


PREDICT THE DIALECT

By Ahmed Waheed

DATA COLLECTION

- ▶ Fetch the data by HTTP request.
- ▶ Clean it by regex.
 - remove English as it is useless in this task.
- ▶ Write it in a new CSV file.

DATA ANALYSIS & PREPROCESSING

- ▶ This corpus have 543645 distinct token.
 - ▶ split words by spaces.
 - ▶ sentence to vector.
 - ▶ padding by zeros.
- 
- A series of several parallel white diagonal lines of varying lengths, located in the bottom right corner of the slide.

SEMANTIC

- * **Word Embedding** : represent word by a vector.

Better to use transfer learning Glove for example,

But due to task case sensitive, Features Matrix isn't important,


As it grow up the semantic but here syntax is the matter.

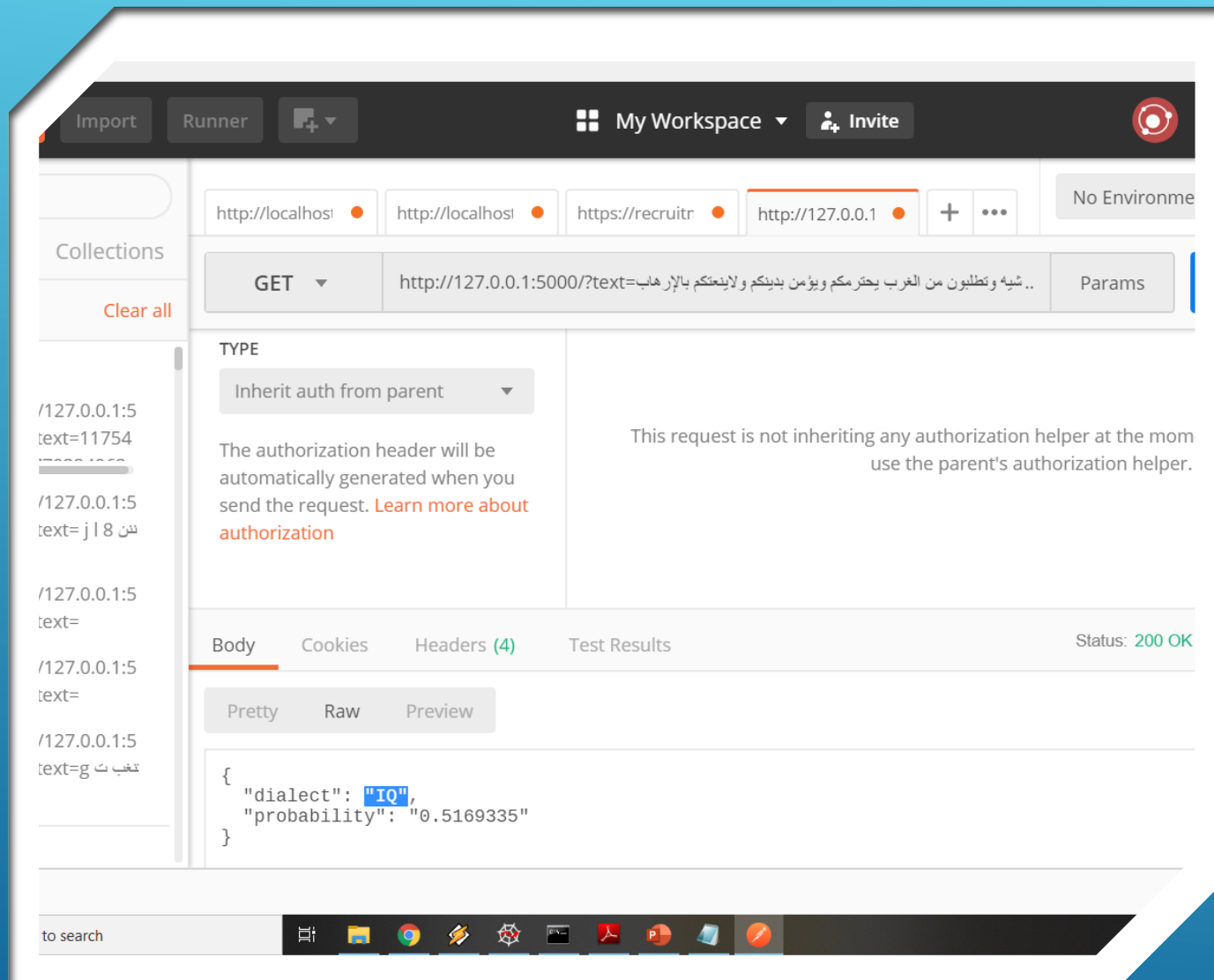
- * For example, by embedding "السلم" is similar to "الدرج"

but we want difference between them.

- * From the other hand we won't use lemma nor stemma
nor any technique that remove prefix & suffix.

THE MODEL

- ▶ LSTM is the best tool dealing with sequences.
 - ▶ Bidirectional NN to extract features from right to left & left to right.
 - ▶ SoftMax activation function to out the probability for each class.
- 
- Several white diagonal lines of varying lengths and thicknesses are positioned in the bottom right corner of the slide, creating a modern, abstract design element.



APP DEPLOYMENT

Using Flask, We made a REST web service that can easily accessed by HTTP request & use its response in any platform.

Very thanks

Ahmed Waheed

Several white lines of varying lengths and thicknesses are positioned in the bottom right corner of the slide, creating a modern, abstract graphic element.