Natural Language Processing: looking for Value in your Unstructured Data

Keywords: #ai #artificialintelligence #ml #machinelearning #python #nlp #csv #gpu #gcp #jupyter #docker #tensorflow

1. Abstract

A frequent business problem is discussed, whereby companies find themselves facing a large base of Unstructured Data, which proves cumbersome to process. A specific case is brought up in which the data arrives continuously from the users and is processed on the fly. A comprehensive approach to processing this Unstructured Data is discussed along with technical considerations.

2. Problem statement

We seek an approach to handling Unstructured Data on the fly with Machine Learning techniques, particularly with Natural Language Processing.

3. Background

Let's imagine you process requests from your users in an on-the-fly fashion and you don't want to force your clients to use a specific format for their requests. Instead they should be able to submit a free-structured CSV (Comma Separated Values) document or even an Excel spreadsheet of their choosing. Your goal is to seamlessly adjust and rework your users' request document so that it matches your back-end's contracted format.

The system's users request data by submitting a CSV file with a list of individuals and their accompanying data. The application then returns some additional data associated with the individuals. But first it needs to identify them.

4. Solution

Let's think about what the back-end expects as input. Suppose it's the following.

a. First name.
b. Middle name.
c. Last name.
d. Address:
    a. Street name.
    b. Street number.
    c. Postcode.

d. State.

e. Phone number.

How do you enable your users to specify these data points in a free format? As you will see, it is possible, but you'll have to go through a number of challenges, such as the following.

i. The user might specify the name in one cell, not splitting it into the first name, the middle name or the last name.
ii. The user might not have or skip the middle name.
iii. The user might specify the address as one line or specify the components of the address in any given order.
iv. The user might specify the headers to their data incorrectly or might not specify them at all.
v. The postcode might mix up with the phone number as their format is similar.
vi. You need to process all of the data on the fly, so you can't go back and correct your model's action subsequently.

These are just a couple of the problems we can run into. The list goes on.

How do you approach this Unstructured Data?

You can do this with various text processing tools, particularly with a method powered by ML (Machine Learning) and specifically by NLP (Natural Language Processing). Broadly speaking, this method enables one to identify the columns, even if the data is incomplete, mislabeled or occasionally incorrect. The great thing about ML is that it learns how to achieve the task on its own, so in a way all you need to do is supervise its progress.

That's exactly what we did: we utilized NLP to do the heavy-lifting for us and arrived at a solution to the problem at hand.

Our NLP enriched approach enabled us to:

I. Split the columns into granular data points if they contain combined data.
II. Reorder the columns.
III. Deal without some of the data points if there was no other way to go.
IV. Differentiate between postcode and phone number based on parts of the address, for example.
V. Parse the address line and split into data points.
VI. Guess the type of the column if no header is provided.

The tech-stack of our solutions included Python, Tensorflow, GPU computing, Google Cloud Platform, Jupyter Notebooks and Docker. We carried out diligent Model Tuning and managed to achieve close to 99% accuracy for the Machine Learning model in question!

5. Conclusion

Unstructured Data throws a lot of seemingly impossible to handle challenges at us in our day-to-day dealing with data. With adequate focus, technology expertise and ML and NLP excellence a lot of these problems can be solved in an automated manner. This paper described some of the problems of this sort along with a description of solutions to these problems.

We delivered this Machine Learning NLP (Natural Language Processing) solution to a client from the data industry based in the US. We carried out diligent Model Tuning and managed to achieve 99% accuracy for the Machine Learning model in question.

Do you process lots of textual or PDF data on a daily basis? Are you looking for a way to automate this process by Classifying, Content-extracting, Understanding, and Question-answering from this textual input? The NLP (Natural Language Processing) solution described above involved an ensemble model that extracted dozens of data points from Unstructured and Semi-structured Data without human intervention. We would be happy to harness this know-how to help you automate your text-based processes with NLP, a Machine Learning technique.

6. References

1. https://docs.python.org/3/library/csv.html
2. https://en.wikipedia.org/wiki/Machine_learning
3. https://en.wikipedia.org/wiki/Natural_language_processing
4. https://www.python.org/
5. https://www.intel.com/content/www/us/en/products/docs/processors/what-is-a-gpu.html
6. https://cloud.google.com/
7. https://jupyter.org/
8. https://www.docker.com/
9. https://www.tensorflow.org/

Would you like to hear more? Please get in touch with us via https://www.jagansolutions.com/contact-us.