Universitat Oberta de Catalunya

uoc.edu

Experiencias del uso de Kaggle en el entorno educativo superior

Dra. Laia Subirats

Investigadora en ciencia de datos @ Eurecat – Centro Tecnológico de Cataluña

Profesora a tiempo parcial @ Universidad Abierta de Cataluña (UOC)

UOC Data Day

Barcelona, 16 de mayo 2018

Índice

1. Introducción a Kaggle

- Fundada en 2010 por el australiano **Anthony Goldbloom**.
- Adquirida en 2017 por **Google**.
- Más de **536 000 usuarios** registrados de más de 190 países.
- Miembros del equipo ganador de Jeopardy de IBM Watson y del equipo que trabaja en DeepMind de Google.
- Publican artículos en **revistas** en función de su desempeño en estas competiciones.

1. Introducción a Kaggle: pasos

- **Lenguaje** de programación  
- **Competición**: destacada, investigación, reclutamiento o iniciación
- **Explorar datos**

1. Introducción a Kaggle: consejos

- **Objetivos** incrementales
- **Kernels** más votados
- **Foros**
- Inicio en **solitario**
- **No preocuparse** por posiciones de la tabla bajas
- **Maximizar aprendizaje** no ganancias

2. Experiencias web scraping

- Trabajo conjunto con
  Dra. Isabel Guitart y Dra. Mireia Calvo
- Estudio de **conceptos básicos**: concepto, legalidad, escala y estructura de la web, errores comunes y características avanzadas
- **Ejemplos** de compañeros del año anterior compartiendo su repositorio **Github**
- **Entrega parcial**
- **Solución**: información de testimonios anonimizados de personas con enfermedades raras

2. Experiencias web scraping

- Título, subtítulo y imagen
- Contexto
- Contenido
- Agradecimientos
- Inspiración
- Licencia
- Código y dataset

- Nota: Debe considerarse la privacidad de los usuarios. General Data Protection Regulation, 25 Mayo 2018.

2. Experiencias análisis de datos

- **Descripción** del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder? *Cambiar de respuesta es evolución. Cambiar de pregunta es revolución* (Jorge Wagensberg). Tipo de pregunta: **descriptiva**, **exploratoria**, **inferencial**, **predictiva**, **causal** o **mecánica** (según Jeffrey Leek).
- **Integración y selección** de los datos interés a analizar.
- **Limpieza de datos**: gestión de elementos vacíos, identificación y tratamiento de valores extremos.

2. Experiencias análisis de datos

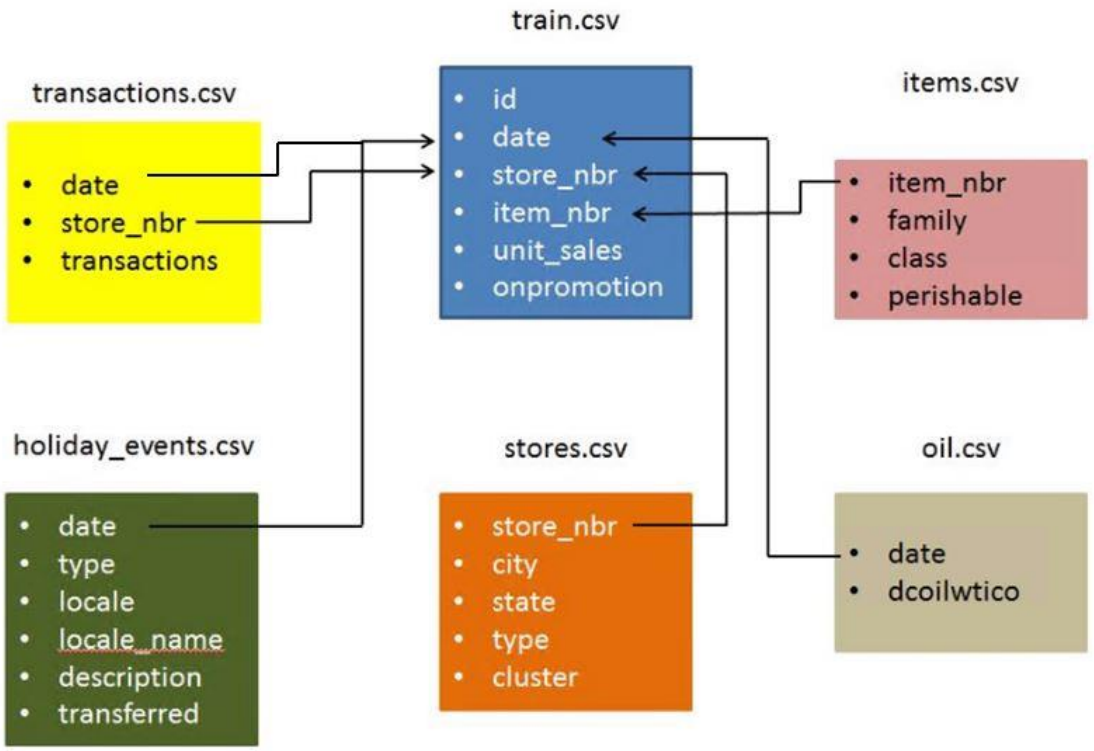
- **Análisis** de los datos: selección grupos de datos, comprobación de la normalidad y homogeneidad de la varianza, aplicación de pruebas estadísticas.
- Representación de los resultados a partir de **tablas** y **gráficas** (boxplot, etc.).
- **Resolución** del problema: ¿conclusiones? ¿los resultados permiten responder al problema?

3. Predicción de ventas de comestibles

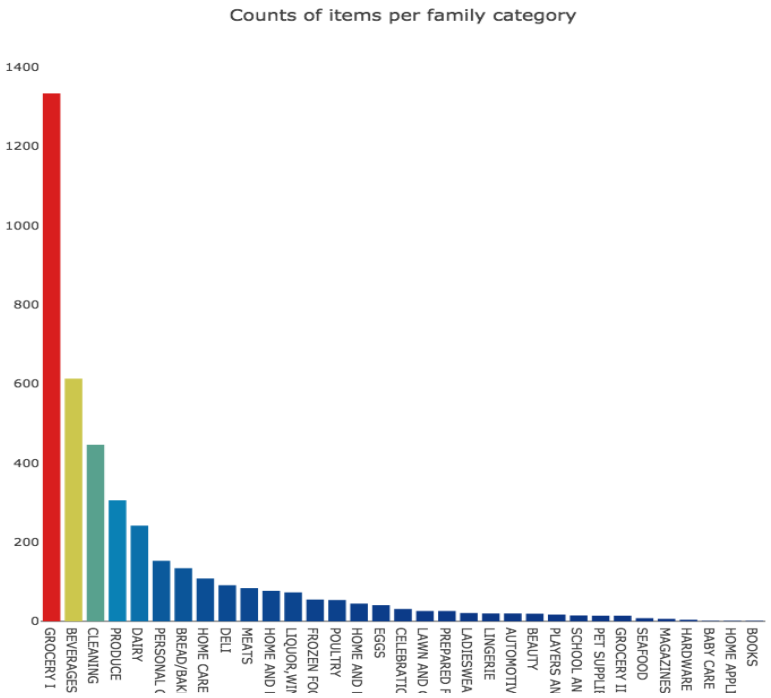
- **Trabajo conjunto con Gabriel Kreplak**
- **Motivación:** competición en la plataforma Kaggle para lograr el mejor modelo de predicción de ventas futuras de cada uno de sus 200.000 productos en cada una de más de 50 tiendas.
- https://www.kaggle.com/c/favorita-grocery-sales-forecasting
- https://www.kaggle.com/gkreplak
- Plazo: del 19 de octubre de 2017 a 15 de enero de 2018
- Premios: 30.000$ a repartir entre los 3 mejores
- Métrica de evaluación: Normalized Weighted Root Mean Squared Logarithmic Error (NWRMSLE)
- Clasificación: 269/1675 (top 17%), NWRMSL: 0,522
- Lenguaje de programación: Python 3.6
- http://hdl.handle.net/10609/74565

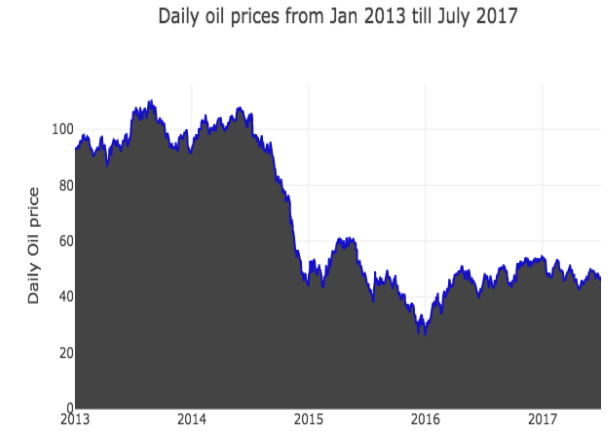
3. Predicción de ventas de comestibles

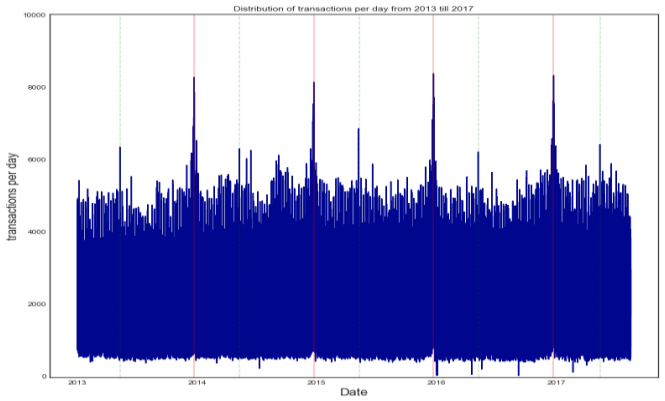
Archivo	Instancias
train	125.497.040
transactions	83.488
items	4.100
holiday events	350
stores	54
oil	1.218



3. Predicción de ventas de comestibles







3. Predicción de ventas de comestibles

Algoritmo **Gradient Boosting Decision Trees** (GBDT): Solución muy popular para problemas de regresión. Implementaciones GBDT:

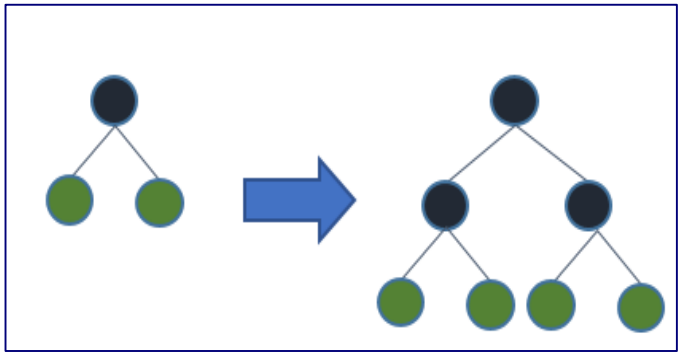
- XGBoost, pGBRT. Eficiencia y escalabilidad comprometida en datasets grandes.
- **Light Gradient Boosting Machine** (LGBM). Mitiga esos problemas. Desarrollado en Microsoft y Universidad de Pekín.

Usa grandes gradientes, excluyendo proporciones significativas de datos y reduciendo atributos mutuamente exclusivos.

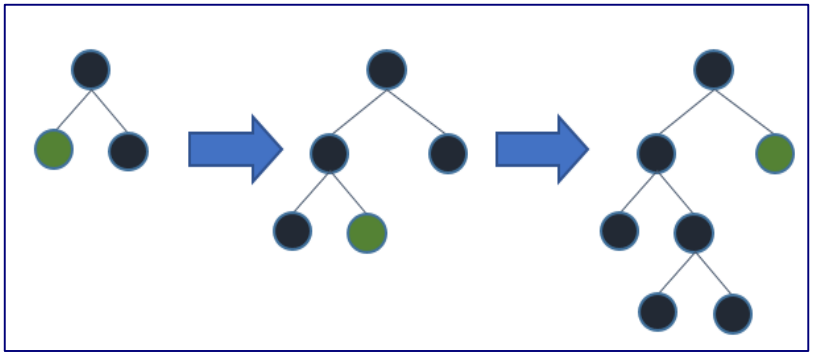
Aplicaciones LGBM: Espectroscopia de suelo, música, física de partículas y Kaggle

3. Predicción de ventas de comestibles

Dos alternativas al crecimiento del árbol de decisión:



Algoritmo standard: Crecimiento por nivel



Algoritmo LGBM: Crecimiento por hojas

Ventajas **Light Gradient Boosting Machine (**LGBM)

- Velocidad
- Precisión
- Posibilidad de proceso paralelo
- Adecuado para grandes datasets

3. Predicción de ventas de comestibles

Algunos parámetros LGBM más importantes:

- **Básicos**: num_leaves
  learning_rate
  max_bin
- **Control del Aprendizaje**:
  min_data_in_leaf
  max_depth
  feature_fraction
  bagging_fraction
  bagging_freq: velocidad

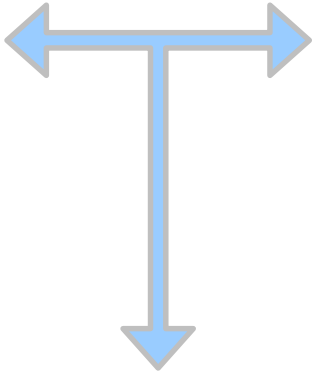
Descripción de la experiencia en el Blog de Ciencia de datos de la UOC http://dataanalysis.blogs.uoc.edu/2018/02/19/proyecto-fin-de-master-de-la-competicion-de-kaggle-prediccion-de-ventas-de-comestibles-corporacion-favorita

4. Predicción de la popularidad de las aerolíneas

- **Trabajo conjunto con Dra. Esther Córdova**
- **Motivación:**

Redes sociales: información en tiempo real

Compañías aéreas: mercado en constante crecimiento con fuerte competencia.



Contar con información volcada en tiempo real sobre la satisfacción de los clientes es un punto clave a tener en cuenta en la estrategia de este tipo de compañías.

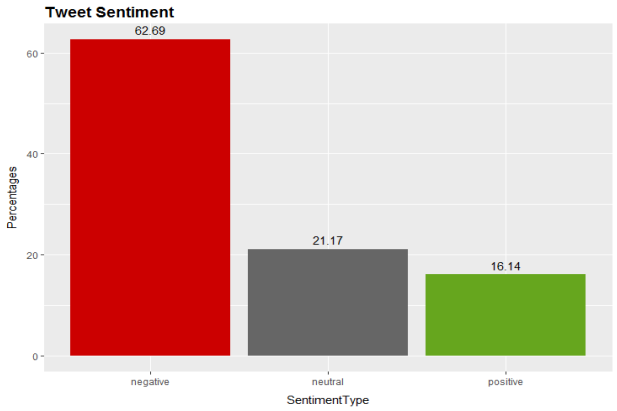
https://www.kaggle.com/crowdflower/twitter-airline-sentiment

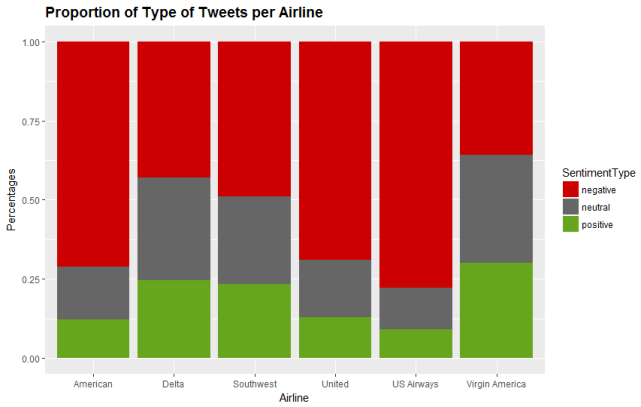
4. Predicción de la popularidad de las aerolíneas

DATASETS

PROCESAMIENTO

- Analizar la **popularidad** de 6 aerolíneas a partir de la polaridad de tweets.
- **Describir temporalmente** el conjunto de datos a analizar: descripción de la evolución de la polaridad de tweets a lo largo del mes y de las diferentes horas del día.
- Analizar la **correlación** entre la polaridad de los tweets y el número de vuelos cancelados o con retaso.
- Analizar el **contenido** de los tweets mediante técnicas de Text Mining.
- Realizar **clusterización** de tweets.
- https://github.com/sther19/UOC_TFM

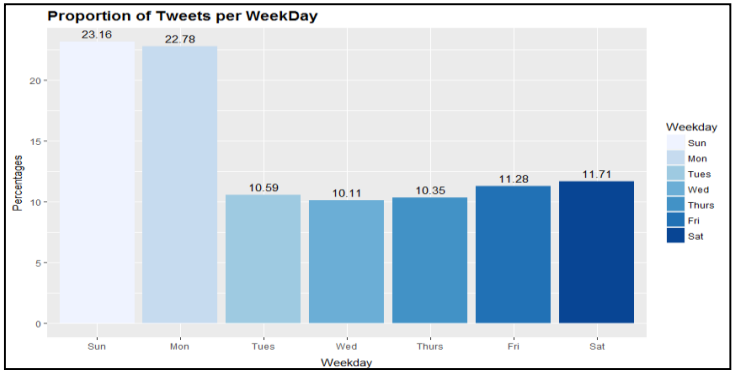


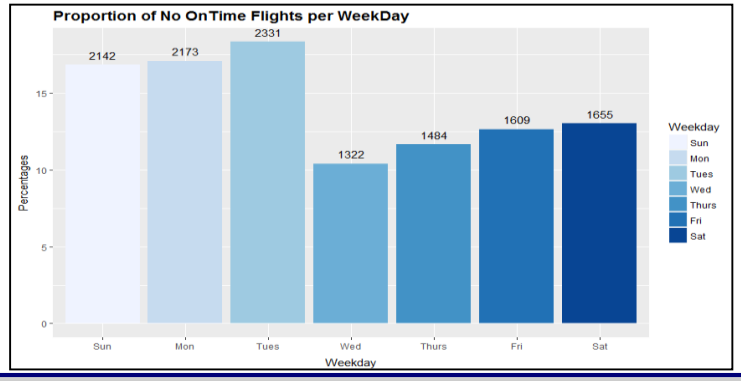


Tweets negativos son más abundantes

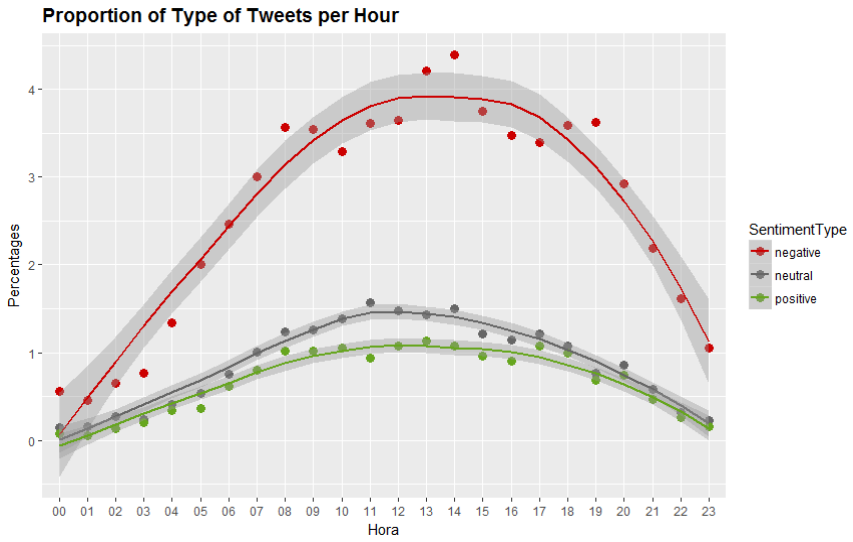
Mayor ratio de tweets Negativos: American, United y US Airways con 13.4%, 18.0% y 15.5%

Mayor ratio de tweets Positivos: Delta, Southwest y VirginAmerica con 3.7%, 3.9% y 1.1%



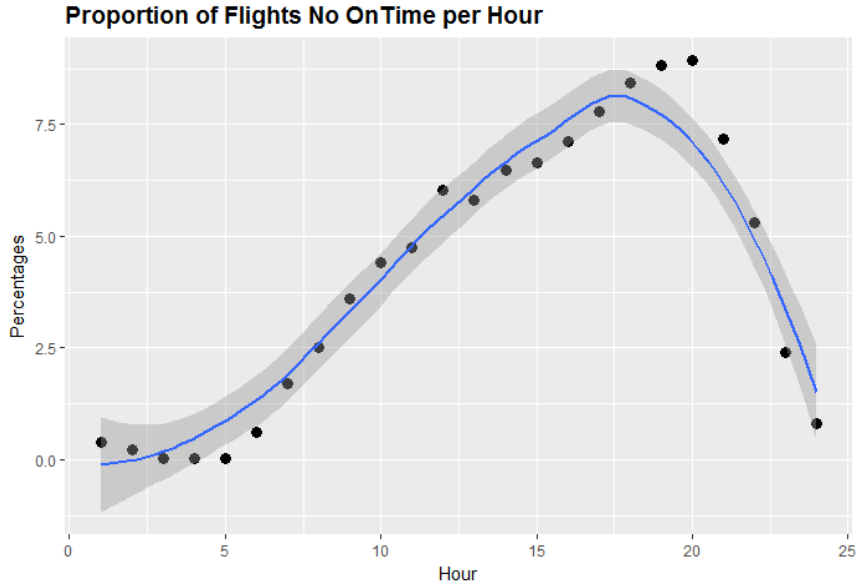


Correlación entre la cantidad de tweets negativos y la cantidad de vuelos retrasados; cancelados; redirigidos. (**r = 0.9155**)

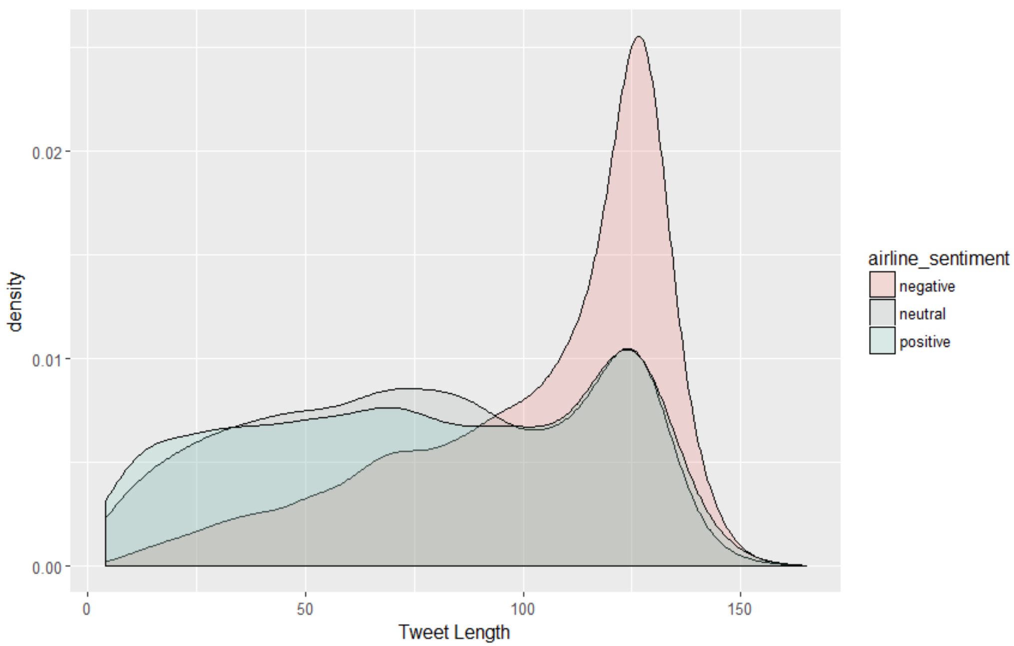


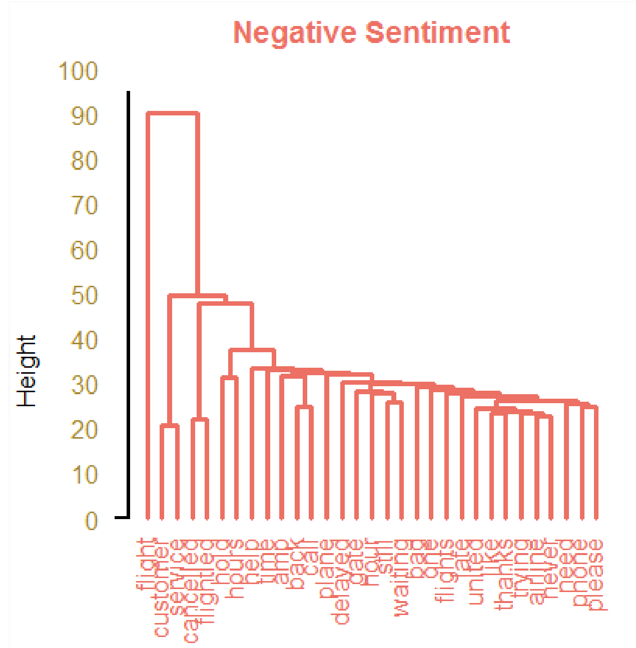
Acumulación horaria de tweets negativos ~ 15:00 pm.

Horario de despegue con mayor retrasos o cancelaciones ~ 17:00 – 18:00 pm.



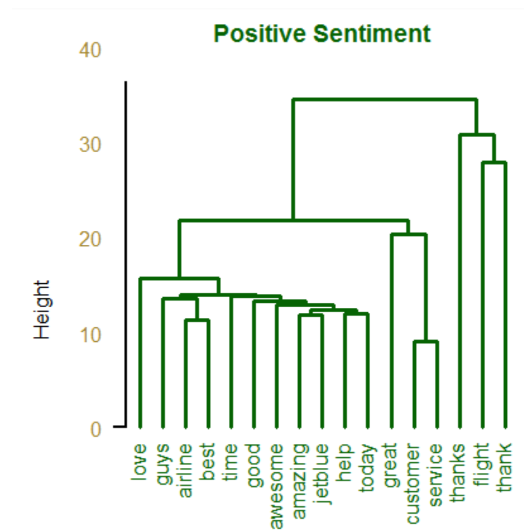
Si tenemos en cuenta que en promedio los usuarios que viajan en avión deben estar como mínimo con 2 horas de antelación en los aeropuertos se puede dilucidar una clara correlación entre el número de tweets negativos y los retrasos de los vuelos.


0.02
0.01
0.00
density
0
50
100
150
Tweet Length
airline_sentiment
negative
neutral
positive


Negative Sentiment
Height
100
90
80
70
60
50
40
30
20
10
0
flight
customer
service
cancelled
flightled
hold
hours
help
time
amp
back
call
plane
delayed
gate
hour
still
waiting
bag
one
flights
late
united
like
thanks
trying
airline
never
need
phone
please


airline
trying
like
united
never
one
customer
waiting
hour
call
help
hours
time
service
flightled
flight
thanks
need
please
cancelled
still
phone
flights
hold
amp
late
gate
plane
delayed
bag
back


amazing
airline
customer
jetblue
best
good
love
guys
service
time
thanks
help
flight
great
today
awesome


Positive Sentiment
Height
40
30
20
10
0
love
guys
airline
best
time
good
awesome
amazing
jetblue
help
today
great
customer
service
thanks
flight
thank

4. Predicción de la popularidad de las aerolíneas: conclusiones

- El número de **tweets negativos es más abundante**, 63%. Los picos horarios con más tweets negativos son las **horas de comer**, mientras que los positivos y neutros se concentran en torno a las **11.00 a.m**.
- **Domingo y lunes** son los días con **más volumen** de tweets, donde el ratio de tweets negativos es mucho mayor. Resultados de esperar, ya tanto lunes como domingo son días de **bastante tráfico aéreo** y hay más posibilidades de que se retrasen vuelos o hayan incidencias, además puesto que el lunes es primer día de la semana laboral esto también podría impactar en los tweets negativos en los usuarios que deben coger vuelos por **razones laborales**. **Domingo, lunes y martes** fueron los días que acumularon más **retrasos**.
- Los tweets con **sentimiento negativo tienden a ser más largos** que los que presentan sentimientos positivos o neutros.
- La palabra que aparece con más frecuencia en los tweets clasificados como **positivos** es *thanks*, mientras que la que aparece más veces en los **negativos** es *flight* seguido de *cancelled*.
- **Trabajo futuro**: realizar gráficos temporales en mapas.

5. Conclusiones Kaggle & entorno educativo superior

- + Consejos **comunidad Kaggle** (empezar en solitario, no preocuparse por posición, etc.)
- Para empezar mejor trabajar **sin deadlines** / escoger una competición que se adecue a la **planificación del trabajo**

- Es aconsejable **experiencia previa** con Kaggle

- **KISS - Keep it sweet & simple, *Simplicity is the ultimate sophistication*** (Leonardo da Vinci)

Universitat Oberta
de Catalunya

facebook
@twitter
instagram

¡Gracias!

https://sites.google.com/site/laiasubirats

UOC.universitat
@UOCuniversitat
UOCuniversitat