Data Science and Visualization

Randbee

Cutting edge technologies and data analysis to support policy-making
on environmental and socioeconomic monitoring

“NOWCASTING”: haciendo predicciones sobre el tiempo presente

Los enfoques tradicionales de “forecasting” usan estadísticas oficiales como indicadores precisos de la actividad social y económica. Sin embargo, muchas de estas estadísticas oficiales sólo están disponibles en frecuencias mensuales, trimestrales o incluso anuales.

Esto significa que pueden ser insuficientes para detectar cambios inesperados durante periodos de gran incertidumbre (e.g., pandemia de COVID-19, guerra de Ucrania).

Los modelos de “nowcasting” difieren de los modelos predictivos, puesto que en lugar de intentar predecir el futuro, se centran en predecir el presente, “rellenando” los huecos de información existentes.

Randbee

NUESTRO DESAFÍO

Generar modelos de “nowcasting” basados en técnicas de ML: e.g., random forests, extreme gradient boosting, stacked ensembles, and neural networks: deep neural network (NN), stacked ensembles regression (SE), random forests (RF), and extreme gradient boosting (XGB), long–short term memory (LSTM) neural network)

Para predecir a tiempo real un conjunto de variables socio-económicas todavía por definir:

- EDUCATION AND TRAINING SERVICES CONSUMED
- CHANGE IN HOUSEHOLD CONSUMPTION PATTERNS
- NUMBER OF SEASONAL WORKERS
- INTEGRATION / ACTIVATION IN THE LABOUR MARKET
- PRIVATE TRANSPORT
- HOMELESSNESS (BROADLY DEFINED)
- HOUSING QUALITY AND COST
- UNMET NEED FOR LONG-TERM CARE

Randbee

CUÁL ES NUESTRA PROPUESTA DE TRABAJO PARA LOS ESTUDIANTES

Generar un modelo de “nowcasting” basado en una técnica de ML para predecir a tiempo real una variable socio-económica a partir de “big data” obtenidos de “Google Trends” y de “GDELT” (‘Global Database on Events Location and Tone’).

Randbee

PRINCIPALES TAREAS A REALIZAR

1. Extraer información sobre vólumenes de consultas de “Google search data” agregadas bajo la forma de “Google Trends” para un conjunto predefinido de “categorías” con una frecuencia mensual para diferentes países

2. Extraer información sobre “sentiment indicators” de una serie de temas a partir de la base de datos GDELT, en forma de “Article tone” y “Topic popularity rate”.

3. Construir (y validar) un modelo de “nowcasting” usando redes neuronales (modelos LSTM) para una de las variables socioeconómicas de interés utilizando los datos the Google Trends y de GDELT previamente extraídos como predictores.

Randbee

CHECK POINTS: reuniones online

• **13 marzo al 1 de abril:** Revisión bibliográfica y evaluación técnica, para familiarizarse con documentos relacionados con el proyecto antes de comenzar.

• **1 al 5 abril (reunión el 5 de abril):** i) Extracción de datos de Google Trend; ii) Extracción de datos GDELT;

• **5 de abril al 10 de abril (reunión el 6 de abril):** Modelos de aprendizaje automático LSTM. Diseño, calibración y validación del modelo LSTM utilizando Google Trends y/o datos GDELT

• **12 de abril:** presentación de resultados, incluyendo un informe breve.

Randbee

SEGUIMIENTO CONTINUADO

Más allá de estos check-points, el equipo de Randbee involucrado en el proyecto estará a disposición de los estudiantes para cualquier consulta adicional que pueda surgir entre “check points”. Estas consultas podrán hacerse por e-mail, chat o convocando reuniones online adicionales.

Randbee

MATERIAL DE APOYO

El equipo de Randbee ha preparado un documento en el que se describe en detalle la metodología que se espera que apliquen los estudiantes para el desarrollo de las tareas.

Este documento contiene además referencias a otras fuentes de información que pueden ser de utilidad y que pueden facilitar ayuda en cuanto a los desafíos técnicos de la propuesta.

Randbee

ALGUNOS DETALLES PARA COMENZAR…

Output (variable respuesta): *Population by sex, age, citizenship and labour status*

Datos anuales, desde 1995

El objetivo es modelar una medida de la integración laboral de la población inmigrante.

Consideraremos únicamente ciudadanos extranjeros que estén en activo (agrupando todas las nacionalidades y clases de edad), distinguiendo hombres y mujeres en la muestra total. Los datos se agrupan por países de la UE.

Fuente de datos: Eurostat (API de Eurostat)

Randbee

ALGUNOS DETALLES PARA COMENZAR…

Inputs (variables predictoras): series temporales derivadas a partir de SVI de *Google Trends* y de número de artículos y tono extraídos de *GDELT* para una serie de temas/categorías de interés.

Datos mensuales, desde 2004 (Google Trends) y desde 2014 (GDELT)

El objetivo es obtener múltiples variables predictoras que expliquen la variación en la integración laboral de la población inmigrante.

Para Google Trends, se extraerán datos mensuales. Los datos diarios disponibles de GDELT se agregarán en datos mensuales para obtener así series temporales mensuales.

Randbee

ALGUNOS DETALLES PARA COMENZAR…

Google Trends

En la extracción, cada punto de datos es filtrado por el rango temporal especificado (mensual en nuestro caso) y geográfico (país de la UE), y se divide por el número total de búsquedas para obtener la medida de popularidad relativa.

Las cifras se basan en una distribución uniforme sobre una muestra aleatoria distribuida de búsquedas de Google actualizada una vez al día desde 2004. Por lo tanto, puede haber diferencias entre extracciones con idénticos parámetros de filtrado.

Randbee

ALGUNOS DETALLES PARA COMENZAR…

Google Trends

Secuencia de seis solicitudes idénticas en términos de área geográfica y periodo temporal (“location” y “timestamp”), en una secuencia de 10 minutos y luego promediarlas. Estas seis muestras de SVIs serán posteriormente promediadas y será sobre este valor promedio con el que se construirá el modelo.

Extracción de los datos por “topic”/“categoría” de la API de Google Trends

Randbee

ALGUNOS DETALLES PARA COMENZAR…

Google Trends: “topics”/categorías

- Crisis/Recession: topic - Economic crisis, topic - Crisis, topic – Recession
- Labour Market: topic - Unemployment benefits, topic - jobs, topic - Unemployment, Welfare & unemployment
- Bankruptcy: topic - Bankruptcy, topic – foreclosure
- Credit, Loans & Personal Finance: topic - Investment, topic - Mortgage, topic - Interest rate, Credit & lending, Investing
- Consumption Items & Services: Food & drink, Vehicle brands, Home & garden, Sports, Autos & vehicles, Grocery & food retailers, Vehicle licensing & registration, Hotels & accommodations
- Jobs & Education: Education, Jobs
- News: Business news, Economy news
- Housing: topic - Affordable housing, topic - House price index
- Business & Industrial Activity: Construction, consulting & contracting, Business services, Transportation & logistics, manufacturing
- Health: Aging & Geriatrics, , Medical Facilities & Services, Mental Health

Randbee

ALGUNOS DETALLES PARA COMENZAR…

GDELT

Global Database of Events, Language, and Tone (GDELT) es una plataforma abierta de *big data* de noticias recopiladas a nivel mundial, que contiene datos estructurados extraídos de fuentes de radiodifusión, impresas y web en más de 65 idiomas. GDELT recopila todas las noticias en línea publicadas cada 15 minutos, traduciéndolas automáticamente al inglés y codificando cada artículo de noticias por tema, sentimiento y tono, ubicación y entidad (organizaciones y personas), entre otros.

Randbee

ALGUNOS DETALLES PARA COMENZAR…

GDELT

Global Database of Events, Language, and Tone (GDELT) es una plataforma abierta de *big data* de noticias recopiladas a nivel mundial, que contiene datos estructurados extraídos de fuentes de radiodifusión, impresas y web en más de 65 idiomas. GDELT recopila todas las noticias en línea publicadas cada 15 minutos, traduciéndolas automáticamente al inglés y codificando cada artículo de noticias por tema, sentimiento y tono, ubicación y entidad (organizaciones y personas), entre otros.

GDELT Global Knowledge Graph (GKG) es una base de datos de artículos de noticias donde se codifican entidades, temas, ubicaciones y tono. De este modo, GDELT permite medir el **volumen y tono de las noticias** sobre un tema en un país.

Randbee

ALGUNOS DETALLES PARA COMENZAR…

GDELT

Usaremos GDELT GKG para filtrar noticias relacionadas con una serie de temas relacionados con la variable repuesta, aunque limitándonos exclusivamente a las noticias que se han publicado sobre los países europeos y periodos de interés de acuerdo con la variable respuesta.

Un enfoque ampliamente utilizado para analizar datos GDELT directamente sin descargarlos es utilizar la interfaz Big Query. Sin embargo, se invita a los estudiantes a explorar otras formas de extraer los datos.

Randbee

ALGUNOS DETALLES PARA COMENZAR…

GDELT

- ✓ Usaremos GDELT GKG para filtrar noticias relacionadas con una serie de temas relacionados con la variable repuesta, aunque limitándonos exclusivamente a las noticias que se han publicado sobre los países europeos y periodos de interés de acuerdo con la variable respuesta.
- ✓ Seleccionaremos solo los artículos que tienen **al menos tres palabras** clave relacionadas con cada uno de los temas especificados y una extensión **mínima de 500 palabras**. Si es posible, solo se tendrán en cuenta aquellas publicaciones que hayan estado funcionando a lo largo de todo el periodo de extracción de la información (mes).

Randbee

ALGUNOS DETALLES PARA COMENZAR…

GDELT

- ✓ Como temas, se utilizarán los extraídos de la **taxonomía temática del World Bank Group**, que se incluye en la GDELT Global Knowledge Graph 2.0. Esta taxonomía cubre una gran variedad de temas, desde agricultura y seguridad alimentaria hasta educación, incluyendo salud, desarrollo social, desarrollo urbano, e incluso cuestiones relacionadas con el agua.
- ✓ Específicamente, pueden extraerse algunas categorías de nivel 2 y 3 definidas en esta taxonomía (alrededor de 40 temas diferentes), y calcular para cada una de ellas el **número de noticias y el tono** (utilizando tres dimensiones: positivo, negativo e incertidumbre), normalizadas para el número total de noticias publicadas en un país.

Randbee

ALGUNOS DETALLES PARA COMENZAR…

Modelo de red neuronal de tipo LSTM

- ✓ Se trata de un tipo de red neuronal recurrente (RNN).
- ✓ Los modelos LSTM pueden tener varias capas, lo que significa que el mismo procedimiento se lleva a cabo varias veces y en múltiples estados de celda y de celda oculta.
- ✓ Siguiendo otros estudios previos, se recomienda usar una variación LSTM de tipo “vanilla”. Este “vanilla” LSTM tiene solo una capa de LSTM seguida de una capa densa, que es una capa completamente conectada que mapea las salidas del LSTM hasta la forma deseada.
- ✓ Este tipo de modelo parece capturar bien las dinámicas de las series temporales.

Randbee

ALGUNOS DETALLES PARA COMENZAR…

Modelo de red neuronal de tipo LSTM

- ✓ Primeramente, construir un modelo anual basado en SVI de Google Trends y número y tono de los artículos de GDELT muestreados con una frecuencia mensual.
- ✓ En segundo lugar, el modelo se aplicará a las series mensuales de Google Trends y de GDELT para obtener predicciones mensuales de la variable respuesta.
- ✓ Esto implica asumir que la relación entre los datos de Google Trends y la variable “output” no dependen de la frecuencia a la que se muestreen. Este enfoque nos permite desagregar la variable de salida a una frecuencia mensual.

Randbee

¿ALGUNA PREGUNTA?

Randbee

Gracias por vuestra atención

Randbee

@RandbeeCo
info@randbee.com
Málaga, Spain
+ 34 676 69 12 29