Introduction to Text Mining

Sains Data Terapan

2022

Politeknik Elektronika Negeri Surabaya
Departemen Teknik Informatika dan Komputer

Pokok Bahasan

- 1. Introduction to Text Mining dan Natural Language Processing
- 2. Data Preparation
- 3. Bag-of-Words
- 4. Word Embeddings
- 5. Text Classification
- 6. Image Captioning
- 7. Text Summarization

Working with Text is… important, under-discussed, and HARD

- We are awash with text, from books, papers, blogs, tweets, news, and increasingly text from spoken utterances.
- Every day, I get questions asking how to develop machine learning models for text data.
- Working with text is hard as it requires drawing upon knowledge from diverse domains such as linguistics, machine learning, statistical natural language processing, and these days, deep learning.

The Problem with Text

- The problem with modeling text is that it is messy, and machine learning algorithms prefer well defined fixed-length inputs and outputs.
- Machine learning algorithms cannot work with raw text directly; the text must be converted into numbers. Specifically, vectors of numbers.
- This is called feature extraction or feature encoding and this is one of the key areas where deep learning is really shaking things up.

UNLOCK Natural Language Processing with Deep Learning

- Classical linguistic methods for natural language processing required experts in language defining rules to cover specific cases. These worked in narrow cases but turned out to be fragile.
- Statistical methods improve upon classical linguistic methods by learning rules and models from data rather than requiring them to be specified in a top-down manner. They result in much better performance, but must still be complemented with hand-crafted augmentations by language experts in order to achieve useful results.
- Often, a pipeline of statistical methods are required to achieve a single modeling outcome, such as in the case of machine translation.
- Deep learning methods are starting to out-compete the statistical methods on some challenging natural language processing problems with singular and simpler models.

The Promise of Deep Neural Networks for NLP

- Deep learning methods are popular, primarily because they are delivering on their promise.
- That is not to say that there is no hype around the technology, but that the hype is based on very real results that are being demonstrated across a suite of very challenging artificial intelligence problems from computer vision and natural language processing.
- Some of the first large demonstrations of the power of deep learning were in natural language processing, specifically speech recognition. More recently in machine translation.

The 5 promises of deep learning for natural language processing:

- **The Promise of Drop-in Replacement Models**. That is, deep learning methods can be dropped into existing natural language systems as replacement models that can achieve commensurate or better performance.
- **The Promise of New NLP Models**. That is, deep learning methods offer the opportunity of new modeling approaches to challenging natural language problems like sequence-to-sequence prediction.
- **The Promise of Feature Learning**. That is, that deep learning methods can learn the features from natural language required by the model, rather than requiring that the features be specified and extracted by an expert.
- **The Promise of Continued Improvement**. That is, that the performance of deep learning in natural language processing is based on real results and that the improvements appear to be continuing and perhaps speeding up.
- **The Promise of End-to-End Models**. That is, that large end-to-end deep learning models can be fit on natural language problems offering a more general and better-performing approach.

Impressive Applications of Deep Learning

- Natural language processing is not "*solved*", but deep learning is required to get you to the state-of-the-art on many challenging problems in the field.
- Let's look at 3 examples to give you a snapshot of the results that deep learning is capable of achieving in the field of natural language processing:

1) Automatic Image Caption Generation	2) Automatic Translation of Text	3) Automatic Text Classification
• Automatic image captioning is the task where, given a photograph, the system must generate a caption that describes the contents of the image.	• Automatic text translation is the task where you are given sentences of text in one language and must translate them into text in another language.	• Automatic text classification is the task of assigning a class label given a text document such as a review, tweet, or email.

Automatic Image Caption Generation

A person is walking along a beach with a big dog

A black and white dog carries a tennis ball in its mouth

A soccer player takes a soccer ball in the grass

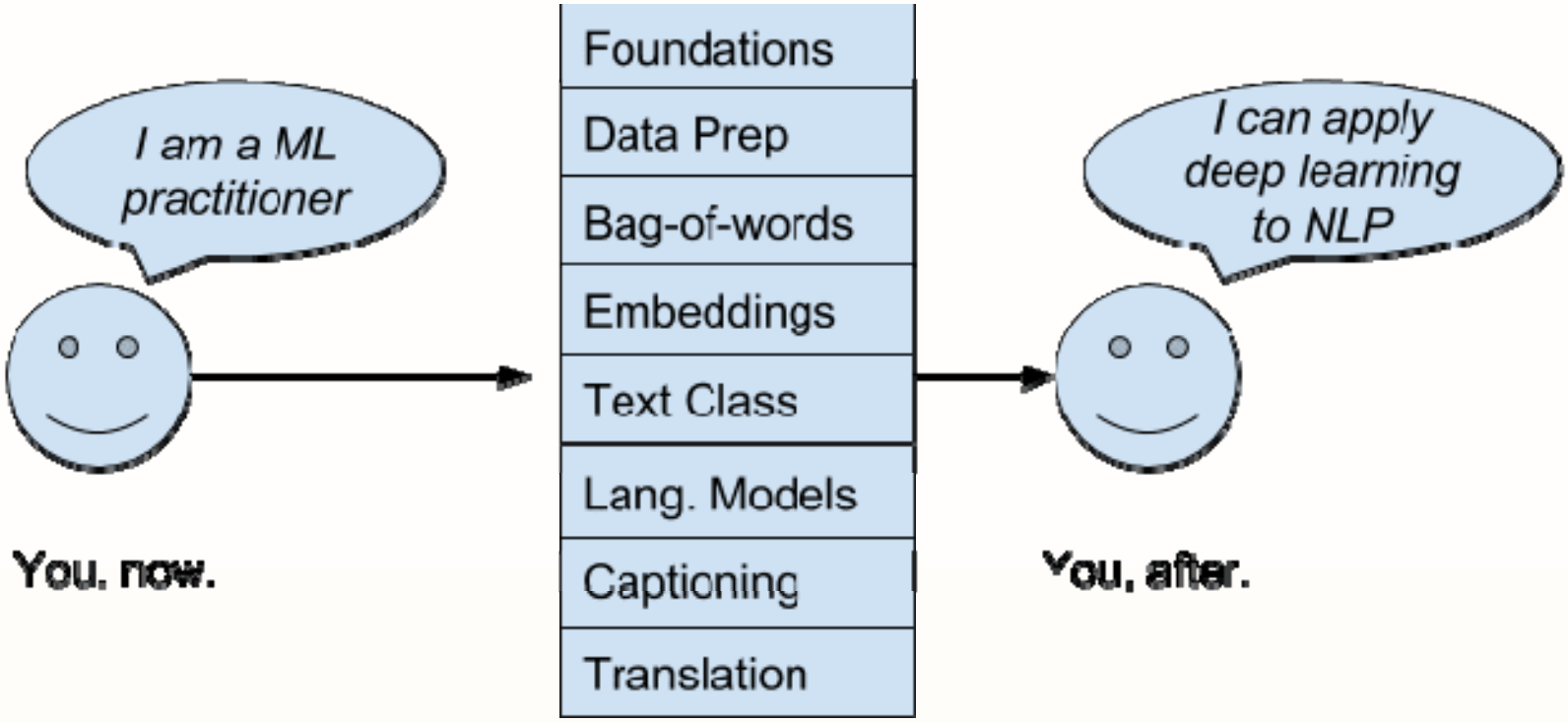
A man is doing a trick on a snowboard

A surfer dives into the ocean

A black and white dog leaps to catch a Frisbee

Evergreen

Bring modern deep learning methods to your natural language processing projects



The lessons assume a few things about you.

You need to know:

- Python and NumPy
- scikit-learn
- Keras for deep learning

You do NOT need to know:

- You **do not** need to be a math wiz!
- You **do not** need to be a deep learning expert!
- You **do not** need to be a master of natural language!

Lessons Outcomes: *you will know*

- What natural language processing is and why it is challenging.
- What deep learning is and how it is different from other machine learning methods, specifically how it is best understood by deep learning experts.
- The promise of deep learning methods for natural language processing problems as defined by experts in the field.
- How to prepare text data for modeling by hand and using best-of-breed Python libraries such as the natural language toolkit or NLTK.
- How to develop and plot distributed representations of text using word embedding models with the Gensim library.
- How to develop a bag-of-words model, a representation technique that can be used for machine learning and deep learning methods.
- How to develop a neural sentiment analysis model for automatically predicting the class label for a text document.
- How to develop a neural language model required for any text generation problem in Keras.
- How to develop a photo captioning system to automatically generate textual descriptions of photographs.
- How to develop a neural machine translation system for translating text from one language to another.

- **Part 1: Foundations**. Discover a gentle introduction to natural language processing, deep learning, and the promise of combining the two, as well as tutorials on how to get started with Keras.
- **Part 2: Data Preparation**: Discover tutorials that show how to clean, prepare and encode text ready for modeling with neural networks.
- **Part 3: Bag-of-Words**. Discover the bag-of-words model, a staple representation for machine learning and a good starting point for neural networks for sentiment analysis.
- **Part 4: Word Embeddings**. Discover a more powerful word representation in word embeddings, how to develop them as standalone models, and how to learn them as part of neural network models.
- **Part 5: Text Classification**. Discover how to leverage word embeddings and convolutional neural networks to learn spatial invariant models of text for sentiment analysis, a successor to the bag-of-words model.
- **Part 6: Image Captioning**. Discover how to combine a pre-trained object recognition model with a language model to automatically caption images.

Develop Practical Skills for NLP That You Can Immediately Apply

You will work through 2 Different NLP Applications

- **Neural Text Classification**. Develop a deep learning model to classify the sentiment of movie reviews as either positive or negative.
- **Neural Photo Captioning**. Develop a model to automatically generate a concise description of ad hoc photographs.

You will work through 3 Different Neural Network Models

- **Neural Bag-of-Words**. Develop neural network models that model text as a bag-of-words where word order is ignored.
- **Neural Word Embedding**. Develop neural network models that model text using a distributed representation.
- **Embedding + CNN**. Develop deep learning models that combine word embedding representations with convolutional neural networks.

Python Technical Details

- **Python Version**: You can use Python 3.
- **SciPy**: You will use NumPy, Pandas and scikit-learn.
- **Keras**: You will need Keras version 2 with either a Theano or TensorFlow backend.
- **Operating System**: You can use Windows, Linux or Mac OS X.
- **Hardware**: A standard modern workstation will do, no GPUs required.
- **Editor**: You can use a text editor and run the example from the command line.

References

- Deep Learning for Natural Language Processing, Jason Brownlee
- Natural Language Processing with Python, Steven Bird, Ewan Klein and Edward Loper
- Taming Text, Grant Ingersoll, Thomas Morton and Drew Farris.
- Text Mining with R, Julia Silge and David Robinson.
- Foundations of Statistical Natural Language Processing, Christopher Manning and Hinrich Schütze.
- Speech and Language Processing, Daniel Jurafsky and James Martin.
- Statistical Machine Translation, Philipp Koehn.
- Statistical Methods for Speech Recognition, Frederick Jelinek.