Resumo do Traballo de Fin de Grao

(Describa brevemente o traballo a desenvolver, xustificando o interese do mesmo, e indicando obxectivos, descrición técnica, proceso de desenvolvemento, e medios empregados. Engada tantas liñas como sexa necesario)

La extracción de información es un campo dentro de las tecnologías de procesamiento del lenguaje natural que se encarga de la identificación y estructuración de información relavante para un dominio a partir del análisis de texto en lenguaje natural. Fundamentalmente se trata de la identificación y categorización de las entidades (nombres de personas, de lugares, organizaciones, etc) mencionadas en el texto analizado y la identificación de relaciones entre las mismas con la finalidad de extraer los hechos y eventos relatados en el texto.

Por otra parte, en el ámbito de la seguridad de la información existen bases de datos y repositorios de recursos (como NVDB https://nvd.nist.gov/, Exploit-DB https://www.exploit-db.com/ o CVE https://cve.mitre.org/) que recopilan y organizan vulnerabilidades identificadas en productos software, donde, junto con los metadatos relativos a las mismas (nivel de peligrosidad, fechas de descubrimiento, versiones afectas, etc) se suelen incluir descripciones textuales con información adicional. La mayor parte de estos repositorios exponen sus colecciones de vulnerabilidades en formatos procesables como XML o JSON o incluso las publican mediante APIs REST.

En este contexto, el objetivo de este TFG es desarrollar una plataforma genérica de extracción de información para procesar estas fuentes de datos empleando técnicas de procesamiento del lenguaje natural para extraer hechos relevantes de estas colecciones de vulnerabilidades software (nombres de fabricantes, nombres de aplicaciones, tipos de vulnerabilidades, versiones, relaciones entre estos elementos, etc). Como parte del desarrollo de este trabajo se definirá el esquema de almacenamiento de la información extraída y se desarrollará un conjunto de extractores de entidades y relaciones adaptados a las fuentes finalmente consideradas.

En el momento de redactar esta propuesta no están totalmente definidas las tecnologías a emplear. En lo que respecta al procesamiento de los textos, hay disponibles herramientas como Stanford CoreNLP (https://stanfordnlp.github.io/CoreNLP/) o similares, que junto con el desarrollo de patrones y/o expresiones regulares propios servirán de base para los componentes encargados del procesamiento de los textos. Por lo que respecto al almacenamiento y estructuración de los hechos y eventos identificados una alternativa a considerar podría ser una base de datos orientada a objetos como Neo4J (https://neo4j.com/), aunque también sería posible el uso de un SGBD convencional.

Respecto al proceso de desarrollo, se pretende seguir el Proceso de Desarrollo Unificado (RUP), dado que el planteamiento iterativo e incremental sobre el que este se asienta encaja a la perfección con la situación de partida de este desarrollo. Inicialmente, están relativamente claras y definidas las funcionalidades básicas que se pretende implementar, pero existe incertidumbre en cuanto a las herramientas que finalmente se van a utilizar y a la disponibilidad de recursos lingüísticos (diccionarios, patrones, etc) susceptibles de ser aprovechados.