Practica 1

Los ficheros correspondientes a la entrega del proyecto se encuentran en el proyecto de Github en el siguiente enlace: https://github.com/carolinekonig2/practica1

La carpeta entrega contiene los ficheros correspondiente a la entrega del proyecto (código R, el dataset y este documento con las respuesta a las preguntas).

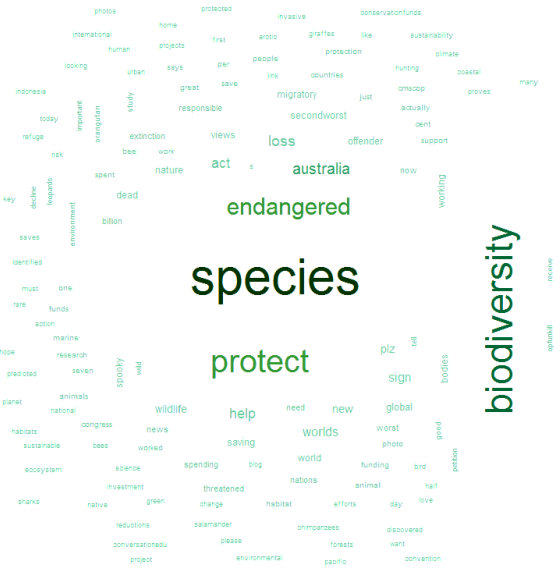
Respuesta a las preguntas de la Practica 1:

1) Titulo del dataset: Temas de la actualidad sobre biodiversidad en twitter.

2) Subtitulo del dataset: Recolección de términos frecuentes y característicos a partir de mensajes de twitter sobre biodiversidad.

Descripción ágil: El dataset proporciona los términos más característicos y frecuentes sobre los mensajes intercambiados en twitter acerca del tema de la 'conservación de biodiversidad'. El dataset por un lado proporciona una lista de los términos más frecuentes y por otro lado los términos más frecuentes de cada subtema (topic) hallado en el conjunto de tweets como también el conjunto de tweets originales.

3) La imagen tipo ' word cloud' identifica el dataset en cuenta que describe los términos más frecuentemente comentados sobre biodiversidad en twitter:



Caroline König

4) Contexto: Materia del conjunto de datos

La conservación de la biodiversidad es en la actualidad un tema de mayor relevancia a consecuencia de la rápida perdida de especies. Por ejemplo el declive de los polinizadores (abejas o insectos en general) es un problema en concreto con gran impacto en los ecosistemas.

En este estudio se pretende analizar los mensajes enviados sobre el término biodiversidad en la red twitter para extraer información sobre los temas relacionados con la conservación de la biodiversidad, que se comenta entre usuarios de la red twitter. La búsqueda se realiza con términos en ingles, i.e. sobre 'biodiversity' y enfocados hacia las iniciativas para proteger o conservar la biodiversidad, por tanto la búsqueda de tweets es sobre los términos 'biodiversity+protect', 'biodiversity+conservation', 'bee+protect', 'species+protect' y 'species+conservation'. El objetivo es recuperar un conjunto de tweets y realizar, en primer lugar, un análisis de los términos más frecuentemente empleados y más descriptivos de los mensajes. Para ello se lleva a cabo un análisis de la frecuencia de términos mediante la construcción de un corpus de texto y una matriz de términos y documento (tdm). Sobre este matriz se extrae el conjunto de términos más frecuentes que se proporciona como información descriptiva sobre los tweets relacionados con la conservación de la biodiversidad. En segundo lugar se analiza el conjunto de tweets en búsqueda de agrupaciones según los términos utilizados más frecuentemente. Se proporciona un fichero que contiene los términos más descriptivos para cada agrupación (cluster) y la lista de tweets que pertenecen a cada agrupación. Esta información tanto sobre los términos más frecuentes como también la aportada sobre las agrupaciones está pensada para aportar conocimiento sobre los términos relacionados con la conservación de la biodiversidad en la actualidad. Esta información por supuesto es puntual, dado que el conjunto de tweets es recién y no abarca un periodo de tiempo demasiado largo (7 días aproximadamente). Para obtener una visión más global sería necesario recolectar tweets en un periodo de tiempo más largo y repetir el análisis.

5) Contenido: Como se han obtenido y a que periodo de tiempo corresponden?

Recolección de tweets

Los tweets se coleccionan utilizando la librería twitterR de RStudio mediante búsqueda por los siguientes términos:

tweets<-searchTwitter('biodiversity+conservation OR biodiversity+protect OR bee+protect OR species+protect OR species+conservation -filter:retweets',n=5000, lang="en")

La búsqueda ha proporcionado 3106 tweets (excluyendo retweets) en ingles sobre los términos buscados. Los tweets recollecionados han sido enviados entre el 25 de octubre y 2 de noviembre.

Descripción de los ficheros y sus campos:

A continuación se describen los diferentes ficheros que componen el dataset ' Temas de la actualidad sobre biodiversidad en twitter ' y sus campos:

- El conjunto de tweets originales:

La búsqueda de tweets ha recuperado un conjunto de 3106 tweets en idioma ingles. El análisis se centra principalmente en el campo 'text' de los tweets. El conjunto de tweets está disponible en formato csv en el fichero tweets.csv.

El fichero contiene 3106 tweets, cada uno descrito mediante 16 campos, que son los habituales

obtenidos mediante búsqueda con la librería twitterR:

"ID","text","favorited","favoriteCount","replyToSN","created","truncated","replyToSID","id","reply ToUID","statusSource","screenName","retweetCount","isRetweet","retweeted","longitude","latit ude"

- Términos frecuentes (de todo el conjunto de mensajes):

Un análisis de los términos más frecuentemente utilizados en los tweets está disponible en el fichero freguTerm.csv. El fichero contiene 175 entradas indicando según un ranking en frecuencia de aparición los términos: cada entrada esta descrito mediante un identificador numérico que indica el puesto en el ranking de frecuencia y la palabra correspondiente. Véase los resultados reportados sobre la extracción de los términos según la frecuencia de aparición en los tweets en la wiki del proyecto: Terminos Frecuentes

La imagen tipo word cloud, utilizado para identificar el conjunto de datos, muestra mediante una nube de palabras de forma grafica los términos más frecuentes y mas descriptivos del conjunto de tweets analizados.

- Términos frecuentes de subtemas:

Proporcionamos el conjunto de mensajes segmentados según una agrupación mediante clustering sobre los términos empleados en el mensaje (Véase los resultados reportados sobre la descripción de temas sobre el texto de los tweets en la wiki del proyecto: Descripción de temas).

En primer lugar el fichero cluster_terms.csv contiene una descripción de las 15 agrupaciones halladas. Cada agrupación (cluster) está identificado mediante un identificador numérico (cluster) y 5 palabras (campos t1 -t5) que son características de la agrupación. En segundo lugar se proporcionan los mensajes correspondientes a cada una de las 15 agrupaciones obtenidos a partir del algoritmo de clustering K-means:

- tweets_c1.csv.
- tweets_c2.csv.
- tweets_c3.csv.
- tweets_c4.csv.
- tweets_c5.csv.
- tweets_c6.csv.
- tweets_c7.csv.
- tweets_c8.csv.
- tweets_c9.csv.
- tweets_c10.csv.
- tweets_c11.csv.
- tweets_c12.csv.
- tweets_c13.csv.
- tweets_c14.csv.
- tweets_c15.csv.

El fichero topicModeling.csv contiene la descripción de cada tema (topic) con 10 términos según el

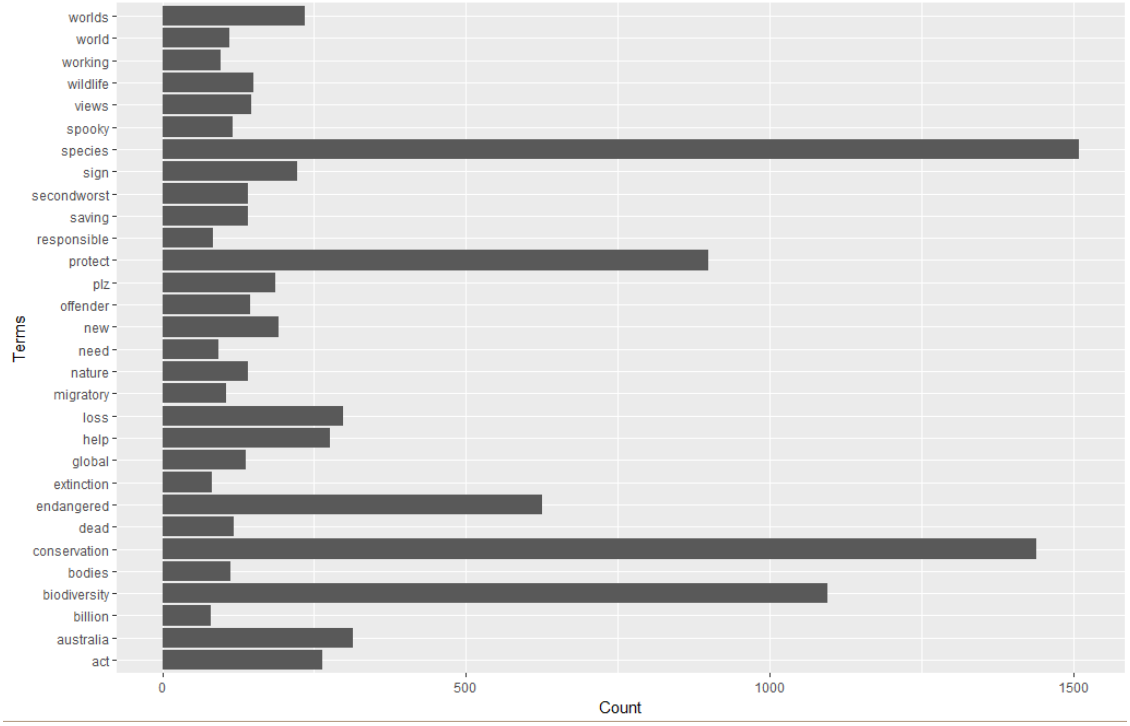
método de topic modeling. El fichero contiene 15 temas identificados por un identificador numérico (id) y 10 palabras características (t1-t10).

6) Agradecimientos: El propietario de los mensajes de twitter originales es la red twitter que proporciona el acceso libre a los datos a través de la API de busqueda (Search API). Por tanto un agradecimiento seria para Twitter Inc. por proporcionar el acceso libre y gratuito a un conjunto de tweets los últimos días y habilitar la función de indexado que permite realizar búsquedas simples. Los otros ficheros con los términos frecuentes o descripción de temas se han creado a partir de los mensajes de twitter y serian de nuestra propia propiedad.

7) Inspiración: Un análisis de los mensajes de twitter enviados en relación a la conservación de biodiversidad es interesante al ser un tema relevante para la sostenibilidad de los ecosistemas, que afecta a las personas de todos los países. Se supone que es un tema comentado a nivel mundial, lo que facilita la recolección de los tweets en el idioma ingles.

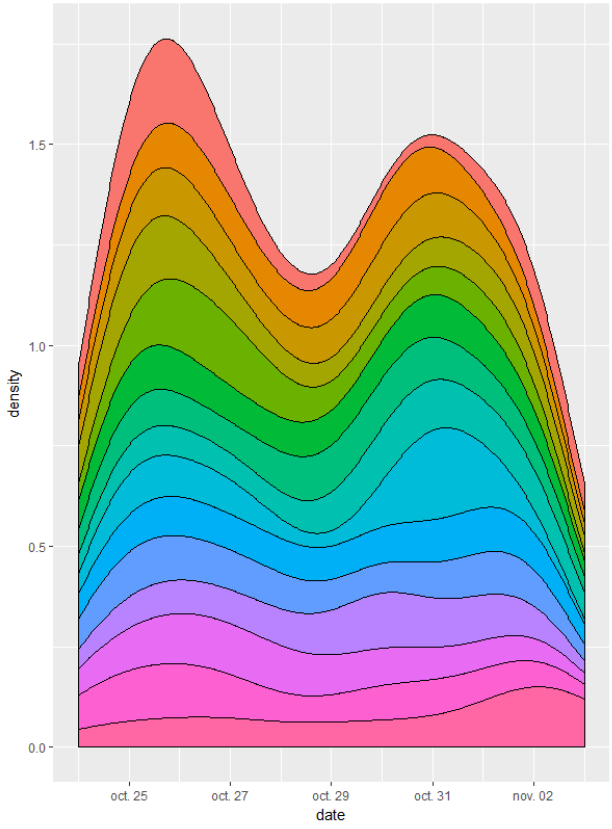
Por un lado es interesante recollecionar los tweets que se intercambian sobre este para valorar la preocupación de los usuarios sobre el tema simplemente analizando el número de mensajes enviados. Hemos contrastado que la cantidad de mensajes es notablemente inferior comparado con las que se intercambian sobre temas políticas actuales, p.ej. la búsqueda sobre temas como 'Trump' proporciona muchos más mensajes para el mismo periodo de tiempo, mientras la búsqueda sobre la 'conservación de la biodiversidad' proporciona 3106 mensajes.

En segundo lugar el resultado del análisis del texto de los mensajes ha proporcionado información tanto sobre los términos más frecuentes como también sobre las agrupaciones de los mensajes en temas. Esta información está pensada para aportar conocimiento sobre los términos relacionados con la conservación de la biodiversidad, es decir los temas relacionados que más impacto tienen en la actualidad (medidos por la frecuencia en la que se comentan en la red twitter). La siguiente imagen resume gráficamente los términos más frecuentes encontrados:



Aunque los terminos 'biodiversity', 'conservation', protect', 'species', utilizados en la búsqueda, resultan los de mayor frecuencia se observa que (entre otros) los terminos 'world', 'loss','help', 'endangered', 'australia', 'act' son frecuentes y descriptivos de los temas comentados en relación a la conservación de la biodiversidad.

El siguiente gráfico muestra la relación entre mensajes y tema (topic). La separación de los mensajes en temas es útil para analizar en más detalle los temas comentados entre los usuarios de twitter. Cada tema esta descrito mediante 10 términos característicos. Se observa, por ejemplo que la noticia que atribuye a australia el segundo puesto en el ranking de países responsable de la pérdida de biodiversidad es un tema muy comentado en la red twitter al aparecer el término 'australia' en el tema 1 y 4.



term[topic]

- biodiversity, australia, worlds, loss, offender, secondworst, worst, conservation, conversationedu, auspol
- biodiversity, conservation, species, protect, blog, funds, photo, wild, need, lost
- biodiversity, conservation, working, views, photo, now, environmental, sustainability, pacific, photos
- biodiversity, responsible, nations, global, countries, conservation, seven, per, cent, australia
- conservation, biodiversity, global, spending, loss, new, species, change, study, predicted
- conservation, nature, biodiversity, species, views, habitat, sustainable, buy, science, image
- conservation, species, migratory, animal, efforts, cmscop, protected, hope, convention, love
- conservation, species, research, funding, new, s, hunting, saves, proves, invasive
- conservation, species, saving, endangered, dead, spooky, bodies, world, billion, spent
- endangered, protect, species, help, act, sign, plz, congress, tell, petition
- protect, species, endangered, must, people, native, wood, detectives, german, planet
- species, protect, biodiversity, conservation, extinction, chimpanzees, leopards, giraffes, protection, threatened
- species, protect, need, just, action, habitats, many, read, first, protectwldlife
- species, protect, one, conservation, work, today, bee, bad, ecology, join
- wildlife, species, protect, new, orangutan, news, environment, great, endangered, forests

En relación a los mensajes sobre la protección de especies las agrupaciones y su descripción mediante términos frecuentes es útil para obtener conocimiento en más detalles, por ejemplo sobre los especies amenazados en concreto (chimpanzees, leopards,giraffes) o medidas propuestas ('need, habitat, wildlife o 'conservation, species, migratory, animal, effort') extraídos a partir de los siguientes subtemas:

- "species, protect, biodiversity, conservation, extinction, chimpanzees, leopards, giraffes, protection, threatened"

- "species, protect, need, just, action, habitats, many, read, first, wildlife"

- "wildlife, species, protect, new, orangutan, news, environment, great, endangered, g

- "conservation, species, migratory, animal, efforts, cmscop, protected, hope, convention, love"

8) Para asegurar el uso libre y la disponibilidad de los datos la licencia ODC OpenDatabase License (ODbL 1.0) seria adecuada, ya que permite compartir (usar, copiar y distribuir) la base de datos, crear nuevos trabajos a partir de la base de dato, adaptar (modificar, transformar y construir) a partir de la base de datos. Estos permisos están sujetos a los siguientes condiciones:

- Atribuir: Indicar la referencia a esta base de datos y la licencia ODbl en caso de redistribución de la base de datos original o trabajos a partir de esta .

- Compartir del mismo modo: En caso de compartir una versión adaptada de la base de datos esta tambien se tiene que publicar con la licencia ODbl.

- Mantener abierta: Si se redistribuye una versión adaptada, esta también tiene que ser libremente accesible.

9) La recolección y el análisis de los tweets se ha llevado a cabo mediante RStudio. La descripción del código en R esta disponible en https://github.com/carolinekonig2/practica1/wiki/R-Code o en el fichero codigo_R.txt en la carpeta Entrega en el proyecto de Github.

10) Los ficheros que componen el dataset se encuentran en el repositorio Github en la carpeta Entrega\dataset:

- fichero con los tweets recoleccionados: tweets.csv

- fichero con los terminos frecuentes: frequTerm.csv

- fichero con los terminos descriptivos de los 15 clusters: fichero cluster_terms.csv y los correspondientes mensajes de tweets segregados por cada cluster: tweets_c1.csv , tweets_c2.csv, tweets_c3.csv , tweets_c4.csv , tweets_c5.csv , tweets_c6.csv , tweets_c7.csv , tweets_c8.csv , tweets_c9.csv , tweets_c10.csv , tweets_c11.csv , tweets_c12.csv, tweets_c13.csv , tweets_c14.csv , tweets_c15.csv.

- El fichero topicModeling.csv con la descripción de cada tema (topic) con 10 términos según el método de topic modeling.