Soñando por un tópico

Grupo 4 - Recuperación y Minería de Texto

Miguel Guerrero, Juan Knebel, Ignacio Chiapella, Estefania De Marzio, Vanesa Copa, Susana Escudero

Universidad de Buenos Aires

June 9, 2020

Contenido

Contenido

Corpus

- Colección de sueños provistos por DreamBank (http://www.dreambank.net/).
- Dividido en 89 grupos de soñadores.
- 43132 sueños, de los cuales 36000 aproximadamente están en inglés que serán los utilizados para este trabajo.
- 7011558 palabras totales.
- 63% pertenecen a soñadores femeninos.
- 44% pertenecen a series de sueños de un mismo individuo.
- 20% pertenecen a grupos de jóvenes entre 18 y 30 años.

Segmentación del corpus

- Realizaremos un análisis sobre dos segmentos diferentes:
  - Género: Masculino / Femenino.
  - Rango Etario: Niños (6-12), Adolescentes (13-18), Adulto Joven (19-29), Adulto (30-49) y Mediana Edad (50-69).
- Compararemos series de sueños de dos individuos:
  - Phil: Profesor de humanidades retirado.
  - Vietnam Vet: Médico de combate en Vietnam y Camboya entre 1969 y 1970 que sufrió de PTSD (Trastorno de estrés postraumático).

Contenido

Qué buscamos

- Descubrir cuáles son los temas en los sueños.
- Analizar cuáles son los temas que predominan según segmentación planteada.
- Analizar cuáles son los sentimientos que predominan según segmentación planteada.
- Analizar si existen sueños recurrentes y diferencias para las series de tiempo de dos individuos.

Qué esperamos

- Segmentación por Rango Etario se espera que:
  - Niños sueñen con temas relacionados con la familia y con temáticas más positivas.
  - Adolescentes con temas de escuela, amor, amigos y con temáticas más positivas..
  - Adulto Joven: deportes, películas, amor, amigos y con una proporción similar de emociones positivas y negativas.
  - Adulto sobre temas de trabajo, familia con una tendencia similar de emociones positivas y negativas.
  - Adulto Media Edad acerca de temáticas relacionadas con pérdidas, familia y se obtenga una proporción similar de emociones positivas y negativas.

Qué esperamos

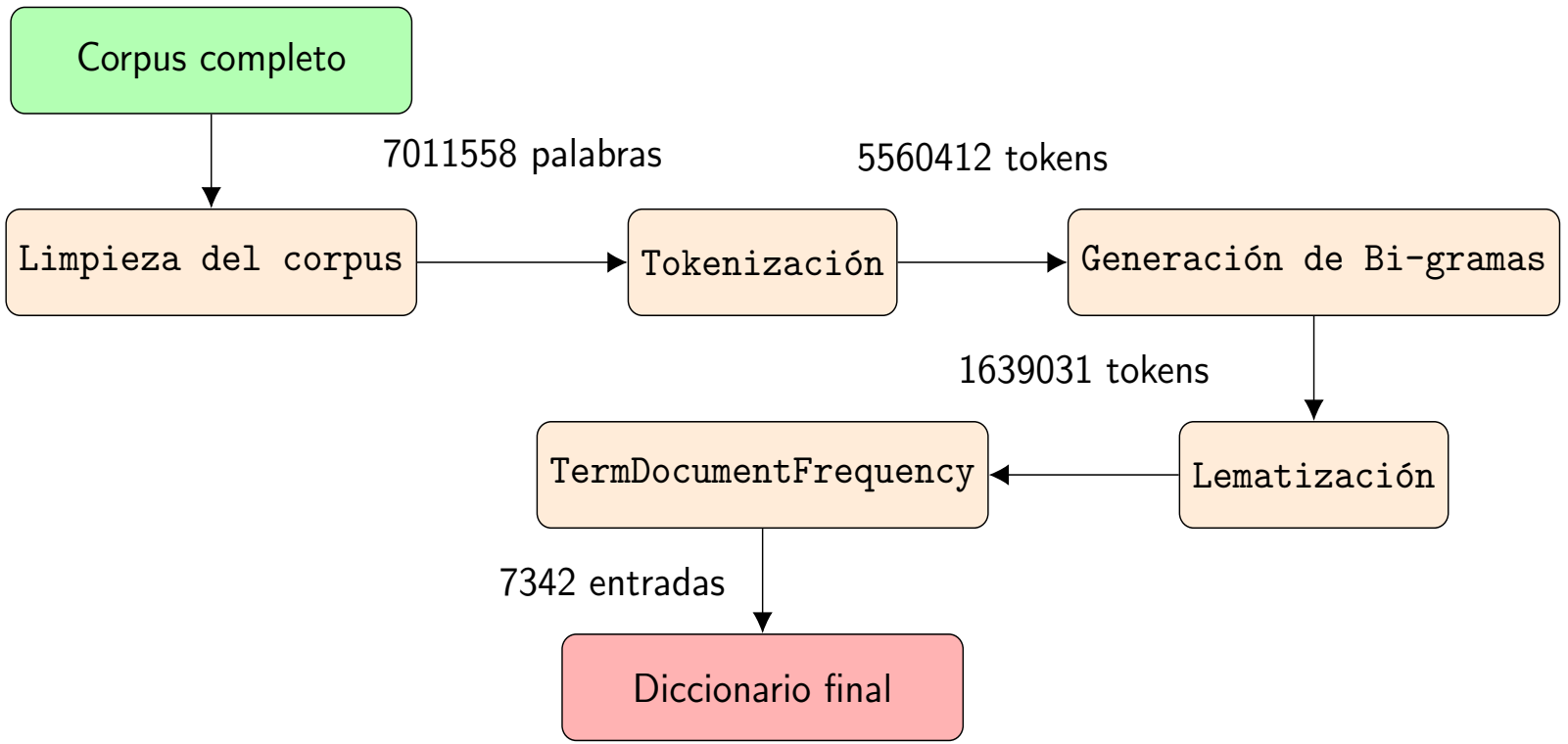
- Phil y Vietnam Vet:
  - La proporción de los tópicos del soñador de Vietnam estén mayormente distribuidos en temas relacionados con pérdida o tristeza, y no en temas más generales.
  - Phil presente sueños relacionados con su actividad, y vida familiar.
  - La presencia de sueños con sesgo negativo sea mayor en Veterano que en Phil.

Contenido

Limpieza de Corpus

- Eliminación de los signos de puntuación, stopwords, palabras de menos de 3 caracteres.
- Transformación del corpus a minúscula.
- Se utilizarán sólo los tokens que aparecen en al menos 10 sueños y que no se repitan en más del 50% de todo el corpus
- Se eliminaron palabras que no fueron significativas como por ejemplo: 'like', 'say', 'remember', 'dream', 'think', 'know', 'could', 'go'.

Proceso


Corpus completo
7011558 palabras
5560412 tokens
Limpieza del corpus
Tokenización
Generación de Bi-gramas
1639031 tokens
TermDocumentFrequency
Lematización
7342 entradas
Diccionario final

Frecuencia de palabras

Table: Top 10 palabras más frecuentes

Palabras	Cantidad
i	2009
and	1371
the	1297
a	1117
to	1089
it	828
s	739
of	631
in	499
m	481

Antes del preprocesamiento

Palabras	Cantidad
like	597
something	517
back	415
know	391
think	372
people	349
going	301
sort	283
around	274
place	273

Después del preprocesamiento

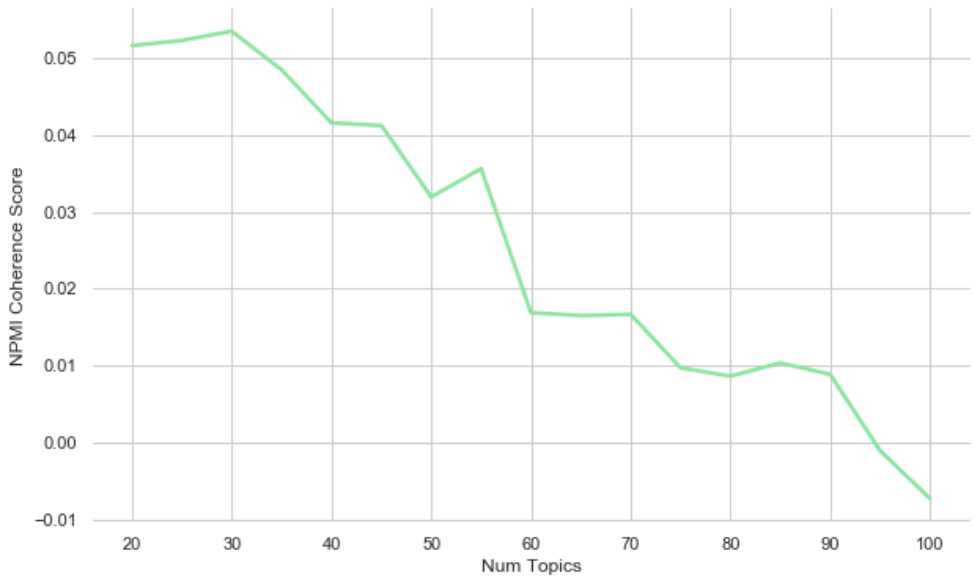
LDA

Se utilizó el modelo **LDA** para el descubrimiento de tópicos

- Se probó entre 20, 25, 30, ..., 100 tópicos.
- Durante el entrenamiento se hicieron 10 pasadas sobre el corpus completo.
- Comparando entre 1000 sueños al menos 1000 veces.
- Para cada una de las pruebas se calculó la coherencia media de los tópicos utilizando:
  - la medida NPMI,
  - considerando sólo 10 keywords,
  - una ventana de vecindad de 30 palabras.

Evaluación de los Tópicos

Para evaluar los distintos modelos de **LDA** utilizamos la medida **Topic Coherence**. Éste mide si las palabras en un tópico tienden a coexistir juntas asignándole un score.
Seleccionamos el modelo de mayor *score de coherencia* en promedio en todos los tópicos.



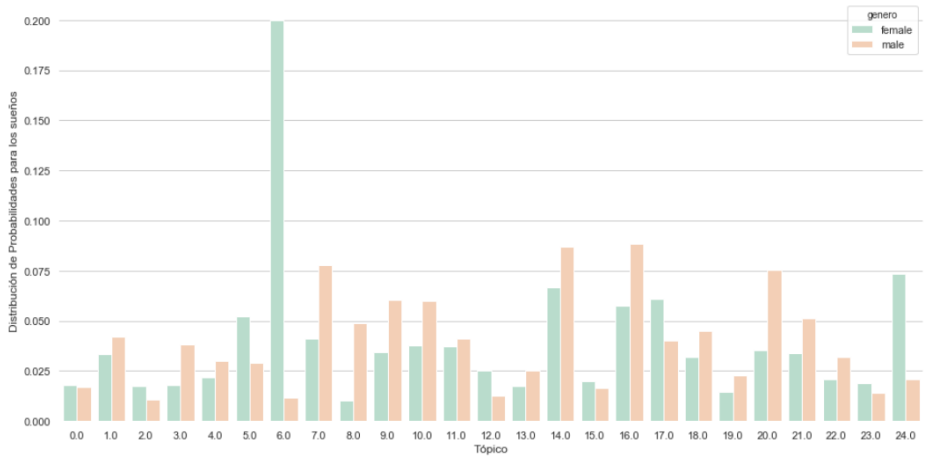
Tópicos descubiertos

Número de topico	Coherence	Palabras	Etiqueta
22	0.14	wear, clothe, bathroom, shirt, shoe, dress, toilet, pant, clean, pair	Vestimenta
7	0.10	car, run, drive, road, plane, gun, fly, shoot, stop, hit	Accidente
15	0.10	play, stage, music, game, show, name, audience, write, piano, singe	Entretenimiento
17	0.09	room, door, bed, open, sleep, leave, window, lock, bedroom, find	Dormitorio
23	0.08	work, paper, computer, write, picture, different, word, information, page, use	Trabajo
20	0.08	car, drive, home, walk, way, road, truck, turn, street, right	Transporte
6	0.07	mom, guy, girl, call, stuff, shop, school, place, later, friend	Amigos, escuela, adolescencia
4	0.07	water, pool, fire, boat, swim, wave, walk, rock, little, big	Deporte acuático
9	0.06	run, tree, horse, jump, fall, dog, big, throw, ball, climb	Deporte al aire libre
12	0.06	phone, call, work, photo, office, find, computer, desk, time, number	Oficina
5	0.06	class, school, sit, teacher, student, room, next, walk, time, give	Colegio
21	0.04	man, walk, woman, hand, kiss, dance, young, old, away, arm	Romance
16	0.04	girl, friend, mother, boy, sister, little, house, time, child, dress	Familia
24	0.04	kind, really, little, stuff, time, wake, bus, sort, sound, sit	Undefined
3	0.03	friend, drink, walk, wedding, eat, bottle, snow, party, dress, find	Fiesta, casamiento
18	0.03	die, kill, dead, cry, happen, brother, man, time, leave, feel	Pérdida, muerte
14	0.02	feel, woman, sit, time, man, answer, question, work, room, give	Relaciones interpersonales
1	0.02	sit, leave, walk, realize, door, room, stand, table, hand, right	En movimiento
0	0.02	work, use, fish, job, test, need, also, different, time, ice	Undefined
13	0.02	baby, walk, sit, woman, stand, table, small, find, man, right	Bebes
8	0.01	man, light, small, large, group, team, building, long, bird, field	Trabajo
11	0.01	leave, feel, mother, find, woman, man, wait, time, help, walk	Relaciones interpersonales
19	0.01	train, boy, chicken, tooth, eat, buy, little, money, work, flower	Amigos, escuela, adolescencia
10	0.01	room, walk, store, time, woman, man, large, house, building, move	En movimiento
2	0.001	movie, doctor, watch, group, film, episode, nurse, show, woman, move	Películas

Distribución de Tópicos por Género

Distribución de los sueños de las Mujeres:

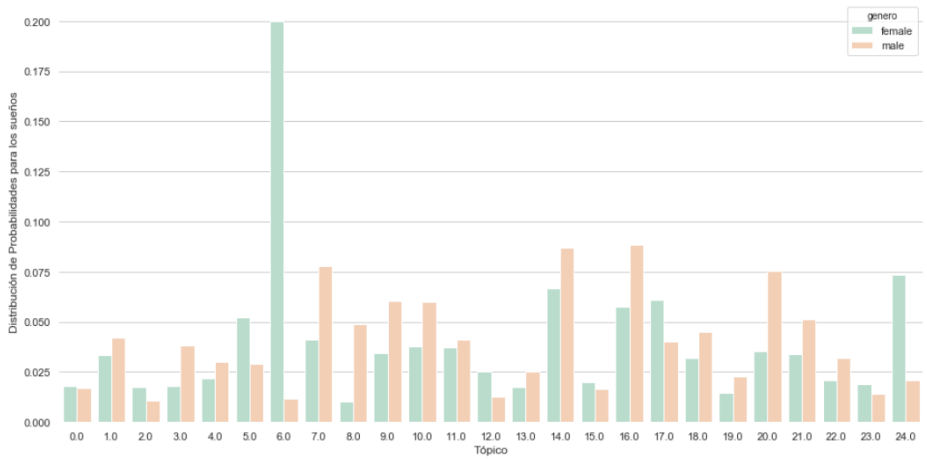
1. 20% en **Amigos/Escuela/ Adolescencia** (6).
2. 7,5% en **Sin definir** (24).
3. 6% para **Relaciones Interpersonales** (14).



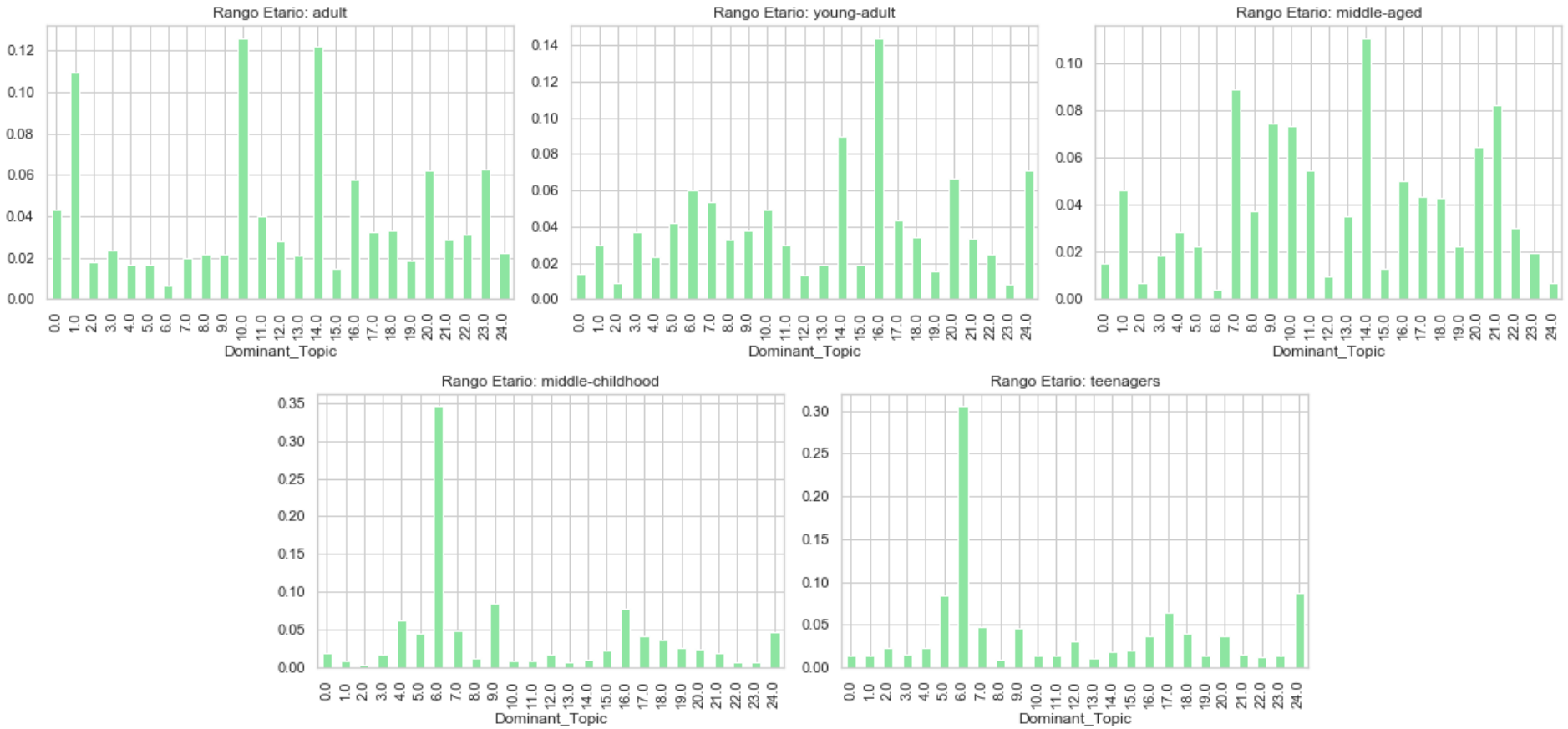
Distribución de Tópicos por Género

Distribución de los sueños de los Hombres:

1. 8-9% en **Relaciones Interpersonales** (14) y **Familia** (16).
2. 7.5% en **Accidente** (7).



Distribución de Tópicos por Rango Etario



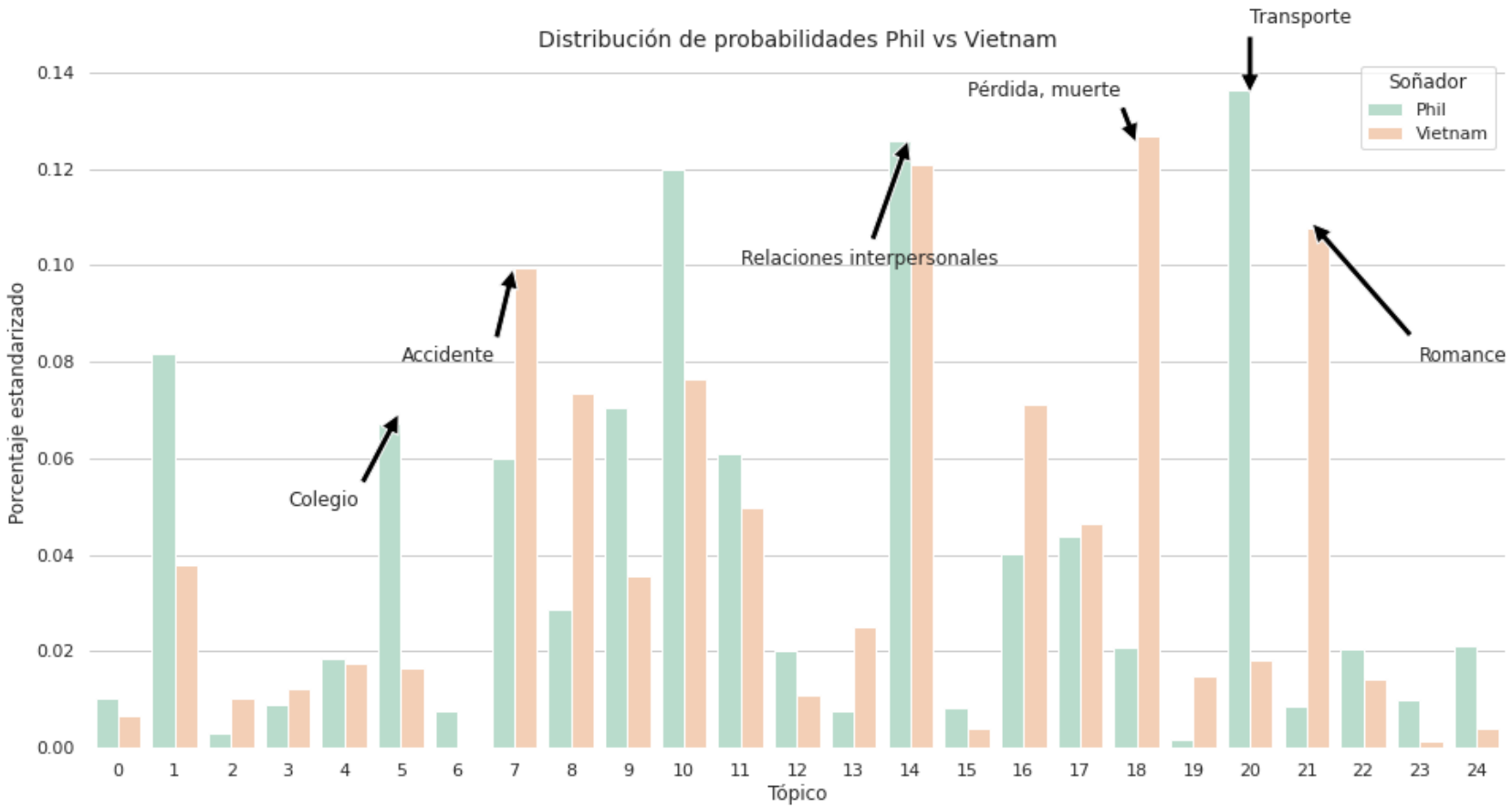
Distribución de Tópicos por Rango Etario

Se analizaron cerca de 19 mil sueños, ya que se dejaron fuera los sueños etiquetados como series (por pertenecer a diferentes grupos etarios) y los sueños registrados por personas al azar.

Distribución de Tópicos por Rango Etario

- *Niños*: Los sueños se distribuyen en un 35% en el Tópico **Amigos/Escuela/Adolescencia** (6).
- *Adolescentes*: se distribuyen en un 30% en el Tópico **Amigos/Escuela/Adolescencia** (6).
- *Adulto Joven*: se distribuyen en un 14% en el Tópico **Familia** (16).
- *Adulto*: se distribuyen en más de un 12% en el Tópico **En Movimiento** (10) y en **Relaciones Interpersonales** 14.
- *Adulto Media Edad*: se distribuyen en más de un 10% en el Tópico **Relaciones Interpersonales** (14) y en más de un 8% en **Accidente** (7).

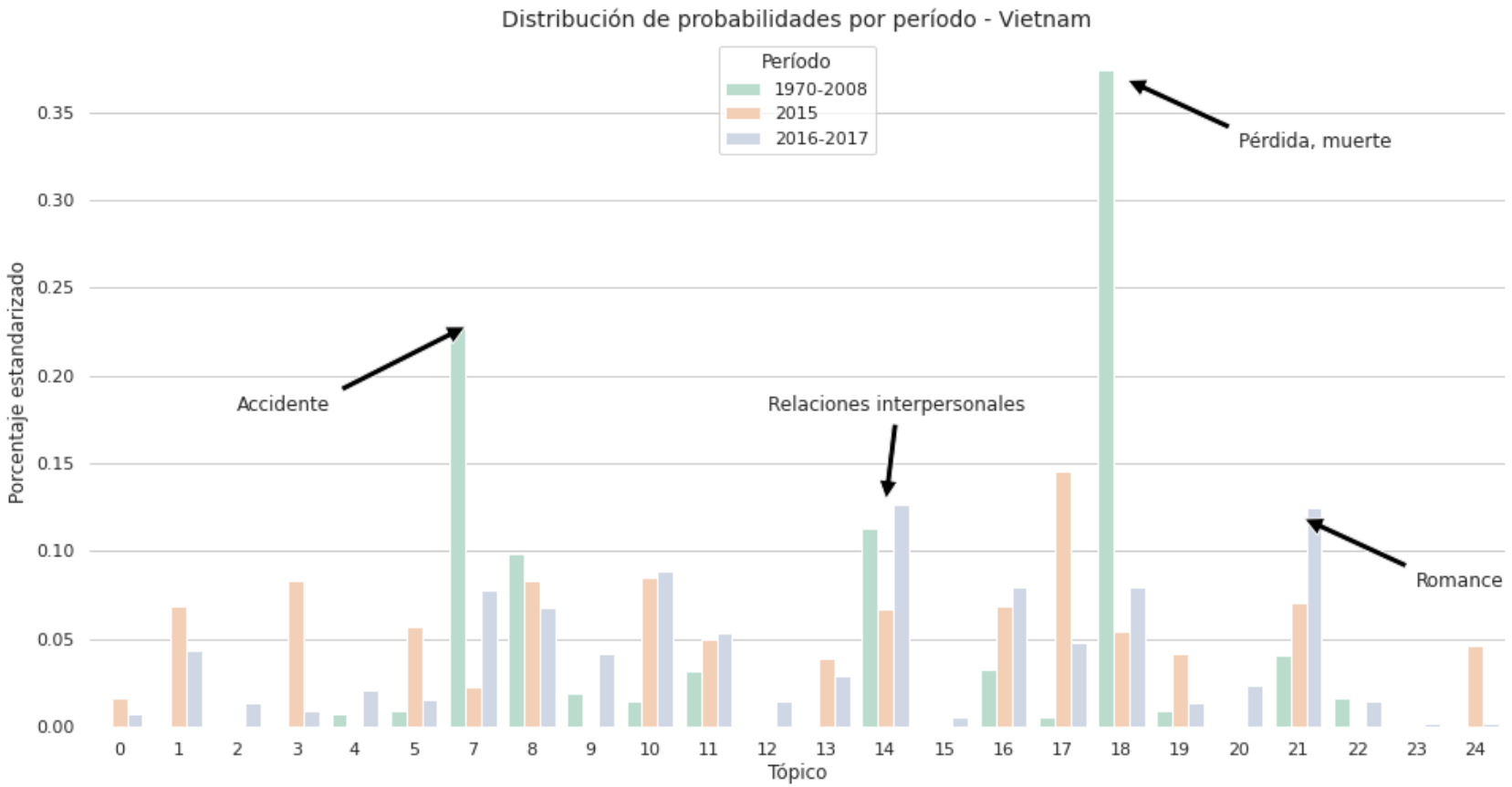
Cómo sueñan Phil y Vietnam Vet



Como sueñan Phil y Vietnam Vet

- En los sueños del veterano de Vietnam se puede ver una marcada diferencia en los tópicos relacionados con *Pérdida*, *Accidente* y *Romance*.
- En los sueños de Phil se ve una clara predominancia del tópico catalogado como *Transporte* y *Amigos*.

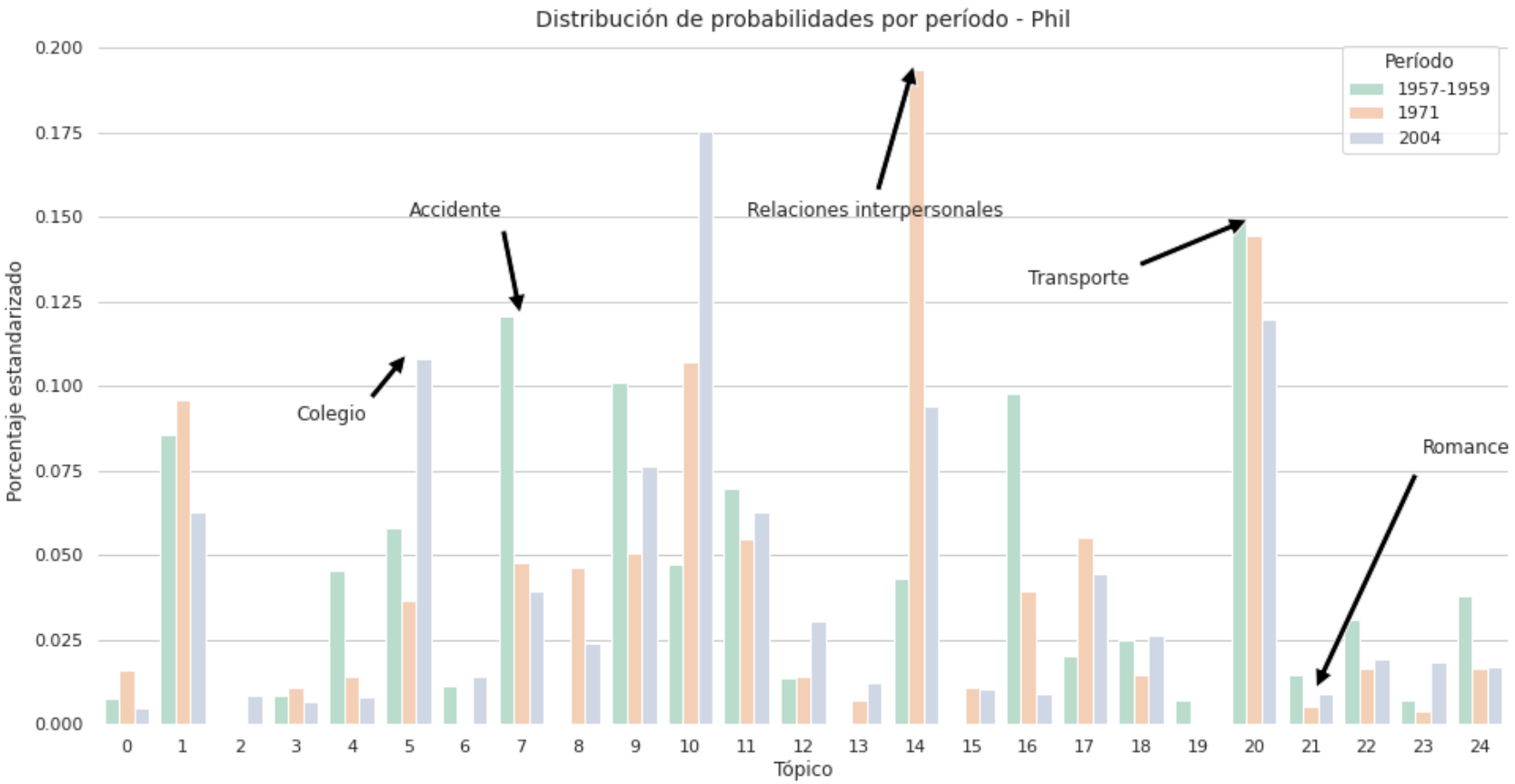
Vietnam Vet y sus etapas



Vietnam Vet y sus etapas

1. En primer lugar hay que notar que el tópico *Amigos, escuela* tiene probabilidad 0 en las 3 etapas de sueños reportados, y el catalogado como *Colegio* tiene muy baja probabilidad.
2. Los tópicos de *Accidente* y *Pérdida, muerte* tiene una alta probabilidad en el primer período reportado, para luego en los subsiguientes disminuir notoriamente.
3. Es interesante ver como el tópico *Romance* incrementa a través de los períodos de forma significativa.

Phil y sus etapas



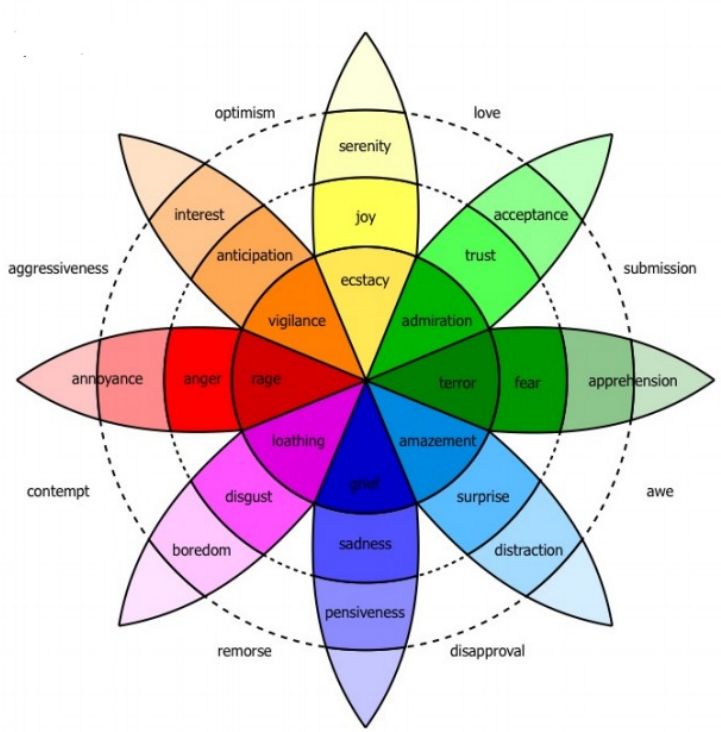
Phil y sus etapas

1. Se puede observar una alta predominancia del tópico *Relaciones interpersonales* en su primer período de sueños.
2. El tópico *Transporte* prácticamente no cambia a lo largo del tiempo.
3. A excepción de algunos pocos tópicos, todos se encuentran con algún porcentaje de participación en los tres períodos reportados.

Contenido

Inducción

- Pre procesamiento
  - Pasar todo el corpus a minúsculas
  - Eliminar caracteres especiales



- Trust
- Joy
- Fear
- Sadness
- Anger
- Anticipation
- Disgust
- Surprise

Lexicones

Word	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	positive	negative
reward	0	1	0	0	1	0	1	1	1	0
worry	0	1	0	1	0	1	0	0	0	1
tenderness	0	0	0	0	1	0	0	0	1	0
sweetheart	0	1	0	0	1	1	0	1	1	0
suddenly	0	0	0	0	0	0	1	0	0	0
thirst	0	1	0	0	0	1	1	0	0	0
garbage	0	0	1	0	0	0	0	0	0	1

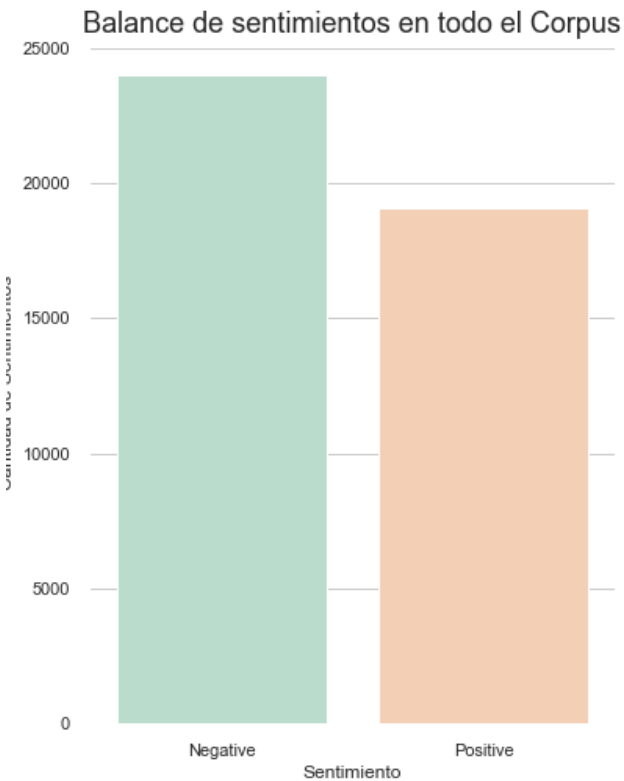
Anger		Fear		Joy		Sadness	
outraged	0.964	horror	0.923	superb	0.864	sad	0.844
violence	0.742	anguish	0.703	cheered	0.773	guilt	0.750
coup	0.578	pestilence	0.625	rainbow	0.531	unkind	0.547
oust	0.484	stressed	0.531	gesture	0.387	difficulties	0.421
suspicious	0.484	failing	0.531	warms	0.391	beggar	0.422
nurture	0.059	confident	0.094	hardship	.031	sing	0.017

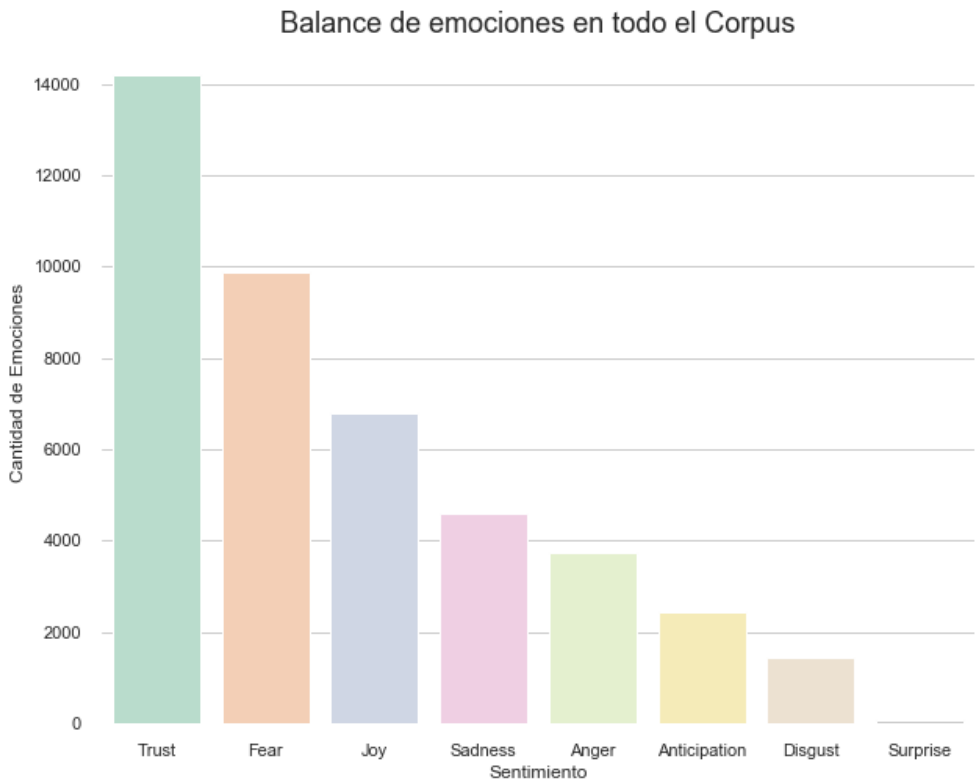
$$f^+ = \sum_{w \text{ s.t. } w \in positivelexicon} \theta_w^+ count(w)$$

$$f^- = \sum_{w \text{ s.t. } w \in negativelexicon} \theta_w^- count(w)$$

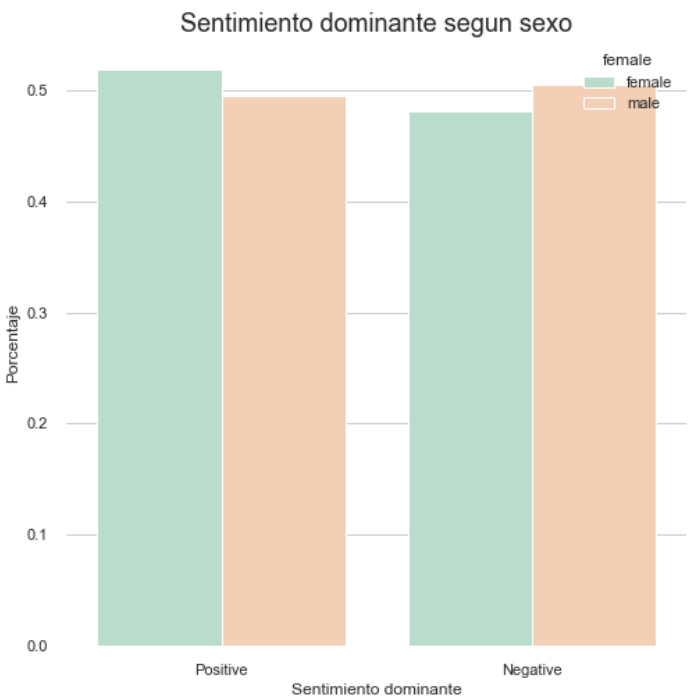
$$sentiment = \begin{cases} + & \text{if } \frac{f^+}{f^-} > \lambda \\ - & \text{if } \frac{f^-}{f^+} > \lambda \\ 0 & \text{otherwise.} \end{cases}$$

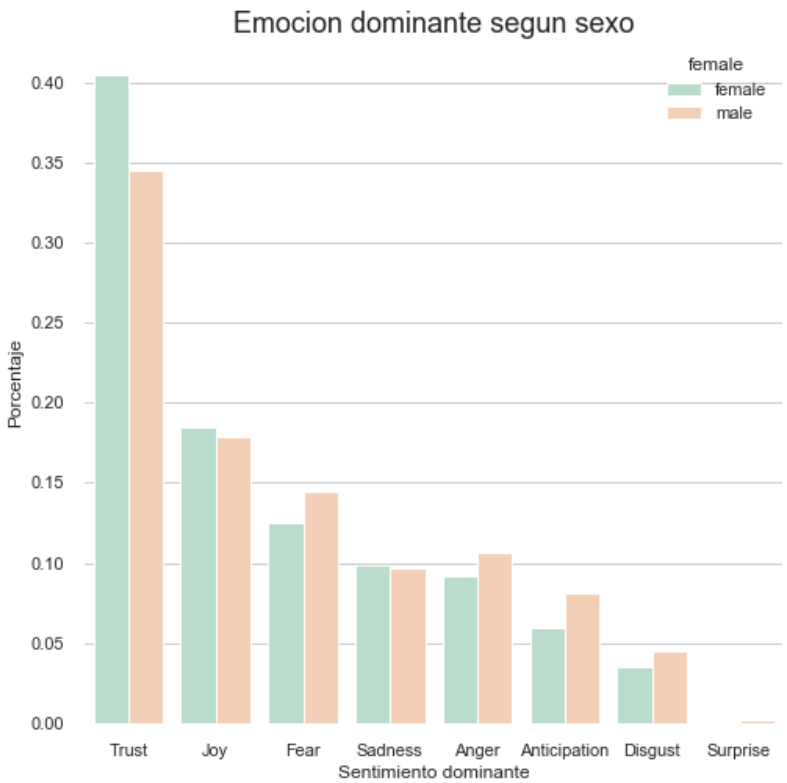
Análisis en todo el corpus



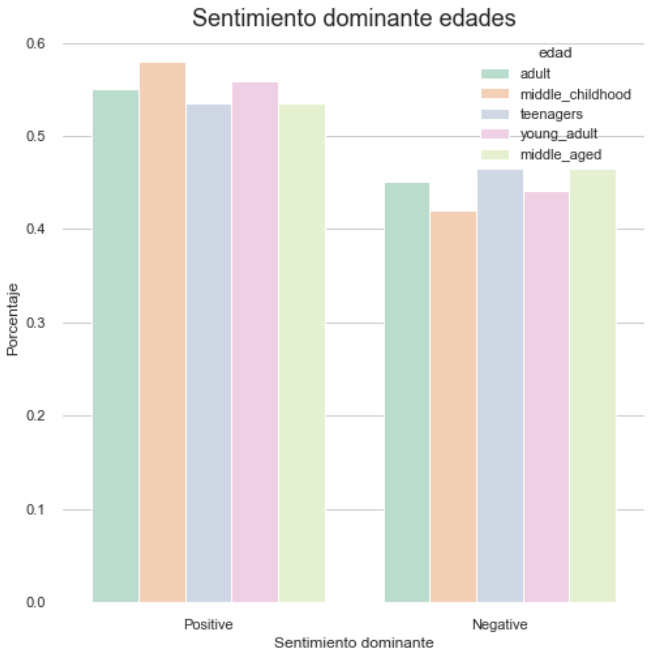


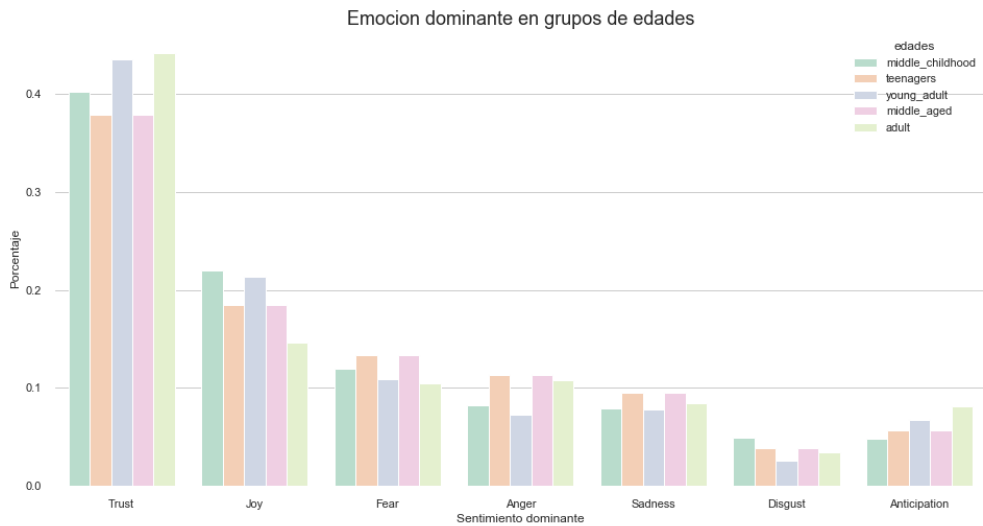
Análisis por sexo



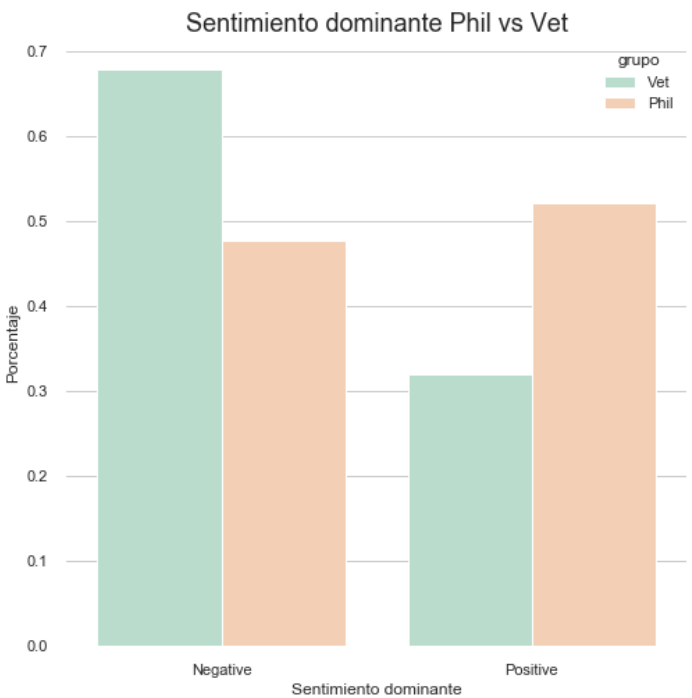


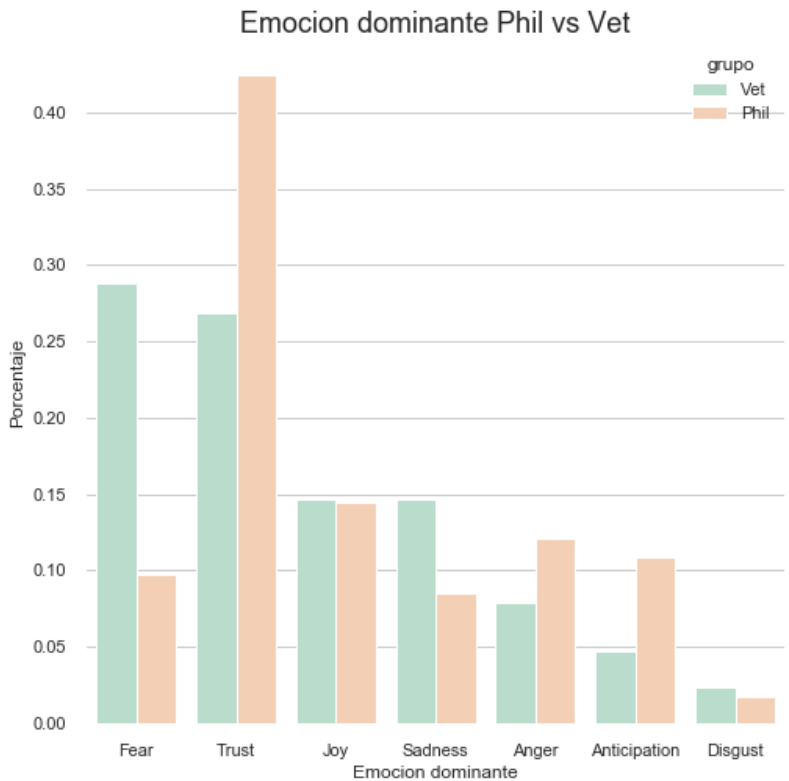
Análisis por franja etaria



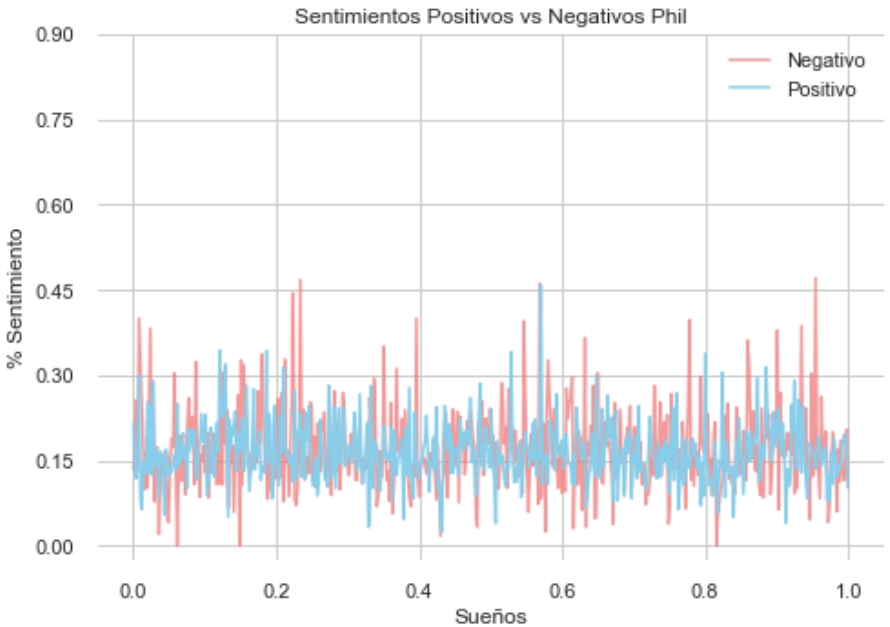


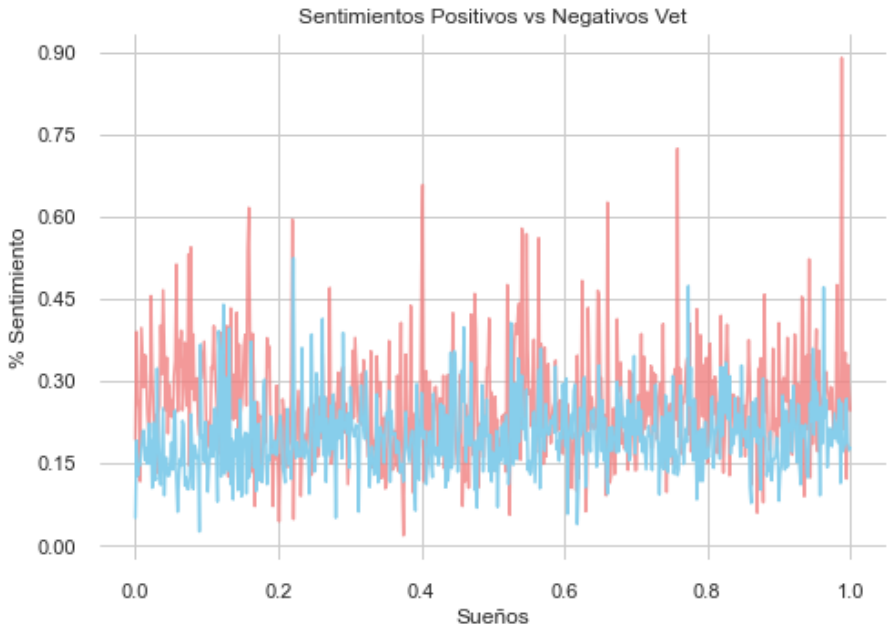
Análisis Phil y Vietnam Vet



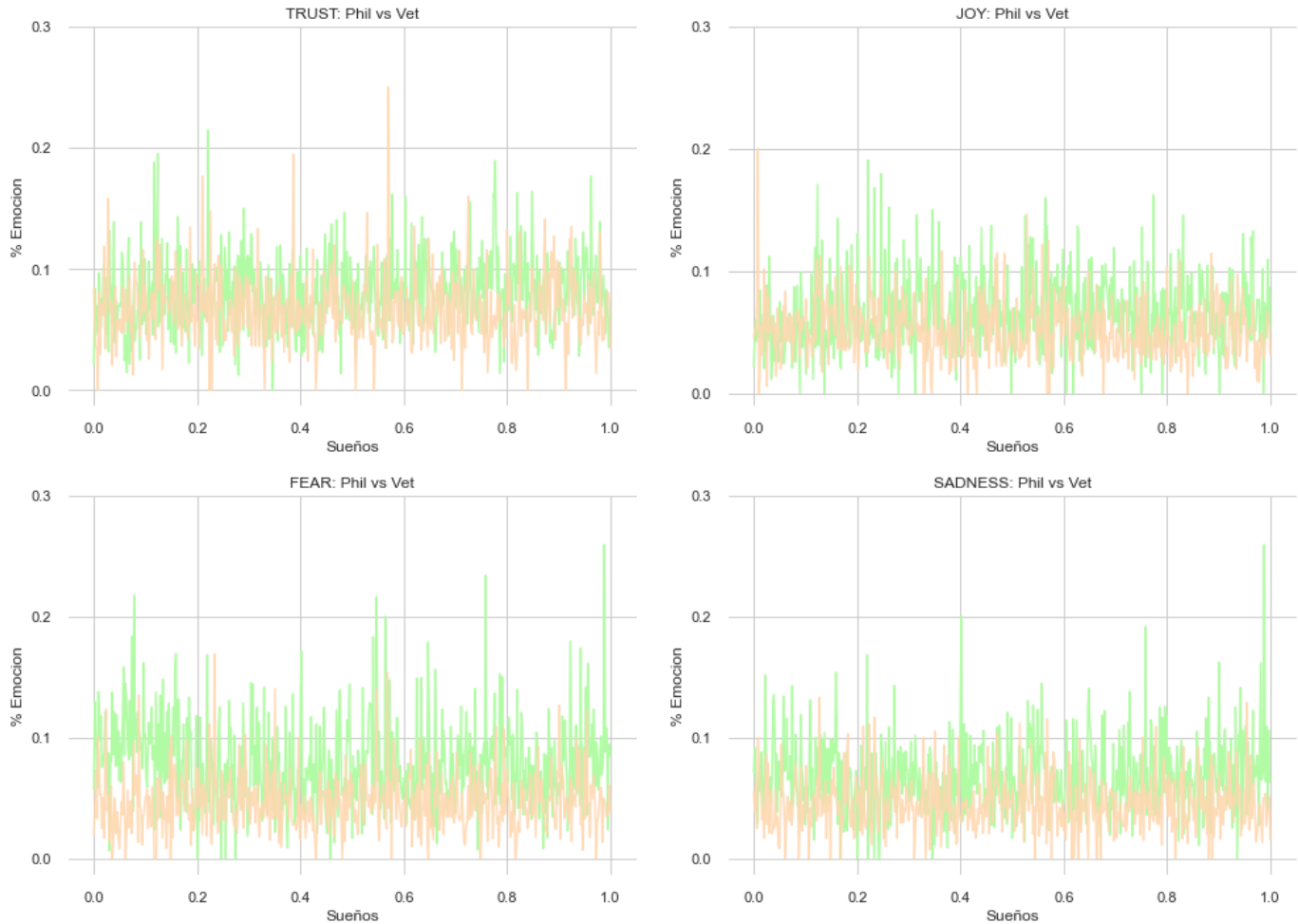


Evolución de emociones en el tiempo Phil y Vietnam Vet



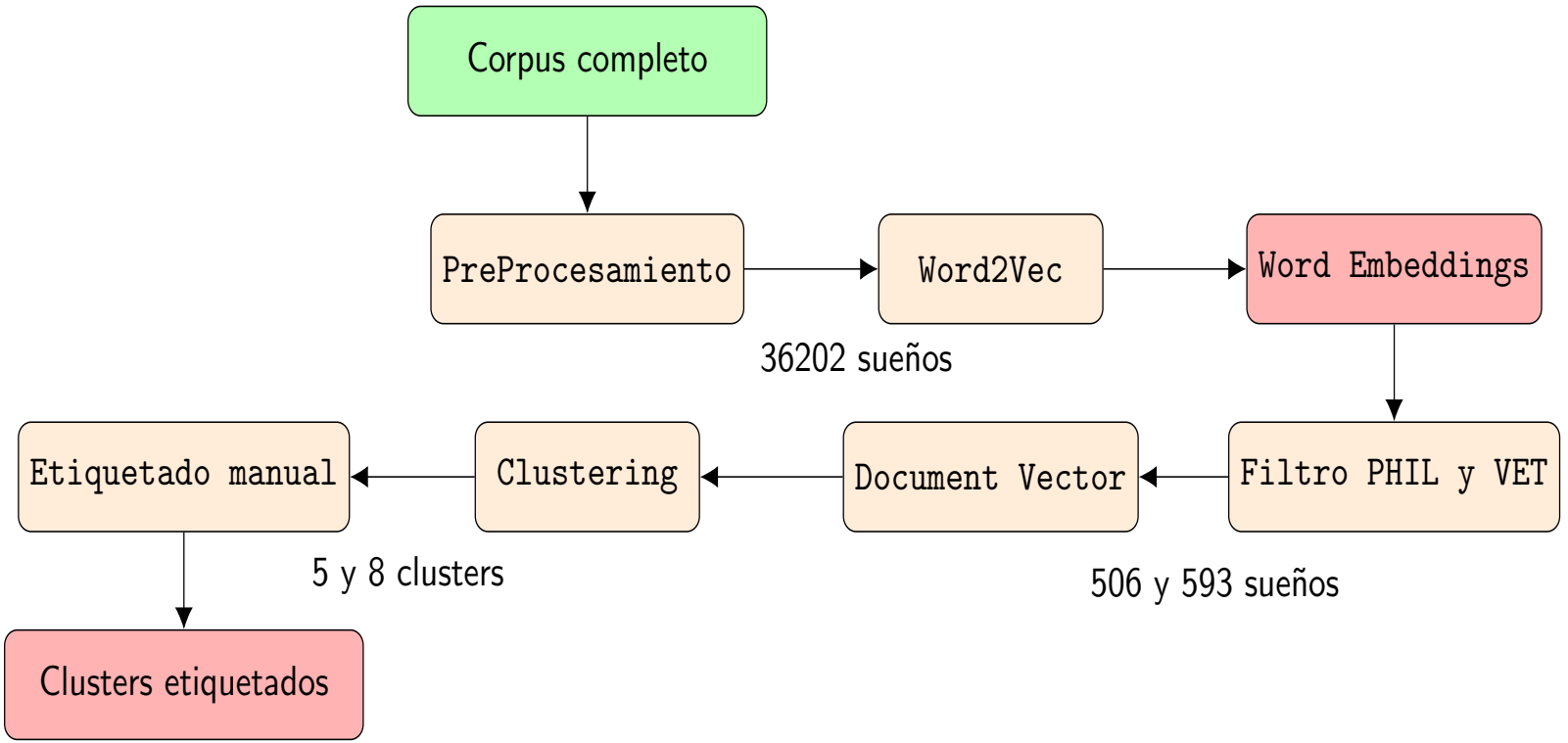


Evolución de emociones en el tiempo Phil y Vietnam Vet



Contenido

Pasos de Procesamiento



Word Embedding y Clustering

Se ejecutó Word2Vec sobre todo el corpus y se calculó el vector promedio por cada sueño de Phil y Vet.

- Word2Vec con dim=20, window=10, y skipgram

Sobre los vectores promedios se hizo clustering para agrupar los sueños por similaridad semántica.

- Clusterización con método jerárquico y kmeans
- Validación de los clusters con coeficiente de Silhouette

Clusters obtenidos

SERIES VET

Etiqueta Manual	Palabras que se destacan
"casa/hotel"	home - kitchen -room – bedroom - hotelroom -bed -house -toilet -furnished - door
"guerra"	bomb - fear - pain - dying - frightened - war - escape run - blood - afraid - explode
"familia/amigos"	girlfriend - fríend - dinner - mother - family - brother – veterans - dog
"huir"	prison - escape - climb - speed - far away - run – hide - accelerate
"guerra"	fear - death - killing - attack - shoot - kick - struggle - fight - shoot - soldier
"mujeres"	woman – attractive - massage - love - girl - feel good - beautiful
"clases/hospital"	medic - class - clinic - students - hospital - teacher- write - tell - psychologist
"miscelaneous"	college - theater - moving - changes - wearing - game - cloth - team

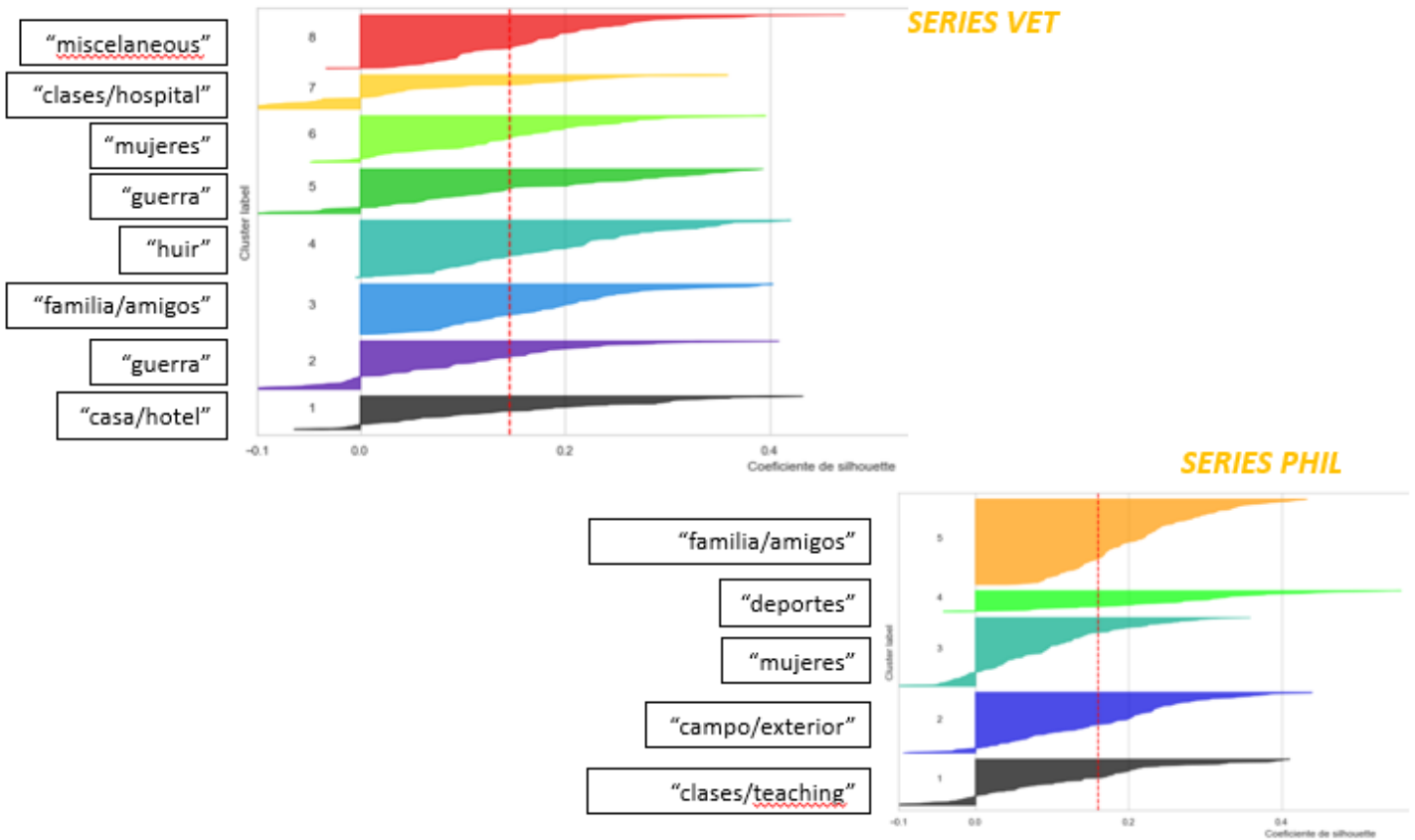
Clusters obtenidos

SERIES PHIL

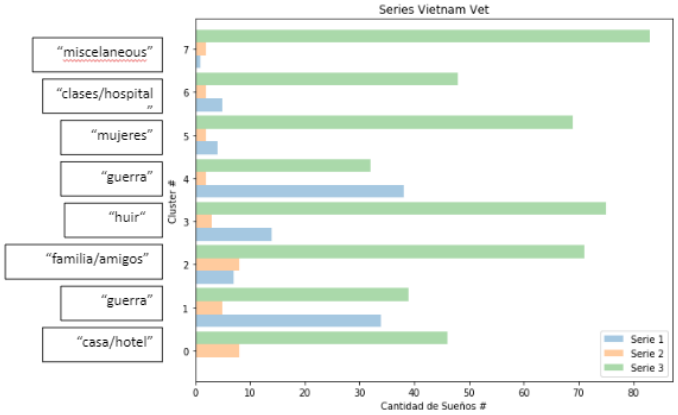
Etiqueta Manual	**Palabras que se destacan**
"clases/teaching"	classes - school - univesity - collegue - job - office - work - friends - semester
"campo/exterior"	river - car - boats - bridge - swim - water - mountain - farm - trees - walking - falling
"mujeres"	relation - kiss - sex - blood - naked
"deportes"	games - football - basketball - handball - league- team
"familia/amigos"	trip - house - home - car - family - wife - drive - bus - leave - go - meeting

Clusters obtenidos

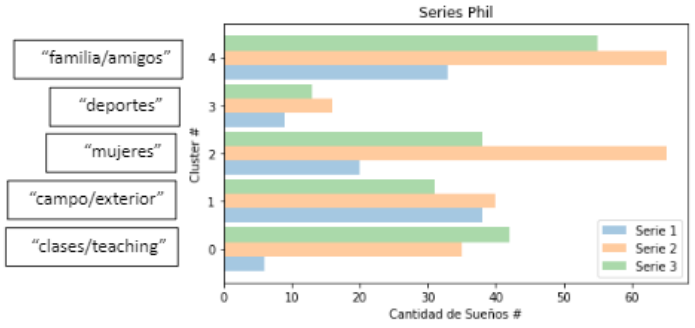
Validación por Silhouette



Distribución de las series en los clusters



VET	Porcentaje
guerra + huir	**41**
miscelaneous	15
familia/amigos	14
mujeres	13
casa/hotel	9
clases/hospital	9



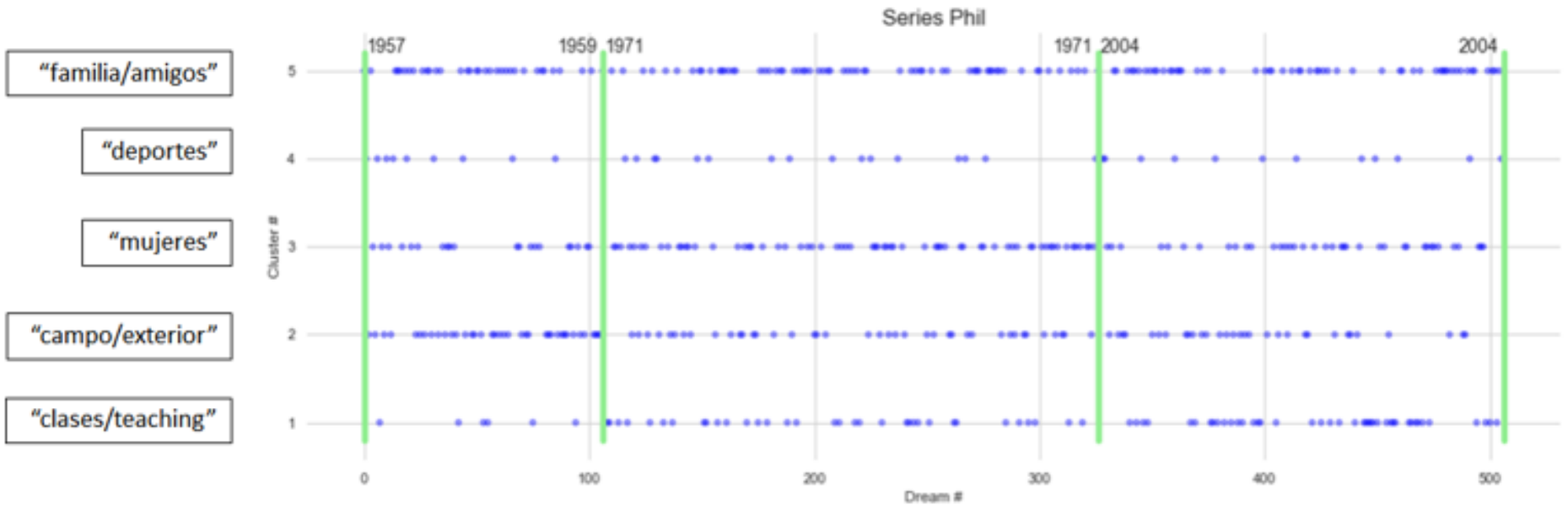
PHIL	Porcentaje
familia/amigos	**30**
mujeres	24
campo/exterior	22
clases/teaching	16
deportes	8

Recurrencia de sueños

Sueños de Vet en distintas etapas de la vida

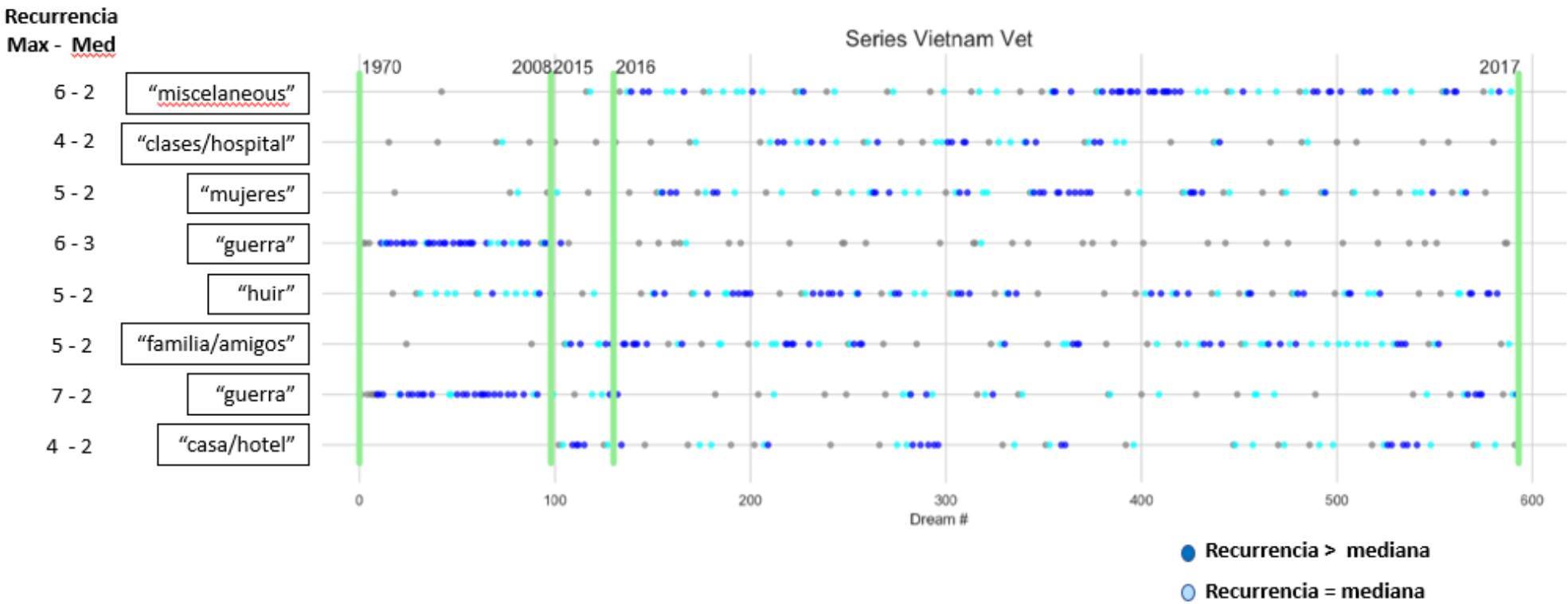
Recurrencia de sueños

Sueños de Phil en distintas etapas de la vida



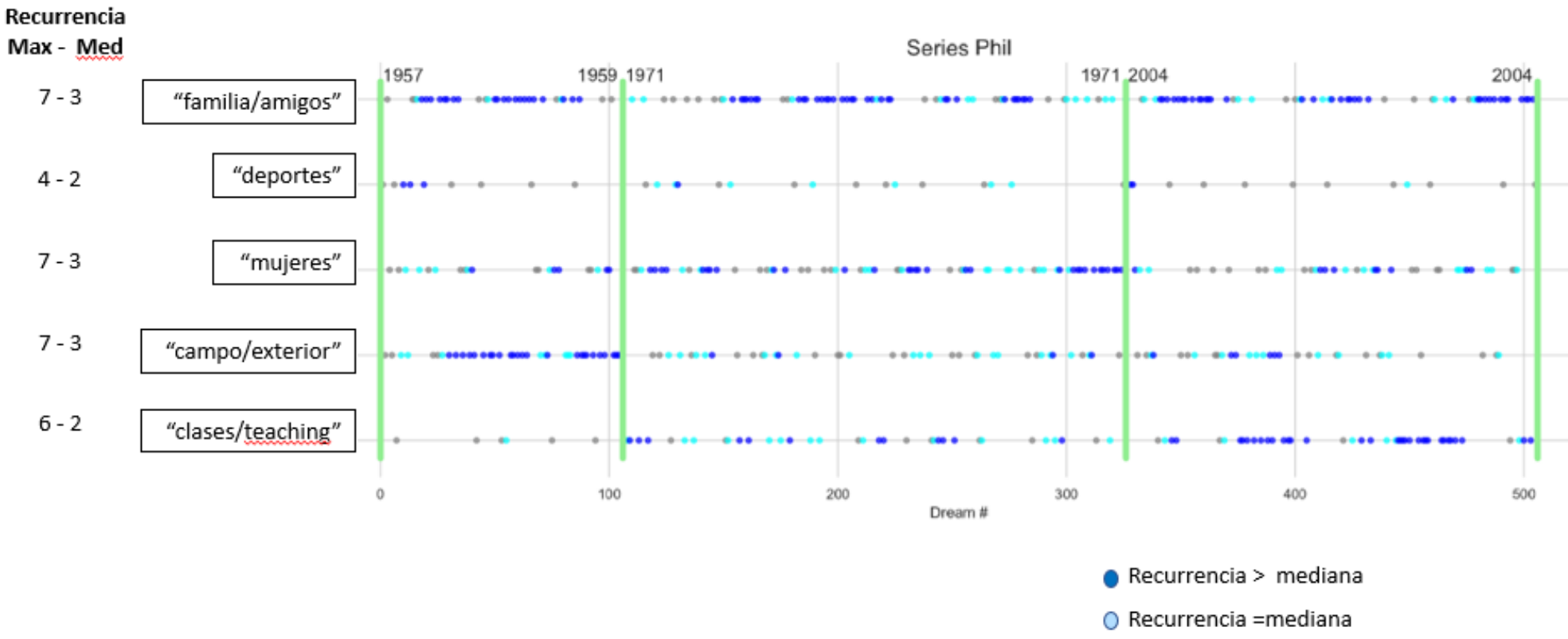
La recurrencia en números

Medida como la cantidad de veces que se repite un cluster en ventana de 10 sueños consecutivos



La recurrencia en números

Calculado con la función rolling() de Pandas

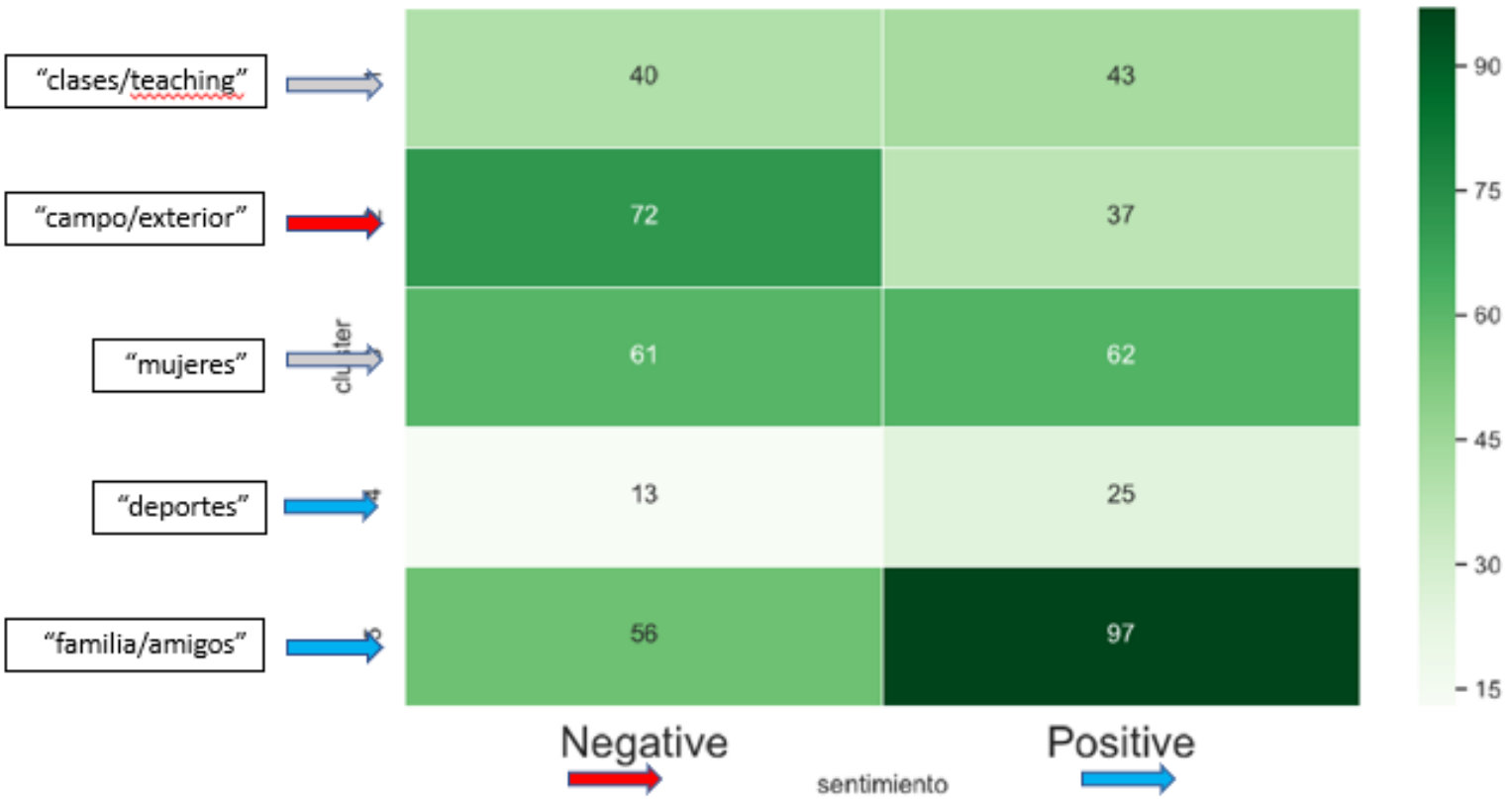


Clustering vs Sentiment Analysis

Sueños de Vet

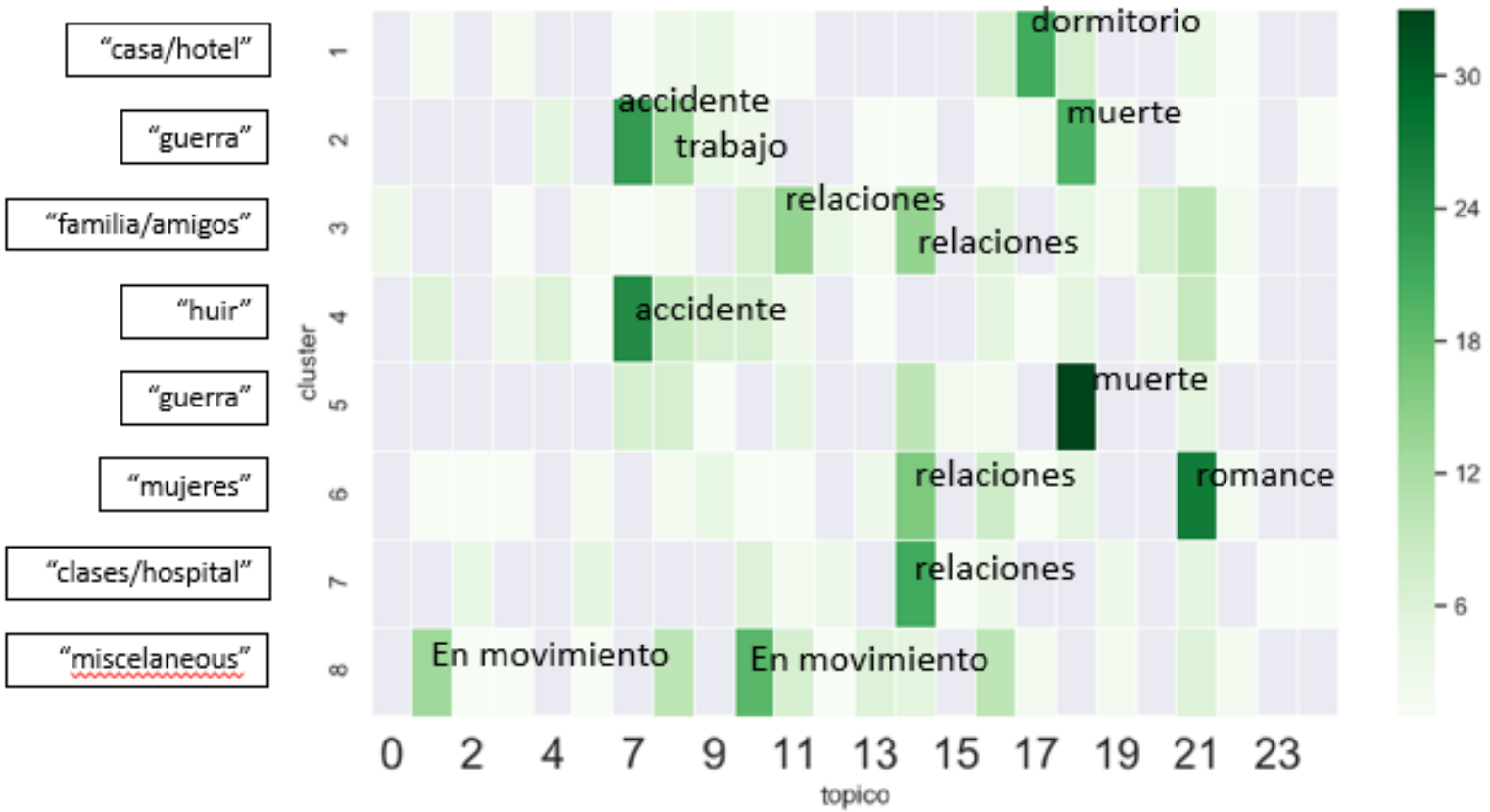
Clustering vs Sentiment Analysis

Sueños de Phil



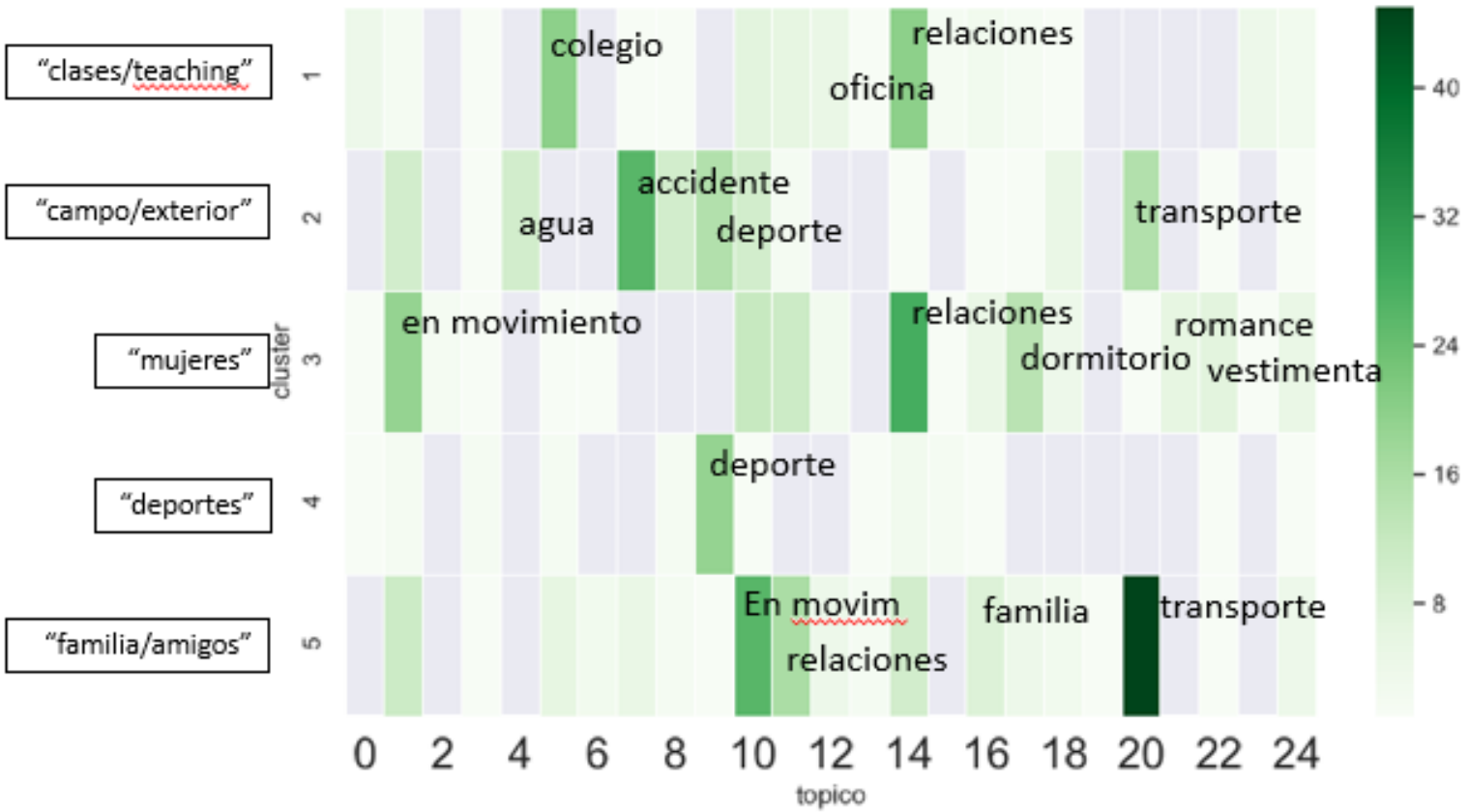
Clustering vs Topic Modelling

Sueños de Vet



Clustering vs Topic Modelling

Sueños de Phil



Contenido

Conclusiones

Se confirmaron las diferencias entre Phil y Vet. Phil tiene una distribución uniforme de los temas de los sueños siendo los principales temas de familia, amigos, trabajo, vida cotidiana (sentimientos positivos). Mientras que Vet tiene mayoría cercana al 41% de sueños relacionados con la guerra y pesadillas de persecución(sentimientos negativos), y recurrentes en las tres series.
En cuanto a los sueños de los distintos grupos de edades no se encontraron diferencias en las emociones reportadas y los temas encontrados para cada uno fueron los esperados.

Preguntas