Sentiment Analysis with Deep Learning

Gerrit Gruben*

Free University Berlin
gerrit.gruben@gmail.com

Abstract

Thomson Reuters successfully transitioned into the information age by the deployment of modern methods of data processing and analytics on huge datasets. Classically, Thomson Reuters is a mass media company, but due to the recent efforts, it can nowadays be called an information firm. Thomson Reuters is in the unique position to combine its formidable resources, flexibility to adapt to the market and technical strategic partnerships to make use of the sizeable in- and out-flow of data and information it has control of. This work focuses on deploying deep learning methods to enable improved sentiment analysis of texts found in Reuter's StreetEvent data set. Deep learning methods have made an astonishing revival by showing successes in several domains such as image classification, pattern detection, speech recognition and natural language processing. They adapt especially well to the variety of data found as they can bring structure into unstructured data in an unsupervised setting.

I. Introduction

1. Summary

This case study targets to lay groundwork for further leverage of big data by Thomson Reuters by structuring text data and extract features out of by using neuronal networks. These features then can be used, together with already discovered features from other sources, to run classification algorithms and/or predictive analysis.

The blog post found in [Thomson Reuters, Sept. 2014] indicates the value of the successful deployment analytical methods for Thomson Reuters to stay ahead of its competition. Obviously, Thomson Reuters already deploys resources to deal with the problem of making use of the data. It is indicated in [Techcrunch, Feb. 2014] that Thomson Reuter's already deploys methods of sentiment analysis to Twitter. It is left unclear on which methods are precisely deployed.

The value added by this report is supposed to be the following:

- Discovering the value of deep learning for sentiment analysis of text to deal with variety of data;
- A component that splits Street Event text data given in XML into chunks of text that can be associated to a specific source ("Who said it and for which company?");
- Code that makes use of Stanford's CoreNLP ([Manning, Christopher D. et. al.]) and instructions on how to use it within the popular Java technology provided by one of Thomson Reuter's strategic partners Oracle ([Forbes, May. 2013]).

2. Use Case

The use case is to feed text data of Street Events into Stanford's CoreNLP. This returns for each statement of every single operator a tree structure that measures the sentiment of the statement. With some work this can be used as a feature for further investigations.

*This article has been created for the second round of the world-wide Texata Data Analytics contest in 2014. The author wants to thank HackerRank for the free invite and Thomson Reuters for the interesting data.

II. Methodology

Scala has been picked as the language to code in. Scala is interopable with Java and runs on the same virtual machine. The project is made with SBT ("Simple Build Tool") as a build tool (similar to Apache Maven). Since Stanford's CoreNLP is written in Java it can be used from this implementation.

The StreetEvent data is given in XML format. In the context of this case study a project has been created that extracts data from this XML and starts to parse the text bodies of Street Events into chunks that is fed into a neuronal network for Sentiment Analysis.

The chunks have the form of tuples of the name of the talker and a text that encodes what he has been saying. For example, talker could be "Lauren Fine, Merrill Lynch" who said "Yes, thank you. Sorry for the background noise. I have a couple of questions. [...]".

III. Outcomes

A Scala program was written that is able to parse Street Event xml data. The program does its job, but due to time constraints there is a lack of visualisation of the results of the sentiment analysis. IntelliJ project files are provided and it is the recommended IDE to open this project.

It can be run via running the SentimentAnalysis.jar with as parameter a path to a directory that contains EventData XML files or a XML file itself. It will then extract everything said by an operator of the conversation and analyse it.

Further analysis should be done on the result found in the class StreetEventXMLReader. Of course it has to be evaluated whether the used training model (the neuronal net) provided by Stanford's CoreNLP is good enough.

References

[Thomson Reuters, Sept. 2014] Big Data: Successfully meeting the latest challenges in financial services *Thomson Reuters Blog* `http://tinyurl.com/ltqaa9h`

[Techcrunch, Feb. 2014] Thomson Reuters Taps Into Twitter For Big Data Sentiment Analysis *Tech Crunch* `http://tinyurl.com/mew5o29`

[Forbes, May. 2013] Thomson Reuters Transforms Big Data Into Big Business *Forbes* `http://tinyurl.com/qypsouq`

[Manning, Christopher D. et. al.] The Stanford CoreNLP Natural Language Processing Toolkit `http://www.aclweb.org/anthology/P/P14/P14-5010`