Guide to Natural Language Processing

The Open Data Science Community's Top **30** Resources

3 top blogs

9 from the experts

11 what's next

6 top ODSC talks

10 editor's note

15 TOP OPEN DATA SCIENCE BLOGS

In the past two years, we have published over 700 articles on data science. NLP is an increasingly hot topic, as organizations are seeking new ways to make sense of their language data.
Here are the 15 most-read articles on NLP to help you thrive in the space.

An Introduction to Natural Language Processing, *Diego Lopez Yse*

NLP is one of the biggest topics in data science now. Need to know more? Read this comprehensive introduction to natural language processing!

Read it here.

Essential NLP Tools, Code, and Tips, *Diego Lopez Yse*

NLP is a complicated beast. Use this guide to gain the essential NLP tools, code, and tips you need to get started.

Read it here.

Intro to Language Processing with the NLTK, *Kailen Santos*

If you have a lot of text-rich data that would be impossible to read, NLP can concentrate that into simple insights. Learn language processing with the NLTK.

Read it here.

20 Open Datasets for Natural Language Processing, *Elizabeth Wallace*

Language is a big part of machine learning, but it requires a lot of data and some good training. Here are 20 open datasets for natural language processing.

Read it here.

15 TOP OPEN DATA SCIENCE BLOGS CONTINUED

Getting to Know Natural Language Understanding, *Elizabeth Wallace*

Let's take a look at natural language understanding - or why computers can win chess matches against world champions but can't grasp sarcasm.

Read it here.

Combining Millions of Products Into One Marketplace Using Computer Vision and NLP, *Ali Vanderveld*

How do online retailers and businesses categorize products without having humans do it manually? With computer vision and NLP, of course!

Read it here.

Using NLP and ML to Analyze Legislative Burdens Upon Businesses, *Serena Peruzzo*

How can businesses and governments leverage NLP and ML for analysis of regulatory burdens and other potential effects of legislation?

Read it here.

Building a Natural Language Question & Answer Search Engine, *Adam Spannbauer*

What high-level steps do you need to take to create a natural language question and answer search engine? Find out right here!

Read it here.

Introduction to Spark NLP: Foundations and Basic Components, *Veysel Kocaman*

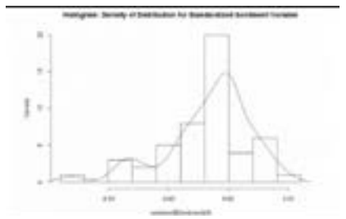
If you're involved in NLP, then you've likely heard of (if not used) the Spark library. What is it, why should you use it, and what are some important foundations that you need to know?

Read it here.

An Introduction to Sentence-Level Sentiment Analysis with sentimentr, *Brandon Dey*

This post explores the basics of sentence level sentiment analysis, unleashing sentimentr on the entire corpus of R-package help-documents on CRAN.

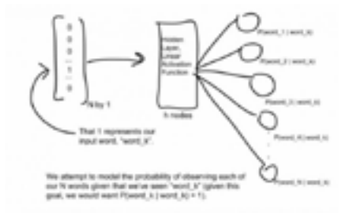
Read it here.



Sentiment Analysis in R Made Simple, *Daniel Gutierrez*

Here's a demonstration of sentiment analysis using the SentimentAnalysis package in R, where you will learn to extract subjective information from textual documents.

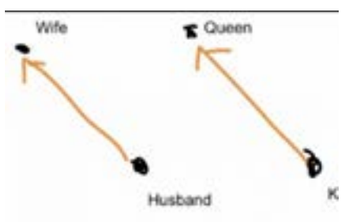
Read it here.



An Idiot's Guide to Word2vec Natural Language Processing, *Spencer Norris*

Word2vec provides vector representations of words, which can help achieve decent performance across tasks that machines have been historically bad at.

Read it here.



Why Word Vectors Make Sense in Natural Language Processing, *Spencer Norris*

Word vectors' primary use is to determine something about what they mean based on how their shapes and orientations relate to each other.

Read it here.

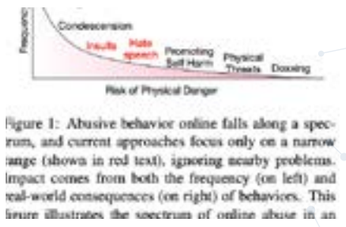


Figure 1: Abusive behavior online falls along a spectrum, and current approaches focus only on a narrow range (shown in red text), ignoring nearby problems. Impact comes from both the frequency (on left) and real-world consequences (on right) of behaviors. This figure illustrates the spectrum of online abuse in an

Best NLP Research of 2019, *Daniel Gutierrez*

Let's help get you up speed with current NLP research efforts by curating a list of the best NLP research of 2019 on arXiv.org

Read it here.

Why NLP is a Great First AI Solution for Businesses, *Alex Amari*

Many execs are afraid that the cost of AI would be too much to make sense. But we need to pinpoint an AI solution for businesses, and NLP is a great option.

Read it here.

15 TOP TALKS FROM ODSC CONFERENCES

Out of the 500+ talks from ODSC conferences in 2019, here are the 15 top-rated sessions covering NLP.

Serena Peruzzo, Daniel Parton, PhD

Lead Data Scientist, Sr. Dara Scientist, East 2019

Analyzing Legislative Burden Upon Businesses Using NLP and ML

Speakers first describe the legislative/business context of detecting parts of the legislature that indicate legislative burden and categorize them, then walk attendees through the technical implementation. The work is conducted by combining various techniques from the NLP toolbox.

Watch here.
Slides.

Frank Zhao

Senior Director Quantamental Research S&P Global, East 2019

NLP: Deciphering the Message Within the Message - Stock Selection Insights Using Corporate Earnings Calls

In this presentation, we explore a number of sentiment- and behavioral-based signals using the content from earnings call transcripts via NLP that have historically demonstrated stock selection power in the U.S. market.

Watch here.
Slides.

David Talby

CTO at Pacific AI, West 2018

State of the Art Natural Language Understanding at Scale

This talk introduces the NLP library for Apache Spark. It natively extends the Spark ML pipeline API's which enables zero-copy, distributed, combined NLP & ML pipelines, leveraging all of Spark's built-in optimizations. David demonstrates using these algorithms to build commonly used pipelines, using PySpark on public notebooks.

Watch here.
Slides.

Madison May

ML Architect, Cofounder at Indico, East 2018

Effective Transfer Learning for NLP

In this talk, we explore parameter and data-efficient mechanisms for transfer learning using sequence representations rather than fixed length document vectors as a medium for communication between models, and show practical improvements on real-world tasks. In addition, we demo the use of Enso, a newly open-sourced library.

Watch here.
Slides.

Andrew Long, PhD

Data Scientist at Fresenius Medical Care, West 2019

Healthcare NLP with a Doctor's Bag of Notes

In this hands-on workshop, viewers have the opportunity to complete a Python NLP project with doctors' discharge summaries to predict unplanned hospital readmission.

Watch here.
Accompanying Blog.

Sudha Subramanian

Data Scientist at Sparkfish, Europe 2019

Identify Heart Disease Risk Factors from Clinical Notes

Stacking embeddings is simply a way to combine multiple embeddings and has demonstrated good results on the i2b2 heart disease risk factors challenge dataset.

Watch here.
Slides.

Veysel Kocaman, PhD

Senior Data Scientist at John Snow Labs, West 2019

Spark NLP for Healthcare: Lessons Learned Building Real-World Healthcare AI Systems

This talk reviews case studies from real-world projects that built AI systems using NLP in healthcare. Veysel covers why and how NLP was used, what deep learning models and libraries were used, and what was achieved.

Watch here.
Slides.

Eric Xing, PhD

Founder & Chief Scientist at Petuum, West 2019

Composable Machine Learning

To break this status quo, ML systems need to become composable, so that ML teams can build applications for a richer spectrum of AI tasks from standardized and reusable building blocks, and take them into scalable production.

Watch here.

Mariana Romanyshyn

Technical Lead at Grammarly Inc., Europe 2018

Linguistics in NLP: why so complex?

In this talk, we examine how linguistic intuition can be formalized and encoded to solve the problem of complex word identification and correction. We investigate word formation, dive into language structures, and learn about language modelling.

Watch here.
Slides.

Ido Shlomo

Senior Data Science Manager at BlueVine, West 2019

Building an Industry Classifier With The Latest Scraping, NLP and Deployment Tools

This presentation will cover the entire development pipeline hands-on: Crowd-sourcing a tagged sample, building a smart and scalable web scraper, prepping and feeding the resulting raw data into BERT, fine tuning the model and finally deploying it as a cloud based service behind an API.

Watch here.

15 TOP TALKS FROM ODSC CONFERENCES CONTINUED

Mariana Romanyshyn

Technical Lead at Grammarly Inc., Europe 2019

Meaning Representation for Natural Language Understanding

In this talk, we discuss the algorithms of building AMR graphs and NLP applications that can benefit from these graphs. Watch here. Slides.

Rosaria Silipo, PhD

Principal Data Scientist at KNIME, Europe 2018

Puzzling Together a Teacher-Bot: Machine Learning, NLP, Active Learning, and Microservices

This talk introduces Emil, a Teacher Bot that was built to answer questions about using the KNIME Analytics Platform. Watch here. Slides.

Sijun He

Machine Learning Engineer at Twitter Cortex, West 2019

Named Entity Recognition At Scale With Deep Learning

Sijun offers insights into how Twitter Cortex built and productionized a DL-based NER system to address those challenges, such as experimentations with state-of-the-art models and learning methods. Watch here. Slides.

Dr. Catherine Havasi

Research Scientist at MIT Media Lab, East 2018

Transfer Learning: Applications for natural language understanding (Accelerate AI)

This talk focuses on language-related use cases for customer service, search, question answer, self-help, and consumer finance. We also have some fun with applications of transfer learning. Watch here. Accompanying Blog.

Sihem Romdhani

Software Engineer at Veeva Systems, East 2018

Machine Learning and Natural Language Processing for Detecting Fake News

Through use cases and examples, this talk discusses the different fake news detection approaches from feature extraction to model construction. We focus on how to leverage NLP to characterize and extract discriminative features of fake news by analyzing its text content. Watch here. Slides.

FROM THE EXPERTS:

What do leading data scientists think about the current state of NLP?

Performance in NLP looks more promising than ever before, and I'm excited to see these advances trickle down into commercial language-based products. We can probably look forward to smarter voice recognition, better autocomplete, smoother translation, and more sensible chatbots—perhaps with more ambitious applications, like free-form dialogue with NPCs in video games.

Natasha Latysheva
Machine Learning Research Engineer at Welocalize

In 2019, the progress in NLP from BERT-variations to GPT-2 and MultiFiT blew me away. Accurate text classification across many languages is becoming common. Performance on other tasks, such as QA and NER, is improving. Text generation is quite impressive and has many practical implications for the world.

Matthew Teschke
Director, Applied Machine Learning at Novetta

> The broadening influence of transformer architectures within deep learning models applied to a range of natural language processing tasks in 2019. Most notably, the GPT-2 model released by researchers at OpenAI is capable of generating remarkably coherent lengthy sequences of text.

Jon Krohn
Chief Data Scientist at Untapt
Author of Deep Learning Illustrated

NLP is where computer vision was 5-6 years ago. 2020 will be the start of the revolution that transforms Enterprise Search, Document Discovery, and numerous other NLP applications. I can't wait to help bring it to life.

Charles Martin
CEO at Calculation Consulting

We have witnessed many exciting breakthroughs in NLP in 2019 and thanks to transformers, we have already surpassed human baselines in several NLU tasks. I believe that NLP will gain even more attention in the healthcare industry.

Veysel Kocaman
Senior Data Scientist at John Snow Labs

I'm excited about the convergence of data and design. Design tools enhanced by AI, and AI apps are becoming more human-centered. Conversational UX is a great example of a field that combines design with the increasing maturity of NLP technology.

Gautam Tambay
CEO & Co-founder at Springboard

Open Data Science HIGHLIGHTS

Blog with the **MOST TRAFFIC:** An Introduction to Natural Language Processing, *Diego Lopez Yse*

Highest-rated **ODSC TALK:** Healthcare NLP with a Doctor's Bag of Notes, *Andrew Long, PhD, ODSC West 2019*

#1 blog on @ODSC Medium: Essential NLP Tools, Code, and Tips, *Diego Lopez Yse*

Learn.AI course with the **HIGHEST ENROLLMENT:** ML and NLP Processing for Detecting Fake News, *Sihem Romdhani* *Watch here*

Letter from the Editor

Maybe I'm biased, but I think language is pretty cool.

Really though, I find NLP to be one of the most important developments under the data science and AI umbrella. Humans speak with words, not numbers, and it's important that we do all we can to teach our machines to understand words and expression.

NLP was probably my first foray into data science. When I was earning my master's degree, I had to take a course on social data analysis, and in the world of media studies, that meant a lot of Twitter data. Considering my background has always been in writing, learning R wasn't the easiest—but I found the results fascinating.

In said course, my team and I examined (an at the time trending piece of) research on the consumption of red meat and cancer. Using sentiment analysis, we wanted to see if the public perception of the term "red meat" had any notable change before and after the World Health Organization (WHO) report on red meat emerged.

As you can imagine, not only did more people discuss "red meat" on Twitter (in reference to food), but also the tone of the discussion became more negative. The results didn't really surprise us, but the process itself was exciting.

Call me crazy, but I loved manual labeling. To teach our machine what to look for, we hand-labeled 200 tweets each, flagging for spam and other non-related content. We performed this research in the summer of 2016, so there were plenty of posts related to "red meat" in the context of politics. It felt oddly therapeutic to clean out all of the unrelated Twitter posts and just leave the relevant ones.

I thank my time at Boston University for making me interested in data science. Since joining ODSC, I've only grown to appreciate it more. It's not just red meat and hand-coding anymore.

When I first learned about NLP, I thought it was just a tool for social media, but I was clearly quite wrong. One of my favorite applications of NLP is definitely in the healthcare setting; being able to develop systems that can diagnose diseases, predict potential risks, and even help with treatment is quite literally a life-saver.

NLP—and of course other AI techniques—will likely never replace humans completely. However, humans must learn to use AI to the best of our abilities. AI is becoming a powerful tool, and with 500 million+ tweets sent daily, we could use all the help we can with interpreting the vast amount of words that we write every day.

-Alex Landa, ODSC Content Manager

BE A PART OF THE ODSC COMMUNITY

There are many ways you can engage with the Open Data Science Community today!

2020 ODSC Events

East: April 13-17
India: September 9-12
Europe: September 14-18
West: October 26-30

More Downloadable Guides

Did you like this guide? We also have downloadable guides for machine learning and deep learning. Download them for free now!

Meetups

We hold meetups in 37 cities around the world, designed to bring data scientists together for education, networking, and even a little fun. See upcoming events here.

Weekly Newsletter

Don't miss any future articles on data science and machine learning! Sign up for our weekly newsletter and get tutorials, insights, and the latest news sent to you directly.

Webinars

We offer free webinars several times a month, covering a variety of topics. Follow this page to learn more about upcoming webinars.

Becoming a Part of ODSC Events

Are you a technical or business expert in the world of data science and AI? Consider speaking at one of our events! Each event has its own speaker submission page:

ODSC East 2020
ODSC Europe 2020

We also offer partnership opportunities! Have your product, service, or research seen by thousands of data scientists at an event. Learn more here.