Practica Calificada 4

Penadillo Lazares Wenses Johan Villarroel Lajo Gerald Takeshi

Universidad Nacional de Ingeniería

12 de junio de 2024

El presente trabajo tiene como propósito resolver la aplicación de consultas y análisis de datos mediante el uso del lenguaje de programación **Scala**, Spark y **Spark SQL** con *UDF* y *MLlib*. Para ello se usara una base de datos de ofertas de trabajo para la carrera de ciencias de la computación en la plataforma de Linkedin obtenido con técnicas de Webscraping.
Se abordarán consultas complejas aplicando distintas funciones de filtrado, agregación, selección, agrupamiento y ordenación con el fin de poder realizar un análisis sobre las tendencias de las ofertas laborales.

Apache Spark es un motor unificado de analíticas para procesar datos a gran escala que integra módulos para SQL, streaming, aprendizaje automático y procesamiento de grafos. Spark se puede ejecutar de forma independiente o en Apache Hadoop, Apache Mesos, Kubernetes, la nube y distintas fuentes de datos [1].

Características:

- Velocidad
- Facilidad de Uso
- Integración con Herramientas [2]

Spark está conformado por: [2]

- **Spark Core:** La base del proyecto Spark, que proporciona gestión de memoria, planificación de tareas, recuperación ante fallos, etc.
- **Spark SQL:** módulo de Spark que permite utilizar datos estructurados, se puede consultar datos estructurados de programas de Spark con SQL o con la API de DataFrame que resulte más cómoda.
- **Spark Streaming:** Dacilita la creación de soluciones de streaming escalables y tolerantes a fallos.
- **MLlib:** Biblioteca escalable de aprendizaje automático de Spark. Contiene numerosos algoritmos de aprendizaje de uso habitual, como clasificación, regresión, recomendación y agrupación en clústeres.
- **GraphX:** API de Spark para grafos y computación en paralelo de grafos. Es flexible y funciona a la perfección tanto con grafos como con colecciones.

- **Consultas SQL:** Permite ejecutar consultas SQL en Spark.
- **DataFrames:** Proporciona una abstracción de datos estructurados similar a una tabla en una base de datos relacional.
- **Compatibilidad con Hive:** Soporta consultas HiveQL y se puede integrar con el metastore de Hive [3].

NOTA: ***(Uso en el Proyecto)***

- **Consulta de Datos:** Para ejecutar consultas SQL en los datos cargados desde un archivo CSV o una base de datos PostgreSQL.
- **Transformaciones:** Para realizar transformaciones en los datos utilizando DataFrames y Datasets.

También conocido por sus siglas en inglés UDF (User-Defined Functions), son funciones definidas por el usuario que permite reutilizar la lógica personalizada en el entorno del usuario.

NOTA: *(Uso en el Proyecto)*

- Personalización: Permite generar piezas de código personalizadas que no está disponible en las funciones integradas de Spark.
  Por ejemplo, cuando se filtran los trabajos que requieren una determinada habilidad y utilizan un conjunto específico de lenguajes de programación.

MLlib es la biblioteca de aprendizaje automático (ML) de Spark. Su objetivo es hacer que el aprendizaje automático práctico sea escalable y sencillo. Entre las herramientas que proporciona tenemos: [4]

- **Algoritmos de aprendizaje automático:** algoritmos de aprendizaje comunes como clasificación, regresión, agrupación y filtrado colaborativo.
- **Caracterización:** extracción de características, transformación, reducción de dimensionalidad y selección.
- **Pipelines:** herramientas para construir, evaluar y ajustar ML Pipelines
- **Persistencia:** guardar y cargar algoritmos, modelos y Pipelines
- **Utilidades:** álgebra lineal, estadística, manejo de datos, etc.

También conocido por sus siglas en inglés RDD (Resilient Distributed Dataset), es una colección de elementos particionados a través de los nodos del clúster que pueden operarse en paralelo. Los RDD se crean a partir de un archivo en el sistema de archivos de Hadoop (o cualquier otro sistema de archivos compatible con Hadoop) o una colección existente de Scala en el programa driver, y se transforman [5].

NOTA: ***(Uso en el Proyecto)***

- Transformaciones y Acciones: Los RDDs soportan dos tipos de operaciones: transformaciones (e.g., map, filter, reduceByKey) y acciones (e.g., collect, count, saveAsTextFile).
- Persistencia y Caché: Los RDDs pueden ser persistidos en memoria o disco para optimizar la reutilización de datos en múltiples operaciones.

Para el presente trabajo usaremos la base de datos [6], la cual fue creada extrayendo datos de ofertas de trabajo de Linkedin usando técnicas de Webscraping.
Descripción de campos relevantes de las bases de datos:

- **job_title:** Titulo de la oferta publicada.
- **job_location:** Ubicación del puesto de trabajo.
- **search_city:** Ciudad usada en la consulta de la búsqueda del trabajo.
- **search_country:** País usada en la consulta de la búsqueda del trabajo.
- **job_level:** Nivel de trabajo pedido para la posición de trabajo.
- **job_type:** Tipo de empleo ofrecido por la oferta de trabajo.
- **job_summary:** Descripción de la oferta de trabajo.
- **job_skills:** Skills solicitadas para la oferta del puesto de trabajo.

Limpieza: El proceso de limpieza consintió en una serie de pasos donde se aplico expresiones regulares, técnicas de NLP y filtrado de palabras clave:

- Reconocimiento de columnas útiles para el desarrollo del presente trabajo (job_title, job_summary, job_skills).
- Eliminación de columnas no útiles para este trabajo.
- Eliminación de caracteres no alfanuméricos del alfabeto ingles.
- Normalización del texto (convertir mayúsculas a minúsculas).
- Eliminar *stopwords* que no agregan información relevante.
- Lematización para reducir el numero de palabras del texto.
- Filtrado de palabras clave para extraer información relevante y crear nuevos campos.
- Guardar todos los registros filtrados en un archivo csv que luego fue almacenado en una base de datos.

Exploración de datos campo *job_title*:

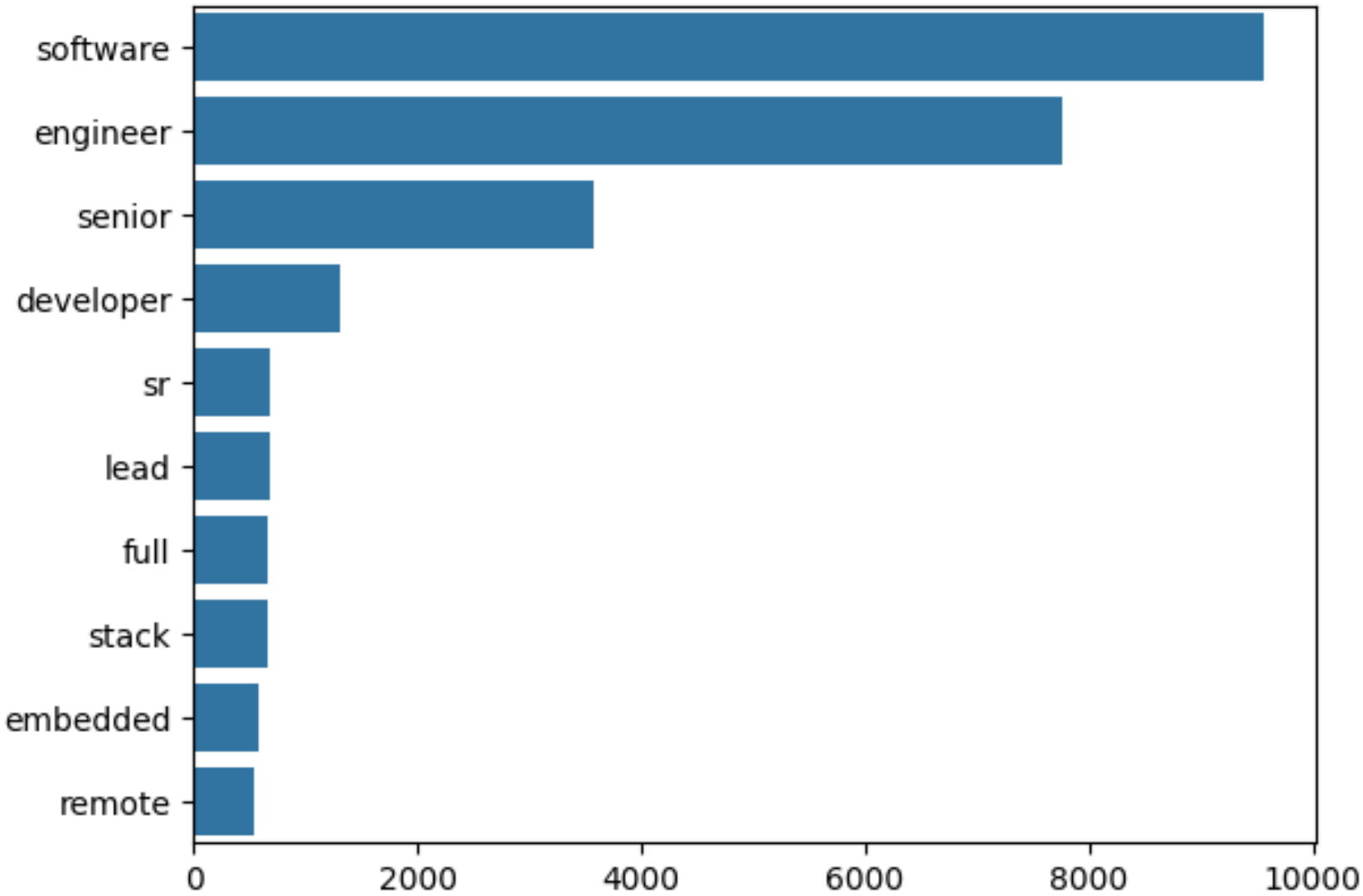


Figura: Palabras mas frecuentes en el campo job_title.

Exploración de datos campo *job_title*:

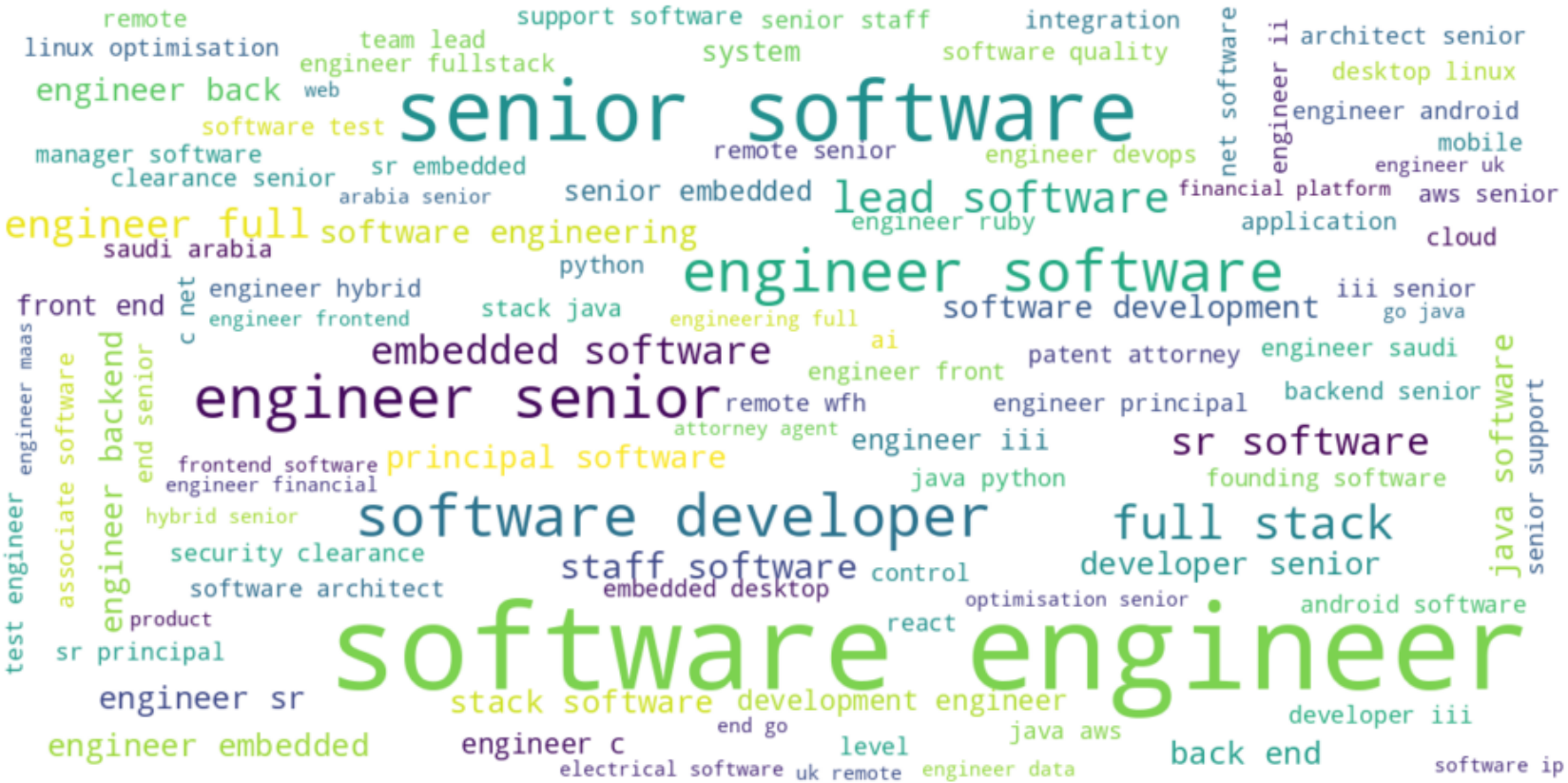


Figura: Nube de palabras del campo job_title.

Exploración de datos campo *job_summary*:

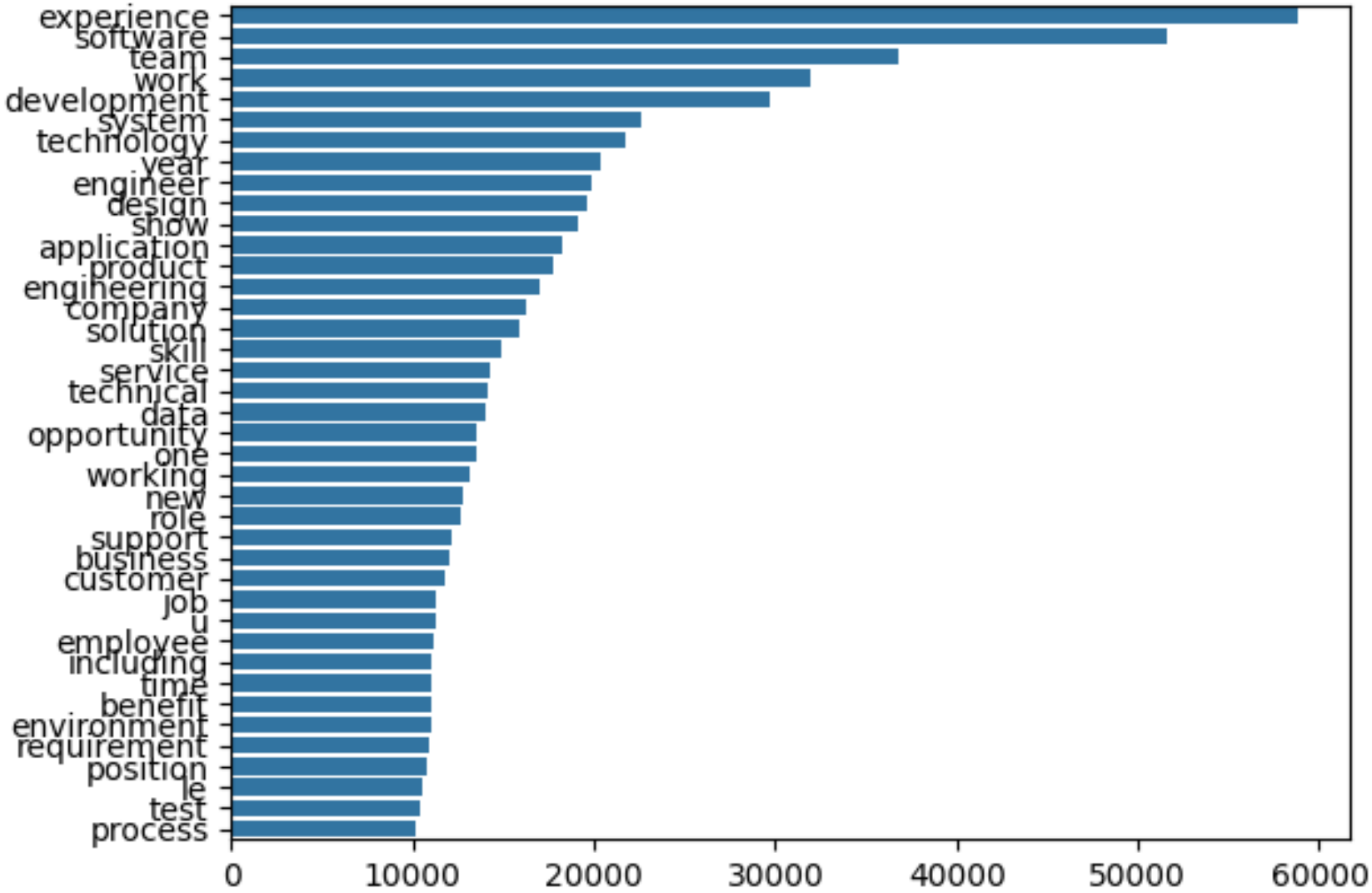


Figura: Palabras mas frecuentes en el campo job_summary.

Exploración de datos campo *job_summary*:

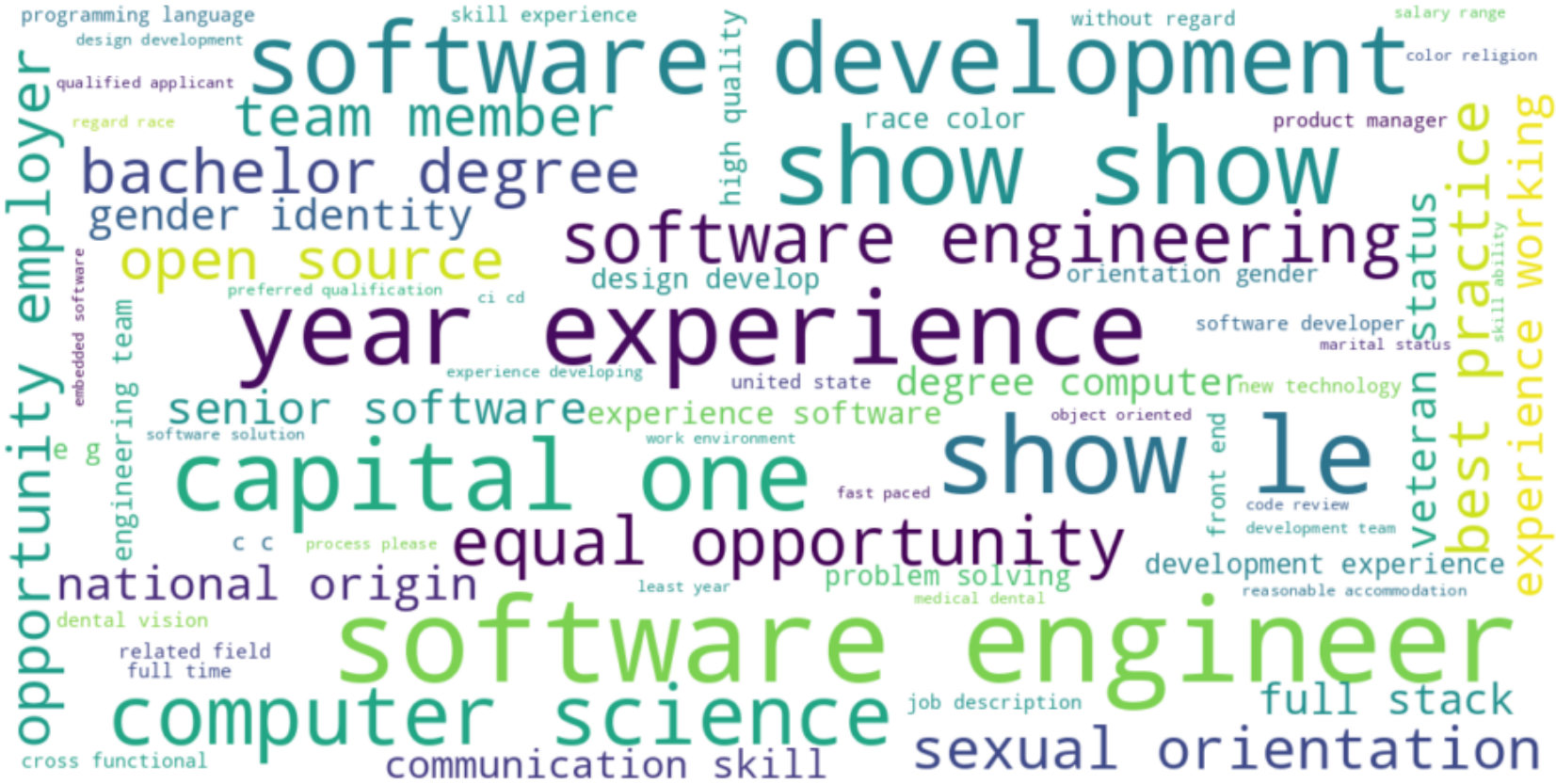


Figura: Nube de palabras del campo job_summary.

Exploración de datos campo *job_skills*:

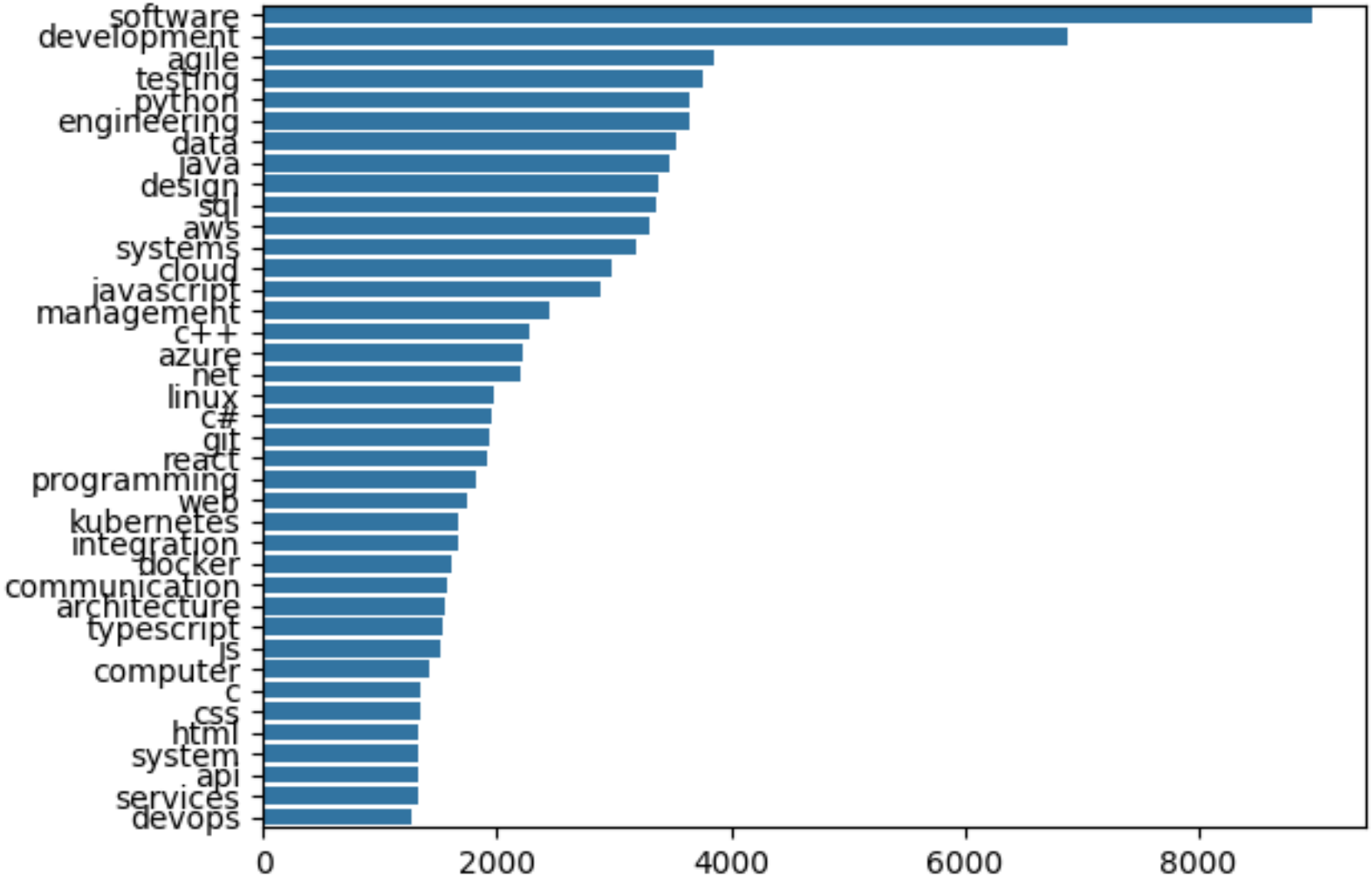


Figura: Palabras mas frecuentes en el campo job_skills.

Exploración de datos campo *job_skills*:

Figura: Nube de palabras del campo job_skills.

Mostrar trabajos que requieren Teamwork como habilidad blanda y utilizan los lenguajes de programación Python y Java.

```
+----------------------+----------------------+----------------------+
|              job_name|programming_languages|           soft_skills|
+----------------------+----------------------+----------------------+
|BackEnd Software  ... | Java-Go-SQL-Scala ... |              Teamwork|
|Full Stack Softwa ... | TypeScript-Java-G ... |              Teamwork|
|Full Stack Softwa ... | Java-JavaScript-P ... |              Teamwork|
|Full Stack Softwa ... | TypeScript-Java-C ... |              Teamwork|
|   Software Engineer| Java-TypeScript-P ... |Problem Solving-O ... |
|Quality Software  ... |      Java-SQL-Python|              Teamwork|
|BackEnd Software  ... | Java-Go-SQL-HTML- ... |              Teamwork|
|BackEnd Software  ... | Java-Go-SQL-Scala ... |Organization-Team ... |
|BackEnd Software  ... | Java-Go-SQL-Scala ... | Teamwork-Creativity|
|Full Stack Softwa ... | TypeScript-Java-C ... |Problem Solving-T ... |
|Full Stack Softwa ... | TypeScript-Java-G ... | Teamwork-Creativity|
|Full Stack Softwa ... | TypeScript-Java-C ... |              Teamwork|
|Full Stack Softwa ... | TypeScript-Java-C ... |Organization-Team ... |
|BackEnd Software  ... | Java-Go-SQL-Scala ... |              Teamwork|
|BackEnd Software  ... | Java-Go-SQL-Scala ... | Teamwork-Creativity|
|Full Stack Softwa ... | TypeScript-Java-C ... |              Teamwork|
|   Software Engineer| Java-Go-SQL-Scala ... |              Teamwork|
|  Software Developer| Java-CSS-HTML-Jav ... |Problem Solving-L ... |
|Full Stack Softwa ... |    JavaScript-Python|Communication-Tea ... |
|   Software Engineer|           Java-Python|Communication-Tea ... |
+----------------------+----------------------+----------------------+
only showing top 20 rows

Execution time: 1681 milliseconds
```

Figura: Respuesta de la query 1 con Spark y Scala.

```
+--------------------+--------------------+--------------------+
|            job_name|programming_languages|         soft_skills|
+--------------------+--------------------+--------------------+
|BackEnd Software  ...| Java-Go-SQL-Scala ...|            Teamwork|
|Full Stack Softwa ...| TypeScript-Java-G ...|            Teamwork|
|Full Stack Softwa ...| Java-JavaScript-P ...|            Teamwork|
|Full Stack Softwa ...| TypeScript-Java-C ...|            Teamwork|
|   Software Engineer| Java-TypeScript-P ...|Problem Solving-O ...|
|Quality Software  ...|      Java-SQL-Python|            Teamwork|
|BackEnd Software  ...| Java-Go-SQL-HTML- ...|            Teamwork|
|BackEnd Software  ...| Java-Go-SQL-Scala ...|Organization-Team ...|
|BackEnd Software  ...| Java-Go-SQL-Scala ...| Teamwork-Creativity|
|Full Stack Softwa ...| TypeScript-Java-C ...|Problem Solving-T ...|
|Full Stack Softwa ...| TypeScript-Java-G ...| Teamwork-Creativity|
|Full Stack Softwa ...| TypeScript-Java-C ...|            Teamwork|
|Full Stack Softwa ...| TypeScript-Java-C ...|Organization-Team ...|
|BackEnd Software  ...| Java-Go-SQL-Scala ...|            Teamwork|
|BackEnd Software  ...| Java-Go-SQL-Scala ...| Teamwork-Creativity|
|Full Stack Softwa ...| TypeScript-Java-C ...|            Teamwork|
|   Software Engineer| Java-Go-SQL-Scala ...|            Teamwork|
|  Software Developer| Java-CSS-HTML-Jav ...|Problem Solving-L ...|
|Full Stack Softwa ...|    JavaScript-Python|Communication-Tea ...|
|   Software Engineer|          Java-Python|Communication-Tea ...|
+--------------------+--------------------+--------------------+
only showing top 20 rows

Execution time: 1661 milliseconds
```

Figura: Respuesta de la query 1 con Spark SQL con UDF.

En cuanto a tiempos de ejecución en este caso con UDF es ligeramente menor.

Mostrar los trabajos que solicitan en Cincinnati con algún tipo de Nivel Académico.

```
+--------------------------+-----------+
|                  job_name|job_count|
+--------------------------+-----------+
|    Software Engineer|          26|
|BackEnd Software  ... |           3|
|Quality Software  ... |           2|
|Full Stack Softwa ... |           1|
|Full Stack Softwa ... |           1|
+--------------------------+-----------+

Execution time: 3634 milliseconds
```

Figura: Respuesta de la query 2 con Spark y Scala.

```
+--------------------------+-----------+
|                  job_name|job_count|
+--------------------------+-----------+
|    Software Engineer|         26|
|BackEnd Software  ... |          3|
|Quality Software  ... |          2|
|Full Stack Softwa ... |          1|
|Full Stack Softwa ... |          1|
+--------------------------+-----------+

Execution time: 3358 milliseconds
```

Figura: Respuesta de la query 2 con Spark SQL con UDF.

En cuanto a tiempos de ejecución en este caso con UDF es ligeramente menor.

Mostrar los trabajos que requieren Web Development como conocimiento y una certificación con la palabra clave "Certified".

```
+----------------------+----------------------+----------------------+
|              job_name|             knowledge|      certifications|
+----------------------+----------------------+----------------------+
|   Software Engineer|Security-Version  ...|Certified Kuberne ...|
|  Software Developer|Testing-Databases ...|Salesforce Certif ...|
|  Software Developer|Testing-Security- ...|Certified Informa ...|
|   Software Engineer|Testing-DevOps-Ve ...|Certified ScrumMa ...|
|   Software Engineer|Testing-DevOps-Ve ...|Certified Informa ...|
|  Software Architect|DevOps-Security-D ...|Certified Informa ...|
|Full Stack Softwa ...|Testing-DevOps-Ve ...|Certified Informa ...|
|   Software Engineer|Testing-Security- ...|Certified Ethical ...|
|  Software Developer|Testing-Security- ...|Certified Informa ...|
|   Software Engineer|DevOps-Databases- ...|Salesforce Certif ...|
|  Software Developer|Agile-Web Develop ...|Certified ScrumMa ...|
|  Software Developer|Agile-Web Develop ...|Certified ScrumMa ...|
|  Software Developer|DevOps-Security-P ...|Certified Informa ...|
|  Software Developer|Testing-DevOps-Da ...|Certified ScrumMa ...|
|  Software Developer|Agile-DevOps-Web  ...|Certified Informa ...|
|   Software Engineer|Security-Mobile D ...|Certified ScrumMa ...|
|  Software Developer|Testing-DevOps-Se ...|Cisco Certified N ...|
|   Software Engineer|Testing-Web Devel ...|Certified ScrumMa ...|
+----------------------+----------------------+----------------------+

Execution time: 1833 milliseconds
```

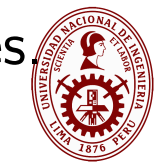
Figura: Respuesta de la query 3 con Spark y Scala.

```
+----------------------+----------------------+----------------------+
|              job_name|             knowledge|        certifications|
+----------------------+----------------------+----------------------+
|    Software Engineer|Security-Version  ... |Certified Kuberne ... |
|   Software Developer|Testing-Databases ... |Salesforce Certif ... |
|   Software Developer|Testing-Security- ... |Certified Informa ... |
|    Software Engineer|Testing-DevOps-Ve ... |Certified ScrumMa ... |
|    Software Engineer|Testing-DevOps-Ve ... |Certified Informa ... |
|   Software Architect|DevOps-Security-D ... |Certified Informa ... |
|Full Stack Softwa ... |Testing-DevOps-Ve ... |Certified Informa ... |
|    Software Engineer|Testing-Security- ... |Certified Ethical ... |
|   Software Developer|Testing-Security- ... |Certified Informa ... |
|    Software Engineer|DevOps-Databases- ... |Salesforce Certif ... |
|   Software Developer|Agile-Web Develop ... |Certified ScrumMa ... |
|   Software Developer|Agile-Web Develop ... |Certified ScrumMa ... |
|   Software Developer|DevOps-Security-P ... |Certified Informa ... |
|   Software Developer|Testing-DevOps-Da ... |Certified ScrumMa ... |
|   Software Developer|Agile-DevOps-Web  ... |Certified Informa ... |
|    Software Engineer|Security-Mobile D ... |Certified ScrumMa ... |
|   Software Developer|Testing-DevOps-Se ... |Cisco Certified N ... |
|    Software Engineer|Testing-Web Devel ... |Certified ScrumMa ... |
+----------------------+----------------------+----------------------+

Execution time: 1833 milliseconds
```

Figura: Respuesta de la query 3 con Spark SQL con UDF.

En cuanto a tiempos de ejecución en este caso vemos que son similares.



Mostrar el promedio de habilidades blandas requerido por nivel de trabajo.

```
+-------------+--------------------------+
| job_level|          avg_skills|
+-------------+--------------------------+
| Associate|3.534536891679749|
|Mid senior|3.477696774990733|
+-------------+--------------------------+

Execution time: 3649 milliseconds
```

Figura: Respuesta de la query 4 con Spark y Scala.

```
+-------------+-------------------------+
| job_level|              avg_skills|
+-------------+-------------------------+
| Associate|3.5345368916679749|
|Mid senior|3.477696774990733|
+-------------+-------------------------+

Execution time: 3089 milliseconds
```

Figura: Respuesta de la query 4 con Spark SQL con UDF.

En cuanto a tiempos de ejecución en este caso con UDF es ligeramente menor.

Mostrar los trabajos que requieren conocimientos en Agile y un grado académico de Bachelor's.

```
+--------------------+--------------------+----------------+--------------+
|            job_name|        job_location|     search_city|search_country|
+--------------------+--------------------+----------------+--------------+
|   Software Engineer|Bristol- England- ... |         Cardiff|United Kingdom|
|   Software Engineer|      Greenville- SC|  South Carolina| United States|
|   Software Engineer|        St Louis- MO|        Ferguson| United States|
|  Software Developer|        St Louis- MO|        Ferguson| United States|
|  Software Developer|        St Louis- MO|        Ferguson| United States|
|   Software Engineer|         Bristol- CT|      Litchfield| United States|
|  Software Developer|Montreal- Quebec- ... |  Côte-Saint-Luc|        Canada|
|   Software Engineer|Montreal- Quebec- ... |  Côte-Saint-Luc|        Canada|
|   Software Engineer|      Pittsburgh- PA|        Brighton| United States|
|   Software Engineer|         Norfolk- VA|         Norfolk| United States|
|Security Software ... |         Norfolk- VA|         Norfolk| United States|
|  Software Developer|Markham- Ontario- ... |          Oshawa|        Canada|
|   Software Engineer|     Westminster- CO|        Longmont| United States|
|   Software Engineer|        Columbus- OH|        Defiance| United States|
|   Software Engineer|        O'Fallon- MO|        Defiance| United States|
|   Software Engineer|California- Unite ... |      California| United States|
|  Software Developer|        St Louis- MO|East Saint Louis| United States|
|  Software Developer|        St Louis- MO|East Saint Louis| United States|
|   Software Engineer|       Cleveland- OH|            Ohio| United States|
|   Software Engineer|Chester- England- ... |       Liverpool|United Kingdom|
+--------------------+--------------------+----------------+--------------+
only showing top 20 rows

Execution time: 2073 milliseconds
```

Figura: Respuesta de la query 5 con Spark y Scala.

```
+--------------------+--------------------+----------------+--------------+
|            job_name|        job_location|     search_city|search_country|
+--------------------+--------------------+----------------+--------------+
|   Software Engineer|Bristol- England- ... |         Cardiff|United Kingdom|
|   Software Engineer|      Greenville- SC|  South Carolina| United States|
|   Software Engineer|        St Louis- MO|        Ferguson| United States|
|  Software Developer|        St Louis- MO|        Ferguson| United States|
|  Software Developer|        St Louis- MO|        Ferguson| United States|
|   Software Engineer|         Bristol- CT|      Litchfield| United States|
|  Software Developer|Montreal- Quebec- ... |  Côte-Saint-Luc|        Canada|
|   Software Engineer|Montreal- Quebec- ... |  Côte-Saint-Luc|        Canada|
|   Software Engineer|      Pittsburgh- PA|        Brighton| United States|
|   Software Engineer|         Norfolk- VA|         Norfolk| United States|
|Security Software ... |         Norfolk- VA|         Norfolk| United States|
|  Software Developer|Markham- Ontario- ... |          Oshawa|        Canada|
|   Software Engineer|     Westminster- CO|        Longmont| United States|
|   Software Engineer|        Columbus- OH|        Defiance| United States|
|   Software Engineer|        O'Fallon- MO|        Defiance| United States|
|   Software Engineer|California- Unite ... |      California| United States|
|  Software Developer|        St Louis- MO|East Saint Louis| United States|
|  Software Developer|        St Louis- MO|East Saint Louis| United States|
|   Software Engineer|       Cleveland- OH|            Ohio| United States|
|   Software Engineer|Chester- England- ... |       Liverpool|United Kingdom|
+--------------------+--------------------+----------------+--------------+
only showing top 20 rows

Execution time: 1699 milliseconds
```

Figura: Respuesta de la query 5 con Spark SQL con UDF.

En cuanto a tiempos de ejecución en este caso con UDF es menor.

Mostrar el lenguaje de programación más solicitado para trabajos presenciales.

```
+--------------------+-----+
|programming_language|count|
+--------------------+-----+
|              Python| 1871|
+--------------------+-----+

Execution time: 4040 milliseconds
```

Figura: Respuesta de la query 6 con Spark y Scala.

```
+--------------------+-----+
|programming_language|count|
+--------------------+-----+
|              Python| 1871|
+--------------------+-----+

Execution time: 3469 milliseconds
```

Figura: Respuesta de la query 6 con Spark SQL con UDF.

En cuanto a tiempos de ejecución en este caso con UDF es ligeramente menor.

Mostrar el numero de trabajos donde solicitan cada área de conocimiento.

```
+--------------------+-----+
|      knowledge_area|count|
+--------------------+-----+
|     Web Development| 5163|
|             Testing| 4701|
|               Agile| 4406|
|           Databases| 4379|
|     Cloud Computing| 4267|
|            Security| 3038|
|        Data Science| 2394|
|     Version Control| 2276|
|Software Architec ... | 1928|
|              DevOps| 1613|
|               CI CD| 1083|
|          Networking|  980|
|  Mobile Development|  925|
|  Project Management|  697|
|                 TDD|  415|
|            No found|  314|
|                 BDD|  166|
|            Mocking|   22|
+--------------------+-----+

Execution time: 3681 milliseconds
```

Figura: Respuesta de la query 7 con Spark y Scala.

```
+----------------------+-----+
|      knowledge_area|count|
+----------------------+-----+
|     Web Development| 5163|
|             Testing| 4701|
|               Agile| 4406|
|           Databases| 4379|
|     Cloud Computing| 4267|
|            Security| 3038|
|        Data Science| 2394|
|     Version Control| 2276|
|Software Architec ... | 1928|
|              DevOps| 1613|
|               CI CD| 1083|
|          Networking|  980|
|  Mobile Development|  925|
|  Project Management|  697|
|                 TDD|  415|
|            No found|  314|
|                 BDD|  166|
|            Mocking|   22|
+----------------------+-----+

Execution time: 3256 milliseconds
```

Figura: Respuesta de la query 7 con Spark SQL con UDF.

En cuanto a tiempos de ejecución en este caso con UDF es ligeramente menor.

```
root
 ├── job_location: string (nullable = true)
 ├── search_city: string (nullable = true)
 ├── search_country: string (nullable = true)
 ├── job_level: string (nullable = true)
 ├── job_type: string (nullable = true)
 ├── job_name: string (nullable = true)
 ├── languages: string (nullable = true)
 ├── certifications: string (nullable = true)
 ├── soft_skills: string (nullable = true)
 ├── programming_languages: string (nullable = true)
 ├── technologies: string (nullable = true)
 ├── academic_degrees: string (nullable = true)
 ├── knowledge: string (nullable = true)
 ├── n_languages: double (nullable = false)
 ├── n_certifications: double (nullable = false)
 ├── n_soft_skills: double (nullable = false)
 ├── n_programming_languages: double (nullable = false)
 ├── n_technologies: double (nullable = false)
 ├── n_knowledge: double (nullable = false)
 ├── n_academic_degrees: double (nullable = false)
 ├── num_job_level: double (nullable = false)
 ├── num_job_type: double (nullable = false)
 ├── num_search_country: double (nullable = false)
```

Figura: Esquema de datos de entrada.

```
Training Data count: 7468
Test Data count: 1899
```

Figura: Creando el train y test data 80 % y 20 %.

```
False positive rate by label:
label 0: 0.0
label 1: 0.0
label 2: 1.0
True positive rate by label:
label 0: 0.0
label 1: 0.0
label 2: 1.0
Precision by label:
label 0: 0.0
label 1: 0.0
label 2: 0.4606320299946438
Recall by label:
label 0: 0.0
label 1: 0.0
label 2: 1.0
F-measure by label:
label 0: 0.0
label 1: 0.0
label 2: 0.6307297396406307
Accuracy: 0.4606320299946438
False positive rate: 0.4606320299946438
True positive rate: 0.4606320299946438
F-measure: 0.2905343203486569
Precision: 0.212181867056986 44
Recall: 0.4606320299946438
```

Figura: Métricas antes del ajuste de parámetros.

```
False positive rate by label:
label 0: 0.005319148936170213
label 1: 0.021261516654854713
label 2: 0.964746772591857
True positive rate by label:
label 0: 0.009528130671506351
label 1: 0.025219298245614034
label 2: 0.9787790697674419
Precision by label:
label 0: 0.42857142857142855
label 1: 0.27710843373493976
label 2: 0.46422170136495244
Recall by label:
label 0: 0.009528130671506351
label 1: 0.025219298245614034
label 2: 0.9787790697674419
F-measure by label:
label 0: 0.018641810918774964
label 1: 0.046231155778894466
label 2: 0.629757785467128
Accuracy: 0.4598286020353508
False positive rate: 0.4511560402182328
True positive rate: 0.4598286020353508
F-measure: 0.3068798823530535
Precision: 0.40799944622380757
Recall: 0.4598286020353508
```

Figura: Métricas después del ajuste de parámetros.

```
Model Parameters Summary:
RegParam: 0.2
ElasticNetParam: 0.0
MaxIter: 20
```

Figura: Resumen de los mejores parámetros encontrados.

```
Metrics before hyperparameter tuning:
Test Error = 0.5660947912694522
Accuracy = 0.4339052087305478
Area Under ROC before tuning: 0.4339052087305478
Area Under PRC before tuning: 0.4339052087305478
```

Figura: Métricas antes del ajuste de parámetros.

```
Metrics after hyperparameter tuning:
Tuned Test Error = 0.5660947912694522
Tuned accuracy = 0.43390520873054783
Area Under ROC after tuning: 0.43390520873054783
Area Under PRC after tuning: 0.43390520873054783
```

Figura: Métricas después del ajuste de parámetros.

```
Model Parameters Summary:
Best maxDepth: 5
Best impurity: gini
```

Figura: Resumen de los mejores parámetros encontrados.

```
Metrics before hyperparameter tuning:
Test Error = 0.18483412322274884
Accuracy = 0.8151658767772512
F1 Score = 0.8151658767772512
Weighted Precision = 0.8151658767772512
Weighted Recall = 0.8151658767772512
```

Figura: Métricas antes del ajuste de parámetros.

```
Metrics after hyperparameter tuning:
Test Error = 0.177461822011585
Accuracy = 0.822538177988415
F1 Score = 0.822538177988415
Weighted Precision = 0.822538177988415
Weighted Recall = 0.822538177988415
```

Figura: Métricas después del ajuste de parámetros.

```
Best Model Parameters:
Number of Trees: 50
Max Depth: 20
Impurity: gini
```

Figura: Resumen de los mejores parámetros encontrados.

- Se logró explorar y visualizar las tendencias de requerimientos en las propuestas de trabajo para la carrera de Ciencia de la computación en la plataforma de LinkedIn mediante consultas con Scala y Spark.
- Se obtuvo que el tiempo de ejecución de las consultas realizadas con Spark SQL y UDF eran en su mayoría *menor* que el de las consultas con Scala y Spark.
- Se reconoció que el modelo de Random Forest de la librería MLlib obtuvo un valor de accuracy bueno al momento de clasificar el valor de *num_search_country* con respecto a los features numéricos generados en la limpieza y manipulación de los datos. Mientras, que en los modelos de Regresión Logística y Decision Tree se obtuvo un valor de accuracy menor al momento de clasificar los valores de *num_job_type* y *num_job_level*.

[1] Apache Spark, *Apache Spark™ - Unified Analytics Engine for Big Data*, Accedido el 12 de Junio del 2024, 2024. dirección: https://spark.apache.org/.

[2] Google Cloud, *¿Qué es Apache Spark?* Accedido el 12 de Junio del 2024, 2024. dirección: https://cloud.google.com/learn/what-is-apache-spark?hl=es.

[3] Apache Spark, *Apache Spark™ - Spark SQL*, Accedido el 12 de Junio del 2024, 2024. dirección: https://spark.apache.org/sql/.

[4] Apache Spark, *Apache Spark™ - MLlib: Machine Learning Guide*, Accedido el 12 de Junio del 2024, 2024. dirección: https://spark.apache.org/docs/latest/ml-guide.html.

[5] Apache Spark, *Apache Spark™ - RDD Programming Guide*, Accedido el 12 de Junio del 2024, 2024. dirección: https://spark.apache.org/docs/latest/rdd-programming-guide.html.

[6] ASANICZKA, *LinkedIn Software Engineering Jobs Dataset*, Accedido el 6 de junio del 2024, 2024. dirección: https://www.kaggle.com/datasets/asaniczka/software-engineer-job-postings-linkedin.