Hitachi Vantara Estágio implementação

1º Passo:

Escolher uma biblioteca de NLP para texto não estruturado.

Escolhi a OpenNLP dado que é das bibliotecas mais estruturadas que suportam processos como Tokenization, Part speech tagging, Language detection, chunking, etc.

Vamos começar por assumir a lingua como o Inglês e portanto a Language detection acaba por ser inútil.

O objetivo o OpenNLP vai ser para numa fase inicial detetar nomes e posteriormente verbos e adjetivos. Se houvesse mais tempo, ou se tivermos tempo, gostariamos também de tentar implementar/explorar a parte das emoções relacionadas com NLP.

2º Passo:

Nesta parte queremos associar os nomes retirados do texto não estruturado a uma rede semântica. Para esta rede semântica vamos usar a API do Wordnet. Existem algumas Tags que saem do OpenNLP que não se ligam diretamente com as classes do wordnet, as preposições e determinantes, mas como vamo focar-nos maioritariamente nos nomes não será um grande problema.

No fim deste passo iremos ter uma clara classificação de cada Tag e a sua respetiva rede semântica.

3º Passo:

Por fim queremos ter tudo integrado com o Pentaho PDI da Hitachi Vantara. Em que o Input é text não estruturado e com um workflow tipo fluxograma sairá uma rede semântica com a respetiva classificação e relações.

Idealmente e se houver tempo para tal queriamos finalizar a integração com o catálogo da Hitachi Vantara e trabalhar o output para ser por exemplo um CSV com as especificações do catálogo relacionadas com o input. Exemplo “Quantas cidades Europeias têm mais de 10 Milhões de Habitantes”.