PEC 1 (20% nota final)

Presentación

En esta Prueba de Evaluación Continuada se trabajan los conceptos generales del ciclo de vida de los datos, y se identifican y revisan sus características. También se trabajan los conceptos fundamentales del Web Scraping.

Competencias

En esta PEC se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento, almacenamiento y administración de datos

Objetivos

Los objetivos concretos de esta Prueba de Evaluación Continua son:

- Conocer el ciclo de vida de los datos y los principales tipos de datos.
- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinarios.
- Desarrollar las habilidades de aprendizaje que permitan continuar estudiando de una manera que tendrá que ser en gran medida autodirigida o autónoma.
- Desarrollar la capacidad de busca, gestión y uso de información y recursos en el ámbito de la ciencia de datos.
- Entender la utilidad, la legalidad y algunas características del web scraping.

Descripción de la PEC a realizar

Ejercicio 1 [70%]

Después de leer el recurso "Calvo, M., Pérez, D., Subirats, L. (2019). Introducción al ciclo de vida de los datos." contesta las siguientes preguntas con tus propias palabras:

1 La clasificación ESCO (European Skills, Competences, Qualifications and Occupations) estandariza los perfiles profesionales. Clasifica en ese estándar los 8 perfiles del ámbito de la ciencia de datos indicando su código. [20%]

Un ejemplo de standard o clasificación es la ESCO (European Skills, Competences, Qualifications and Occupations). A partir de la URL https://esco.ec.europa.eu/en/classification/occupation_main se pueden buscar el código de los diferentes perfiles de ciencia de datos:

Perfil	ESCO
Científico de datos	2511.4
Analista de datos	2511.3
Arquitecto de datos	2511.14, 2521.2 y 2511.14
Ingeniero de datos	2529.9, 2511.4 y 2529.9
Estadístico	2120.6
Administrador de base de datos	2521.1
Analista de negocio	2511.9
Líder de ciencia de datos	1330.1 y 2511.8

2 Pon como ejemplo dos ofertas de trabajo (p.ej., Linkedin o Indeed) que correspondan a diferentes perfiles profesionales dentro del ámbito de la ciencia de datos, y justifica por qué. Indica el enlace de la oferta y una captura de pantalla de la misma. Si es necesario, indica cuáles serían las correcciones pertinentes a las ofertas que has encontrado. (Máximo 200 palabras). [10%]

Las dos referencias que se proponen son correctas:

- Machine Learning Scientist a Amazon. Es una oferta para data scientist donde requieren conocimientos de aprendizaje automático, estadística y programación

(lenguajes como Python, Spark e Hive). Piden además Java, que no está en la lista de lenguajes pero que también es interesante porque Spark provee Apis con Java, Scala, Python y R.

Machine Learning Scientist

Amazon Barcelona

Description

Are you interested in building state-of-the-art machine learning systems for the most complex, and fastest growing, transportation network in the world? If so, Amazon has the most exciting, and never-before-seen, challenges at this scale (including those in sustainability, e.g. how to reach net zero carbon by 2040).

Amazon's transportation systems get millions of packages to customers worldwide faster and cheaper while providing world class customer experience – from online checkout, to shipment planning, fulfillment, and delivery. Our software systems include services that use tens of thousands of signals every second to make business decisions impacting billions of dollars a year, that integrate with a network of small and large carriers worldwide, that manage business rules for millions of unique products, and that improve experience of over hundreds of millions of online shoppers.

As part of this team you will focus on ML methods and algorithms for core planning systems, as well as for other applications within Amazon Transportation Services. Current research includes machine learning forecast, online reinforcement learning, and anomaly detection models, among others.

We are looking for a Machine Learning Scientist with a strong academic background, and expertise in either of the following:

- *Graph Neural Networks (GNNs), Temporal Graph Networks (TGNs), and/or Graph Deep Learning (GDL)*
- *Probabilistic Machine Learning*

At Amazon, we strive to be the most customer-centric company and the best employer on Earth. To continue improving, we need exceptionally talented, bright, and driven people. If you'd like to help us build the place to find and buy anything online, and deliver in the most efficient and greenest way possible, this is your chance to make history.

Basic Qualifications

- *PhD in Computer Science, AI, Mathematics, or Statistics with specialization in ML (alternatively, MSc. and 3+ years in a ML scientist role).*
- *Deep knowledge of fundamentals, and the state-of-the art, in relevant areas of ML.*

- *2+ years of hands-on experience in ML research and ML systems.*
- *Expertise, in any of:*
- *Graph Neural Networks (GNNs), Temporal Graph Networks (TGNs), and/or Graph Deep Learning (GDL);*
- *Probabilistic Machine Learning.*
- *Strong coding skills in Python.*

Preferred Qualifications

- *Experience with cloud computing services such as AWS.*
- *Experience with programming languages such as Java, Scala, and/or others.*
- *Experience working effectively with research science, data engineering, and software engineering teams.*
- *Proven track record of innovation in creating novel algorithms and applying the state-of the-art.*
- *Strong verbal and written communication skills.*
- *Strong publication/scientific track record.*

Company - Amazon EU SARL (Spain Branch). Job ID: A1701627

- Data Engineer Data platform a Glovo. Es una oferta para data engineer que permite trabajar con Amazon Web Services (AWS), Spark, Hive y Python. También habla que tendrá interacción con Data Scientists and Business Analysts.

Data Engineer - Data Platform (Madrid)

Glovo Madrid

About Glovo

We're a Barcelona-based startup and the fastest-growing delivery player in Europe, Hispanic America and Africa. With food at the core of the business, Glovo delivers any product within your city at any time of day.

Our vision and ambition are not only to make everything immediately available in your city but it is also to offer our employees the job of their lives. A job where you'll be challenged and have the most fun working in through tech-enabled experiences. Your Work-life Opportunity

We are looking for talented and passionate Data Engineer to join the Data Engineering team in our Madrid office.

Glovo has a culture of data-driven decision-making, and demands data that is timely, accurate, and actionable. We grow really fast collecting Terabytes of data from tens of data sources and providing interfaces for our internal customers to access and query the data hundreds of thousands of times per day.

As a Data Engineer you will be building and constantly improving Glovo's reliable and scalable Big Data Platform using technologies like Amazon Web Services (AWS), Spark, Python and many more. Joining us your work will have an immediate influence on the shape of data consumed by teams across Glovo including Central BI, Data Scientists and Business Analysts.

Your depth of experience and past achievements will speak for itself having helped deliver on data platform wide projects that have had significant impact. You have developed complex, scalable and well designed data pipelines and defined engineering standards which have helped different data teams achieve efficiency while providing mentoring and technical leadership where necessary. You will be at times working independently and at other times within a team to achieve a given goal. Others respect you and you are a sought out Data Engineer because of your expertise and depth of knowledge.

Be a Part Of a Team Where You Will

- *Design, implement and keep improving Glovo Data Platform*
- *Build scalable data pipelines using different technologies*
- *Participate in the development of Data Lake, Data Warehouse, different methods of data ingestion and Self Service ETL tools using best architectural practices.*
- *Mentor & share technical expertise with data engineers, data scientists, BI analysts and other technology colleagues*
- *Be on top of new technologies and industry trends*

You Have

- *At least 3+ years of software/data engineering experience*
- *At least 2+ years experience in Python and Spark*
- *Professional experience building complex ETLs/data pipelines*
- *Working experience with Amazon Web Services / Google Cloud Platform*
- *Experience with task orchestration tools (Airflow, Luigi)*
- *Cloud Data Warehousing experience in Redshift or another distributed platform (e.g. Hadoop + Hive/Presto, BigQuery or Snowflake)*
- *Experience in Data Streaming (Spark, Flume, Kafka, Kinesis, Flink, etc.)*
- *Import and transform data from many third-party APIs*
- *Strong analytical and problem-solving skills*
- *Very good English*

Nice To Have

- *Experience working with AWS EMR, AWS Glue, Databricks*
- *Experience with Docker, Kubernetes*
- *Experience in building Data Lake*
- *Orchestration of Machine Learning pipelines*

Experience Our Glovo Life Benefits

- *Enticing Phantom Shares plan*
- *Attractive Relocation package*
- *Comprehensive Private Health Insurance*
- *Cobee discounts on kindergarten, transportation, and food*
- *Free monthly Glovo credits to spend on our restaurant products (and zero Glovo delivery fee on all Glovo orders!)*
- *Cool perks such as fresh fruit and healthy snacks every day, beers on Fridays, Culture Days every 2 months!*
- *Discounted Gym memberships*
- *Flexible working environment*

What You'll Find When Working At Glovo

- *Gas: We work hard with energy and passion for what we do.*
- *Care: We act in the best interest of a sustainable future.*
- *Good vibes: We always see the positive side in every situation and act with fairness and honesty with everyone.*
- *Stay Humble: We embrace mistakes and feedback to learn from them.*
- *Glownership: We roll up our sleeves and get work done no matter our position and level.*

If you believe you match these values, we look forward to meeting you!

At Glovo we believe that diversity adds incredible value to our teams, our products, and our culture. We know that the best ideas and solutions come by bringing together people from all over the world and by fostering a culture of inclusion where everyone feels heard and has the chance to make a real impact. It's because of this that we are committed to providing equal opportunities to talent from all backgrounds.

Wanna take a peek into what it's like to work at Glovo? Follow us on Instagram and like us on Facebook!

Glovo is transforming the way consumers access local goods, enabling anyone to get almost any product delivered in minutes. Our on-demand logistics connect customers with independent local couriers who acquire goods from any restaurant or store in a city, as well as deliver urgent packages for a variable fee. As of September 30, 2019, we're currently present in more than 26 countries across Europe, Latin America, Africa, and Asia.

3 Además de las habilidades técnicas, las soft skills o habilidades blandas son muy importantes en la vida laboral. ¿Cómo harías para averiguar cuáles son las soft skills requeridas habitualmente en los perfiles de ciencia de datos? Indica tres de las soft skills que has encontrado. (Máximo 100 palabras). [10%]

Se podría realizar una búsqueda mediante web scraping de portales de búsqueda de trabajo como Indeed con diferentes soft skills (comunicación, adaptabilidad, trabajo en equipo, liderazgo, etc.) y posteriormente mediante procesado del lenguaje natural se miraría qué ofertas corresponden a los perfiles de ciencia de datos. Tres ejemplos de soft skills serían comunicación, adaptabilidad y trabajo en equipo.

4 Pon un ejemplo de datos restringidos, privados y públicos. [10%]

Restringidos: Datos de los hospitales por ejemplo, se debe firmar un acuerdo de confidencialidad para poder acceder a ellos.
Privados: Datos de la bolsa tras realizar web scraping.
Públicos: Datos de la iniciativa Open Data Barcelona.

5 Lista los diferentes factores que influyen en la calidad de los datos y pon un ejemplo que no sea el que se explica en el material de la asignatura. (Máximo 300 palabras). [10%]

- Exactitud: Un economista introduce los datos con puntos indicando miles, pero va a Estados Unidos y se interpreta como un decimal.
- Completitud: Un vendedor de casas introduce los datos que describen la casa que quiere vender, sin dejar ningún valor en blanco.
- Consistencia: Si se mide el índice de masa corporal (IMC), tiene que estar en consonancia con peso/altura^2.
- Puntualidad: Diferencia entre la fecha de un examen, el resultado de la nota, y la introducción al expediente académico.
- Unicidad: En un examen de una asignatura, se comprueba que no hay una misma persona con dos notas.
- Validez: En una base de datos hay un valor de una persona que pesa 1000 kg. Este dato no es válido porque se encuentra fuera de rango, una persona no puede pesar 1000 kg.

6 Explica cuáles son los beneficios de las bases de datos no relacionales y pon 3 ejemplos de estas bases de datos (es suficiente con citar el nombre). (Máximo 100 palabras). [10%]

No es necesario conocer a priori el que se quiere almacenar y son más flexibles puesto que se puede almacenar cualquier tipo de dato sin importar la estructura. Tres ejemplos de estas bases de datos son: MongoDB, Redis y Cassandra.

Ejercicio 2 [30%]

Después de leer el recurso “Subirats, L., Calvo, M. (2019). Web Scraping”, capítulos 1 y 7. Contesta las siguientes preguntas con tus propias palabras:

1 Cita el nombre de tres librerías útiles para un proyecto de web scraping (Máximo 100 palabras). [10%]

Se pueden utilizar librerías como Request y BeautifulSoup, aunque hay otras librerías utilizadas como Scrapy.

2 Explica dos buenas prácticas, a la hora de realizar web scraping, para evitar ser bloqueado (Máximo 100 palabras). [10%]

No saturar el servidor con peticiones web. Esto puede evitarse colocando retardos definidos entre las peticiones. De esta manera, no se sobrecarga el servidor y no levanta alarmas que resulten en un bloqueo.

Modificar el *user-agent*. Al poseer un *user-agent* propio dentro de la cabecera se evita los bloqueos por default de parte del *webmaster*

3 Pon un ejemplo donde realizar web scraping podría ser útil para un negocio. (Máximo 100 palabras). [10%]

Si imaginamos que tenemos una empresa de zapatos y queremos saber los precios de la competencia, haríamos web scraping para obtener esta información y poder fijar nuestros precios según nos convenga.

Recursos

Los siguientes recursos son de utilidad para la realización de la PEC:

Básicos

- Calvo, M., Pérez, D., Subirats, L. (2019). Introducción al ciclo de vida de los datos. Editorial UOC.
- Subirats, L., Calvo, M. (2019). Web Scraping. Editorial UOC.

Complementarios

- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Capítulo 1. Introduction to Web Scraping.
- Minguillón, J. (2016). Fundamentos de data Science. Editorial UOC.

Criterios de valoración

La ponderación de los ejercicios es la siguiente:

- Ejercicio 1.1: 20%
- Ejercicio 1.2: 10%
- Ejercicio 1.3: 10%
- Ejercicio 1.4: 10%
- Ejercicio 1.5: 10%
- Ejercicio 1.6: 10%
- Ejercicio 2.1: 10%
- Ejercicio 2.2: 10%
- Ejercicio 2.3: 10%

Se evaluará la precisión de los ejemplos así como lo respecto al número de palabras máximo establecido para cada pregunta. La idoneidad y claridad de las respuestas también será evaluada.

Formato y fecha de entrega

Hay que librar un único documento Word, Open Office o PDF (**preferiblemente este último**) con las respuestas a las preguntas.

Este documento se tiene que librar en el espacio de Entrega y Registro de AC del aula antes de las **23:59** del día **18 de octubre**. No se aceptarán entregas fuera de plazo.