento fueron: Francia como país de control, no esperamos asociaciones con mafia ni con comunismo, China que esperaríamos ver algo de asociación con comunismo, pero no con mafia, e Italia, la inversa. De esta forma podemos efectivamente ver si las asociaciones existen y cuán fuerte es la relación para este país contra otros estereotipos posibles.

En la figura 4.4 observamos el PPMI de comunismo con los cuatro países mencionados, y el resultado fue que, efectivamente, la asociación de Rusia con comunismo es mayor que la de los otros tres países, donde China es el que tiene más relación.

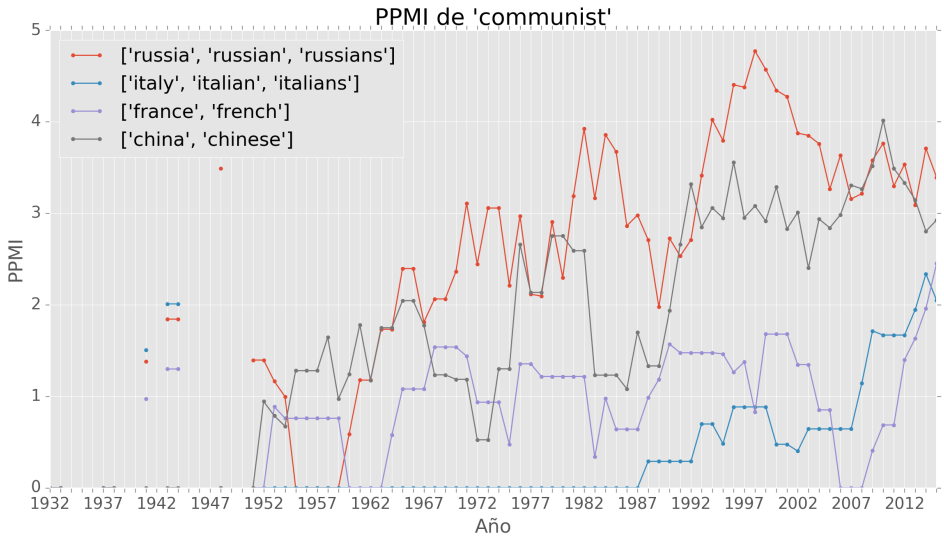


Figura 4.4: PPMI entre los contextos de comunismo y los de países a investigar (smoothing 3)

Luego observamos a los cuatro países en su relación con mafia y vemos en la figura 4.5 que Rusia tiene la relación con mafia observada que comienza en los años 90, mientras que previo a eso, la representación venía del lado de la mafia italiana. La asociación italiana continúa luego de la aparición rusa, pero ésta la supera. Los otros dos países no muestran una asociación relevante.

Habiendo visto estas relaciones es claro que el cine estadounidense parecería hablar mucho de Rusia en relación al comunismo y la Guerra Fría, y es comprensible que así sea, dado que durante muchos años esa fue su relación con este país. La aparición de otros personajes cuenta también que este relato no es el único en el que se involucra, sino que el más novedoso de la mafia también es prevalente.

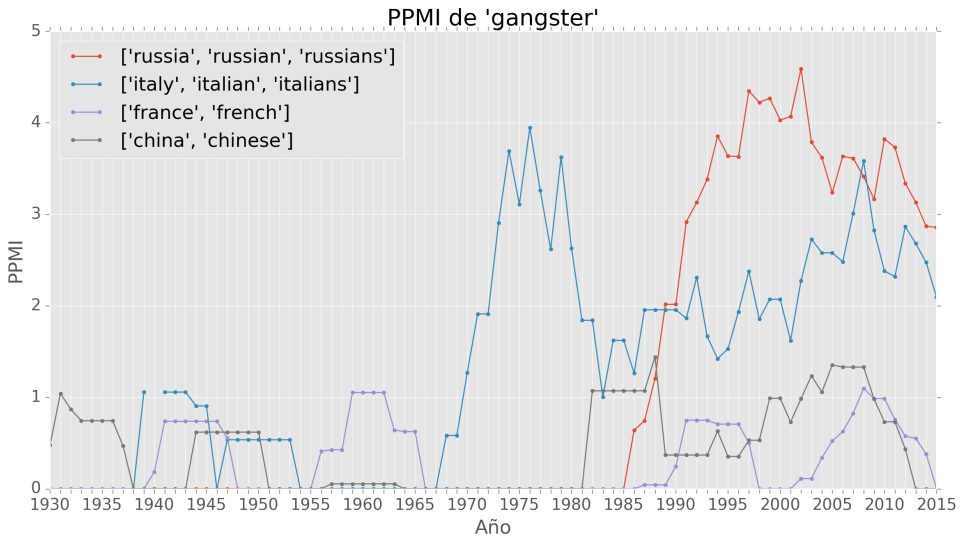


Figura 4.5: PPMI entre los contextos de mafia y lo de los países a investigar (smoothing 3)

4.2 Percepción de la cultura rusa

4.2.1 Introducción

Para terminar con el análisis de la visión Rusa decidimos observar detalles culturales del país, y tratar de dilucidar si las películas que hablan sobre Rusia también incluyen algo de historia o contexto socio-cultural. De las secciones anteriores sabemos que gran parte del cine que habla sobre Rusia también habla sobre comunismo, lo que presenta un momento histórico de la Rusia soviética.

En esta sección la intención es observar si al hablar de Rusia (y de comunismo) el cine muestra la situación en Rusia, explica algo de lo que sucede políticamente, o se involucra de alguna forma en la cultura rusa fuera del villano comunista por oposición al representante estadounidense.

4.2.2 Hipótesis

Para evaluar si el cine estadounidense profundiza históricamente en Rusia vamos a utilizar dos ejes:

- Comida tradicionalmente rusa, como forma de saber si se visualiza la cotidianidad rusa en una película donde se habla del país
- Vocabulario específico relacionado con la revolución rusa, dado que si se está hablando de la revolución con cierto análisis histórico, hay vocabulario que resulta fundamental y esperaríamos encontrarlo en ese contexto

Procedemos a testear si algunas de esas palabras aparecen en el cine, y con cuánta variedad.

4.2.3 Resultados y discusión

El vocabulario original elegido para representar comida rusa era: Beef Stroganoff, Bliny, Caviar, Chicken Kiev, Coulibiac, Dressed herring, Golubtsy, Guriev porridge, Kasha, Kissel, Knish, Kholodets, Kulich, Medovukha, Mimosa salad, Oladyi, Olivier salad, Paskha, Pelmeni, Pirog, Pirozhki, Pozharsky cutlet, Rassolnik, Sbiten, Shchi, Solyanka, Sorrel soup, Syrniki, Ukha, Vatrushka, Veal

Orlov, Vinegret y Zakuski[1], pero no todas las palabras tenían menciones en el corpus. Habiendo sacado las que no aparecían nunca, y modificado los platos de más de una palabra para reconocer sólo la menos común, la lista de platos quedó: (Beef) Stroganoff, Caviar, Kasha, Kissel, Knish, Pirozhki, Sorrel (soup) y (Veal) Orlov, también eliminamos Kiev por ser el nombre de una ciudad.

En segundo lugar, las palabras asociadas a la revolución rusa elegidas fueron aquellas con significado político, y exclusivo de esta época[2]. El resultado fue: comrade, glasnost, gulag, intelligentsia, perestroika, politburo, tzar, tsar (dos formas de escribir lo mismo), commissar, apparatchik, agitprop. Ninguna de estas tiene traducción directa a inglés. Las únicas dos que están en inglés son "comrade" y "tsar", pero son conceptos relativamente genéricos de la política soviética: por ejemplo gulag es una prisión para enemigos políticos; y perestroika es una política económica adoptada por el gobierno soviético. Todas se encuentran al menos una vez en el corpus.

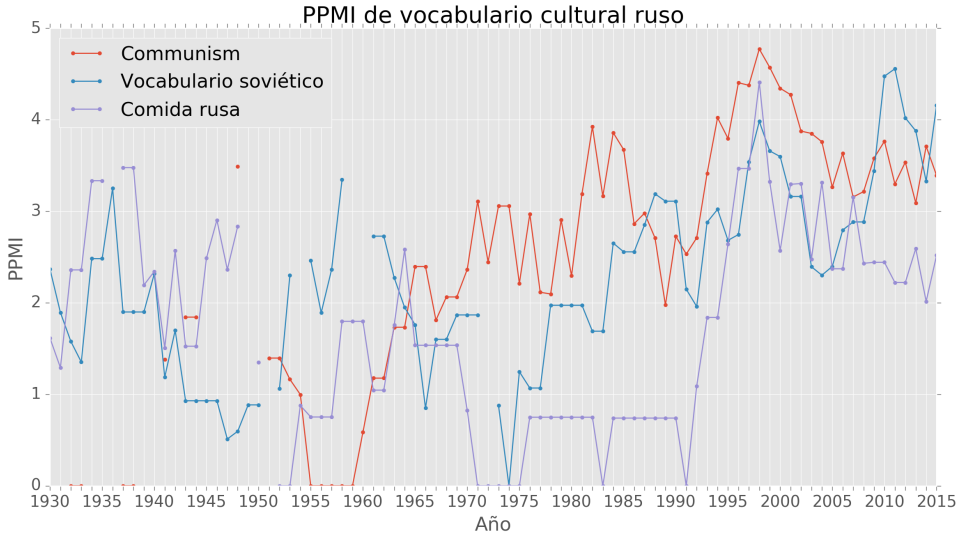


Figura 4.6: PPMI entre Rusia, el vocabulario de la Revolución Rusa y comida tradicional (smoothing 3)

Para evaluar si efectivamente estas palabras figuran en relación directa con Rusia, y comparar la fuerza de las asociaciones con aquellas que ya vimos, en la figura 4.6 utilizamos toda la lista de comidas como un concepto, la lista de palabras relacionadas a la revolución como otro concepto, y tomamos comunismo por comparación. Vemos que existen asociaciones y que las tres siguen tendencias muy parecidas a lo largo de los años.

Para indagar un poco más sobre el resultado verificamos en las figuras 4.7 y 4.8 la frecuencia de las palabras dentro de cada concepto.

En la figura 4.7 se observa la frecuencia de cada una de las comidas, las apariciones de casi todas son esporádicas, con algunas subiendo de frecuencia en años particulares (probablemente debido a una única película donde tienen varias apariciones), para luego volver a desaparecer. La única excepción a esto es "caviar" que se encuentra mucho más presente y, en promedio, tiene siempre mayor frecuencia que el resto.

[1]https://en.wikipedia.org/wiki/List_of_Russian_dishes
[2]https://www.vocabulary.com/lists/421968

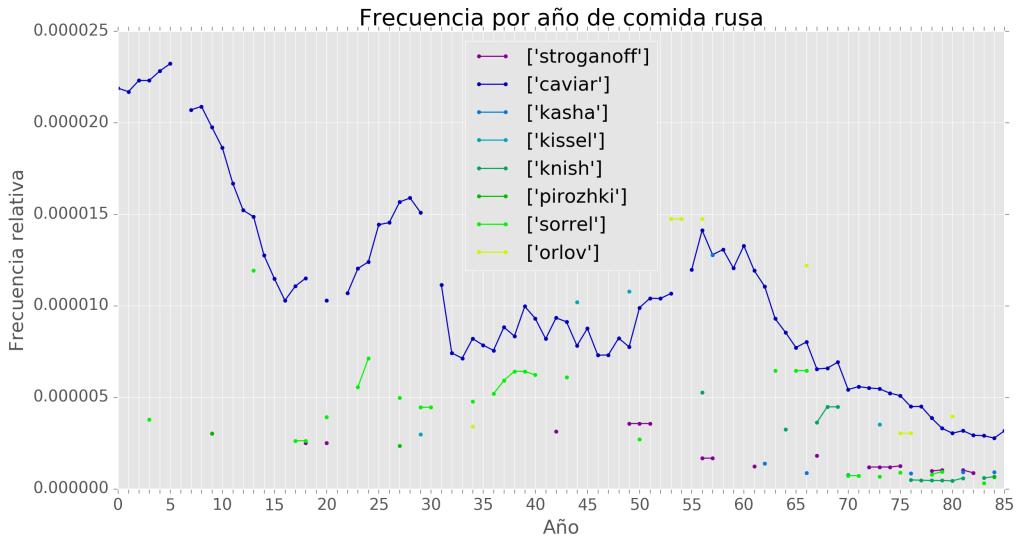


Figura 4.7: Frecuencia de cada plato en la lista de comida rusa (smoothing 3)

Observando la figura 4.8 se destaca que el fenómeno se repite con el vocabulario político: casi todas las palabras tienen apariciones escasas y esporádicas a lo largo de los años, excepto comrade que aparece consistentemente y en mucha mayor proporción.

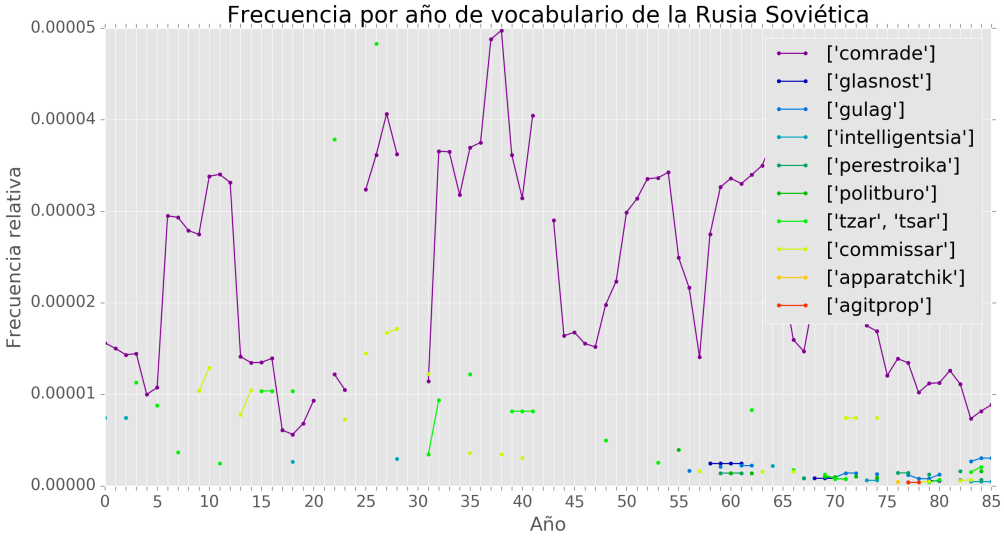


Figura 4.8: Frecuencia de cada palabra en el vocabulario de la Revolución Rusa (smoothing 3)

Frente a estos resultados volvimos a elaborar la comparación de los conceptos culturales rusos y la asociación a comunismo y el resultado se presenta en la figura 4.9.

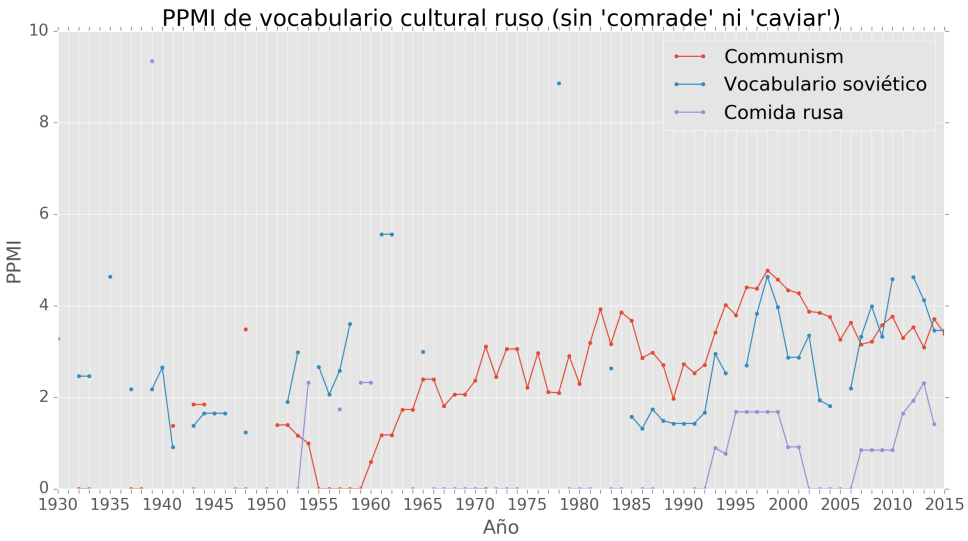


Figura 4.9: PPMI entre Rusia, el vocabulario de la Revolución Rusa y comida tradicional habiendo sacado 'caviar' y 'comrade' (smoothing 3)

Habiendo sacado caviar del concepto, las asociaciones a comidas rusas prácticamente desaparecen hasta principio de los años 90. También se ve en esta figura que el vocabulario relacionado con la revolución tiene asociaciones relevantes antes y después del momento crítico de la Guerra Fría: hay asociaciones previas a los años 60 y luego reaparecen a fin de los 80, ya llegando al final de la Guerra y la estabilización de la situación internacional. De esto podemos decir que hay evidencia de que el cine estadounidense estuvo interesado en la situación rusa, pero mientras duró el conflicto, la imagen presentada desde el entretenimiento fue más estereotípica, dado que en la figura 4.6 sí existían asociaciones relativamente fuertes: es decir, las menciones a Rusia en conexión con el comunismo sabíamos que existían, y además se presentan relaciones entre comrade y caviar, pero se ve poca profundización más allá de esa mirada.

Por otro lado, y contra lo esperado, parecería que en retrospectiva, Hollywood comenzó a profundizar e interiorizarse más con la situación real en Rusia, dado que comienza a haber asociaciones con vocabulario mucho más específico y cercano a ocurrencias históricas.

4.3 Conclusiones

Los resultados de esta sección mostraron que aquellos estereotipos esperados efectivamente tienen presencia en el cine, y las asociaciones son fuertes. Pero no encontramos evidencia de los estereotipos secundarios: el país empobrecido, y la moda de las novias por encargo.

Además, observamos vocabulario más cercano a la cultura rusa, y extrajimos que una vez finalizado el conflicto, el cine estadounidense parece haberse adentrado en el análisis histórico de la Guerra Fría desde el punto de vista ruso, pero no mientras el conflicto estaba sucediendo.

Con respecto a las metodologías, al usar PPMI nos encontramos con las limitaciones que propone la escasez de este vocabulario. Pero también con la limitación que ofrece la herramienta que estamos usando para compensarlo: cuando agregamos los resultados de varias palabras entendiéndolas como un único concepto, es posible que alguna de esas palabras esté sumando un significado que las demás no tienen, o que no coincida con la semántica buscada. En particular, el agregado del vocabulario de

revolución rusa nos estaba ocultando que había una palabra con mucho más peso que las demás, y eso puede generar, en este caso y en otros más sutiles, una interpretación menos exacta del resultado.

Capítulo 5

Conclusiones

5.1 Conclusiones

A partir de las investigaciones realizadas para este trabajo, vemos que la búsqueda de tendencias culturales, estereotipos y lugares comunes en el cine es realizable con técnicas conocidas de análisis de texto. Estos métodos ya existentes, usados y probados en otros corpus pudieron ser aplicados en el nuestro, y en general, logramos reproducir resultados existentes parcial o totalmente.

Viendo lo estudiado en los capítulos referentes a Rusia y terrorismo, nos parece que para captar el significado de un término en un cierto momento histórico es necesario aplicar herramientas que no tengan los sesgos del día de hoy: el PPMI como herramientas de similaridad semántica funcionó mucho mejor en relaciones históricas que el word2vec entrenado con un corpus reciente; aún con las falencias que tiene el PPMI para palabras con pocas apariciones.

Queremos destacar la importancia de no desestimar los sesgos existentes en cualquier corpus de entrenamiento. Si el foco de estudio es ese corpus, herramientas como word2vec son una forma de hacer un foco interesante, pero el análisis de un texto a partir de un corpus externo, por amplio y completo que sea, viene con su propio conjunto de preconceptos cuyo aislamiento es un problema completamente distinto.

Este corpus cuenta con una dimensión más allá del texto: la temporalidad, utilizar ésta como delimitador de ventana permitió detectar aquellos sesgos que estábamos buscando. Esta innovación en sí misma es un dato de interés para seguir involucrando en la investigación del cine.

Las herramientas utilizadas resultaron bastante poderosas para la investigación de este corpus, pero es necesario que quién las utilice tenga conocimiento en los campos a estudiar para poder generar resultados nuevos, y realizar un análisis más profundo sobre las asociaciones resultantes.

5.2 Trabajo futuro

A partir de este trabajo vemos que hay amplias posibilidades de continuar explorando tendencias en el cine. En particular una continuación lógica sería aplicar estas herramientas más a fondo:

- Utilizar LSA como continuación a los análisis de PPMI, para profundizar en asociaciones sin necesidad de coocurrencia, pero manteniendo el aislamiento histórico
- Entrenar word2vec con este corpus, para buscar directamente la similaridad entre vectores, en lugar de a través de un entrenamiento externo con sus propios sesgos
- Agregar al corpus etiquetas sintácticas para evaluar mejor los casos de uso de cada término

- Extraer películas cuyos sesgos sean particularmente fuertes y aislar los diálogos que generan ese efecto

Por otro lado, el uso de un ventana de 10 segundos resultó efectivo en este trabajo, pero para poder saber realmente qué longitud representa el largo de una conversación sobre un tema habría que investigar más a fondo con los tamaños de ventana sobre un corpus con anotaciones pertinentes.

Por último, sería interesante explorar cómo y cuánto afectan estos sesgos y estereotipos que se detectaron en el corpus a cada persona. A partir de la extracción de escenas particularmente estereotípicas se podría evaluar cuál es el impacto que genera consumir estas ideas a través de una película.

Bibliografía

Altszyler, E., Sigman, M., y Slezak, D. F. (2016). Comparative study of lsa vs word2vec embeddings in small corpora: a case study in dreams database. *arXiv preprint arXiv:1610.01520*.

Bem, S. L. (1979). Theory and measurement of androgyny: A reply to the pedhazur-tetenbaum and locksley-colten critiques.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., y Kalai, A. (2016). Quantifying and reducing stereotypes in word embeddings. *arXiv preprint arXiv:1606.06121*.

Boyeva-Omelechko, N. y Posternyak, K. (2016). The image of russia in the british everyday discourse. *SCIEURO*, (1):96–101.

Cejka, M. A. y Eagly, A. H. (1999). Gender-stereotypic images of occupations correspond to the sex segregation of employment. *Personality and social psychology bulletin*, 25(4):413–423.

Charlesworth, A. y Glantz, S. A. (2005). Smoking in the movies increases adolescent smoking: A review. *Pediatrics*, 116(6):1516–1528.

Church, K. W. y Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Cochrane, K. (2013). *All the rebel women: The rise of the fourth wave of feminism*, volumen 8. Guardian Books.

Diuk, C. G., Slezak, D. F., Raskovsky, I., Sigman, M., y Cecchi, G. A. (2012). A quantitative philology of introspection. *Frontiers in integrative neuroscience*, 6.

Donnelly, K., Twenge, J. M., Clark, M. A., Shaikh, S. K., Beiler-May, A., y Carter, N. T. (2016). Attitudes toward women's work and family roles in the united states, 1976–2013. *Psychology of Women Quarterly*, 40(1):41–54.

Dunn, E. W., Moore, M., y Nosek, B. A. (2005). The war of the words: How linguistic differences in reporting shape perceptions of terrorism. *Analyses of social issues and public policy*, 5(1):67–86.

Fedorov, A. (2013a). The image of russia on the western screen: the present stage (1992-2013). *European researcher. Series A*, (4-3):1051–1064.

Fedorov, A. (2013b). Western audiovisual stereotypes of russian image: the ideological confrontation epoch (1946-1991). *European researcher. Series A*, (5-4):1565–1580.

Fisher, R. A. (1922). On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94.

Grayson, S., Mulvany, M., Wade, K., Meaney, G., y Greene, D. (2017). Exploring the role of gender in 19th century fiction through the lens of word embeddings. En *International Conference on Language, Data and Knowledge*, pp. 358–364. Springer.

Hamilton, W. L., Leskovec, J., y Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Jurafsky, D. y Martin, J. H. (2014). *Speech and language processing*, volumen 3. Pearson London.

Kulkarni, V., Al-Rfou, R., Perozzi, B., y Skiena, S. (2015). Statistically significant detection of linguistic change. En *Proceedings of the 24th International Conference on World Wide Web*, pp. 625–635. International World Wide Web Conferences Steering Committee.

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

Landauer, T. K., Foltz, P. W., y Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Lenton, A. P., Sedikides, C., y Bruder, M. (2009). A latent semantic analysis of gender stereotype-consistency and narrowness in american english. *Sex Roles*, 60(3-4):269–278.

Lichtblau, E. (2015). Crimes against muslim americans and mosques rise sharply.

Linz, D., Donnerstein, E., y Penrod, S. (1984). The effects of multiple exposures to filmed violence against women. *Journal of Communication*, 34(3):130–147.

Madi, M. (2017). Reality check: Is islamophobia on the rise?

Metzl, J. M. y MacLeish, K. T. (2015). Mental illness, mass shootings, and the politics of american firearms. *Journal Information*, 105(2).

Michel, Jean-Baptiste and Shen, Yuan Kui and Aiden, Aviva Presser and Veres, Adrian and Gray, Matthew K and Pickett, Joseph P and Hoiberg, Dale and Clancy, Dan and Norvig, Peter and Orwant, Jon and others (2011). Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Mikolov, T., Chen, K., Corrado, G., y Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nurullah, A. S. (2010). Portrayal of muslims in the media:"24" and the 'othering'process. *International Journal of Human Sciences*, 7(1):1020–1046.

Ouyang, X., Zhou, P., Li, C. H., y Liu, L. (2015). Sentiment analysis using convolutional neural network. En *Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM), 2015 IEEE International Conference on*, pp. 2359–2364. IEEE.

Ramakrishna, A., Malandrakis, N., Staruk, E., y Narayanan, S. S. (2015). A quantitative analysis of gender differences in movies using psycholinguistic normatives. En *EMNLP*, pp. 1996–2001.

Riegler, T. (2010). Through the lenses of hollywood: depictions of terrorism in american movies. *Perspectives on Terrorism*, 4(2).

Sagi, E., Diermeier, D., y Kaufmann, S. (2013). Identifying issue frames in text. *PLoS one*, 8(7):e69185.

Sendén, M. G., Sikström, S., y Lindholm, T. (2015). "she" and "he" in news media messages: pronoun use reflects gender biases in semantic contexts. *Sex Roles*, 72(1-2):40–49.

Silva, D. (2017). The othering of muslims: Discourses of radicalization in the new york times, 1969–2014. En *Sociological Forum*, volumen 32, pp. 138–161. Wiley Online Library.

Steuter, E. y Wills, D. (2009). Discourses of dehumanization: Enemy construction and canadian media complicity in the framing of the war on terror. *Global Media Journal: Canadian Edition*, 2(2).

Team, B. I. (2015). New study analyzes media coverage of islam over time.

Turney, P. D. y Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Twenge, J. M., Campbell, W. K., y Gentile, B. (2012). Male and female pronoun use in us books reflects women's status, 1900–2008. *Sex roles*, 67(9-10):488–493.

van Erp, M. y Vossen, P. (2016). Entity typing using distributional semantics and dbpedia. En *International Semantic Web Conference*, pp. 102–118. Springer.

Wijaya, D. T. y Yeniterzi, R. (2011). Understanding semantic change of words over centuries. En *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web*, pp. 35–40. ACM.

Apéndice A

Vocabulario por género

A.1 Vocabulario del BSRI (Bem, 1979)

A.1.1 Asignado a femineidad

- yielding
- cheerful
- shy
- affectionate
- flatterable
- loyal
- feminine (no incluído)
- sympathetic
- sensitive
- understanding
- compassionate
- eager to soothe hurt feelings (no incluído porque no hay traducción de una palabra)
- soft
- warm
- tender
- gullible
- childlike
- sweet (insertado en lugar de: does not use harsh language)
- loving
- gentle

A.1.2 Asignado a masculinidad

- selfreliant
- defends own beliefs (no incluído porque no hay traducción de una palabra)
- independent
- athletic
- assertive
- strong
- forceful
- analytical
- leadership ability (reemplazado por leader más abajo)
- brave (insertado en lugar de: willing to take risks)
- decisive
- selfsufficient
- dominant
- masculine (no incluído)
- assured (insertado en lugar de: willing to take a stand)
- aggressive
- leader (insertado en lugar de: acts as a leader)
- individualistic
- competitive
- ambitious

A.1.3 Asignado a neutral

- helpful
- moody
- conscientious
- theatrical
- happy
- unpredictable
- reliable
- jealous
- truthful

- secretive
- sincere
- conceited
- likable
- solemn
- friendly
- inefficient
- adaptable
- unsystematic
- tactful
- conventional

A.2 Vocabulario de roles en Lenton *et al.* (2009)

A.2.1 Asignados a femineidad

- beautician
- caregiver
- cheerleader
- dancer
- decorator
- designer
- dietician
- florist
- hairdresser
- homemaker
- housekeeper
- model
- nanny
- nurse
- receptionist
- stylist
- typist

A.2.2 Asignados a masculinidad

- architect
- carpenter
- coach
- contractor
- detective
- electrician
- engineer
- farmer
- firefighter
- gambler
- inventor
- machinist
- mechanic
- officer
- physicist
- pilot
- programmer
- rancher
- sheriff
- soldier

A.2.3 Asignados como neutrales

- assistant
- cashier
- clerk
- doctor
- editor
- lawyer
- poet
- reporter
- servant
- worker

Apéndice B

Tamaños de ventana

B.1 Ejemplo de PPMI de contextos de pronombres de ambos géneros y atributos femeninos según Bem (1979)

En este ejemplo vemos que la curva de pronombres masculinos se suaviza con el aumento de la ventana, no así la femenina. En líneas generales, las tendencias son las mismas más allá del tamaño de ventana. En las secciones anteriores, todos los gráficos utilizan una ventana de 10 segundos con la intención de no perder coocurrencias interesantes achicando demasiado la ventana, pero no tenemos evidencia específica de que sea el tamaño ideal.

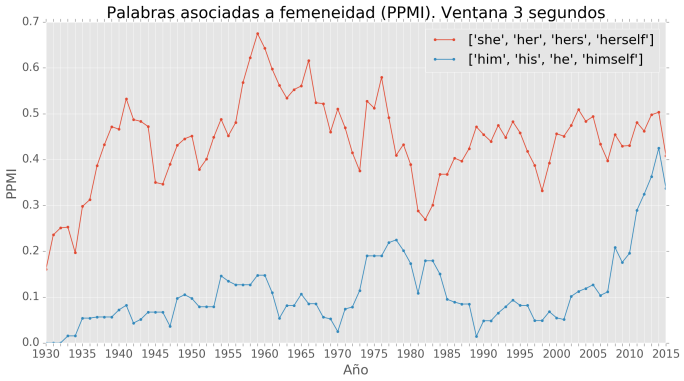


Figura B.1: Ventana de 3 segundos

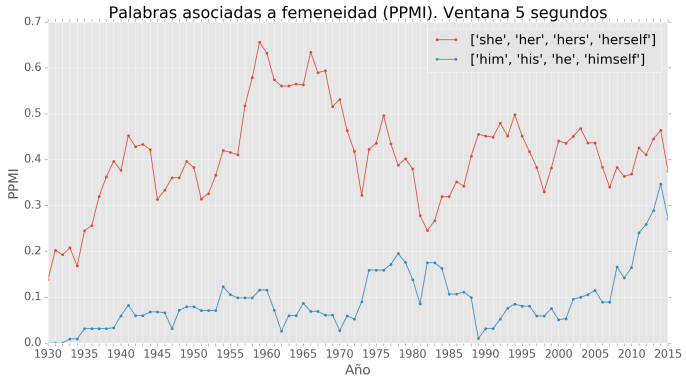


Figura B.2: Ventana de 5 segundos

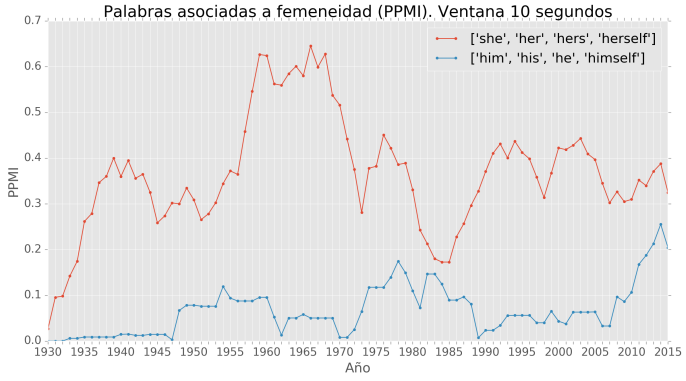


Figura B.3: Ventana de 10 segundos

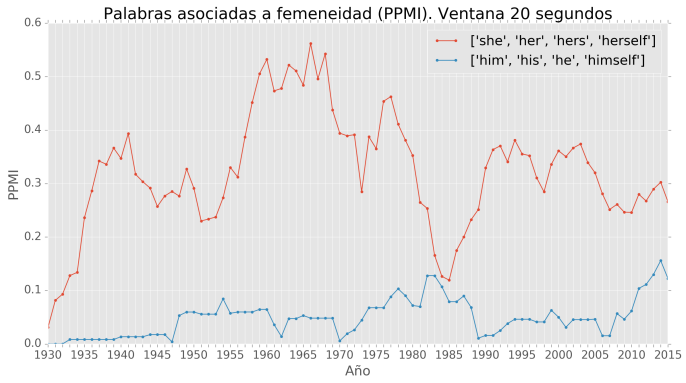


Figura B.4: Ventana de 20 segundos

Apéndice C

PPMI de vocabulario con escasas coocurrencias

C.1 PPMI de vocabulario deshumanizador en conjunción con terrorismo e islam

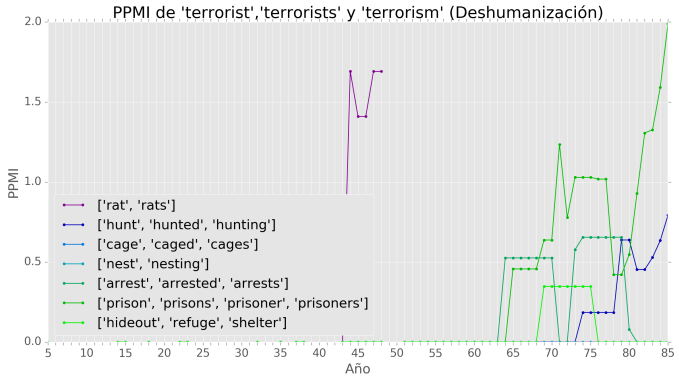


Figura C.1: PPMI entre contextos de terrorista y lenguaje deshumanizador visto en Steuter y Wills (2009)

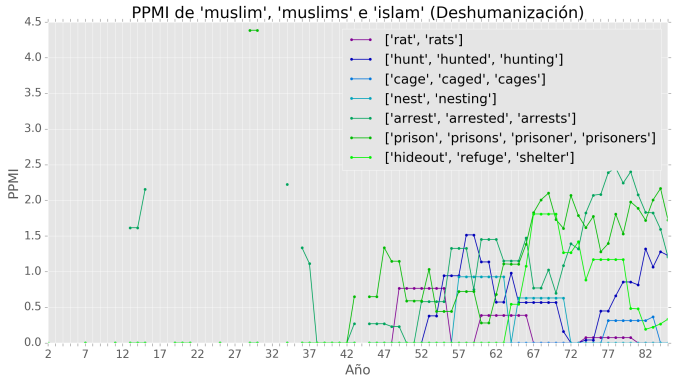


Figura C.2: PPMI entre contextos de terrorista y lenguaje deshumanizador visto en Steuter y Wills (2009)