

Proyecto Google Play Store Apps

1st Esneider Pantoja
Facultad de ciencias naturales
e ingeniería
Universidad De Bogota
Jorge Tadeo lozano
Bogota D.C
esneider.pantojad@utadeo.edu.co

2nd Juan Naranjo
Facultad de ciencias naturales
e ingeniería
Universidad De Bogota
Jorge Tadeo lozano
Bogota D.C
juanf.naranjog@utadeo.edu.co

3rd Cristian Hernández
Facultad de ciencias naturales
e ingeniería
Universidad De Bogota
Jorge Tadeo lozano
Bogota D.C
cristianc.hernandezs@utadeo.edu.co

Resumen

En el presente proyecto se contempla el proceso de análisis datos, basados en Google Play Store App. Una base de datos en la cual nos entregan 10 mil datos acerca de apps que ofrece play store. Partiendo de una depuración y organización de datos utilizando las herramientas de python como pandas, se logra una mejor visualización de los resultados que se presentan en el siguiente reporte.

1. Introducción

Si bien muchos conjuntos de datos públicos (en Kaggle y similares) proporcionan datos de la App Store de Apple, no hay muchos conjuntos de datos equivalentes disponibles para las aplicaciones de Google Play Store en la web. Al profundizar, descubrí que la página de la App Store de iTunes implementa una estructura similar a un apéndice bien indexada para permitir un raspado web simple y fácil. Por otro lado, Google Play Store utiliza sofisticadas técnicas modernas (como la carga dinámica de páginas) utilizando JQuery, lo que hace que el scraping sea más desafiante.

2. Método de visualización

Partiendo de la necesidad de tener resultados coherentes y acordes a las necesidades del entorno se trabajó mediante la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) para evaluar la trazabilidad de los datos entregados por la página Kaggle: Google play store. Para ello se aplicaron los 6 pasos guía de dicha metodología (comprensión del problema, comprensión de datos, preparación de datos, modelado, evaluación del modelo e implementación del mismo).

Con un proceso inicial de limpieza, transformación y revisión de los datos usamos las herramientas entregadas en la clase para tal fin. Con PYTHON como herramienta principal se usó (NUMPY) con su extensión (PANDAS), para lograr un tratamiento de los datos entregados y así obtener una visualización coherente y óptima de los datos utilizables y operacionales para un análisis postumo, logrando una toma de decisiones acertiva.

3. Análisis de los datos

3.1. Comprensión del problema:

Se hizo entrega de una base de datos para nuestro archivo, que indica un total de 10841 registros divididos en 13 columnas. Características diferentes para un total de 140933 entradas. Es evidente la necesidad de aplicación de técnicas de Machine learning, para la depuración de datos obsoletos o que estén fuera de contexto para las necesidades de nuestro proyecto.

Nuestro objetivo es identificar el tipo de aplicaciones más demandadas dentro de la plataforma Google Play, bajo limitaciones específicas de rangos de público, como rango de edades de los usuarios, zonas geográficas de mayor demanda y aplicaciones de preferencia.

3.2. Comprensión de datos:

Nos dirigimos a la página Kaggle: Google play store. Descargamos el conjunto de datos referentes al proyecto en la opción 'Download' (2 MB), la cual contiene los datos del proyecto. Conjunto de datos: -googleplaystore.csv (1.3 MB). Este archivo contiene los detalles de las aplicaciones en Google Play, con 13 características que describen una aplicación dada. -googleplaystore_users_reviews.csv (7,31 MB). Este archivo contiene las primeras 100 reseñas más relevantes de cada aplicación. Cada texto

3.3. Preparación de datos:

Con la implementacion de Python lo primero en hacerse fue la importacion de las librerias: numpy es una libreria que da soporte para crear vectores y matrices grandes multidimensionales, junto con una gran colección de funciones matemáticas de alto nivel para operar con ellas. pandas Es una libreria como extensión de NumPy para manipulación y análisis de datos, en particular ofrece estructuras de datos y operaciones para manipular tablas numéricas y series temporales.

-Para examinar el contenido del archivo ejecutamos la funcion head() que nos indica los 5 primeros registros del dataSet.

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

Figura 1. “5 primeros registros del dataSet.” [?]

-Para saber todas las columnas de nuestro archivo ejecutamos el siguiente atributo de dataset columns esto nos permitira conocer todos las caracteristicas de nuestro archivo.

```
Index(['App', 'Category', 'Rating', 'Reviews', 'Size', 'Installs', 'Type',
       'Price', 'Content Rating', 'Genres', 'Last Updated', 'Current Ver',
       'Android Ver'],
      dtype='object')
```

Figura 2. “columnas del archivo.” [?]

-Para conocer el tipo de datos de las columnas ejecutamos info().

-Agrupamos nuestro conjunto de datos por Category: Una vez agrupado nuestro conjunto de datos, generamos los datos estadisticos para aquellas categorias donde se pueda aplicar, esto se logra con describe() Con include = ‘all’ obtenemos el conjunto estadistico de todas las columnas np.number incluye solo las columnas numericas del DataFrame

	Rating							
	count	mean	std	min	25 %	50 %	75 %	max
Category								
ART AND DESIGN	62.0	4.358.065	0.358297	3.2	4.100	4.4	4.700	5.0
AUTO AND VEHICLES	73.0	4.190.411	0.543692	2.1	4.000	4.3	4.600	4.9
BEAUTY	42.0	4.278.571	0.362603	3.1	4.000	4.3	4.575	4.9
BOOKS AND REFERENCE	178.0	4.346.067	0.429046	2.7	4.100	4.5	4.600	5.0
BUSINESS	303.0	4.121.452	0.624422	1.0	3.900	4.3	4.500	5.0
COMICS	58.0	4.155.172	0.537758	2.8	3.825	4.4	4.500	5.0
COMMUNICATION	328.0	4.158.537	0.426192	1.0	4.000	4.3	4.400	5.0
DATING	195.0	3.970.769	0.630510	1.0	3.700	4.1	4.400	5.0
EDUCATION	155.0	4.389.032	0.251894	3.5	4.200	4.4	4.600	4.9
ENTERTAINMENT	149.0	4.126.174	0.302556	3.0	3.900	4.2	4.300	4.7
EVENTS	45.0	4.435.556	0.419499	2.9	4.200	4.5	4.700	5.0
FAMILY	1747.0	4.192.272	0.508026	1.0	4.000	4.3	4.500	5.0
FINANCE	323.0	4.131.889	0.642108	1.0	4.000	4.3	4.500	5.0
FOOD AND DRINK	109.0	4.166.972	0.548070	1.7	4.000	4.3	4.500	5.0
GAME	1097.0	4.286.326	0.365375	1.0	4.100	4.4	4.500	5.0
HEALTH AND FITNESS	297.0	4.277.104	0.617822	1.4	4.100	4.5	4.600	5.0
HOUSE AND HOME	76.0	4.197.368	0.368411	2.8	4.000	4.3	4.500	4.8
LIBRARIES AND EMO	65.0	4.178.462	0.378522	3.1	4.000	4.2	4.400	5.0
LIFESTYLE	314.0	4.094.904	0.693907	1.5	3.800	4.2	4.600	5.0
MAPS AND NAVIGATION	124.0	4.051.613	0.519926	1.9	3.775	4.2	4.400	4.9
MEDICAL	350.0	4.189.143	0.663581	1.0	4.000	4.3	4.600	5.0
NEWS AND MAGAZINES	233.0	4.132.189	0.536707	1.7	3.900	4.2	4.500	5.0
PARENTING	50.0	4.300.000	0.517845	2.0	4.100	4.4	4.675	5.0
PERSONALIZATION	314.0	4.335.987	0.352732	2.5	4.200	4.4	4.600	5.0
PHOTOGRAPHY	317.0	4.192.114	0.462896	2.0	4.000	4.3	4.500	5.0
PRODUCTIVITY	351.0	4.211.396	0.504931	1.0	4.100	4.3	4.500	5.0
SHOPPING	238.0	4.259.664	0.404577	1.6	4.100	4.3	4.500	5.0
SOCIAL	259.0	4.255.598	0.413809	1.9	4.100	4.3	4.500	5.0
SPORTS	319.0	4.223.511	0.427857	1.5	4.100	4.3	4.500	5.0
TOOLS	734.0	4.047.411	0.616143	1.0	3.800	4.2	4.400	5.0
TRAVEL AND LOCAL	226.0	4.109.292	0.504691	2.2	3.900	4.3	4.400	5.0
VIDEO PLAYERS	160.0	4.063.750	0.551098	1.8	3.800	4.2	4.400	4.9
WEATHER	75.0	4.244.000	0.331353	3.3	4.050	4.3	4.500	4.8

-Visualizacion de los datos

- aplicaciones con calificacion de 5.

Category	Installs	App	Rating	
739	BUSINESS	5+	EB Cash Collections	5.0
740	SOCIAL	50+	UP EB Bill Payment	Details 5.0
741	SOCIAL	10+	DN Blog	5.0
742	SOCIAL	5+	CB Heroes	5.0
743	GAME	50+	Axe Champs! Wars	5.0
...
1003	GAME	10+	211:CK	5.0
1004	DATING	500+	Spine- The dating app	5.0
1005	DATING	100+	Online Girls Chat Group	5.0
1006	TOOLS	100+	BK Formula Calculator	5.0
1007	TOOLS	100+	Jabbla BT	5.0

- tipos de personas que califican las apps:

- Apps por genero

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10841 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column            Non-Null Count Dtype  
--- 
 0   App               10841 non-null   object  
 1   Category          10841 non-null   object  
 2   Rating             9367 non-null   float64 
 3   Reviews            10841 non-null   object  
 4   Size               10841 non-null   object  
 5   Installs           10841 non-null   object  
 6   Type               10840 non-null   object  
 7   Price              10841 non-null   object  
 8   Content Rating    10840 non-null   object  
 9   Genres             10841 non-null   object  
 10  Last Updated       10841 non-null   object  
 11  Current Ver        10833 non-null   object  
 12  Android Ver        10838 non-null   object  
dtypes: float64(1), object(12)
memory usage: 1.1+ MB

```

Figura 3. “tipos de datos.” [?]

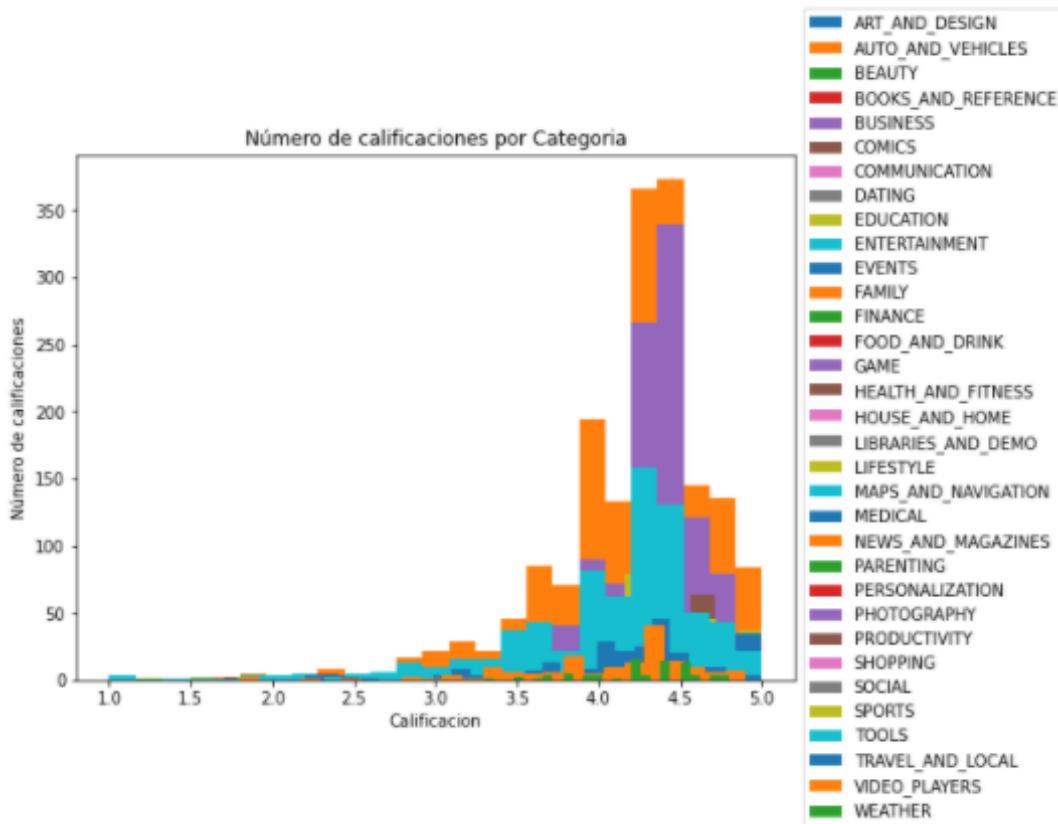


Figura 4. “numero de calificaciones por categoria.” [?]

4. Conclusiones

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

Referencias

- [1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.
- [2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- [3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
- [4] K. Elissa, “Title of paper if known,” unpublished.
- [5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
- [6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
- [7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

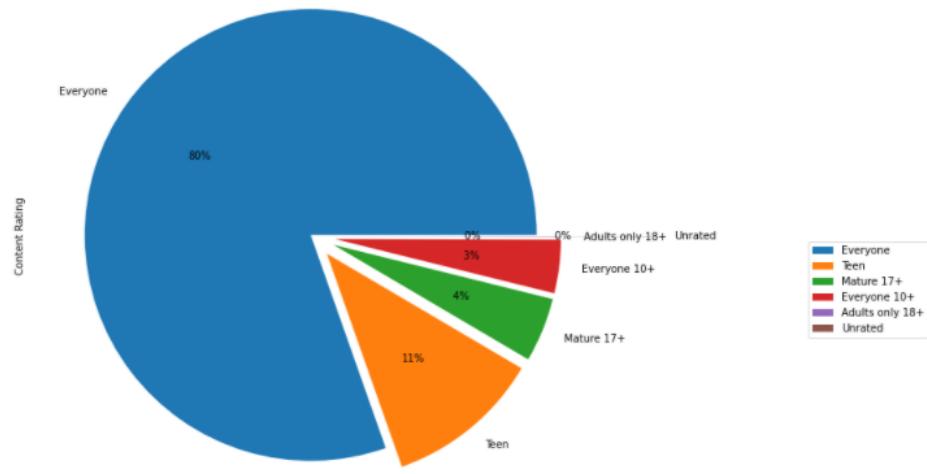


Figura 5. “tipos de personas que califican las apps.” [?]

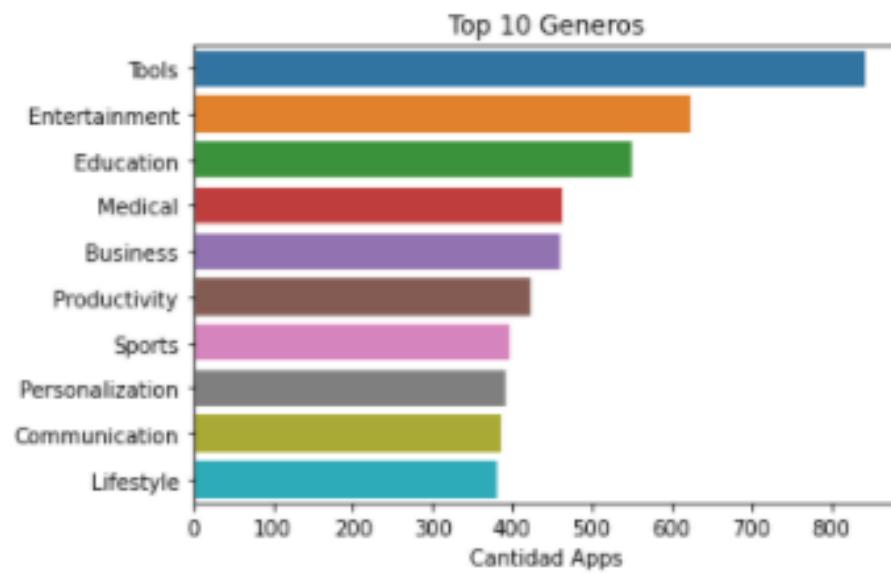


Figura 6. “Apps por genero.” [?]