# Extracción de información a partir de descripciones de cursos

Se comunica que la propuesta de trabajo 7351 - 'Extracción de información a partir de descripciones de cursos', del plan 10II ha sido asignada a RUTH VERONICA OCAMPO PRADO (rv.ocampo) con mail rv.ocampo@alumnos.upm.es

**Planes para los que la propuesta es válida:**

**10MI Matemáticas e Informática 10ID Administración y Dirección de Empresa 10II Ingeniería Informática**
**¿Puede este trabajo también formar parte de un doble trabajo conjunto para alumnos de doble grado de Ingeniería Informática y Administración y Dirección de Empresas?**
**No**
**Nombre del trabajo*:**
Extracción de información a partir de descripciones de cursos

**Resumen general del trabajo*:**
En el contexto del proyecto europeo AI4LABOUR, necesitamos extraer las habilidades y tareas que cubren distintos cursos de plataformas MOOC y guías docentes universitarias. Partiendo de distintas bases de datos de habilidades y tareas, el alumno deberá identificar identificar las habilidades y tareas similares (por ejemplo, en una base de datos una habilidad puede ser “Programación en C” y en otra “Lenguaje C”, pero realmente son la misma) y localizar las apariciones de las mismas en descripciones de texto usando técnicas de similaridad semántica, sistemas de reglas u otras técnicas de procesamiento del lenguaje.

Se proporcionarán al alumno scripts previos de similaridad semántica como punto de partida (en Python).

**Lista de objetivos concretos del trabajo*:**
- Diseño e implementación de algoritmo de identificación de habilidades y tareas similares de distintas bases de datos.
- Diseño e implementación de extracción de dichas tareas y habilidades a partir de descripciones textuales de cursos.
- Evaluación de los algoritmos.

**Desglose de la dedicación total del trabajo en horas (297 horas en los Grados)*:**
- Análisis de scripts previos, datos disponibles y estado del arte (50h)
- Diseño de algoritmo de identificación de habilidades/tareas (entidades) similares (10h)
- Implementación del algoritmo (40h)
- Evaluación del algoritmo (20h)
- Diseño del algoritmo de extracción de entidades a partir de descripciones de cursos (10h)
- Implementación del algoritmo (87h)
- Evaluación del algoritmo (20h)
- Documentación y escritura del TFG (40h)

- Tutorías/reuniones progreso (20h)

**Lista de conocimientos previos recomendados para realizar el trabajo*:**
- Programación (preferiblemente Python)
- Procesamiento del lenguaje natural

Cualquiera de estos conocimientos puede sustituirse por voluntad de aprenderlo.

18/09/2023

# Primera reunión (online)

- Intro proyecto
- Repasar tareas
- Mandaré algunos resultados de MOOCs (json)
- Presentar ESCO y O*NET
- Breve estado del arte de normalización de skills

TODO:
- Ver el estado del arte
- Ideas de mejora/probar alguno (SkillGPT)
- Reunión semana que viene/dentro de dos semanas

María: mandar scripts, bases de datos, paper ai4labour

# SotA skill extraction/normalization/ESCO link

Por orden de relevancia:

- SkillGPT: a RESTful API service for skill extraction and standardization using a Large Language Model (2023) Interesante, tiene un github pero de momento ha dado problemas instalarlo… está ITCL con ello https://github.com/aida-ugent/SkillGPT
- Job Posting-Enriched Knowledge Graph for Skills-based Matching (2021): usan Jaccard con threshold=0,66 para similaridad de skills. Interesante para construir KG, usan link prediction, ejemplos de usos…
- A Survey on Skill Identification From Online Job Ads | IEEE Journals & Magazine (2021): interesante como survey de papers, ya han hecho varias veces lo de recomendar MOOCs.
- Large Language Models as Batteries-Included Zero-Shot ESCO Skills Matchers (2023)
    - Usan prompts en GPT3.5/4 para crear dataset a partir de ESCO
    - Embeddings + clasificador binario por skill + mock Python Programming
    - hallucia skills a veces, en método python no
    - negative sampling

- ESCOXLM-R: Multilingual Taxonomy-driven Pre-training for the Job Market Domain (2023) Tiene varios datasets interesantes
- Kompetencer: Fine-grained Skill Classification in Danish Job Postings via Distant Supervision and Transfer Learning (2022) [SUBRAYADO EN CARPETA] Para el danés (https://github.com/jjzha/kompetencer). Usan Levenshtein para enlazar con ESCO (ponen el algoritmo). Conclusiones interesantes sobre la extracción:
    - Ejemplos en idiomas distintos ayudan
    - Dicen que domain adaptative es mejor
    - Faltan skills técnicas (eg TensorFlow)
- Extreme Multi-Label Skill Extraction Training using Large Language Models (2023) : generan un dataset sintético de 183K con GPT (público). luego hacen contrastive learning (https://lilianweng.github.io/posts/2021-05-31-contrastive/). + augmentation
    - dicen que ESCo skills tienen una media de 7 sinónimos en inglés
    - prompt:
      System: Respond with sentences from hypothetical job ads that require a certain skill, as asked by the user.
      User: Number of sentences: 2
      Skill: Java
      Definition: The techniques and principles {...} in Java.
      Assistant: - experience with Java development, preferably web-based
      - looking for a Java programmer this summer
      User: Number of sentences: 2
      Skill: project management
      Definition: Understanding project management and {...} events.
      Assistant: - successful project managers are able to manage multiple tasks and deadlines simultaneously
      - being able to effectively manage projects can give you valuable experience and skills
      User: Number of sentences: 10
      Skill: skill
      Definition: skill description
- SKILL: A System for Skill Identification and Normalization (2015): usan wikipedia para desambiguar/normalizar, poco explícito.
- SkillNER: Mining and mapping soft skills from any text (2021): extraen soft skills (transversal), no existe ya la demo, no útil.