Proyecto Final Big Data

David González

12/12/2017

FaMAF - Universidad Nacional de Córdoba

Contenido

- Introducción
  - Marco
  - Idea y Motivación
- Proceso
  - Herramientas
  - Preproceso
  - Clustering
  - Clasificación de sentimientos
- Visualización
- Resultados

Introducción

Marco:

La siguiente presentación se realizó como tarea final de la materia **Programación Distribuida Sobre Grandes Volúmenes de Datos.** El proyecto consiste en reproducir, utilizando Spark, el estudio realizado en la tesis de grado de Martín Becerra.

Teniendo como referencia dicha tesis se decidió una arquitectura del sistema y se realizaron los programas necesarios para implementarla.

Introducción

Idea y Motivación:

La idea central se basa en utilizar un dataset de más de 122mil tweets con el hashtag #Oscars del año 2016, para hacer clustering y así agrupar aquellos tweets que estén relacionados en cuanto a su contenido, así como también hacer clasificación sobre ellos para analizar qué tweets son negativos y cuales positivos, para poder tener una apreciación general de las opiniones de la gente.

Esto cae en lo que se denomina minería de texto, más particularmente en la minería de opiniones o análisis de sentimientos, la cual se enfoca en el tratamiento automático de textos en los cuales se ven reflejados la opinión, los sentimientos, las emociones y las actitudes de las personas hacia ciertos temas y sus aspectos.

Proceso

Herramientas:

Para el preprocesamiento de los datos, clustering y análisis de sentimientos se utilizó la librería de Spark org.apache.spark.ml:

- org.apache.spark.ml.feature.{RegexTokenizer, StopWordsRemover, NGram, HashingTF, CountVectorizer, IDF, VectorAssembler, StringIndexer}
- org.apache.spark.ml.clustering.{KMeans, LDA, BisectingKMeans}
- org.apache.spark.ml.classification.{NaiveBayes, DecisionTreeClassifier}
- org.apache.spark.ml.regression.LinearRegression
- org.apache.spark.ml.tuning.{CrossValidator, ParamGridBuilder}
- org.apache.spark.ml.{Pipeline, PipelineModel}

Para la parte gráfica se utilizó d3js

Proceso

Preproceso:

- Reemplazar números por “NUM”
- Eliminar links y Stopwords (tanto stopwords “default” como stopwords “handcrafted”)
- Creación de unigramas, bigramas y trigramas
- Vectorizar cada uno de los anteriores
- Normalizar los vectores
- “Ensamblar” los vectores

Proceso

Clustering:

Se utilizaron los siguientes algoritmos:

- KMeans: El algoritmo de clustering más común. Agrupa los datos en un número establecido de clusters.
- BisecticKMeans: Clustering jerárquico usando un divisivo. Todos los datos empiezan en un mismo cluster y luego se hacen divisiones recursivamente.
- LDA: Da probabilidades de pertenencia para cada cluster posible.

El que mejores resultados obtuvo fue BisecticKMeans

Proceso

Clasificación de sentimientos:

Se utilizó un dataset de tweets anotados manualmente para entrenar el clasificador NaiveBayes, el cual asume independencia entre features y aplica así el teorema de Bayes para computar la distribución de probabilidad condicional de una clase dada una observación y usarla para predecir.

Luego se utilizó el modelo entrenado para clasificar los tweets de los Oscars.

Visualización

Para visualizar el contenido de cada cluster se usaron nubes de burbujas.

Cada burbuja contiene una palabra, y abajo de ésta un número que indica la cantidad de veces que tal palabra aparece entre los tweets que conforman dicho cluster. El tamaño de la burbuja depende de dicho número. El color de la burbuja será verde si la palabra apareció más veces en tweets con sentimientos positivos que negativos, o será roja si se da lo contrario.

Por otro lado, al seleccionar una burbuja se mostrará, por defecto, cinco tweets positivos y cinco tweets negativos, donde aparece la palabra de la burbuja seleccionada.

Las burbujas pueden arrastrarse para cambiarlas de lugar.

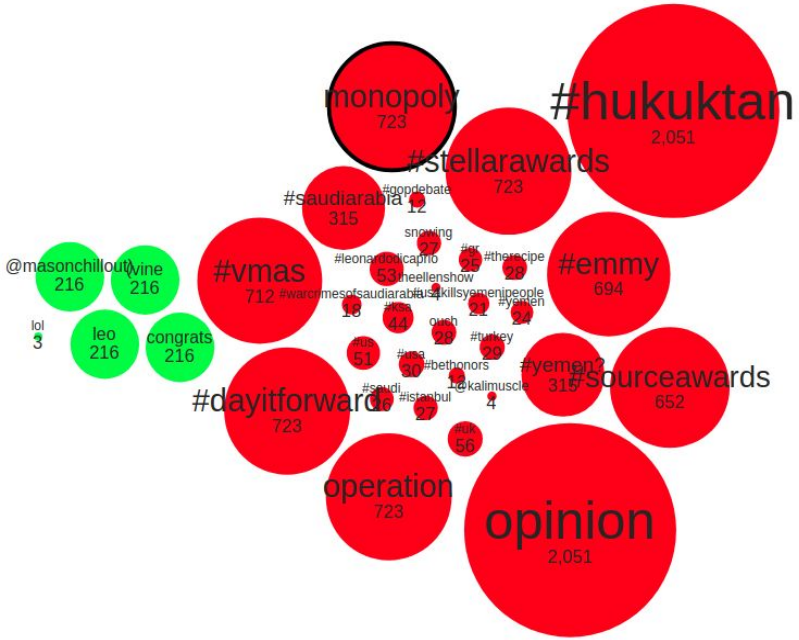
Resultados

Distribución de tweets por cluster:

cluster	tweets_amount
0	3305
1	22323
2	4679
3	2763
4	2272
5	964
6	912
7	837

Resultados

Cluster **número 1 (incluyendo tabla de tweets):**



Positive tweets (Showing 5 maximum)

Friend til the end .. Lol Monopoly and operation . #dayitforward #stellarawards #oscars #vmas… https://t.co/Sv19iMv4FX

Prince and michael jackson Monopoly and operation . #dayitforward #stellarawards #oscars #vmas… https://t.co/r7FiyFssxC

I wont check on you ever again ... Monopoly and operation . #dayitforward #stellarawards #oscars… https://t.co/oiv6Fev5JY

1 day and learned something and left ..... Monopoly and operation . #dayitforward #stellarawards #oscars #vmas #emm… https://t.co/iY3rnGKkDJ

I was born in that spot ... Monopoly and operation . #dayitforward #stellarawards #oscars #vmas… https://t.co/pDcuOHtKBt

Negative tweets (Showing 5 maximum)

Monopoly and operation . #dayitforward #stellarawards #oscars #vmas #emmy #sourceawards… https://t.co/NTJVHxRsrm

Monopoly and operation . #dayitforward #stellarawards #oscars #vmas #emmy #sourceawards… https://t.co/kPqqItuNMT

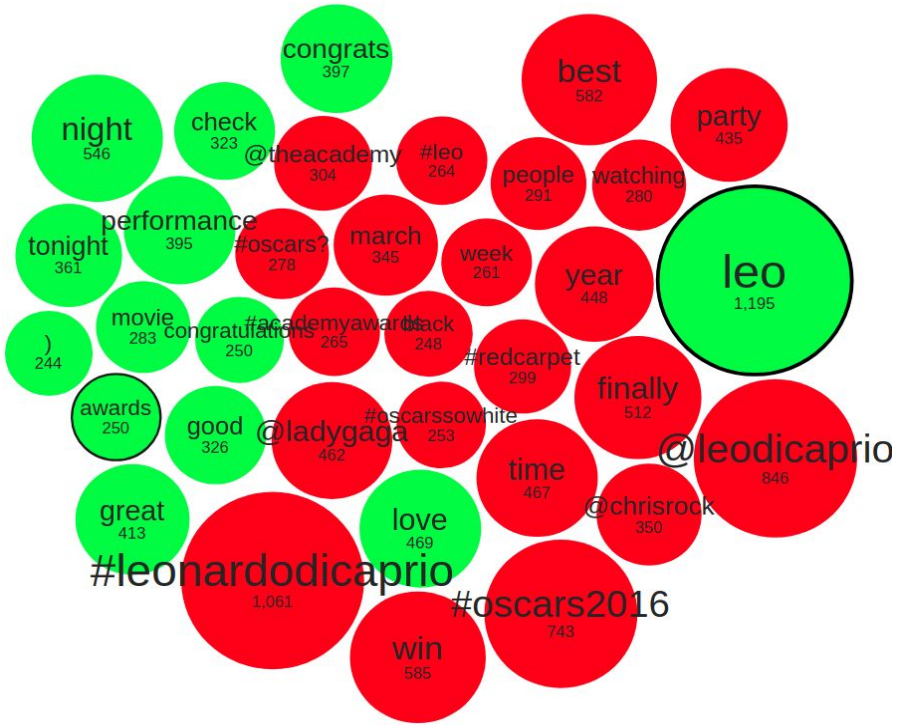
Monopoly and operation . #dayitforward #stellarawards #oscars #vmas #emmy #sourceawards… https://t.co/RLBJEOZgOa

Monopoly and operation . #dayitforward #stellarawards #oscars #vmas #emmy #sourceawards… https://t.co/OAuKWIeZDk

Monopoly and operation . #dayitforward #stellarawards #oscars #vmas #emmy #sourceawards… https://t.co/HA9f3HxPGg

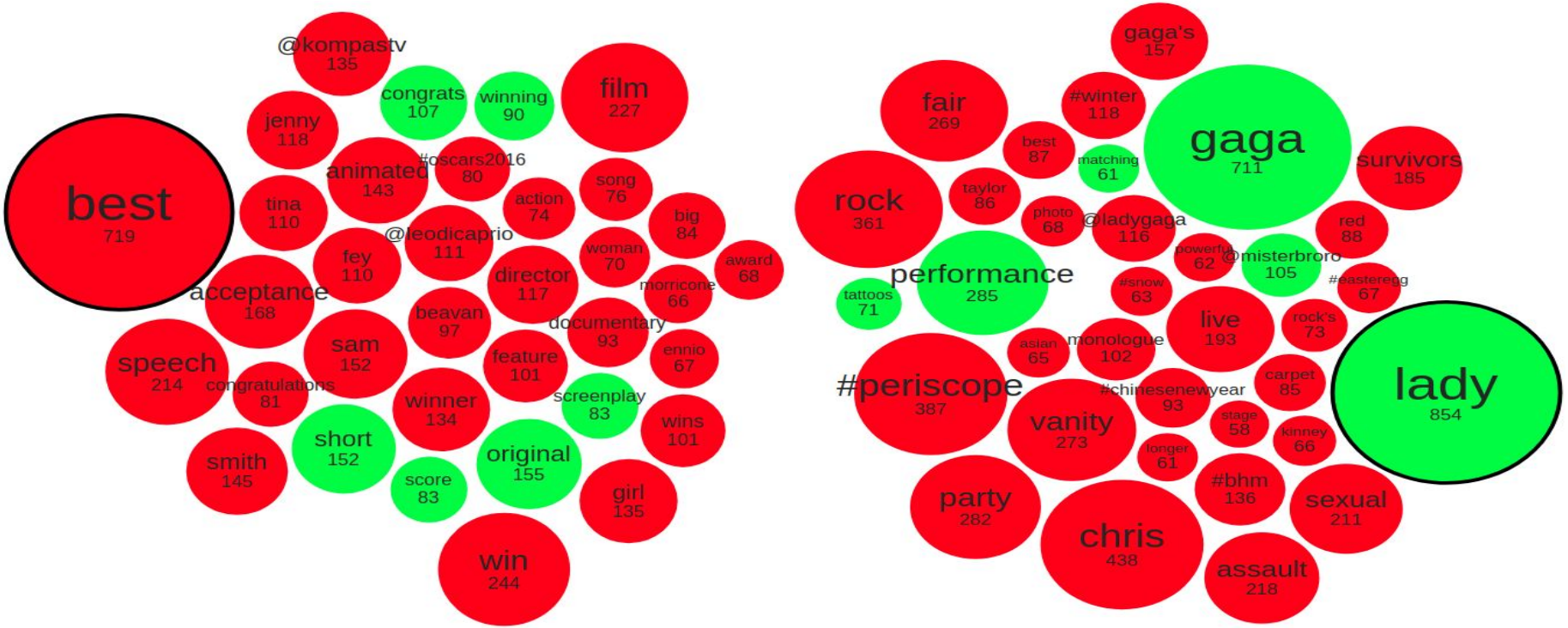
Resultados

Cluster **número 2 y 3:**



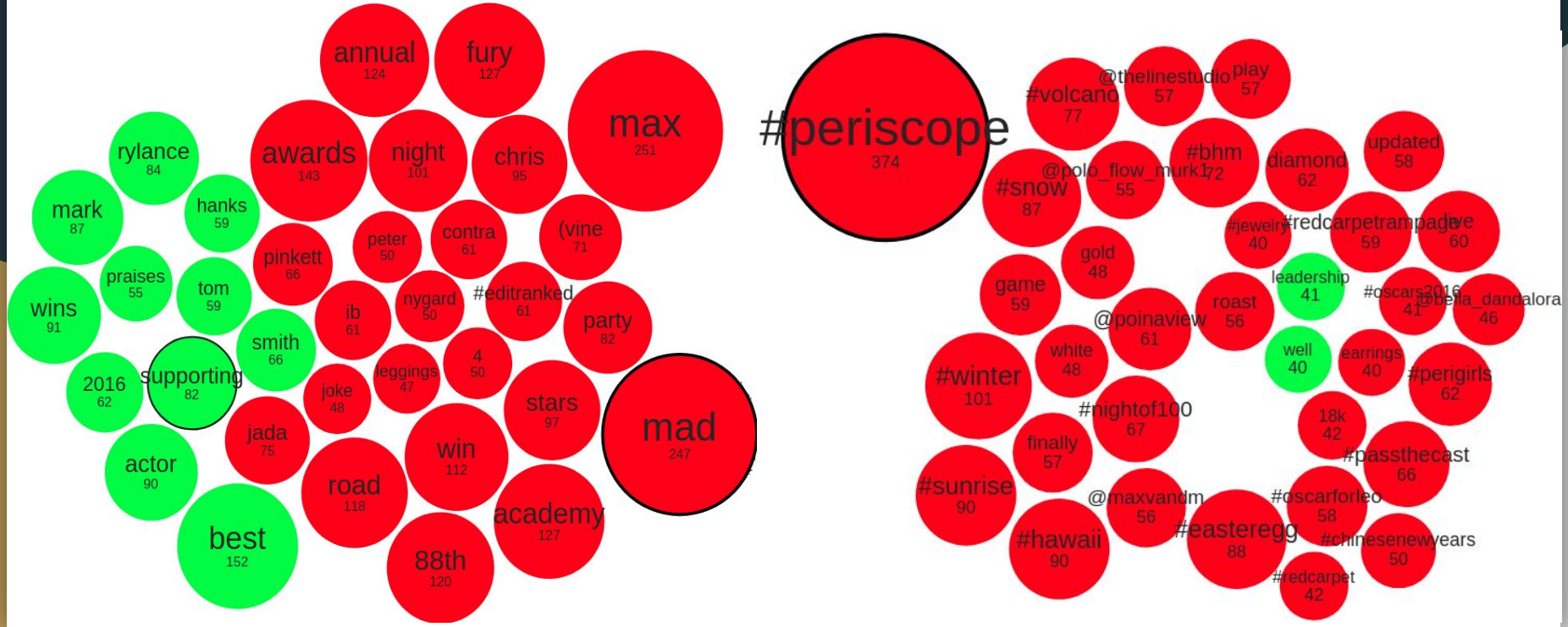
Resultados

Clusters **número 4 y 5:**



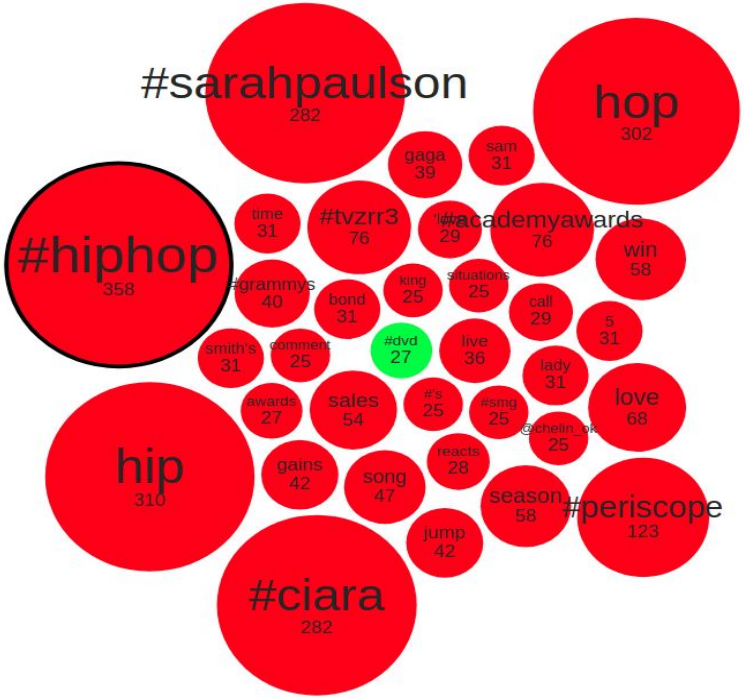
Resultados

Clusters **número 6 y 7:**



Resultados

Cluster **número 8 (incluyendo tablas de tweets):**



Positive tweets (Showing 5 maximum)

Iggy Azalea Covers Elle Canada, Talks Hip-Hop, #hiphop #Oscars #AcademyAwards: #Tvzrr3 https://t.co/lkZp9EPKjH https://t.co/Dwry4S3Ilw

Stevie J Involved In Brawl Outside North Carolina #hiphop #Oscars #AcademyAwards: #Tvzrr3 https://t.co/SnRLoQQsgO https://t.co/qUcVZpkO5h

NO LIMIT TO REUNITE AT BOOM 94.5 FM'S EVOLUTION OF #hiphop #Oscars #Ciara: #SarahPaulson https://t.co/TT9ECbmcnG https://t.co/gqQgx1qhrV

Stevie J Involved In Brawl Outside North Carolina #hiphop #Oscars #AcademyAwards: #Tvzrr3 https://t.co/Rw7Wq9fjls https://t.co/81pzbLLKg3

25 of Lil Duval's Best Hip-Hop Tweets #hiphop #Oscars #AcademyAwards: #Tvzrr3 https://t.co/yzQJXGwuKP https://t.co/TjpoHzsDLK

Negative tweets (Showing 5 maximum)

'Love And Hip Hop Atlanta' Cast; Moniece Slaughter On #hiphop #Oscars #Ciara: #SarahPaulson https://t.co/fEx8mNndEE https://t.co/ys43x7SBgE

Hip Hop Single Sales: Rihanna, G-Eazy & Yo Gotti #hiphop #Oscars #Ciara: #SarahPaulson https://t.co/zQ5C3bi3et https://t.co/s7CSzJdWHL

Love & Hip Hop New York' Season 6 Spoilers: Remy Ma #hiphop #Oscars #Ciara: #SarahPaulson https://t.co/XeOxUph3Ol https://t.co/oSnz0UaZNx

Hip-Hop Reacts To The 88th Academy Awards #hiphop #Oscars #Ciara: #SarahPaulson https://t.co/g7m1V5FUMx https://t.co/PKg4x1KVM6

Love & Hip Hop: New York' Stars Remy Ma And Papoose #hiphop #Oscars #Ciara: #SarahPaulson https://t.co/fenM1eMXpm https://t.co/QRmaHaOA3j