## Descrizione del progetto di ricerca

Il progetto di ricerca, in cui il lavoro oggetto della presente si inserisce, consiste nell'analisi dei DDT, degli attuali dipendenti o pervenuti da possibili candidati per le posizioni aperte, e nell'estrazione delle informazioni rilevanti in essi contenute.

L'analisi dei documenti (strutturati e non) e la relativa estrazione delle informazioni è uno dei task tradizionali nell'ambito del Natural Language Processing. Attualmente si stanno affermando approcci in grado di estrarre informazioni da documenti sfruttando il deep learning per l'analisi dei testi in essi contenuti. Tali approcci si avvalgono delle reti neurali (ANN) per consentire l'analisi di un largo insieme di documenti per addestrare un sistema all'estrazione automatica delle informazioni da essi. In particolare trovano ampia applicazione ai task di NLP le reti neurali di tipo word embedding. Per word embedding si intende una rappresentazione vettoriale che permette di memorizzarne sia le informazioni semantiche che sintattiche partendo da un corpus non annotato e costruendo uno spazio vettoriale in cui i vettori delle parole sono più vicini se le parole occorrono negli stessi contesti linguistici, cioè se sono riconosciute come semanticamente più simili (secondo l'ipotesi della semantica distribuzionale).

A valle dell'estrazione delle informazioni si rende necessario per la realizzazione del task di importazione dei DDT ricevuti, l'importazione delle informazioni in essi contenute in una struttura dati predefinita e integrata nel sistema Myquipu.

## Obiettivo della tesi

Analisi dello stato dell'arte degli approcci basati sull'impiego delle reti neurali (ANN) nello sviluppo di sistemi per il Natural Language Processing e individuazione delle tipologie di ANN maggiormente performanti in tale ambito, con particolare attenzione agli approcci basati su word embedding.

Inoltre la tesi verterà sullo sviluppo di uno o più dei seguenti componenti:

1. Implementazione di un componente basato su rete neurale in grado di suddividere il documento in aree informative (Layout analysis i.e., Clusterizzazione delle aree informative identificate in gruppi omogenei tramite CNN).
2. Ottimizzazione del testo estratto mediante l'utilizzo di sistemi OCR attraverso l'applicazione di reti neurali (word2vec).
3. Ideazione e implementazione di una metodologia di ottimizzazione di reti neurali esistenti basata sulla tecnica dell'Hyperparameter Optimization.
4. Implementazione del componente per l'inserimento dei dati nella base di dati.

## Tecnologie da utilizzare

- Python 3.7 o superiori in ambiente Anaconda.
- Tensor Flow e relative librerie (ad esempio Keras, gensim).