Text Analytics

The concept and use of text analytics and sentiment analysis is fascinating to me. It is almost like it is and English and grammar class (which is my least favorite subject) come back to haunt me. However, I find the field of text analytics, sentiment analysis, and NLP in general quite interesting.

The first topic is Latent Dirichlet Allocation or LDA for short. LDA is, I would say, a form of automatic feature engineering in that it automatically discovers topics in sentences in each corpus. LDA categorizes each sentence into a specific topic first and then describes the topic once all the sentences have been read. From my understanding LDA is a form of token exploration tied to a very specific topic mixture. Then once the topic mixture has been determined then you use it on documents and assign words to each topic. Then to improve you would calculate the probability of that word in that topic and possibly move it to another topic that makes more sense.

Sentiment analysis refers to gaining insights into public opinion via social media monitoring. Sentiment analysis is necessary for companies now to have any competitiveness whatsoever. Since it readily deals with social media and peoples and customers opinion, being able to accurately understand attitudes can make other thing have a bigger impact, such as a more targeted marketing campaign. The though part of sentiment analysis is they way humans write. Humans can interpret the tone of writing or someone saying something much easier than a machine can, so teaching it to properly know how to understand that is the trick to winning with sentiment analysis.

One of the reading addresses the issue with sentiment analysis doing a poor job at reading and understanding sentiment. It is called a Recursive Neural Tensor Network and it is trained on a Sentiment Treebank that consists of 215,154 phrases in trees of 11,855 decision tree style trees. At the time of writing this model was the only word model to accurately capture contrastive conjunctions and negation for both positive and negative phrases.

So great, we have a model that can do all these things, but how long does it take to do that. Speed is pertinent for these models. Documents can be very large, like internet size large, so how can we effectively make this happen in real time. IBM’s Watson has done all of this and used it to win against real human opponents in Jeopardy. IBM advanced the field of text analytics with Watson because it was able to recognize and discard irrelevant and misleading information while learning from mistake all in real time. Watson also has a capability of hypothesizing differential diagnoses in patients to assists doctors with faster treatment times. This does not mean that Watson will be the doctor because there is still a large dependence on humans for their intuition and understanding, but with real time text analytics some things can be made easier.