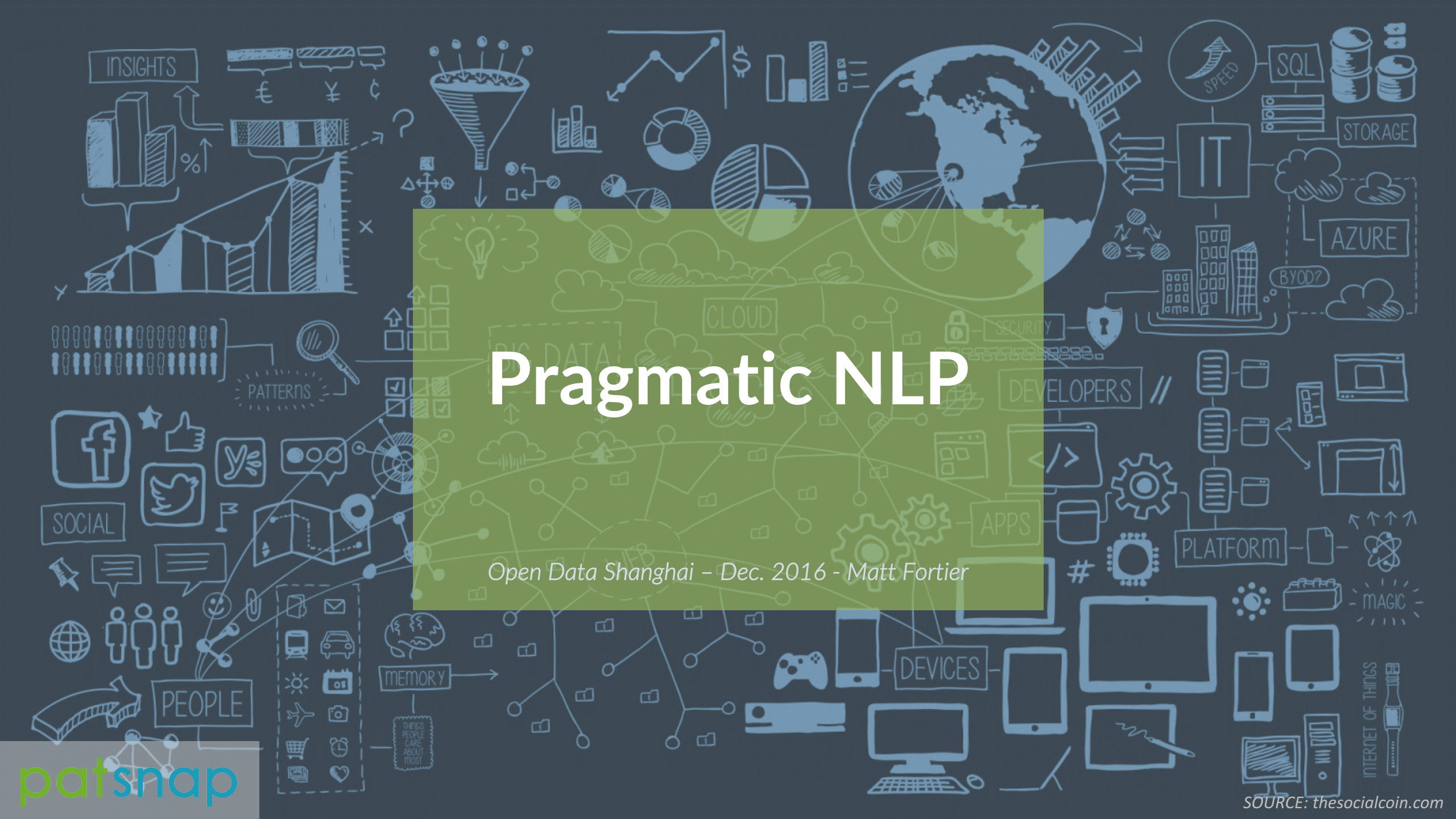


Pragmatic NLP

Open Data Shanghai – Dec. 2016 - Matt Fortier

patsnap

SOURCE: thesocialcoin.com

CONTENTS

NATURAL LANGUAGE PROCESSING?

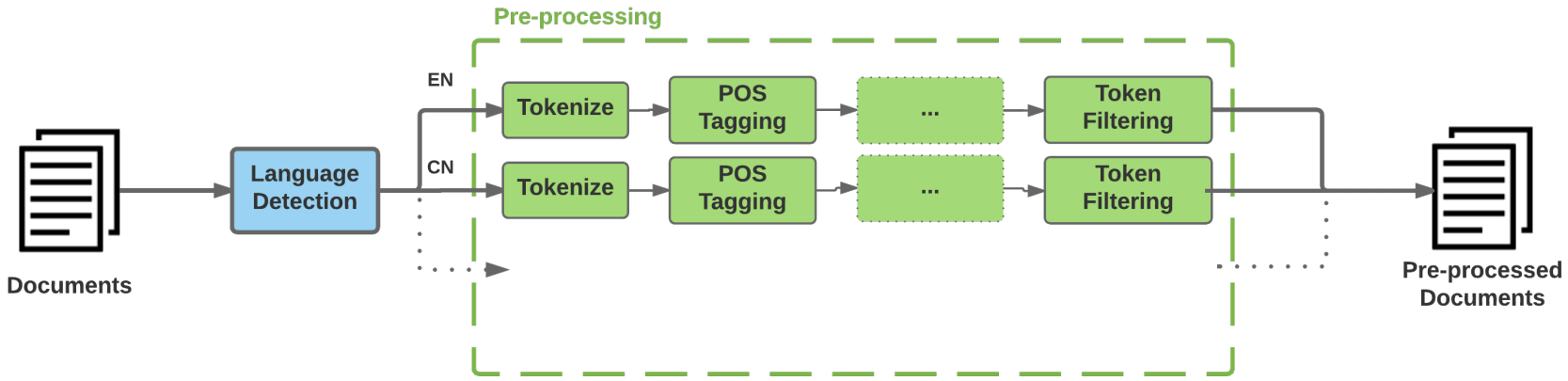

NATURAL LANGUAGE PROCESSING
SO HOT RIGHT NOW
memegenerator.net

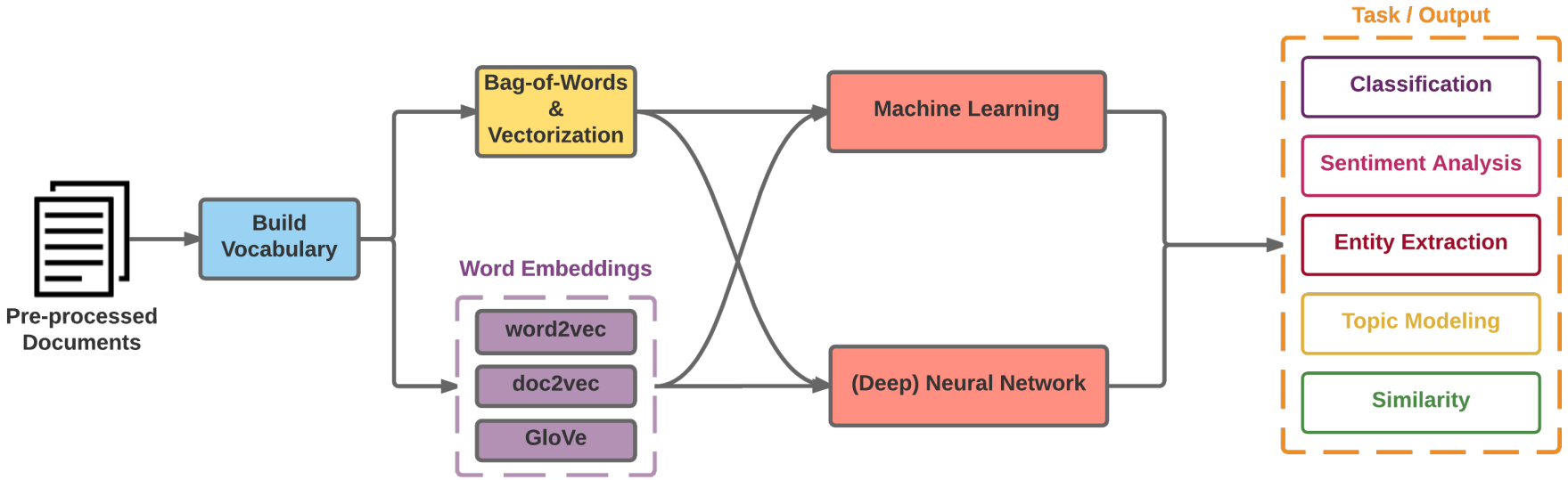
Purpose

The goal of NLP is to extract (meaningful) information out of text documents by applying statistical analysis and machine learning.

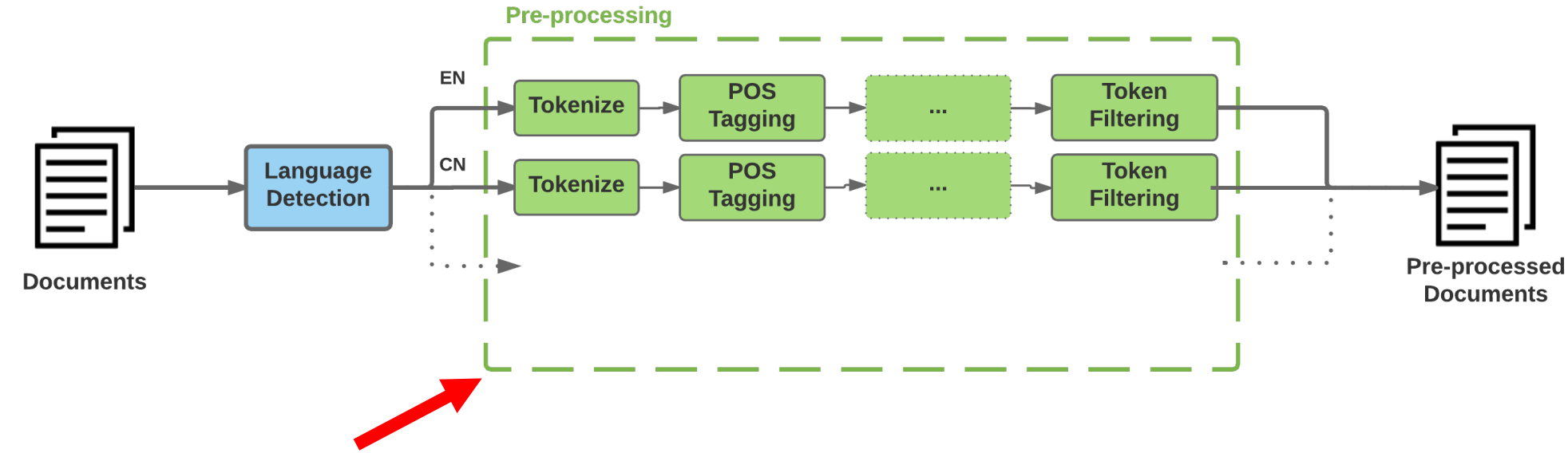
This can allow you to quickly understand a set of thousands or millions of documents without having to ever read them.

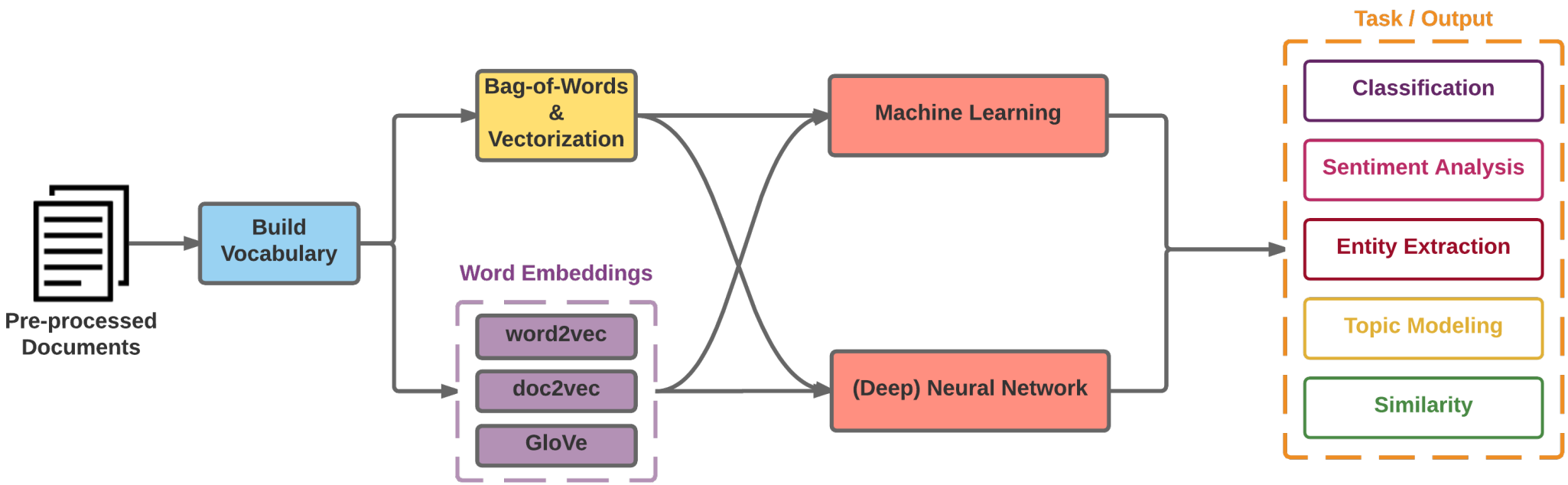
The Modern NLP Pipeline





THE MODERN NLP PIPELINE


Pre-processing
EN
CN
Tokenize
POS Tagging
...
Token Filtering
Tokenize
POS Tagging
...
Token Filtering
Documents
Language Detection
Pre-processed Documents


Task / Output
Classification
Sentiment Analysis
Entity Extraction
Topic Modeling
Similarity
Bag-of-Words & Vectorization
Machine Learning
Word Embeddings
word2vec
doc2vec
GloVe
(Deep) Neural Network
Pre-processed Documents
Build Vocabulary

Preprocessing

Here is were most of the “linguistic” work happens: We try to cleanup and put text into a proper form for modeling stages.

- Tokenization > Split a sentence into words

‘The black cat sleeps’ -> [‘The’, ‘black’, ‘cat’, ‘sleeps’]

- Stem/Lemmatize > Reduce words to their root

‘sleeps’ -> ‘sleep’ ; ‘told’ -> ‘tell’

- Filtering > Reject tokens via set of rules (format/stopwords/etc.)

Cooler Tricks

Part-of-Speech Tagging

'be' -> 'VB'
'held' -> 'VBN'
'accountable' -> 'JJ'

Dependency Parsing

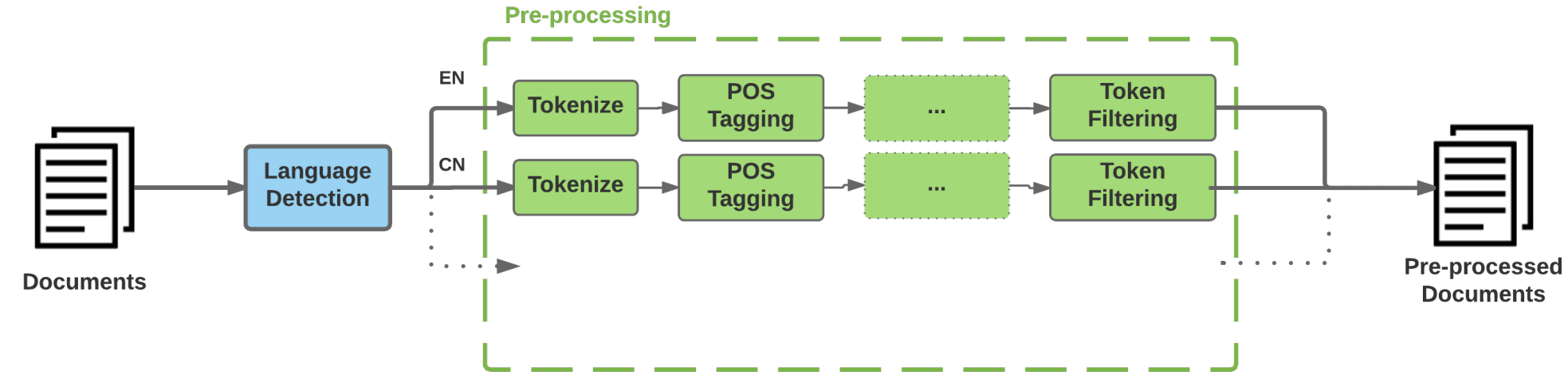
Get back the relationship & syntactic role of each token in the sentence.

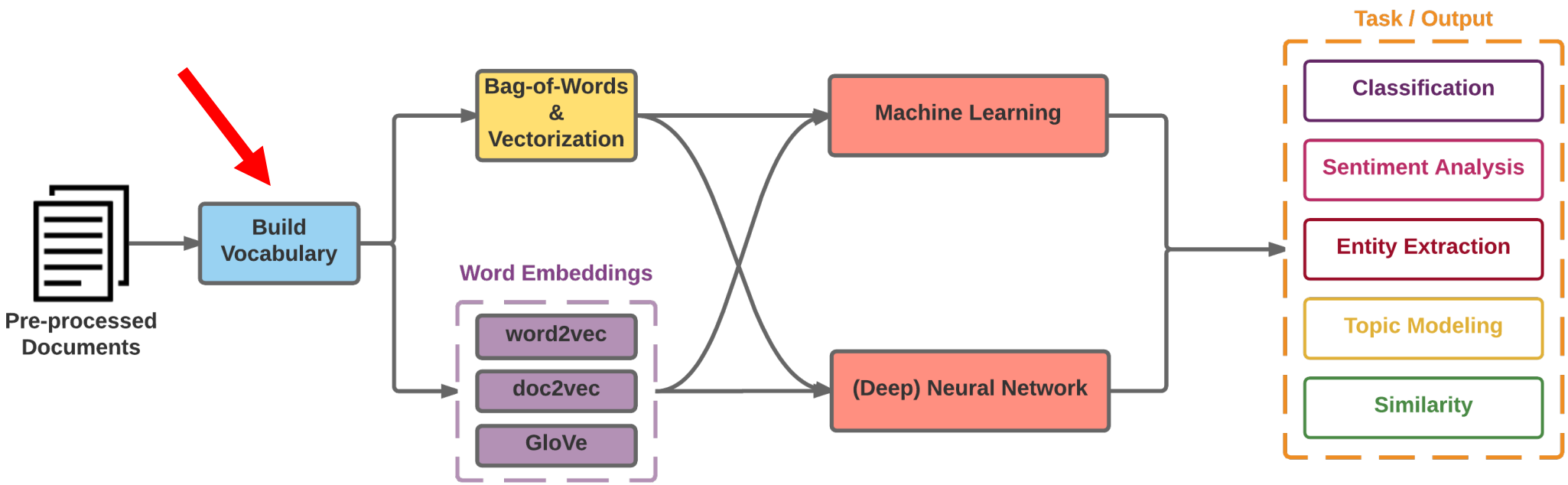
N-grams

'new', 'york' -> 'new york'

Spell Correction

'shangai' -> 'shanghai'

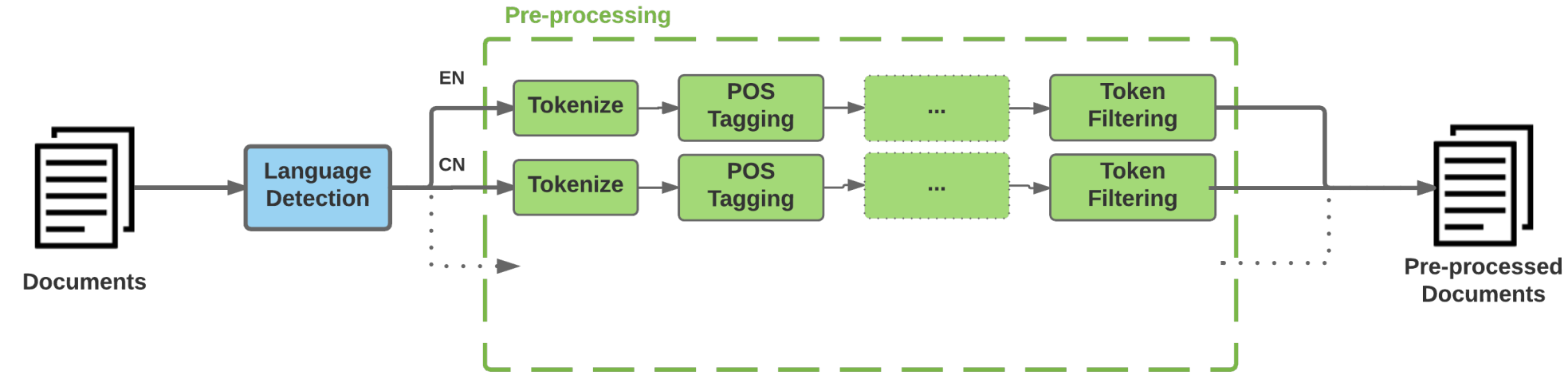

Pre-processing
EN
CN
Tokenize
POS Tagging
...
Token Filtering
Tokenize
POS Tagging
...
Token Filtering
Documents
Language Detection
Pre-processed Documents

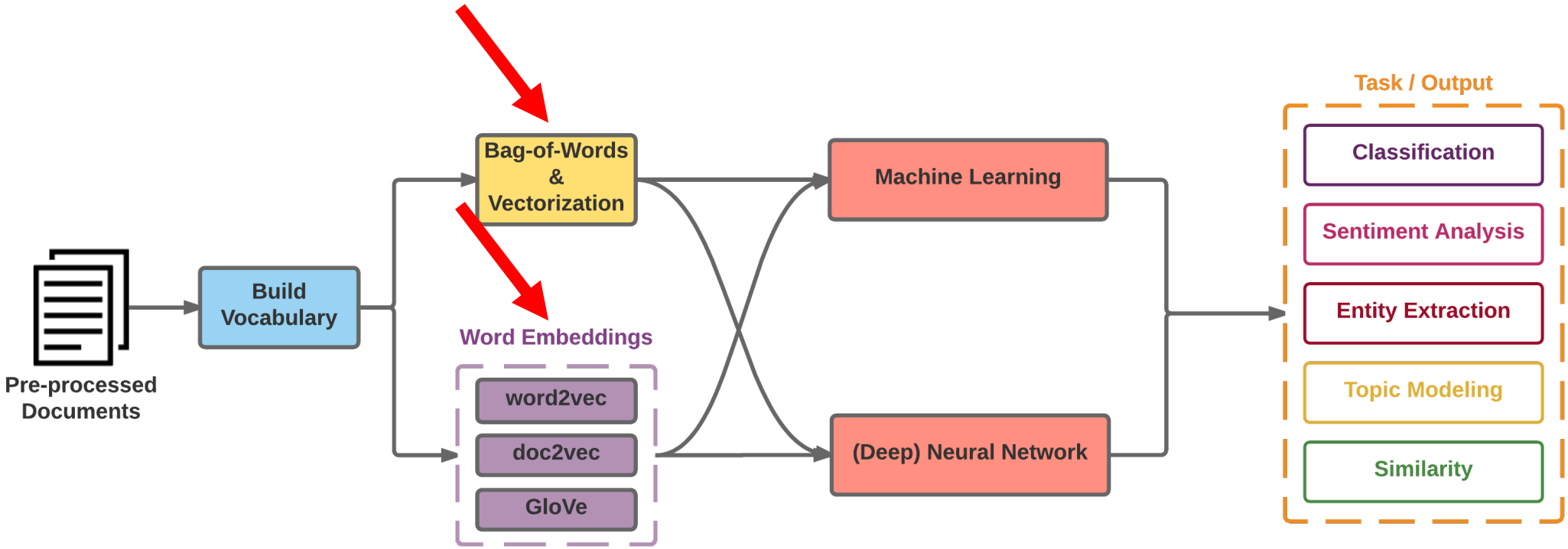

Task / Output
Classification
Sentiment Analysis
Entity Extraction
Topic Modeling
Similarity
Bag-of-Words & Vectorization
Machine Learning
Word Embeddings
word2vec
doc2vec
GloVe
(Deep) Neural Network
Pre-processed Documents
Build Vocabulary

Building Vocabulary

We assign an ID number to each word we encounter in the preprocessed data. We call this mapping “ID -> word” a vocabulary/dictionary.

You can also apply more clever filtering rules to have better control over your vocabulary’s content.


Pre-processing
EN
CN
Tokenize
POS Tagging
...
Token Filtering
Tokenize
POS Tagging
...
Token Filtering
Documents
Language Detection
Pre-processed Documents


Bag-of-Words & Vectorization
Word Embeddings
word2vec
doc2vec
GloVe
Machine Learning
(Deep) Neural Network
Pre-processed Documents
Build Vocabulary
Task / Output
Classification
Sentiment Analysis
Entity Extraction
Topic Modeling
Similarity

Vector Space Representation

Often, the first goal is to represent your words/sentences/documents in a mathematical vector space.

This then allows you to do document similarity, clustering, classification, etc.

Bag-of-Words

We transform a document into an array of (word id, frequency in document), which we commonly call a bag-of-words. Context/Order is lost.

‘The black cat sleeps’ -> [(13,1), (100,1), (2642,1), (543,1)]

TF-IDF

Statistical transformations like TF-IDF helps you assign an “importance score” to your tokens.

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

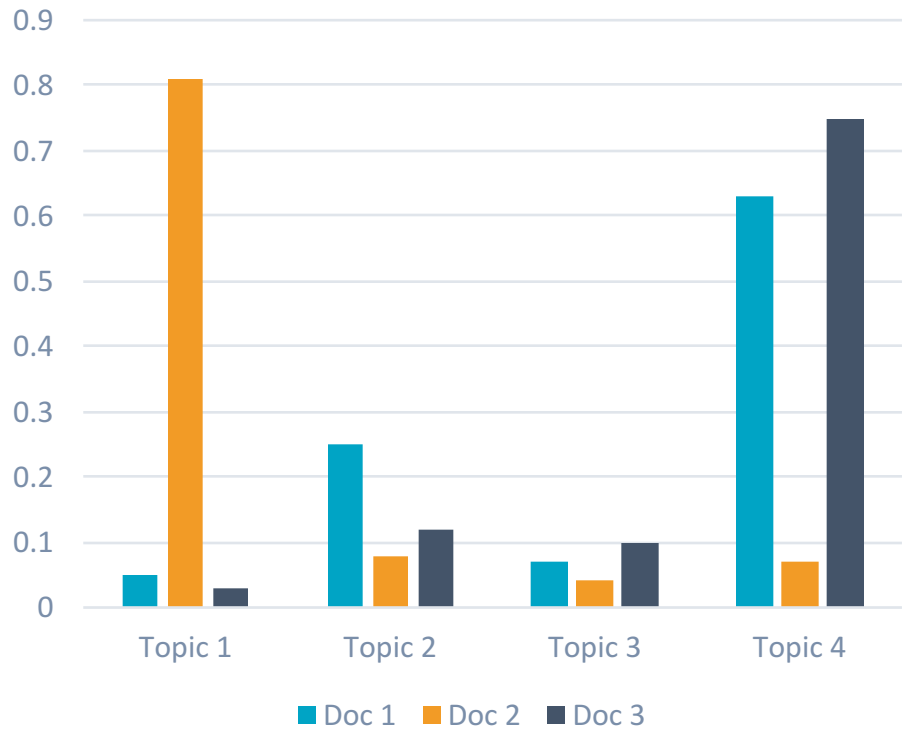
df_x = number of documents containing x

N = total number of documents

LDA

Latent Dirichlet Allocation is a probabilistic technique that models documents as a mixture of topics, and in turn topics as a mixture of words.

It's a very good exploratory tool for textual data.



Topic 1	Topic 2	Topic 3	Topic 4
computer 0.04	gene 0.09	math 0.07	carbon 0.03
cpu 0.03	data 0.06	matrix 0.05	water 0.03
data 0.02	bio 0.03	vector 0.04	atom 0.02

word2vec (2013)

Dense Vector Representation

+

Context-aware

=

Game-changer for NLP

Efficient Estimation of Word Representations in Vector Space

Tomas Mikolov
Google Inc., Mountain View, CA
tmikolov@google.com

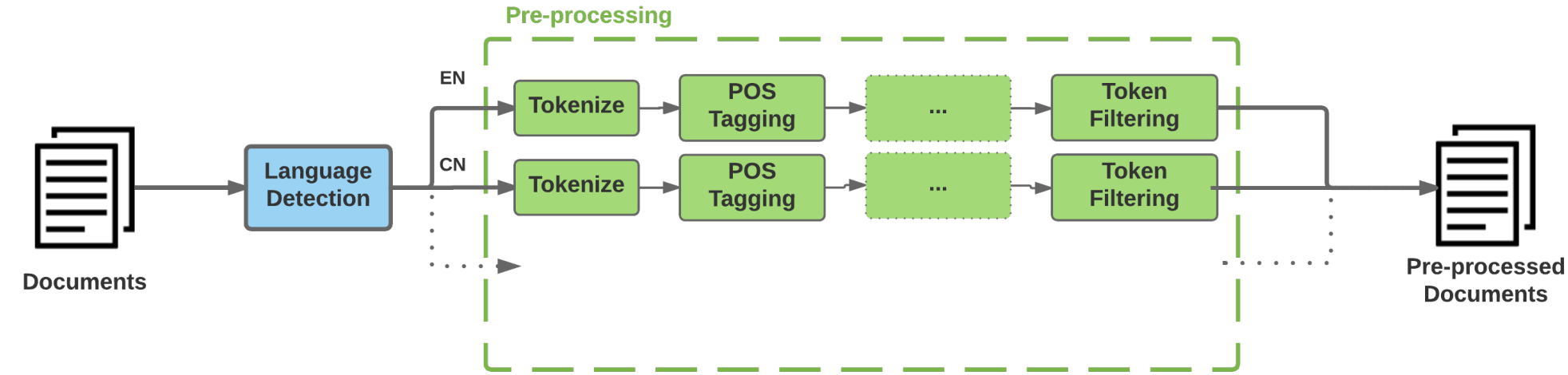
Kai Chen
Google Inc., Mountain View, CA
kaichen@google.com

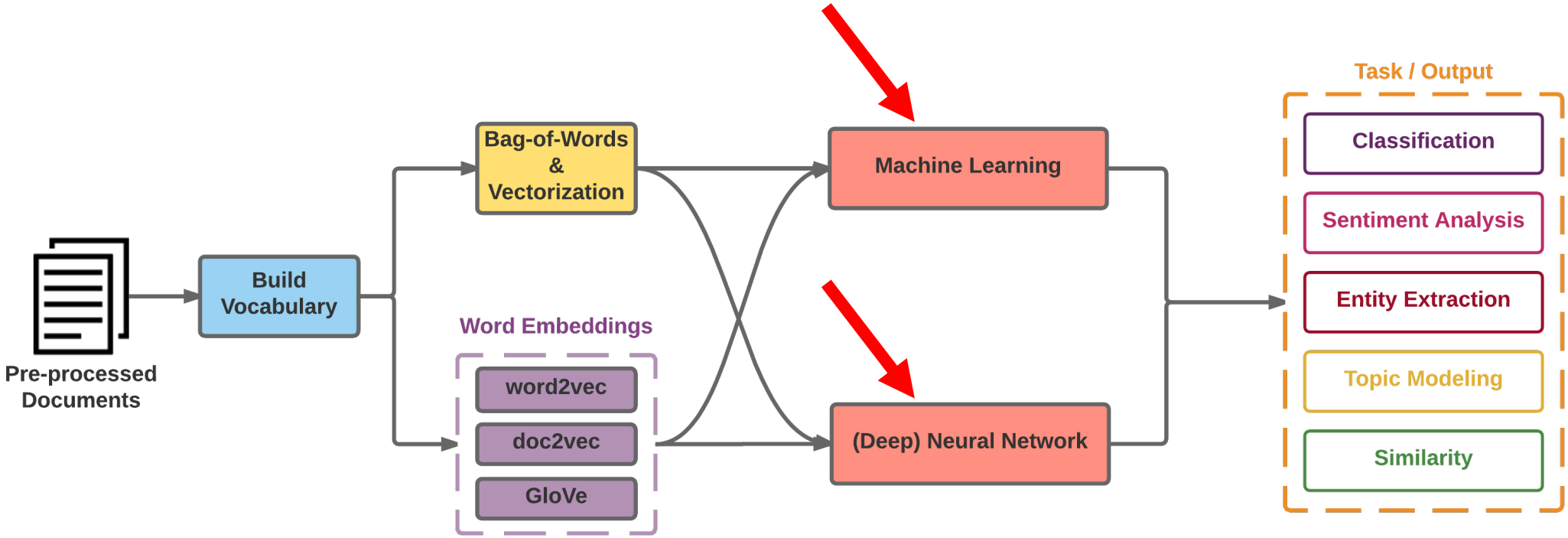
Greg Corrado
Google Inc., Mountain View, CA
gcorrado@google.com

Jeffrey Dean
Google Inc., Mountain View, CA
jeff@google.com

Abstract

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.


Pre-processing
EN
CN
Tokenize
POS Tagging
...
Token Filtering
Tokenize
POS Tagging
...
Token Filtering
Language Detection
Documents
Pre-processed Documents


Task / Output
Classification
Sentiment Analysis
Entity Extraction
Topic Modeling
Similarity
Machine Learning
(Deep) Neural Network
Bag-of-Words & Vectorization
Word Embeddings
word2vec
doc2vec
GloVe
Build Vocabulary
Pre-processed Documents

Congratulations, you now have represented your textual data as meaningful, weighted vectors / matrices.

Now get creative!

Classification | Labeling

Naïve Bayes / SVM / etc

Clustering

K-means / Hierarchical Dirichlet Process / etc.

Text Summarization

TextRank

Similarity

Cosine Distance / Hellinger Distance / etc.

PYTHON
NLP TOOLBOX

Preprocessing

TextBlob

http://textblob.readthedocs.io/en/latest/

https://github.com/sloria/textblob

NLTK

http://www.nltk.org/

jieba

https://github.com/fxsjy/jieba

SnowNLP

https://github.com/isnowfy/snownlp

Modeling

gensim

Advanced/robust/efficient NLP library.

This is a very potent solution for your NLP production needs.

https://radimrehurek.com/gensim/

scikit-learn

Most of the algorithms you can think about, they have…

Visualization

wordcloud

https://github.com/amueller/word_cloud

pyLDAvis

https://github.com/bmabey/pyLDAvis

DEMO TIME !

SOURCE: thesocialcoin.com

WRAP UP

Get to know your data / Explore.

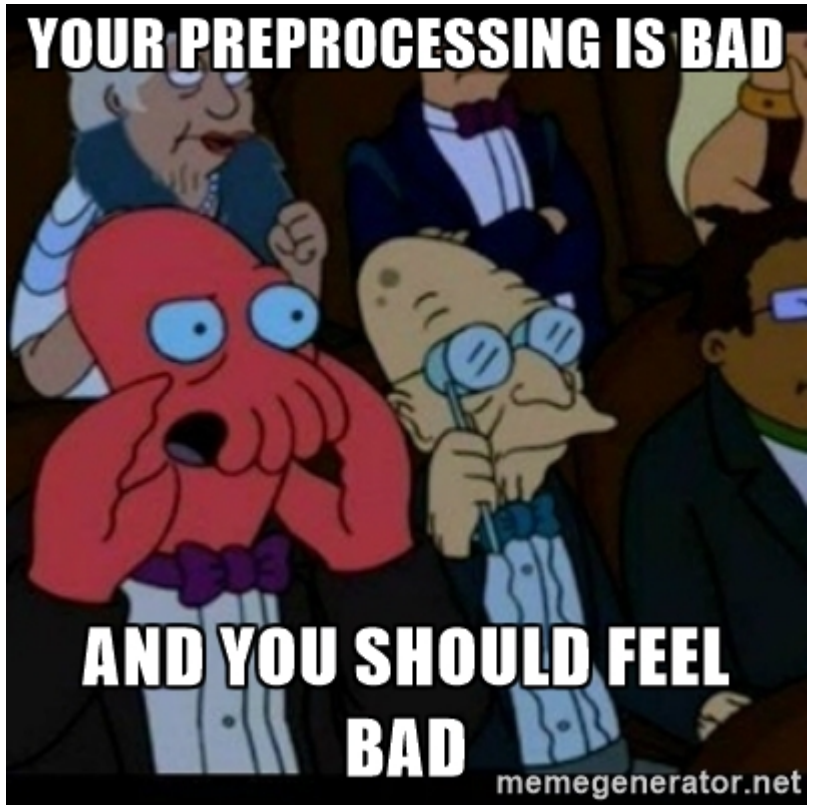
Come up with a good preprocessing strategy.

Start simple, build up in complexity.

Get creative!

Remember…

If all you are doing to prepare your data is str.split(‘ ‘) then…



What Next?

Expand You Knowledge - Good Resources

https://github.com/gutfeeling/beginner_nlp

www.reddit.com/r/LanguageTechnology/

http://nlp.stanford.edu/

Kaggle!

Experiment, Experiment, Experiment

Try different kind of text data, different formats, etc.

patsnap

SEQUOIA

SUMMIT PARTNERS

SHUNWEI

Applying Big Data Analytics / Data Science / Machine Learning to Innovation & IP.

Largest IP dataset in the World*:

- 120M+ Patents
- Grants
- Litigation
- Chemical/Biochemical

Technology Fast 500
Asia Pacific 2016 Ranking

500 Technology Fast 500 2016 APAC

Ranked 44th in Deloitte's Technology Fast 500 Asia Pacific 2016 Ranking.

Thanks !

谢谢！

patsnap

SOURCE: thesocialcoin.com