Introduction to Natural Language

Matt Kiser

2016

What is Natural Language Processing?

NLP is a way for computers to analyze, understand, and derive meaning from human language in a smart and useful way. By utilizing NLP, developers can organize and structure knowledge to perform tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation.

“Apart from common word processor operations that treat text like a mere sequence of symbols, NLP considers the hierarchical structure of language: several words make a phrase, several phrases make a sentence and, ultimately, sentences convey ideas,” John Rehling, an NLP expert at Meltwater Group, said in *How Natural Language Processing Helps Uncover Social Media Sentiment*. “By analyzing language for its meaning, NLP systems have long filled useful roles, such as correcting grammar, converting speech to text and automatically translating between languages.”

NLP is used to analyze text, allowing machines to understand how human’s speak. This human-computer interaction enables real-world applications like automatic text summarization, sentiment analysis, topic extraction, named entity recognition, parts-of-speech tagging, relationship extraction, stemming, and more. NLP is commonly used fortext mining, machine translation, and automated question answering.

NLP is characterized as a hard problem in computer science. Human language is rarely precise, or plainly spoken. To understand human language is to understand not only the words, but the concepts and how they’re linked together to create meaning. Despite language being one of the easiest things for humans to learn, the ambiguity of language is what makes natural language processing a difficult problem for computers to master.

What Can Developers Use NLP Algorithms For?

NLP algorithms are typically based on machine learning algorithms. Instead of hand-coding large sets of rules, NLP can rely on machine learning to automatically learn these rules by analyzing a set of examples (i.e. a large corpus, like a book, down to a collection of sentences), and making a statical inference. In general, the more data analyzed, the more accurate the model will be.

- **Summarize blocks of text** using Summarizer to extract the most important and central ideas while ignoring irrelevant information.
- Create a **chat bot** using Parsey McParseface, a language parsing deep learning model made by Google that uses Point-of-Speech tagging.
- **Automatically generate keyword tags** from content using AutoTag, which leverages LDA, a technique that discovers topics contained within a body of text.
- **Identify the type of entity extracted**, such as it being a person, place, or organization using Named Entity Recognition.
- Use Sentiment Analysis to **identify the sentiment of a string of text**, from very negative to neutral to very positive.
- **Reduce words to their root**, or stem, using PorterStemmer, or **break up text into tokens** using Tokenizer.

Open Source NLP Libraries

These libraries provide the algorithmic building blocks of NLP in real-world applications. Algorithmia provides a free API endpoint for many of these algorithms, without ever having to setup or provision servers and infrastructure.

- Apache OpenNLP: a machine learning toolkit that provides tokenizers, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, coreference resolution, and more.
- Natural Language Toolkit (NLTK): a Python library that provides modules for processing text, classifying, tokenizing, stemming, tagging, parsing, and more.
- Standford NLP: a suite of NLP tools that provide part-of-speech tagging, the named entity recognizer, coreference resolutionsystem, sentiment analysis, and more.
- MALLET: a Java package that provides Latent Dirichlet Allocation, document classification, clustering, topic modeling, information extraction, and more.

A Few NLP Examples

- Use **Summarizer** to automatically summarize a block of text, exacting topic sentences, and ignoring the rest.
- Generate keyword topic tags from a document using **LDA** (Latent Dirichlet Allocation), which determines the most relevant words from a document. This algorithm is at the heart of the **Auto-Tag** and **Auto-Tag URL** microservices.
- **Sentiment Analysis**, based on StanfordNLP, can be used to identify the feeling, opinion, or belief of a statement, from very negative, to neutral, to very positive. Often, developers with use an algorithm to identify the sentiment of a term in a sentence, or use sentiment analysis to analyze social media.

NLP algorithms can be extremely helpful for web developers, providing them with the turnkey tools needed to create advanced applications, and prototypes.

Example Natural Language Processing Use Cases

NLP algorithms are typically based on machine learning algorithms. Instead of hand-coding large sets of rules, NLP can rely on machine learning to automatically learn these rules by analyzing a set of examples (i.e. a large corpus, like a book, down to a collection of sentences), and making a statical inference. In general, the more data analyzed, the more accurate the model will be.

Social media analysis is a great example of NLP use. Brands track conversations online to understand what customers are saying, and glean insight into user behavior.

“One of the most compelling ways NLP offers valuable intelligence is by tracking sentiment — the tone of a written message (tweet, Facebook update, etc.) — and tag that text as positive, negative or neutral,” Rehling said.

Build your own social media monitoring tool

1. Start by using the algorithm Retrieve Tweets With Keyword to capture all mentions of your brand name on Twitter. In our case, we search for mentions of Algorithmia.
2. Then, pipe the results into the Sentiment Analysis algorithm, which will assign a sentiment rating from 0-4 for each string (Tweet).

Similarly, Facebook uses NLP to track trending topics and popular hashtags.

“Hashtags and topics are two different ways of grouping and participating in conversations,” Chris Struhar, a software engineer on News Feed, said in *How Facebook Built Trending Topics With Natural Language Processing*. “So don’t think Facebook won’t recognize a string as a topic without a hashtag in front of it. Rather, it’s all about NLP: natural language processing. Ain’t nothing natural about a hashtag, so Facebook instead parses strings and figures out which strings are referring to nodes — objects in the network. We look at the text, and we try to understand what that was about.”

It’s not just social media that can use NLP to it’s benefit. Publishers are hoping to use NLP to improve the quality of their online communities by leveraging technology to “auto-filter the offensive comments on news sites to save moderators from what can be an ‘exhausting process’,” Francis Tseng said in *Prototype winner using ‘natural language processing’ to solve journalism’s commenting problem*.

Other practical uses of NLP include monitoring for malicious digital attacks, such as phishing, or detecting when somebody is lying.

Use NLP to build your own RSS reader

You can build a machine learning RSS reader in less than 30-minutes using the follow algorithms:

1. ScrapeRSS to grab the title and content from an RSS feed.
2. Html2Text to keep the important text, but strip all the HTML from the document.
3. AutoTag uses Latent Dirichlet Allocation to identify relevant keywords from the text.
4. Sentiment Analysis is then used to identify if the article is positive, negative, or neutral.
5. Summarizer is finally used to identify the key sentences.

Recommended NLP Books for Beginners

- Speech and Language Processing: “The first of its kind to thoroughly cover language technology – at all levels and with all modern technologies – this book takes an empirical approach to the subject, based on applying statistical and other machine-learning algorithms to large corporations.”
- Foundations of Statistical Natural Language Processing: “This foundational text is the first comprehensive introduction to statistical natural language processing (NLP) to appear. The book contains all the theory and algorithms needed for building NLP tools. It provides broad but rigorous coverage of mathematical and linguistic foundations, as well as detailed discussion of statistical methods, allowing students and researchers to construct their own implementations. The book covers collocation finding, word sense disambiguation, probabilistic parsing, information retrieval, and other applications.”
- Handbook of Natural Language Processing: “The Second Edition presents practical tools and techniques for implementing natural language processing in computer systems. Along with removing outdated material, this edition updates every chapter and expands the content to include emerging areas, such as sentiment analysis.”
- Statistical Language Learning (Language, Speech, and Communication): “Eugene Charniak breaks new ground in artificial intelligenceresearch by presenting statistical language processing from an artificial intelligence point of view in a text for researchers and scientists with a traditional computer science background.”

- Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit
“*This is a book about Natural Language Processing. By “natural language” we mean a language that is used for everyday communication by humans; languages like English, Hindi or Portuguese. At one extreme, it could be as simple as counting word frequencies to compare different writing styles*.
- Speech and Language Processing, 2nd Edition 2nd Edition
“*An explosion of Web-based language techniques, merging of distinct fields, availability of phone-based dialogue systems, and much more make this an exciting time in speech and language processing. The first of its kind to thoroughly cover language technology – at all levels and with all modern technologies – this text takes an empirical approach to the subject, based on applying statistical and other machine-learning algorithms to large corporations. The authors cover areas that traditionally are taught in different courses, to describe a unified vision of speech and language processing*.”
- Introduction to Information Retrieval
“*As recently as the 1990s, studies showed that most people preferred getting information from other people rather than from information retrieval systems. However, during the last decade, relentless optimization of information retrieval effectiveness has driven web search engines to new quality levels where most people are satisfied most of the time, and web search has become a standard and often preferred source of information finding. For example, the 2004 Pew Internet Survey (Fallows, 2004) found that 92% of Internet users say the Internet is a good place to go for getting everyday information.” To the surprise of many, the field of information retrieval has moved from being a primarily academic discipline to being the basis underlying most people’s preferred means of information access*.”

NLP Tutorials

- Natural Language Processing Tutorial: “We will go from tokenization to feature extraction to creating a model using a machine learning algorithm. You can get the source of the post from github.”
- Basic Natural Language Processing: “In this tutorial competition, we dig a little “deeper” into sentiment analysis. People express their emotions in language that is often obscured by sarcasm, ambiguity, and plays on words, all of which could be very misleading for both humans and computers.“
- An NLP tutorial with Roger Ebert: “Natural Language Processing is the process of extracting information from text and speech. In this post, we walk through different approaches for automatically extracting information from text—keyword-based, statistical, machine learning—to explain why many organizations are now moving towards the more sophisticated machine-learning approaches to managing text data.”