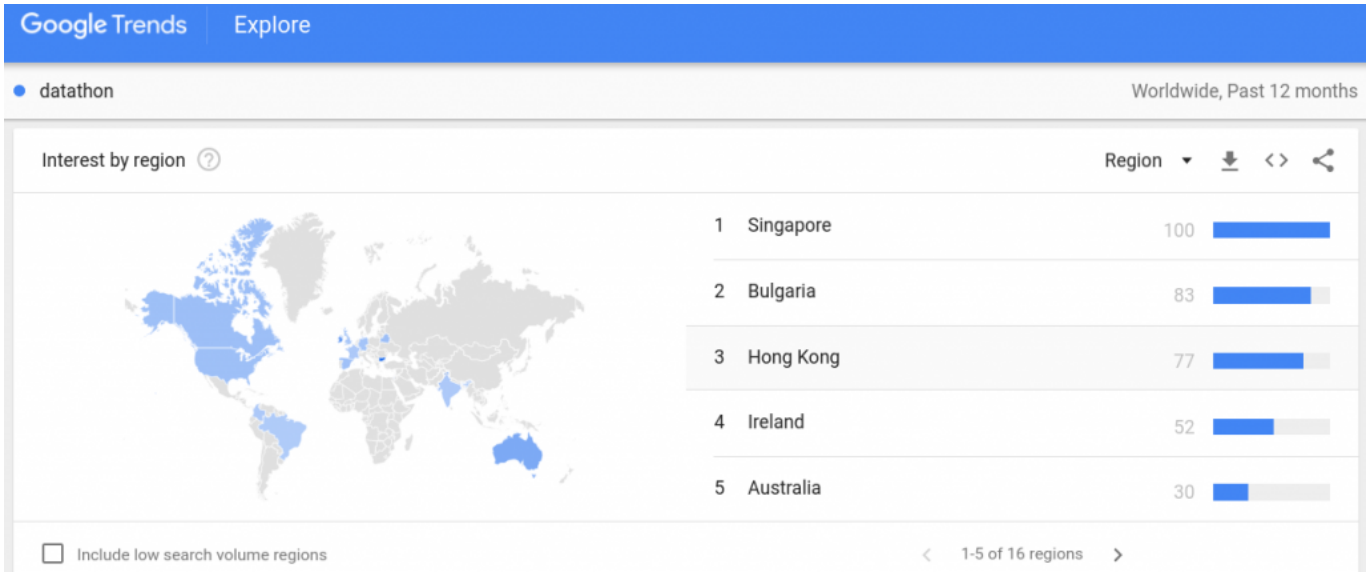
Many data scientists in Sofia and over the world didn't sleep much during the weekend of February 9-11. The reason? This year, the Data Science Society (DSS) organized a data challenge that was met with unprecedented enthusiasm and quickly went beyond the borders of Bulgaria. Thanks to the remarkable efforts of the DSS – this is their third consecutive datathon – Bulgaria is now on the map of data science (see Google Trends for the past 12 months).



Nine industry companies, including Ontotext, presented interesting and difficult data analysis problems relevant to their business. Their cases competed for attracting the skills and creativity of teams of data scientists. In their turn, the teams couldn't wait to discover hidden correlations and patterns as well as how well their models could fit unseen test data.

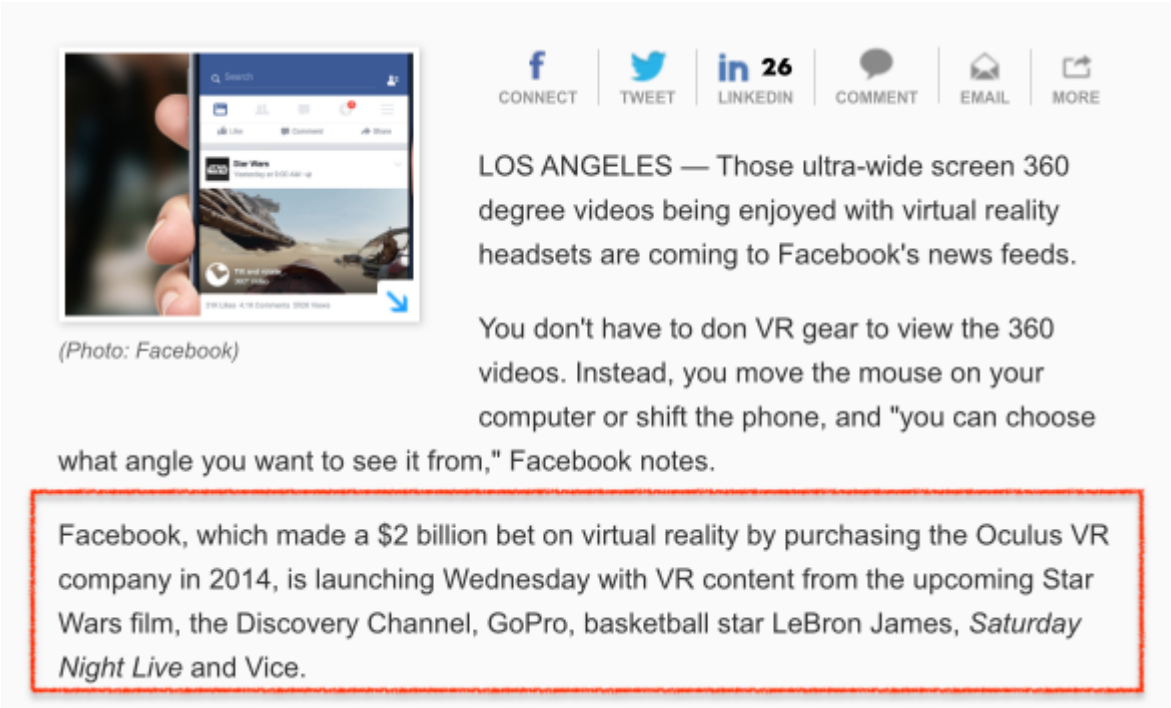
The Case

Ontotext presented a large set of news articles about companies. The challenge was to identify parent-subsidiary pairs of companies by training models on already annotated examples.

Read our White Paper: Text Analytics for Enterprise Use!

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

For instance, Facebook is the parent of Oculus VR and the relation between the two companies is expressed in the following news excerpt:



CONNECT TWEET 26 LINKEDIN COMMENT EMAIL MORE

(Photo: Facebook)

LOS ANGELES — Those ultra-wide screen 360 degree videos being enjoyed with virtual reality headsets are coming to Facebook's news feeds.

You don't have to don VR gear to view the 360 videos. Instead, you move the mouse on your computer or shift the phone, and "you can choose what angle you want to see it from," Facebook notes.

Facebook, which made a $2 billion bet on virtual reality by purchasing the Oculus VR company in 2014, is launching Wednesday with VR content from the upcoming Star Wars film, the Discovery Channel, GoPro, basketball star LeBron James, *Saturday Night Live* and Vice.

This task is called relation extraction and is not new in the field of Natural Language Processing (NLP). However, traditional models don't perform well and are often restricted to specific domains. The bottleneck is the small amount of manually annotated examples versus the great flexibility of human language, which can express one relation in many arbitrary ways. The revolution in Artificial Intelligence brought by Deep Learning holds many promises. But in the NLP area, Neural Networks need even more annotated examples than the traditional methods.

This is why the R&D team at Ontotext implemented a distant supervision paradigm in order to obtain a large set of annotated examples by using parent-subsidiary relations from Linked Open Data (in this case, DBpedia). Ontotext's FactForge supplied annotated news articles spanning years back. In this way, all key components came into place for a nice dataset of about 90K text snippets expressing parent-subsidiary relations.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

ontotext

The Teams

Two teams worked hard for almost 48 hours to find solutions to Ontotext's data challenge. The members were mostly Bulgarian, but we were also joined by enthusiasts from Spain, Greece and Belgium. The participants' background with data science was very mixed. We were happy to see very experienced coders, team members with solid theoretical knowledge as well as juniors who were eager to learn from the experts.

During the Datathon, we provided our continuous support. The performance of a data scientist depends heavily on how well he or she understands the domain of the data. So we answered a lot of questions about the data and offered some suggestions for the technical methods.

The Results

Ontotext couldn't be happier about the outcome of the challenge. Both teams presented solutions that exceeded expectations. Within such a narrow timeframe, they tried basic methods that were easy to implement and test as well as some advanced Neural Networks approaches that lead to very good results.

For the curious, logistic regression methods were used as simple baseline approaches with modest performance. Long-Short-Term-Memory networks, which are essentially recurrent neural networks that can "remember" context and are often applied to solve NLP tasks, gave much better performance, an F1 score of 79%. A similar performance was achieved by a state-of-the-art attention network, which combines all evidence available for a given company pair to make predictions.

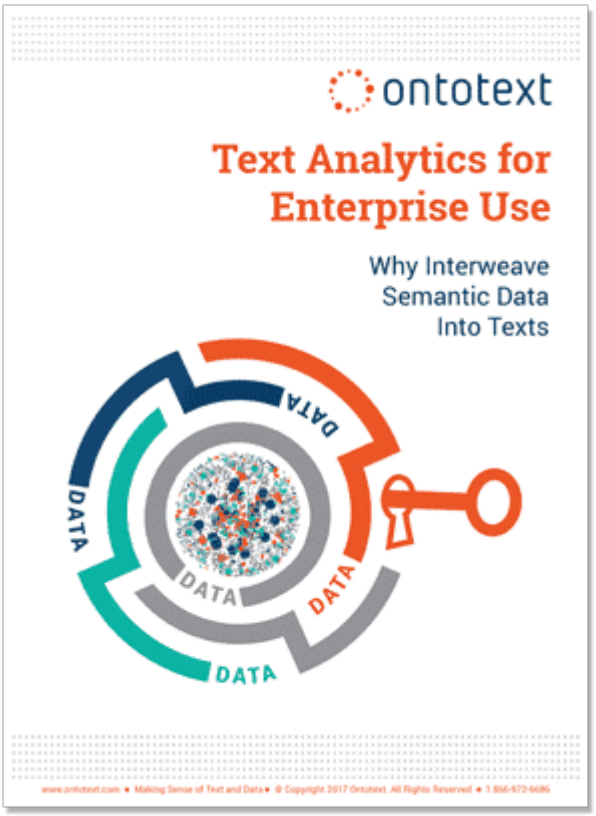
Importantly, after only 48 hours the teams showed in-depth understanding about the data, the biases, the noise, the limitations of their methods and had ideas about possible directions of improvement. Despite the lack of sleep, they managed to communicate their results with rigour, enthusiasm, humor and self-criticism. And perhaps most impressively, they said they'd like to continue the experiments.

Ontotext would like to thank **Team_A** and team **Centroida** for their work and hope we'd meet again at future Data Science events.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

ontotext

Want to learn more about text analytics and relation extraction?



White Paper: Text Analytics for Enterprise Use
Use the power of text analytics for your enterprise

Download Now

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com