CAPSTONE 2 PROJECT PROPOSAL

* **What is the problem you want to solve?**

The European Parliament conducts business in many languages due to the huge differences in languages and cultures that make up its members. The transcriptions of its meetings need to be translated into a multitude of languages. One of these is French.

* **Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn’t have done otherwise?**

Translations Services Inc (TSI) has been contracted by the European Parliament to translate conference transcriptions from English to French on a continual basis. Due to the overwhelming number of documents to translate, TSI hired me as their Data Scientist to build an NLP model to automatically translate the documents from English to French

* **What data are you using? How will you acquire the data?**

The data I will use to build the NLP model is data from European Parliament: <http://www.statmt.org/europarl/>, more precisely [parallel corpus French-English](http://www.statmt.org/europarl/v7/fr-en.tgz), 194 MB, 04/1996-11/2011

* **Briefly outline how you’ll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.**

I will build a Neural Network NLP model using the Keras package with Seq2Seq.

1. I will first do some data Preprocessing.
   1. I will Tokenize and Pad both the input and output sentences (documents)
   2. I will then use GloVe word embeddings on the input (English) and custom word embeddings on the output (French). Then I will create a dictionary where words are the keys and the corresponding vectors are values.
2. Next I will build the model
   1. I will be encoding the input and then decoding the output using lstm
   2. Next I will train the model.
3. Then I will modify the model for predictions
4. I will then make predictions
5. I will finally test the model by randomly choosing a sentence from the input

* **What are your deliverables? Typically, this includes code, a paper, or a slide deck.**

The deliverables will be the code, and a paper which will be placed in <https://github.com/Omolinari10/Capstone-Project-2-NLP>