人工智慧文本分析
(Artificial Intelligence for Text Analytics)
文本分析的基礎：自然語言處理
(Foundations of Text Analytics: Natural Language Processing; NLP)

1082AITA02
MBA, IMTKU (M2455) (8410) (Spring 2020)
Wed 8, 9 (15:10-17:00) (B605)

Min-Yuh Day
戴敏育
Associate Professor
副教授
Dept. of Information Management, Tamkang University

http://mail.tku.edu.tw/myday/
2020-03-11
週次 (Week)  日期 (Date)  內容 (Subject/Topics)
1  2020/03/04  人工智慧文本分析課程介紹
(Course Orientation on Artificial Intelligence for Text Analytics)
2  2020/03/11  文本分析的基礎：自然語言處理
(Foundations of Text Analytics: Natural Language Processing; NLP)
3  2020/03/18  Python自然語言處理
(Python for Natural Language Processing)
4  2020/03/25  處理和理解文本
(Processing and Understanding Text)
5  2020/04/01  文本表達特徵工程
(Feature Engineering for Text Representation)
6  2020/04/08  人工智慧文本分析個案研究 I
(Case Study on Artificial Intelligence for Text Analytics I)
課程大綱 (Syllabus)

週次 (Week)   日期 (Date)   內容 (Subject/Topics)
7   2020/04/15 文本分類 (Text Classification)
8   2020/04/22 文本摘要和主題模型
    (Text Summarization and Topic Models)
9   2020/04/29 期中報告 (Midterm Project Report)
10 2020/05/06 文本相似度和分群 (Text Similarity and Clustering)
11 2020/05/13 語意分析和命名實體識別
    (Semantic Analysis and Named Entity Recognition; NER)
12 2020/05/20 情感分析 (Sentiment Analysis)
<table>
<thead>
<tr>
<th>週次 (Week)</th>
<th>日期 (Date)</th>
<th>內容 (Subject/Topics)</th>
</tr>
</thead>
</table>
| 13         | 2020/05/27 | 人工智慧文本分析個案研究 II  
(Case Study on Artificial Intelligence for Text Analytics II) |
| 14         | 2020/06/03 | 深度學習和通用句子嵌入模型  
(Deep Learning and Universal Sentence-Embedding Models) |
| 15         | 2020/06/10 | 問答系統與對話系統  
(Question Answering and Dialogue Systems) |
| 16         | 2020/06/17 | 期末報告 I (Final Project Presentation I) |
| 17         | 2020/06/24 | 期末報告 II (Final Project Presentation II) |
| 18         | 2020/07/01 | 教師彈性補充教學 |
Outline

• Text Analytics

• Natural Language Processing (NLP)
Text Analytics (TA)
Natural Language Processing (NLP)
Artificial Intelligence (AI)
Evolution of Computerized Decision Support to Analytics/Data Science

The timeline in Figure 1.8 shows the terminology used to describe analytics since the 1970s. During the 1970s, the primary focus of information systems support for decision making focused on providing structured, periodic reports that a manager could use for decision making (or ignore them). Businesses began to create routine reports to inform decision makers (managers) about what had happened in the previous period (e.g., day, week, month, quarter). Although it was useful to know what had happened in the past, managers needed more than this: They needed a variety of reports at different levels of granularity to better understand and address changing needs and challenges of the business. These were usually called management information systems (MIS).

In the early 1970s, Scott-Morton first articulated the major concepts of DSS. He defined DSSs as “interactive computer-based systems, which help decision makers utilize data and models to solve unstructured problems” (Gorry and Scott-Morton, 1971). The following is another classic DSS definition, provided by Keen and Scott-Morton (1978):

Decision support systems couple the intellectual resources of individuals with the capabilities of the computer to improve the quality of decisions. It is a computer-based support system for management decision makers who deal with semistructured problems.

Note that the term decision support system, like management information system and several other terms in the field of IT, is a content-free expression (i.e., it means different things to different people). Therefore, there is no universally accepted definition of DSS.

During the early days of analytics, data was often obtained from the domain experts using manual processes (i.e., interviews and surveys) to build mathematical or knowledge-based models to solve constrained optimization problems. The idea was to do the best with limited resources. Such decision support models were typically called operations research (OR). The problems that were too complex to solve optimally (using linear or nonlinear mathematical programming techniques) were tackled using heuristic methods such as simulation models. (We will introduce these as prescriptive analytics later in this chapter and in a bit more detail in Chapter 6.)

In the late 1970s and early 1980s, in addition to the mature OR models that were being used in many industries and government systems, a new and exciting line of models had emerged: rule-based expert systems. These systems promised to capture experts’ knowledge in a format that computers could process (via a collection of if–then–else rules or heuristics) so that these could be used for consultation much the same way that one...

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
Three Types of Analytics

Business Analytics

Descriptive

- What happened?
- What is happening?
- ✓ Business reporting
  ✓ Dashboards
  ✓ Scorecards
  ✓ Data warehousing

Outcomes

- Well-defined business problems and opportunities

Predictive

- What will happen?
- Why will it happen?
- ✓ Data mining
  ✓ Text mining
  ✓ Web/media mining
  ✓ Forecasting

Outcomes

- Accurate projections of future events and outcomes

Prescriptive

- What should I do?
- Why should I do it?
- ✓ Optimization
  ✓ Simulation
  ✓ Decision modeling
  ✓ Expert systems

Outcomes

- Best possible business decisions and actions
Text Analytics and Text Mining

Text Mining “Knowledge Discovery in Textual Data”

- Document Matching
- Link Analysis
- Information Retrieval
- Search Engines
- POS Tagging
- Lemmatization
- Word Disambiguation

Web Mining
- Web Content Mining
- Web Structure Mining
- Web Usage Mining

Data Mining
- Classification
- Clustering
- Association

Statistics
Machine Learning
Management Science
Artificial Intelligence
Computer Science
Other Disciplines

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
AI
Definition of Artificial Intelligence (A.I.)
Artificial Intelligence

“... the science and engineering of making intelligent machines”

(John McCarthy, 1955)
Artificial Intelligence

“... technology that thinks and acts like humans”
Artificial Intelligence

“... intelligence exhibited by machines or software”
## 4 Approaches of AI

<table>
<thead>
<tr>
<th>Thinking Humanly</th>
<th>Thinking Rationally</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acting Humanly</td>
<td>Acting Rationally</td>
</tr>
</tbody>
</table>

4 Approaches of AI

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Thinking Rationally: The “Laws of Thought” Approach</td>
<td></td>
</tr>
<tr>
<td>4. Acting Rationally: The Rational Agent Approach</td>
<td></td>
</tr>
</tbody>
</table>

AI Acting Humanly: The Turing Test Approach
(Alan Turing, 1950)

- Natural Language Processing (NLP)
- Knowledge Representation
- Automated Reasoning
- Machine Learning (ML)
- Computer Vision
- Robotics

Can a robot pass a university entrance exam?
Noriko Arai at TED2017

Noriko Arai at TED2017

Can a robot pass a university entrance exam?

https://www.ted.com/talks/noriko_arai_can_a_robot_pass_a_university_entrance_exam
https://www.youtube.com/watch?v=XQZjkPyJ8KU
Artificial Intelligence (A.I.) Timeline

1950
TURING TEST
Computer scientist Alan Turing proposes a test for machine intelligence. If a machine can trick humans into thinking it is human, then it has intelligence.

1955
A.I. BORN
Term ‘artificial intelligence’ is coined by computer scientist John McCarthy to describe "the science and engineering of making intelligent machines.”

1961
UNIMATE
First industrial robot, Unimate, goes to work at GM replacing humans on the assembly line.

1964
ELIZA
Pioneering chatbot developed by Joseph Weizenbaum at MIT. ELIZA holds conversations with humans.

1966
SHAKEY
The ‘first electronic person’ from Stanford, Shakey is a general-purpose mobile robot that reasons about its own actions.

A.I. WINTER
Many false starts and dead-ends leave A.I. out in the cold.

1997
DEEP BLUE
Deep Blue, a chess-playing computer from IBM defeats world chess champion Garry Kasparov.

1998
KISMET
Cynthia Breazeal at MIT introduces Kismet, an emotionally intelligent robot insofar as it detects and responds to people’s feelings.

1999
AIBO
Sony launches first consumer robot pet dog AIBO (A.I. robot) with skills and personality that develop over time.

2002
ROOMBA
First mass-produced autonomous robotic vacuum cleaner from iRobot learns to navigate and clean homes.

2011
SIRI
Apple integrates Siri, an intelligent virtual assistant with a voice interface, into the iPhone 4S.

2011
WATSON
IBM’s question answering computer Watson wins first place on popular $1M prize television quiz show Jeopardy.

2014
EUGENE
Eugene Goostman, a chatbot passes the Turing Test with a third of judges believing Eugene is human.

2014
ALEXA
Amazon launches Alexa, an intelligent virtual assistant with a voice interface that completes shopping tasks.

2016
TAY
Microsoft’s chatbot Tay goes rogue on social media making inflammatory and offensive racist comments.

2017
ALPHAGO
Google’s A.I. AlphaGo beats world champion Ke Jie in the complex board game of Go, notable for its vast number (2170) of possible positions.

Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

AI, ML, DL

Artificial Intelligence (AI)

Machine Learning (ML)

- Supervised Learning
- Unsupervised Learning
- Semi-supervised Learning
- Reinforcement Learning

Deep Learning (DL)

- CNN
- RNN
- LSTM
- GRU
- GAN
Text Analytics and Text Mining
Text Analytics

• **Text Analytics** = Information Retrieval + Information Extraction + Data Mining + Web Mining

• **Text Analytics** = Information Retrieval + Text Mining

Emotions

- Love
- Joy
- Surprise
- Anger
- Sadness
- Fear

Example of Opinion: review segment on iPhone

“I bought an iPhone a few days ago. It was such a nice phone. The touch screen was really cool. The voice quality was clear too. However, my mother was mad with me as I did not tell her before I bought it. She also thought the phone was too expensive, and wanted me to return it to the shop. ...”

“(1) I bought an iPhone a few days ago.
(2) It was such a nice phone.
(3) The touch screen was really cool.
(4) The voice quality was clear too.
(5) However, my mother was mad with me as I did not tell her before I bought it.
(6) She also thought the phone was too expensive, and wanted me to return it to the shop. ...”

A Multistep Process to Sentiment Analysis

1. **Step 1**
   - Calculate the O–S Polarity
   - **Lexicon**

2. **Step 2**
   - Calculate the N–P Polarity of the sentiment
   - **Lexicon**

3. **Step 3**
   - Identify the target for the sentiment

4. **Step 4**
   - Record the Polarity, Strength, and the Target of the sentiment
   - **Target**
   - Tabulate & aggregate the sentiment analysis results

Sentiment Analysis

**Tasks**

- Subjectivity Classification
- Sentiment Classification
- Review Usefulness Measurement
- Opinion Spam Detection
- Lexicon Creation
- Aspect Extraction
- Polarity Determination
- Vagueness resolution in opinionated text
- Multi- & Cross-Lingual SC
- Cross-domain SC

**Approaches**

- Machine Learning based
- Lexicon based
- Hybrid approaches
- Ontology based
- Non-Ontology based

Sentiment Classification Techniques

Sentiment Analysis

- Machine Learning Approach
  - Supervised Learning
    - Decision Tree Classifiers
    - Linear Classifiers
    - Rule-based Classifiers
    - Probabilistic Classifiers
  - Unsupervised Learning
    - Support Vector Machine (SVM)
    - Deep Learning (DL)
    - Naïve Bayes (NB)
    - Bayesian Network (BN)
    - Maximum Entropy (ME)

- Lexicon-based Approach
  - Dictionary-based Approach
    - Statistical
  - Corpus-based Approach
    - Semantic

Natural Language Processing (NLP)
Natural Language Processing (NLP)

- Natural language processing (NLP) is an important component of text mining and is a subfield of artificial intelligence and computational linguistics.

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
Natural Language Processing (NLP) and Text Mining

- Raw text
- Sentence Segmentation
- Tokenization
- Part-of-Speech (POS)
- Stop word removal
- Stemming / Lemmatization
- String Metrics & Matching

word’s stem: am → am
word’s lemma: am → be
having → hav
having → have

Source: Nitin Hardeniya (2015), NLTK Essentials, Packt Publishing; Florian Leitner (2015), Text mining - from Bayes rule to dependency parsing
Text Summarization

Text Input
- Pre-processing
- Text Structure Analysis
- Word Segmentation
- POS Tagging
- Occurrence Statistic
- Keyword Extraction
- Weigh Words & Sentences
- Sentences Selection
- Rough Summary Generation
- Smoothing
- Summary Output

Dictionary / Thesaurus

Topic Modeling

Seeking Life's Bare (Genetic) Necessities

Cold Spring Harbor, New York—How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough. Although the numbers don't match precisely, those predictions are not all that far apart, especially in comparison to the 75,000 genes in the human genome, notes Svante Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an *Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.*

SCIENCE • VOL. 272 • 24 MAY 1996

Natural Language Processing (NLP)

- Part-of-speech tagging
- Text segmentation
- Word sense disambiguation
- Syntactic ambiguity
- Imperfect or irregular input
- Speech acts

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
NLP Tasks

- Question answering
- Automatic summarization
- Natural language generation
- Natural language understanding
- Machine translation
- Foreign language reading
- Foreign language writing.
- Speech recognition
- Text-to-speech
- Text proofing
- Optical character recognition

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
Modern NLP Pipeline

Pre-processing

Documents

Language Detection

EN

Tokenize

POS Tagging

... Token Filtering

CN

Tokenize

POS Tagging

... Token Filtering

Pre-processed Documents

Pre-processed Documents

Build Vocabulary

Bag-of-Words & Vectorization

Machine Learning

Task / Output

Classification

Sentiment Analysis

Entity Extraction

Topic Modeling

Similarity

Word Embeddings

word2vec

doc2vec

GloVe

(Deep) Neural Network

Modern NLP Pipeline

Source: http://mattfortier.me/2017/01/31/nlp-intro-pt-1-overview/
Deep Learning NLP

- Documents
  - Preprocessing
  - Dense Word Embeddings
  - Deep Neural Network

Task / Output
- Classification
- Sentiment Analysis
- Entity Extraction
- Topic Modeling
- Document Similarity

Source: http://mattfortier.me/2017/01/31/nlp-intro-pt-1-overview/
BERT:
Pre-training of Deep Bidirectional Transformers for Language Understanding

**BERT**

Bidirectional Encoder Representations from Transformers

**Pre-training model architectures**

**BERT** uses a bidirectional Transformer.
**OpenAI GPT** uses a left-to-right Transformer.
**ELMo** uses the concatenation of independently trained left-to-right and right-to-left LSTM to generate features for downstream tasks.
Among three, only BERT representations are jointly conditioned on both left and right context in all layers.

The input embeddings is the sum of the token embeddings, the segmentation embeddings and the position embeddings.
BERT Sequence-level tasks

(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks:
SST-2, CoLA

BERT Token-level tasks

(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

General Language Understanding Evaluation (GLUE) benchmark

GLUE Test results

<table>
<thead>
<tr>
<th>System</th>
<th>MNLI-(m/mm)</th>
<th>QQP</th>
<th>QNLI</th>
<th>SST-2</th>
<th>CoLA</th>
<th>STS-B</th>
<th>MRPC</th>
<th>RTE</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-OpenAI SOTA</td>
<td>80.6/80.1</td>
<td>66.1</td>
<td>82.3</td>
<td>93.2</td>
<td>35.0</td>
<td>81.0</td>
<td>86.0</td>
<td>61.7</td>
<td>74.0</td>
</tr>
<tr>
<td>BiLSTM+ELMo+Attn</td>
<td>76.4/76.1</td>
<td>64.8</td>
<td>79.9</td>
<td>90.4</td>
<td>36.0</td>
<td>73.3</td>
<td>84.9</td>
<td>56.8</td>
<td>71.0</td>
</tr>
<tr>
<td>OpenAI GPT</td>
<td>82.1/81.4</td>
<td>70.3</td>
<td>88.1</td>
<td>91.3</td>
<td>45.4</td>
<td>80.0</td>
<td>82.3</td>
<td>56.0</td>
<td>75.2</td>
</tr>
<tr>
<td>BERTBASE</td>
<td>84.6/83.4</td>
<td>71.2</td>
<td>90.1</td>
<td>93.5</td>
<td>52.1</td>
<td>85.8</td>
<td>88.9</td>
<td>66.4</td>
<td>79.6</td>
</tr>
<tr>
<td>BERTLARGE</td>
<td><strong>86.7/85.9</strong></td>
<td>72.1</td>
<td>91.1</td>
<td><strong>94.9</strong></td>
<td><strong>60.5</strong></td>
<td><strong>86.5</strong></td>
<td><strong>89.3</strong></td>
<td><strong>70.1</strong></td>
<td><strong>81.9</strong></td>
</tr>
</tbody>
</table>

**MNLI**: Multi-Genre Natural Language Inference  
**QQP**: Quora Question Pairs  
**QNLI**: Question Natural Language Inference  
**SST-2**: The Stanford Sentiment Treebank  
**CoLA**: The Corpus of Linguistic Acceptability  
**STS-B**: The Semantic Textual Similarity Benchmark  
**MRPC**: Microsoft Research Paraphrase Corpus  
**RTE**: Recognizing Textual Entailment

NLP Libraries and Tools
Natural Language Processing with Python
– Analyzing Text with the Natural Language Toolkit

Steven Bird, Ewan Klein, and Edward Loper

This version of the NLTK book is updated for Python 3 and NLTK 3. The first edition of the book, published by O'Reilly, is available at http://nltk.org/book_1ed/. (There are currently no plans for a second edition of the book.)

0. Preface
1. Language Processing and Python
2. Accessing Text Corpora and Lexical Resources
3. Processing Raw Text
4. Writing Structured Programs
5. Categorizing and Tagging Words (minor fixes still required)
6. Learning to Classify Text
7. Extracting Information from Text
8. Analyzing Sentence Structure
9. Building Feature Based Grammars
10. Analyzing the Meaning of Sentences (minor fixes still required)
11. Managing Linguistic Data (minor fixes still required)
12. Afterword: Facing the Language Challenge

Bibliography
Term Index

This book is made available under the terms of the Creative Commons Attribution Noncommercial No-Derivative-Works 3.0 US License. Please post any questions about the materials to the nltk-users mailing list. Please report any errors on the issue tracker.

http://www.nltk.org/book/
spaCy

Industrial-Strength Natural Language Processing in Python

Fastest in the world
spaCy excels at large-scale information extraction tasks. It’s written from the ground up in carefully memory-managed Cython. Independent research has confirmed that spaCy is the fastest in the world. If your application needs to process entire web dumps, spaCy is the library you want to be using.

Get things done
spaCy is designed to help you do real work — to build real products, or gather real insights. The library respects your time, and tries to avoid wasting it. It's easy to install, and its API is simple and productive. I like to think of spaCy as the Ruby on Rails of Natural Language Processing.

Deep learning
spaCy is the best way to prepare text for deep learning. It interoperates seamlessly with TensorFlow, Keras, Scikit-Learn, Gensim and the rest of Python's awesome AI ecosystem. spaCy helps you connect the statistical models trained by these libraries to the rest of your application.

https://spacy.io/
Gensim is a FREE Python library

- Scalable statistical semantics
- Analyze plain-text documents for semantic structure
- Retrieve semantically similar documents

https://radimrehurek.com/gensim/
TextBlob

TextBlob: Simplified Text Processing

Release v0.12.0. [Changelog](https://textblob.readthedocs.io)

TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

```python
from textblob import TextBlob

text = '''
The titular threat of The Blob has always struck me as the ultimate movie monster: an insatiably hungry, amoeba-like mass able to penetrate virtually any safeguard, capable of—as a doomed doctor chillingly describes it—"assimilating flesh on contact. Snide comparisons to gelatin be damned, it's a concept with the most devastating of potential consequences, not unlike the grey goo scenario proposed by technological theorists fearful of artificial intelligence run rampant.
'''

blob = TextBlob(text)

blob.tags  # [('The', 'DT'), ('titular', 'JJ'),
             # ('threat', 'NN'), ('of', 'IN'), ...]

blob.noun_phrases  # WordList(['titular threat', 'blob',
             #   'ultimate movie monster',
             #   'amoeba-like mass', ...])

for sentence in blob.sentences:
    polarity = sentence.sentiment.polarity
    print(sentence, polarity)
```

https://textblob.readthedocs.io
Welcome to polyglot's documentation!

polyglot

Polyglot is a natural language pipeline that supports massive multilingual applications.

- Free software: GPLv3 license

Features

- Tokenization (165 Languages)
- Language detection (196 Languages)
- Named Entity Recognition (40 Languages)
- Part of Speech Tagging (16 Languages)
- Sentiment Analysis (136 Languages)
- Word Embeddings (137 Languages)
- Morphological analysis (135 Languages)
- Transliteration (69 Languages)

https://polyglot.readthedocs.io/
scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

- Identifying to which category an object belongs to.
- **Applications**: Spam detection, Image recognition.
- **Algorithms**: SVM, nearest neighbors, random forest, ...
- — Examples

Regression

- Predicting a continuous-valued attribute associated with an object.
- **Applications**: Drug response, Stock prices.
- **Algorithms**: SVR, ridge regression, Lasso, ...
- — Examples

Clustering

- Automatic grouping of similar objects into sets.
- **Applications**: Customer segmentation, Grouping experiment outcomes
- **Algorithms**: k-Means, spectral clustering, mean-shift, ...
- — Examples

Dimensionality reduction

- Reducing the number of random variables to consider.
- **Applications**: Visualization, Increased efficiency

Model selection

- Comparing, validating and choosing parameters and models.
- **Goal**: Improved accuracy via parameter tuning

Preprocessing

- Feature extraction and normalization.
- **Application**: Transforming input data such as text for use with machine learning algorithms.
- **Modules**: preprocessing, feature extraction

http://scikit-learn.org/
The Stanford NLP Group makes parts of our Natural Language Processing software available to everyone. These are statistical NLP toolkits for various major computational linguistics problems. They can be incorporated into applications with human language technology needs.

All the software we distribute here is written in Java. All recent distributions require Oracle Java 6+ or OpenJDK 7+. Distribution packages include components for command-line invocation, jar files, a Java API, and source code. A number of helpful people have extended our work with bindings or translations for other languages. As a result, much of this software can also easily be used from Python (or Jython), Ruby, Perl, Javascript, and F# or other .NET languages.

Supported software distributions

This code is being developed, and we try to answer questions and fix bugs on a best-effort basis.

All these software distributions are open source, licensed under the GNU General Public License (v2 or later). Note that this is the full GPL, which allows many free uses, but does not allow its incorporation into any type of distributed proprietary software, even in part or in translation. Commercial licensing is also available; please contact us if you are interested.

Stanford CoreNLP
An integrated suite of natural language processing tools for English and (mainland) Chinese in Java, including tokenization, part-of-speech tagging, named entity recognition, parsing, and coreference. See also: Stanford Deterministic Coreference Resolution, and the online CoreNLP demo, and the CoreNLP FAQ.

Stanford Parser
Implementations of probabilistic natural language parsers in Java: highly optimized PCFG and dependency parsers, a lexicalized PCFG parser, and a deep learning reranker. See also: Online parser demo, the Stanford Dependencies page, and Parser FAQ.

Stanford POS Tagger
A maximum-entropy (CMM) part-of-speech (POS) tagger for English,
Stanford University is located in California. It is a great university.

Part-of-Speech:

1. Stanford University is located in California.
2. It is a great university.

Named Entity Recognition:

1. Stanford University is located in California.
2. It is a great university.

Coreference:

1. Stanford University is located in California.
2. It is a great university.
Stanford University is located in California. It is a great university.
Stanford University is located in California. It is a great university.
Stanford University is located in California. It is a great university.
Stanford University is located in California. It is a great university.
Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

Collapsed dependencies:

1. Stanford University is located in California.
2. It is a great university.

Collapsed CC-processed dependencies:

1. Stanford University is located in California.
2. It is a great university.

Visualisation provided using the brat visualisation/annotation software. Copyright © 2011, Stanford University, All Rights Reserved.
Stanford University is located in California. It is a great university.

```
<table>
<thead>
<tr>
<th>Id</th>
<th>Word</th>
<th>Lemma</th>
<th>Char begin</th>
<th>Char end</th>
<th>POS</th>
<th>NER</th>
<th>Normalized NER</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Stanford</td>
<td>Stanford</td>
<td>0</td>
<td>8</td>
<td>NNP</td>
<td>ORGANIZATION</td>
<td></td>
<td>PER0</td>
</tr>
<tr>
<td>2</td>
<td>University</td>
<td>University</td>
<td>9</td>
<td>19</td>
<td>NNP</td>
<td>ORGANIZATION</td>
<td></td>
<td>PER0</td>
</tr>
<tr>
<td>3</td>
<td>is</td>
<td>be</td>
<td>20</td>
<td>22</td>
<td>VBZ</td>
<td>O</td>
<td></td>
<td>PER0</td>
</tr>
<tr>
<td>4</td>
<td>located</td>
<td>located</td>
<td>23</td>
<td>30</td>
<td>JJ</td>
<td>O</td>
<td></td>
<td>PER0</td>
</tr>
<tr>
<td>5</td>
<td>in</td>
<td>in</td>
<td>31</td>
<td>33</td>
<td>IN</td>
<td>O</td>
<td></td>
<td>PER0</td>
</tr>
<tr>
<td>6</td>
<td>California</td>
<td>California</td>
<td>34</td>
<td>44</td>
<td>NNP</td>
<td>LOCATION</td>
<td></td>
<td>PER0</td>
</tr>
<tr>
<td>7</td>
<td>.</td>
<td>.</td>
<td>44</td>
<td>45</td>
<td>.</td>
<td>O</td>
<td></td>
<td>PER0</td>
</tr>
</tbody>
</table>
```

Parse tree:
(ROOT (S (NP (NNP Stanford) (NNP University)) (VP (VBZ is) (ADJP (JJ located) (PP (IN in) (NP (NNP California)))))) (. .))
Stanford University is located in California. It is a great university.
Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

Stanford University is located in California. It is a great university.
Stanford CoreNLP
http://nlp.stanford.edu:8080/corenlp/process

Stanford University is located in California. It is a great university.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Head</th>
<th>Text</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>Stanford University</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>It</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>a great university</td>
<td></td>
</tr>
</tbody>
</table>
Stanford University is located in California.

It is a great university.
Stanford University is located in California. It is a great university.
Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan  @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

NEW YORK (CNNMoney)

For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder.
In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million.

That puts him behind Microsoft's former CEO Steve Ballmer who owns 333 million shares.
Related: Gates reclaims title of world's richest billionaire
Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires.
It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation.
The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.
Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

NEW YORK (CNNMoney) For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.

Potential tags:
LOCATION
TIME
PERSON
ORGANIZATION
MONEY
PERCENT
DATE

Copyright © 2011, Stanford University, All Rights Reserved.
Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

NEW YORK (CNNMoney) — For the first time in Microsoft’s history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft, bringing his total to roughly 330 million. That puts him behind Steve Ballmer, who owns 333 million shares. Related: Gates reclaims title of world’s richest billionaire

It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates Foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.
Stanford Named Entity Tagger

Classifier: english.muc.7class.distsim.crf.ser.gz
Output Format: xml
Preserve Spacing: yes

Please enter your text here:

Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

Submit Clear
Stanford Named Entity Tagger (NER)

http://nlp.stanford.edu:8080/ner/process

Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

Copyright © 2011, Stanford University, All Rights Reserved.
Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

NEW YORK (CNNMoney) For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer, who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire

Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires. It's a passing of the torch for Gates, who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.
Stanford Named Entity Tagger (NER)

http://nlp.stanford.edu:8080/ner/process

Stanford Named Entity Tagger

Classifier: english.all.3class.distsim.crf.ser.gz

Output Format: highlighted

Preserve Spacing: yes

Please enter your text here:

Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET
Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

NEW YORK (CNNMoney) For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.

Potential tags:

LOCATION
ORGANIZATION
PERSON

Copyright © 2011, Stanford University, All Rights Reserved.
Bill Gates no longer Microsoft's biggest shareholder By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET Bill Gates sold nearly 8 million shares of Microsoft over the past two days. NEW YORK (CNNMoney) For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.

Potential tags:
LOCATION
TIME
PERSON
ORGANIZATION
MONEY
PERCENT
DATE
Bill Gates no longer Microsoft's biggest shareholder. By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET Bill Gates sold nearly 8 million shares of Microsoft over the past two days. For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.
Bill Gates no longer Microsoft's biggest shareholder

By Patrick Sheridan @CNNTech
May 2, 2014 5:46 PM ET

NEW YORK - CNNMoney

For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder.

In the past two days, Gates has sold nearly 8 million shares of Microsoft, bringing his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer, who owns 333 million shares.

Related: Gates reclaims title of world's richest billionaire

Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires.

It's a passing of the torch for Gates, who has always been the largest single owner of his company's stock.

Gates now spends his personal fortune helping run the Bill & Melinda Gates Foundation, which has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.
Vector Representations of Words

Word Embeddings

Word2Vec

GloVe
Modern NLP Pipeline

Pre-processing

Documents

Language Detection

EN

Tokenize

POS Tagging

... Token Filtering

CN

Tokenize

POS Tagging

... Token Filtering

Pre-processed Documents

Pre-processed Documents

Build Vocabulary

Bag-of-Words & Vectorization

Machine Learning

Word Embeddings

(word2vec, doc2vec, GloVe)

(Deep) Neural Network

Task / Output

Classification

Sentiment Analysis

Entity Extraction

Topic Modeling

Similarity

Pre-trained word vectors
Word2Vec
wiki.zh.vec (861MB)
332647 word
300 vec

Pre-trained word vectors for 90 languages, trained on Wikipedia using fastText.

These vectors in dimension 300 were obtained using the skip-gram model with default parameters.

https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

Facebook Research FastText

Word2Vec: wiki.zh.vec

(861MB) (332647 word 300 vec)

https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md
Word Embeddings in LSTM RNN

Time Expanded LSTM Network

LSTM Internal States

Word Embeddings

Input Question: Is this person dancing?

Fixed length question vector encoded by the LSTM

Source: https://avisingh599.github.io/deeplearning/visual-qa/
## NLP Tools: spaCy vs. NLTK

<table>
<thead>
<tr>
<th>Feature</th>
<th>spaCy</th>
<th>SyntaxNet</th>
<th>NLTK</th>
<th>CoreNLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy installation</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Python API</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Multi-language support</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Tokenization</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Part-of-speech tagging</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Sentence segmentation</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Dependency parsing</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Entity Recognition</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Integrated word vectors</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sentiment analysis</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Coreference resolution</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

Source: [https://spacy.io/docs/api/](https://spacy.io/docs/api/)
Natural Language Processing (NLP)
spacy

1. Tokenization
2. Part-of-speech tagging
3. Sentence segmentation
4. Dependency parsing
5. Entity Recognition
6. Integrated word vectors
7. Sentiment analysis
8. Coreference resolution

Source: https://spacy.io/docs/api/
### spaCy:
**Fastest Syntactic Parser**

<table>
<thead>
<tr>
<th>SYSTEM</th>
<th>LANGUAGE</th>
<th>ACCURACY</th>
<th>SPEED (WPS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>spaCy</td>
<td>Cython</td>
<td>91.8</td>
<td>13,963</td>
</tr>
<tr>
<td>ClearNLP</td>
<td>Java</td>
<td>91.7</td>
<td>10,271</td>
</tr>
<tr>
<td>CoreNLP</td>
<td>Java</td>
<td>89.6</td>
<td>8,602</td>
</tr>
<tr>
<td>MATE</td>
<td>Java</td>
<td>92.5</td>
<td>550</td>
</tr>
<tr>
<td>Turbo</td>
<td>C++</td>
<td>92.4</td>
<td>349</td>
</tr>
</tbody>
</table>

Source: https://spacy.io/docs/api/
# Processing Speed of NLP libraries

<table>
<thead>
<tr>
<th>SYSTEM</th>
<th>ABSOLUTE (MS PER DOC)</th>
<th>RELATIVE (TO SPACY)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TOKENIZE</td>
<td>TAG</td>
</tr>
<tr>
<td>spaCy</td>
<td>0.2ms</td>
<td>1ms</td>
</tr>
<tr>
<td>CoreNLP</td>
<td>2ms</td>
<td>10ms</td>
</tr>
<tr>
<td>ZPar</td>
<td>1ms</td>
<td>8ms</td>
</tr>
<tr>
<td>NLTK</td>
<td>4ms</td>
<td>443ms</td>
</tr>
</tbody>
</table>

Source: [https://spacy.io/docs/api/](https://spacy.io/docs/api/)
## Google SyntaxNet (2016): Best Syntactic Dependency Parsing Accuracy

<table>
<thead>
<tr>
<th>SYSTEM</th>
<th>NEWS</th>
<th>WEB</th>
<th>QUESTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>spaCy</td>
<td>92.8</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Parsey McParseface</td>
<td>94.15</td>
<td>89.08</td>
<td>94.77</td>
</tr>
<tr>
<td>Martins et al. (2013)</td>
<td>93.10</td>
<td>88.23</td>
<td>94.21</td>
</tr>
<tr>
<td>Zhang and McDonald (2014)</td>
<td>93.32</td>
<td>88.65</td>
<td>93.37</td>
</tr>
<tr>
<td>Weiss et al. (2015)</td>
<td>93.91</td>
<td>89.29</td>
<td>94.17</td>
</tr>
<tr>
<td>Andor et al. (2016)</td>
<td>94.44</td>
<td>90.17</td>
<td>95.40</td>
</tr>
</tbody>
</table>

Source: https://spacy.io/docs/api/
# Named Entity Recognition (NER)

<table>
<thead>
<tr>
<th>SYSTEM</th>
<th>PRECISION</th>
<th>RECALL</th>
<th>F-MEASURE</th>
</tr>
</thead>
<tbody>
<tr>
<td>spaCy</td>
<td>0.7240</td>
<td>0.6514</td>
<td>0.6858</td>
</tr>
<tr>
<td>CoreNLP</td>
<td>0.7914</td>
<td>0.7327</td>
<td>0.7609</td>
</tr>
<tr>
<td>NLTK</td>
<td>0.5136</td>
<td>0.6532</td>
<td>0.5750</td>
</tr>
<tr>
<td>LingPipe</td>
<td>0.5412</td>
<td>0.5357</td>
<td>0.5384</td>
</tr>
</tbody>
</table>
Text Classification

Source: https://developers.google.com/machine-learning/guides/text-classification/
Text Classification Workflow

- Step 1: Gather Data
- Step 2: Explore Your Data
- Step 2.5: Choose a Model*
- Step 3: Prepare Your Data
- Step 4: Build, Train, and Evaluate Your Model
- Step 5: Tune Hyperparameters
- Step 6: Deploy Your Model

Source: https://developers.google.com/machine-learning/guides/text-classification/
Text Classification Flowchart

Text Classification S/W<1500: N-gram

Text Classification S/W>=1500: Sequence

Select top_k features [freq]

min(top_1K, 2K,... 15K) 20K 25K, ... 90K all)

Normalization mode

samplewise None featurewise

Embeddings

S/W < 15K

Yes

Fine-tuned pre-trained embedding

No

Frozen pre-trained embedding

Embeddings learned from scratch

Build model

RNN stacked RNN CNN-RNN sepCNN CNN

Hyperparameter tuning

Step 2.5: Choose a Model

**Samples/Words < 1500**

**150,000/100 = 1500**

IMDb review dataset, the samples/words-per-sample ratio is ~ 144
Step 2.5: Choose a Model

Samples/Words < 15,000

1,500,000/100 = 15,000

Step 3: Prepare Your Data

Texts:
T1: 'The mouse ran up the clock'
T2: 'The mouse ran down'

Token Index:
{'the': 1, 'mouse': 2, 'ran': 3, 'up': 4, 'clock': 5, 'down': 6}.

NOTE: 'the' occurs most frequently, so the index value of 1 is assigned to it. Some libraries reserve index 0 for unknown tokens, as is the case here.

Sequence of token indexes:
T1: 'The mouse ran up the clock' = [1, 2, 3, 4, 1, 5]
T1: 'The mouse ran down' = [1, 2, 3, 6]
# One-hot encoding

'The mouse ran up the clock' =

<table>
<thead>
<tr>
<th>Word</th>
<th>Index</th>
<th>Encoding</th>
</tr>
</thead>
<tbody>
<tr>
<td>The</td>
<td>1</td>
<td>[0, 1, 0, 0, 0, 0, 0, 0]</td>
</tr>
<tr>
<td>mouse</td>
<td>2</td>
<td>[0, 0, 1, 0, 0, 0, 0, 0]</td>
</tr>
<tr>
<td>ran</td>
<td>3</td>
<td>[0, 0, 0, 1, 0, 0, 0, 0]</td>
</tr>
<tr>
<td>up</td>
<td>4</td>
<td>[0, 0, 0, 0, 1, 0, 0, 0]</td>
</tr>
<tr>
<td>the</td>
<td>1</td>
<td>[0, 1, 0, 0, 0, 0, 0, 0]</td>
</tr>
<tr>
<td>clock</td>
<td>5</td>
<td>[0, 0, 0, 0, 0, 0, 1, 0]</td>
</tr>
</tbody>
</table>

[0, 1, 2, 3, 4, 5, 6]
Word embeddings

Male-Female
Verb Tense
Country-Capital

Source: https://developers.google.com/machine-learning/guides/text-classification/step-3
Word embeddings

The mouse ran up the clock

<table>
<thead>
<tr>
<th>the</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>mouse</td>
<td>2</td>
</tr>
<tr>
<td>ran</td>
<td>3</td>
</tr>
<tr>
<td>up</td>
<td>4</td>
</tr>
<tr>
<td>clock</td>
<td>5</td>
</tr>
<tr>
<td>down</td>
<td>6</td>
</tr>
</tbody>
</table>

[1, 2, 3, 4, 1, 5]

The mouse ran down

<table>
<thead>
<tr>
<th>the</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>mouse</td>
<td>2</td>
</tr>
<tr>
<td>ran</td>
<td>3</td>
</tr>
<tr>
<td>up</td>
<td>4</td>
</tr>
<tr>
<td>clock</td>
<td>5</td>
</tr>
<tr>
<td>down</td>
<td>6</td>
</tr>
</tbody>
</table>

[1, 2, 3, 6]

Embedding layer (output dim = 4)

[[0.236, -0.141, 0.000, 0.045], [0.006, 0.652, 0.270, -0.556], [0.305, 0.569, -0.028, 0.496], [0.421, 0.195, -0.058, 0.477], [0.236, -0.141, 0.000, 0.045], [0.844, -0.001, 0.763, 0.201]]

Source: https://developers.google.com/machine-learning/guides/text-classification/step-3
Sequence to Sequence (Seq2Seq)
Transformer (Attention is All You Need)
(Vaswani et al., 2017)

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT (Bidirectional Encoder Representations from Transformers)

Overall pre-training and fine-tuning procedures for BERT

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT (Bidirectional Encoder Representations from Transformers)

BERT input representation

BERT, OpenAI GPT, ELMo

Fine-tuning BERT on Different Tasks

(a) Sentence Pair Classification Tasks:
- MNLI, QQP, QNLI, STS-B, MRPC,
- RTE, SWAG

(b) Single Sentence Classification Tasks:
- SST-2, CoLA

(c) Question Answering Tasks:
- SQuAD v1.1

(d) Single Sentence Tagging Tasks:
- CoNLL-2003 NER

Pre-trained Language Model (PLM)

Source: https://github.com/thunlp/PLMpapers
Turing Natural Language Generation (T-NLG)

Transformers

State-of-the-art Natural Language Processing for TensorFlow 2.0 and PyTorch

• Transformers
  – pytorch-transformers
  – pytorch-pretrained-bert

• provides state-of-the-art general-purpose architectures
  – (BERT, GPT-2, RoBERTa, XLM, DistilBert, XLNet, CTRL...)
  – for Natural Language Understanding (NLU) and Natural Language Generation (NLG)
  with over 32+ pretrained models
  in 100+ languages
  and deep interoperability between TensorFlow 2.0 and PyTorch.

Source: https://github.com/huggingface/transformers
Transfer Learning in Natural Language Processing

# NLP Benchmark Datasets

<table>
<thead>
<tr>
<th>Task</th>
<th>Dataset</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WMT 2014 EN-FR</td>
<td></td>
</tr>
<tr>
<td><strong>Text Summarization</strong></td>
<td>CNN/DM</td>
<td><a href="https://cs.nyu.edu/~kcho/DMQA/">https://cs.nyu.edu/~kcho/DMQA/</a></td>
</tr>
<tr>
<td></td>
<td>Newsroom</td>
<td><a href="http://summarisi.es/">http://summarisi.es/</a></td>
</tr>
<tr>
<td></td>
<td>Gigaword</td>
<td><a href="https://catalog.ldc.upenn.edu/LDC2012T21">https://catalog.ldc.upenn.edu/LDC2012T21</a></td>
</tr>
<tr>
<td><strong>Reading Comprehension</strong></td>
<td>ARC</td>
<td><a href="http://data.allenai.org/arc/">http://data.allenai.org/arc/</a></td>
</tr>
<tr>
<td></td>
<td>CliCR</td>
<td><a href="http://aclweb.org/anthology/N18-1140">http://aclweb.org/anthology/N18-1140</a></td>
</tr>
<tr>
<td></td>
<td>CNN/DM</td>
<td><a href="https://cs.nyu.edu/~kcho/DMQA/">https://cs.nyu.edu/~kcho/DMQA/</a></td>
</tr>
<tr>
<td></td>
<td>NewsQA</td>
<td><a href="https://datasets.maluuba.com/NewsQA">https://datasets.maluuba.com/NewsQA</a></td>
</tr>
<tr>
<td></td>
<td>RACE</td>
<td><a href="http://www.qizhexie.com/data/RACE_leaderboard">http://www.qizhexie.com/data/RACE_leaderboard</a></td>
</tr>
<tr>
<td></td>
<td>SQuAD</td>
<td><a href="https://rajpurkar.github.io/SQuAD_leaderboard/">https://rajpurkar.github.io/SQuAD_leaderboard/</a></td>
</tr>
<tr>
<td><strong>Question Answering</strong></td>
<td>Story Cloze Test</td>
<td><a href="http://aclweb.org/anthology/W17-0906.pdf">http://aclweb.org/anthology/W17-0906.pdf</a></td>
</tr>
<tr>
<td></td>
<td>NarrativeQA</td>
<td><a href="https://github.com/deepmind/narrativeqa">https://github.com/deepmind/narrativeqa</a></td>
</tr>
<tr>
<td></td>
<td>Quasar</td>
<td><a href="https://github.com/nyu-dl/Quasar">https://github.com/nyu-dl/Quasar</a></td>
</tr>
<tr>
<td></td>
<td>SearchQA</td>
<td></td>
</tr>
<tr>
<td><strong>Question Generation</strong></td>
<td>AMR parsing</td>
<td><a href="https://amr.isi.edu/index.html">https://amr.isi.edu/index.html</a></td>
</tr>
<tr>
<td></td>
<td>ATIS (SQL Parsing)</td>
<td><a href="https://github.com/jkkummerfeld/text2sql-data/tree/master/data">https://github.com/jkkummerfeld/text2sql-data/tree/master/data</a></td>
</tr>
<tr>
<td></td>
<td>WikiSQL (SQL Parsing)</td>
<td><a href="https://github.com/salesforce/WikiSQL">https://github.com/salesforce/WikiSQL</a></td>
</tr>
<tr>
<td><strong>Semantic Parsing</strong></td>
<td>IMDB Reviews</td>
<td><a href="http://ai.stanford.edu/~amaas/data/sentiment/">http://ai.stanford.edu/~amaas/data/sentiment/</a></td>
</tr>
<tr>
<td></td>
<td>SST</td>
<td><a href="https://nlp.stanford.edu/sentiment/index.html">https://nlp.stanford.edu/sentiment/index.html</a></td>
</tr>
<tr>
<td></td>
<td>Yelp Reviews</td>
<td><a href="https://www.yelp.com/dataset/challenge">https://www.yelp.com/dataset/challenge</a></td>
</tr>
<tr>
<td></td>
<td>Subjectivity Dataset</td>
<td><a href="http://www.cs.cornell.edu/people/pabo/movie-review-data/">http://www.cs.cornell.edu/people/pabo/movie-review-data/</a></td>
</tr>
<tr>
<td><strong>Sentiment Analysis</strong></td>
<td>AG News</td>
<td><a href="http://www.di.unipi.it/~gulli/AG_corpus_of_news_articles.html">http://www.di.unipi.it/~gulli/AG_corpus_of_news_articles.html</a></td>
</tr>
<tr>
<td></td>
<td>DBpedia</td>
<td><a href="https://wiki.dbpedia.org/Datasets">https://wiki.dbpedia.org/Datasets</a></td>
</tr>
<tr>
<td></td>
<td>TREC</td>
<td><a href="https://trec.nist.gov/data.html">https://trec.nist.gov/data.html</a></td>
</tr>
<tr>
<td><strong>Text Classification</strong></td>
<td>SNLI Corpus</td>
<td><a href="https://nlp.stanford.edu/projects/snli/">https://nlp.stanford.edu/projects/snli/</a></td>
</tr>
<tr>
<td></td>
<td>MultiNLI</td>
<td><a href="https://www.nyu.edu/projects/bowman/multinli/">https://www.nyu.edu/projects/bowman/multinli/</a></td>
</tr>
<tr>
<td></td>
<td>SciTail</td>
<td><a href="http://data.allenai.org/scitail/">http://data.allenai.org/scitail/</a></td>
</tr>
<tr>
<td><strong>Natural Language Inference</strong></td>
<td>Proposition Bank</td>
<td><a href="http://propbank.github.io/">http://propbank.github.io/</a></td>
</tr>
<tr>
<td></td>
<td>OneNotes</td>
<td><a href="https://catalog.ldc.upenn.edu/LDC2013T19">https://catalog.ldc.upenn.edu/LDC2013T19</a></td>
</tr>
</tbody>
</table>

A High-Level Depiction of DeepQA Architecture

Chatbots

Bot Maturity Model

Customers want to have simpler means to interact with businesses and get faster response to a question or complaint.

Dialogue on Airline Travel Information System (ATIS)
The ATIS (Airline Travel Information System) Dataset


Training samples: 4978
Testing samples: 893
Vocab size: 943
Slot count: 129
Intent count: 26

https://www.kaggle.com/siddhadev/atis-dataset-from-ms-cntk
SF-ID Network (E et al., 2019)
Slot Filling (SF)
Intent Detection (ID)

A Novel Bi-directional Interrelated Model for Joint Intent Detection and Slot Filling

Intent Detection on ATIS
State-of-the-art

<table>
<thead>
<tr>
<th>RANK</th>
<th>METHOD</th>
<th>ACCURACY</th>
<th>PAPER TITLE</th>
<th>YEAR</th>
<th>PAPER</th>
<th>CODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SF-ID</td>
<td>0.9776</td>
<td>A Novel Bi-directional Interrelated Model for Joint Intent Detection and Slot Filling</td>
<td>2019</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Capsule-NLU</td>
<td>0.950</td>
<td>Joint Slot Filling and Intent Detection via Capsule Neural Networks</td>
<td>2018</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: [https://paperswithcode.com/sota/intent-detection-on-atis](https://paperswithcode.com/sota/intent-detection-on-atis)
Slot Filling on ATIS

State-of-the-art

### Slot Filling on ATIS

<table>
<thead>
<tr>
<th>Rank</th>
<th>Method</th>
<th>F1</th>
<th>Paper Title</th>
<th>Year</th>
<th>Paper</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SF-ID</td>
<td>0.958</td>
<td>A Novel Bi-directional Interrelated Model for Joint Intent Detection and Slot Filling</td>
<td>2019</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Capsule-NLU</td>
<td>0.952</td>
<td>Joint Slot Filling and Intent Detection via Capsule Neural Networks</td>
<td>2018</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: [https://paperswithcode.com/sota/slot-filling-on-atis](https://paperswithcode.com/sota/slot-filling-on-atis)
TensorFlow NLP Examples

• Basic Text Classification (Text Classification) (46 Seconds)

• NMT with Attention (20-30 minutes)
Text Classification

IMDB Movie Reviews

[Source: https://colab.research.google.com/github/tensorflow/docs/blob/master/site/en/tutorials/keras/basic_text_classification.ipynb]
Summary

• Text Analytics

• Natural Language Processing (NLP)
References


• Charu C. Aggarwal (2018), Machine Learning for Text, Springer.

• Gabe Ignatow and Rada F. Mihalcea (2017), An Introduction to Text Mining: Research Design, Data Collection, and Analysis, SAGE Publications.

