PROYECTO PARCIAL
ANÁLISIS EXPLORATORIO DE DATOS
FIEC – ESPOL
2017 – 1

Profesor: Carmen Vaca, PhD.

Project: Valuable insight from spatio temporal data

Introducción

En su proyecto parcial de análisis exploratorio de datos usted debe demostrar la habilidad de:
a.) Recolectar datos de plataformas online,
b.) Procesar datasets combinando diferentes fuentes de datos,
c.) Limpiar datasets detectando ruido,
d.) Discretizar data para hacer cálculos por diferentes grupos encontrados en su dataset
e.) Analizar datos empleando métricas propuestas por otros (citando los papers).
f.) Comunicar visualmente los hallazgos encontrados en su dataset.
g.) Utilizar alguna herramienta no vista en clases para el análisis
h.) Escribir un artículo sobre el análisis exploratorio realizado

Dataset

- Su dataset debe tener al menos algunos cientos de miles de registros.
- Su dataset debe tener al menos un campo con información temporal y resulta recomendable tener al menos un campo con información espacial. Use datetime para convertir la columna fecha,
- Escoja nombres adecuados para las columnas de su dataset

Artículo

El artículo debe contener las siguientes secciones:
1. Introducción.
- Qué problema busca resolver en su análisis?
- Qué se ha hecho antes (al menos dos referencias bibliográficas)?
- Cuál es su aporte (sueñen con los aportes a realizar)?

2. Dataset
- El tiempo de recolección de datos uando delays de al menos un segundo entre requerimientos si fuera un crawler o un delay apropiado de acuerdo al rate limit del API utilizado.
- Descripción del método de colección de datos para cada dataset utilizado. Recuerde que usted debe utilizar al menos dos datasets.

- Descripción del filtrado aplicado a los datos (e.g. remoción de data no relevante).
- Estadísticas del dataset original y el dataset del filtrado.
- Periodo de colección de los datos.
- Pre-procesamiento para remoción de outliers, ruido.
- Ipython notebook con el código usado para la extracción de datos
- Link al dataset.
- Ejemplo: Ver Apéndice A
-

3. Methods
- Métodos para categorizar datos (discretización) y razones por las cuales escogieron ese método. De preferencia una referencia a un paper que haga algo parecido.
- Dos métricas extraídas de un paper y aplicadas a sus datos. Si usted tiene una métrica complicada de implementar puede justificar el haber usado una sola métrica. (cada dos miembros del grupo deben encargarse de una métrica).
- Métodos de análisis. Por ejemplo: análisis de covarianza, de correlación, análisis de distribución de los datos.
- Transformaciones aplicadas a los datos (por ejemplo: logaritmo para aproximar una distribución normal).
- Ipython notebook con el código usado para el análisis de datos
- Métodos para dividir los datos en zonas geográficas. Por ejemplo:

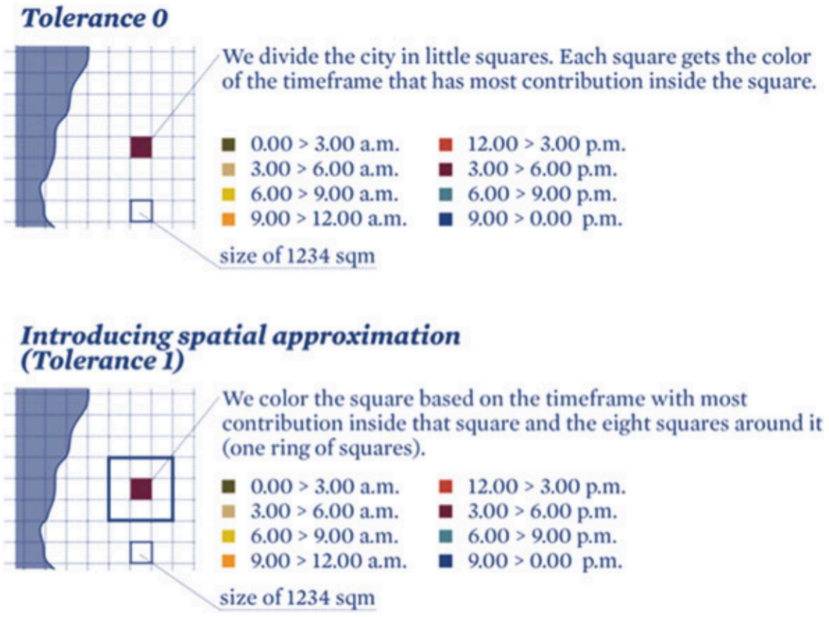

Fig. 5.13 Geographies of Time, New York, an explanation of the process. Available online on Springer Extra Materials (http://extras.springer.com/)

4. Findings/Discussion

- Análisis sobre los findings en su dataset. Esta sección es un componente importante de su reporte. Se esperan unas mínimas conclusiones sobre hipótesis planteadas en secciones anteriores. Estas hipótesis pueden haber sido extraídas de la literatura.
- En los findings es importante describir los insights obtenidos para diferentes momentos de tiempo en su dataset o para diferentes áreas geográficas. Adicionalmente, explicar las diferencias encontradas en diferentes grupos detectados en su data.

5. Visualización

Usted debe incluir visualizaciones que considere útiles para comunicar hallazgos en el análisis realizados.

- No use pie charts o gráficos producidos en Excel.

La calificación de visualizaciones tendrá tres niveles asociados a los puntos que usted obtendrá. Cada nivel asigna puntos incluyendo los requisitos del nivel anterior. A continuación se detallan los niveles

Nivel básico:

- Distribución
- Al menos 2 graficos de análisis univariado comparando diferentes grupos
- Al menos 2 gráficos de análisis multivariado

Nivel moderado:

- Uso de facets para determinar las relaciones que aparecen cuando se incluye una variable nominal en el análisis.
- Time series
- Bubble charts: https://plot.ly/pandas/bubble-charts/

Nivel excelencia

- Small multiples https://flowingdata.com/2014/10/15/linked-small-multiples/
- Geographic visualization: Choroplet, http://bl.ocks.org/mbostock/4060606

Las visualizaciones deben estar comentadas en el artículo

Apéndice A: Ejemplos

Data Collection and Analysis

We first identified sixteen social media sources that citizens use to discuss Humboldt Park (HP) by 1) using search engines and 2) asking HP residents which websites they use to discuss neighborhood issues. **We developed data scrapers in Python to collect online conversations,** which were typically forum posts in the form of username, subject, message, date, and URL**.** ***We scraped nearly 12,000 messages*** in total, reading each and **removing those that were unrelated to the neighborhood** (e.g., random advertisements, spam, trolling messages). As shown **in Table 1, we collected 10,472 messages from various sources that spanned dates from July 2005 to August 2014**. To analyze the data, we first created a codebook with 102 codes based on a similar project conducted with in-person interviews and

observations [9]. Each of the four team members (two professors and two graduate students) coded a random 10% (roughly 1000 messages) of the data based on the initial codebook. We met weekly to discuss modifi- cations to the codebook, which included adding and deleting codes as well as redefining existing codes. After four weeks of iterating on the codebook based on examining the data, two graduate students were tasked with coding the entire dataset by first calculating interrater reliability (Cohen's Kappa 81%) and then splitting the data in half, each receiving half the messages from each source. The team met weekly to discuss the progress and challenges faced …[1]

[1] Erete, Sheena, et al. "That Neighborhood is Sketchy!: Examining Online Conversations about Social Disorder in Transitioning Neighborhoods."*Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2016.

Facebook was chosen as the main social media platform because of its popularity. Pew Research Center reported Facebook as the top social media platform with 71% of U.S. online adults (or 58% of all U.S. adults) as its users [1]. None of the other popular social media, including LinkedIn, Pinterest, Twitter, and Instagram, exceeded 28% market share. In the same study, Facebook users also reported the highest level of use frequency, with 70% checking into Facebook on a daily basis. Overall, Facebook is positioned to be a primary choice for businesses to communicate and interact with users online. The sample of small business was drawn from a mid-size city in north Texas with a population of approximately 379577 [19]. Companies with a corporate office in the chosen city and less than 50 employees qualified for the study. **A total of 12,904 small local businesses in this city were identified via a commercial source, ReferenceUSA Online Database in March 2014**. This population includes businesses that have multiple locations of operations, such as chain convenience stores and insurance agencies. For the purpose of this study, each of these franchise stores and service providers is treated as an independent member of the population. From this population, 480 businesses were randomly selected to form the sample of this study. Based on the North American Industry Classification System (NAICS), entries of this sample were categorized to 22 different industries. Health care and social assistance, retail trade, and professional, scientific, and technical services were the top three industries comprising 42.5% of the sample. The distribution of all industries in this sample is included in Table 1. 11 industries, each representing less than 4% of the sample, were combined and reported in the "other" category. Each entry of this sample was then searched on both Facebook and Google to find their Facebook pages. [2]

[2] Chyng-yang Jang. 2015. The (lack of) use of Facebook by small businesses. In Proceedings of the 2015 International Conference on Social Media & Society (SMSociety '15), Anatoliy Gruzd, Jenna Jacobson, Philip Mai, and Barry Wellman (Eds.). ACM, New York, NY, USA, , Article 16 , 6 pages.

Methods

4.2 Message Types

To explore whether different message types show different diffusion patterns, three types of messages were defined: ad-hoc reporting, situation verifying information, and action-supportive messages, developed based on the previous literature [25,26]. We selected 5% of collected data by daily-volume based stratified sampling and had three coders categorize the selected tweets' message type, with Cohen kappa ranged between .709 and .732 (Table 1)

Entregables

- Artículo en inglés.
- Ipython notebook para las secciones en que haya sido requerida

- Bitácora de trabajo en equipo:
    - Pueden usar Teams de Microsoft, un repositorio de código online o la herramienta que ustedes escojan para documentar la contribución de cada uno.
    - Debe ser visible las contribuciones individuales por fecha
    - Deben tener al menos tres discusiones (chats documentados) sobre el proyecto