自然語言處理核心技術與文字探勘
(Core Technologies of Natural Language Processing and Text Mining)

Time: 2020/05/15 (Fri) (9:10 -12:00)
Place: 國立臺北護理健康大學 (台北市明德路365號) G210
Host: 祝國忠 院長 (健康科技學院院長)

Min-Yuh Day
戴敏育
Associate Professor
副教授
Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系

http://mail.tku.edu.tw/myday/
2020-05-15
(Min-Yuh Day, Ph.D.)

Publications Co-Chairs, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013- )

Program Co-Chair, IEEE International Workshop on Empirical Methods for Recognizing Inference in TExt (IEEE EM-RITE 2012- )

Publications Chair, The IEEE International Conference on Information Reuse and Integration (IEEE IRI)
1. Core Technologies of Natural Language Processing and Text Mining
2. Artificial Intelligence for Text Analytics: Foundations and Applications
3. Feature Engineering for Text Representation
4. Semantic Analysis and Named Entity Recognition; NER
5. Deep Learning and Universal Sentence-Embedding Models
6. Question Answering and Dialogue Systems
Outline

- Text Analytics and Text Mining
- Natural Language Processing (NLP)
- Text Analytics with Python
Text Analytics (TA)
Text Mining (TM)
Natural Language Processing (NLP)
Artificial Intelligence (AI)
Text Analytics and Text Mining

Text Mining “Knowledge Discovery in Textual Data”

TEXT ANALYTICS

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

Data Mining
- Classification
- Clustering
- Association

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

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
Evolution of Decision Support, Business Intelligence, and Analytics

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...
AI Dialogue System
AIWISFIN
AI Conversational Robo-Advisor
(人工智能對話式理財機器人)
First Place, InnoServe Awards 2018

https://www.youtube.com/watch?v=sEhmyoTXmGk
2018 The 23\textsuperscript{th} International ICT Innovative Services Awards (InnoServe Awards 2018)

- Annual ICT application competition held for university and college students
- The largest and the most significant contest in Taiwan.
- More than ten thousand teachers and students from over one hundred universities and colleges have participated in the Contest.

https://innoserve.tca.org.tw/award.aspx

https://innoserve.tca.org.tw/award.aspx
IMTKU
Emotional Dialogue System
for
Short Text Conversation
at
NTCIR-14 STC-3 (CECG) Task
IMTKU Textual Entailment System for Recognizing Inference in Text at NTCIR-9 RITE

Department of Information Management
Tamkang University, Taiwan

Min-Yuh Day
Chun Tu
myday@mail.tku.edu.tw

NTCIR-9 Workshop, December 6-9, 2011, Tokyo, Japan
IMTKU Textual Entailment System for Recognizing Inference in Text at NTCIR-10 RITE-2

Department of Information Management
Tamkang University, Taiwan

Min-Yuh Day
Chun Tu
Hou-Cheng Vong
Shih-Wei Wu
Shih-Jhen Huang

myday@mail.tku.edu.tw
IMTKU Textual Entailment System for Recognizing Inference in Text at NTCIR-11 RITE-VAL

Tamkang University 2014

Min-Yuh Day, Ya-Jung Wang, Che-Wei Hsu, En-Chun Tu

Huai-Wen Hsu, Yu-An Lin, Shang-Yu Wu, Yu-Hsuan Tai, Cheng-Chia Tsai

NTCIR-11 Conference, December 8-12, 2014, Tokyo, Japan
IMTKU Question Answering System for World History Exams at NTCIR-12 QA Lab2

Department of Information Management
Tamkang University, Taiwan

myday@mail.tku.edu.tw
IMTKU Question Answering System for World History Exams at NTCIR-13 QALab-3

Department of Information Management
Tamkang University, Taiwan

myday@mail.tku.edu.tw

NTCIR-13 Conference, December 5-8, 2017, Tokyo, Japan
IMTKU Emotional Dialogue System for Short Text Conversation at NTCIR-14 STC-3 (CECG) Task

Department of Information Management
Tamkang University, Taiwan

myday@mail.tku.edu.tw

NTCIR-14 Conference, June 10-13, 2019, Tokyo, Japan
IMTKU System Architecture for NTCIR-13 QALab-3

Question (XML)

- Question Analysis
  - JA&EN Translator
  - Stanford CoreNLP
  - Wikipedia

Document Retrieval

Answer Extraction

Answer Generation

Word Embedding
- Wiki Word2Vec

Answer (XML)

Complex Essay
Simple Essay
True-or-False
Factoid
Slot-Filling
Unique

NTCIR-13 Conference, December 5-8, 2017, Tokyo, Japan
System Architecture of Intelligent Dialogue and Question Answering System


Deep Learning → TensorFlow

Python → NLTK

AIML KB

Cloud Resource

RNN LSTM GRU

Document Retrieval → Answer Extraction → Answer Generation 

Deep Learning → IR

Dialogue KB
IMTKU Emotional Dialogue System Architecture

1. Retrieval-Based Model
2. Generation-Based Model
3. Emotion Classification Model
4. Response Ranking
The system architecture of IMTKU retrieval-based model for NTCIR-14 STC-3

Retrieval-Based Model

1. Post
2. Word Segmentation
   - Keyword Boolean Query
3. Corpus
   - Building Index
4. Solr Matching
5. Distinct Result Data
   - Emotion Matching
   - Emotion Classification
6. Word2Vec Similarity Ranking
7. Retrieval-Based Response

NTCIR-14 Conference, June 10-13, 2019, Tokyo, Japan
The system architecture of IMTKU generation-based model for NTCIR-14 STC-3

**Generation-Based Model**

1. Training Data
2. Building Word Index
3. Word Embedding
4. Training Data Seq2seq model
5. Post
6. Word Segmentation
7. Short Text Emotion Classifier
8. Trained Model
9. Emotion Matching
10. Word2Vec Similarity Ranking
11. Generation-Based Response

**Generative Model**
The system architecture of IMTKU emotion classification model for NTCIR-14 STC-3

Emotion Classification Model

1. Corpus
2. Emotion Classification
3. Training Dataset
4. MLP LSTM BiLSTM
5. Testing Dataset
6. Emotion Classification Model
7. Emotion Prediction
The system architecture of IMTKU Response Ranking for NTCIR-14 STC-3

Response Ranking

1. STC3 Corpus
2. Chinese Segmentation using Jieba
3. Stop Words Removal
4. Word2Vec
5. 1.2 million data (300 dimensions)
6. Vector of Corpus
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>1. Acting Humanly: The Turing Test Approach (1950)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Thinking Humanly: The Cognitive Modeling Approach</td>
</tr>
<tr>
<td>3. Thinking Rationally: The “Laws of Thought” Approach</td>
</tr>
<tr>
<td>4. Acting Rationally: The Rational Agent Approach</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

Text Analytics and Text Mining
Dipanjan Sarkar (2019),

Text Analytics with Python:
A Practitioner’s Guide to Natural Language Processing,

Source: https://www.amazon.com/Text-Analytics-Python-Practitioners-Processing/dp/1484243536

Source: https://www.amazon.com/Applied-Text-Analysis-Python-Language-Aware/dp/1491963042
Charu C. Aggarwal (2018),
Machine Learning for Text,
Springer

Source: https://www.amazon.com/Machine-Learning-Text-Charu-Aggarwal/dp/3319735306
Gabe Ignatow and Rada F. Mihalcea (2017),

An Introduction to Text Mining: Research Design, Data Collection, and Analysis, SAGE Publications.

Source: https://www.amazon.com/Introduction-Text-Mining-Research-Collection/dp/1506337007
Rajesh Arumugam (2018), Hands-On Natural Language Processing with Python:
A practical guide to applying deep learning architectures to your NLP applications, Packt

Text Analytics

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

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

Text Mining

• Text Data Mining

• Knowledge Discovery in Textual Databases

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

Statistics

Database Systems

Natural Language Processing

Information Retrieval

Machine Learning

Applications

Pattern Recognition

Visualization

Algorithms

High-performance Computing

Adapted from: Jiawei Han and Micheline Kamber (2011), Data Mining: Concepts and Techniques, Third Edition, Elsevier
Application Areas of Text Mining

• Information extraction
• Topic tracking
• Summarization
• Categorization
• Clustering
• Concept linking
• Question answering

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
Text-Based Deception-Detection Process

1. Statements Transcribed for Processing
2. Cues Extracted & Selected
3. Text Processing Software-Generated Quantified Cues
4. Classification Models Trained and Tested on Quantified Cues
5. Statements Labeled as Truthful or Deceptive by Law Enforcement

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
Multilevel Analysis of Text for Gene/Protein Interaction Identification

... expression of Bcl-2 is correlated with insufficient white blood cell death and activation of p53.

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
Context Diagram for the Text Mining Process

Unstructured data (text) → Extract knowledge from available data sources
Structured data (databases) → Context-specific knowledge

Software/hardware limitations
Privacy issues
Linguistic limitations

Domain expertise
Tools and techniques

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
The Three-Step/Task Text Mining Process

**Task 1: Establish the Corpus**
Collect and organize the domain-specific unstructured data.
- The inputs to the process include a variety of relevant unstructured (and semi-structured) data sources such as text, XML, HTML, etc.
- The output of Task 1 is a collection of documents in some digitized format for computer processing.

**Task 2: Create the Term-Document Matrix**
Introduce structure to the corpus.
- Feedback
- The output of Task 2 is a flat file called a term-document matrix where the cells are populated with the term frequencies.

**Task 3: Extract Knowledge**
Discover novel patterns from the T-D matrix.
- Feedback
- The output of Task 3 is a number of problem-specific classification, association, clustering models and visualizations.

# Term–Document Matrix

| Documents       | Investment Risk | Project Management | Software Engineering | Development | SAP | ...
|-----------------|-----------------|--------------------|----------------------|-------------|-----|------
| Document 1      | 1               |                    | 1                    |             |     |      
| Document 2      |                 | 1                  |                      |             |     |      
| Document 3      |                 |                    | 3                    | 1           |     |      
| Document 4      |                 |                    | 1                    |             |     |      
| Document 5      |                 |                    | 2                    | 1           |     |      
| Document 6      |                 |                    | 1                    |             |     |      
| …               |                 |                    |                      |             |     |      

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

Example of Opinion:
review segment on iPhone

“(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. **Calculate the O−S Polarity**
   - Lexicon
   - A statement

2. **Calculate the N−P Polarity of the sentiment**
   - Lexicon
   - No
   - Yes

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

4. **Record the Polarity, Strength, and the Target of the sentiment**
   - O−S Polarity measure
   - N−P Polarity

5. **Tabulate & aggregate the sentiment analysis results**

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
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)
      - Neural Network (NN)
      - Deep Learning (DL)
      - Naïve Bayes (NB)
      - Bayesian Network (BN)
      - Maximum Entropy (ME)
  - Lexicon-based Approach
    - Dictionary-based Approach
    - Corpus-based Approach
  - Statistical
  - Semantic

P–N Polarity and S–O Polarity Relationship

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

Web Mining
- Data Mining
- Text Mining

Web Content Mining
Source: unstructured textual content of the Web pages (usually in HTML format)

Web Structure Mining
Source: the unified resource locator (URL) links contained in the Web pages

Web Usage Mining
Source: the detailed description of a Web site’s visits (sequence of clicks by sessions)

Search Engines
- Page Rank
- Search Engine Optimization
- Marketing Attribution

Sentiment Analysis
- Information Retrieval
- Social Network Analysis
- Customer Analytics

Semantic Webs
- Graph Mining
- Social Media Analytics
- 360 Customer View

Web Analytics
- Social Analytics
- Clickstream Analysis
- Weblog Analysis
- Voice of the Customer

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), *Business Intelligence, Analytics, and Data Science: A Managerial Perspective*, 4th Edition, Pearson
Structure of a Typical Internet Search Engine

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

• **Web usage mining (Web analytics)** is the extraction of useful information from data generated through Web page visits and transactions.

• **Clickstream Analysis**

Extraction of Knowledge from Web Usage Data

Preprocess Data
- Collecting
- Merging
- Cleaning
- Structuring
  - Identify users
  - Identify sessions
  - Identify page views
  - Identify visits

Extract Knowledge
- Usage patterns
- User profiles
- Page profiles
- Visit profiles
- Customer value

How to better the data
How to improve the Web site
How to increase the customer value

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

• Social analytics is defined as monitoring, analyzing, measuring and interpreting digital interactions and relationships of people, topics, ideas and content.

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

Social Analytics

Social Network Analysis (SNA)

Social Media Analytics

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

Natural Language Processing
(NLP)
Text Mining Concepts

• 85-90 percent of all corporate data is in some kind of unstructured form (e.g., text)
• Unstructured corporate data is doubling in size every 18 months
• Tapping into these information sources is not an option, but a need to stay competitive
• Answer: text mining
  – A semi-automated process of extracting knowledge from unstructured data sources
  – a.k.a. text data mining or knowledge discovery in textual databases

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text mining

Text Data Mining

Intelligent Text Analysis

Knowledge-Discovery in Text (KDT)

Text Mining
(text data mining)

the process of deriving high-quality information from text

http://en.wikipedia.org/wiki/Text_mining
Text Mining: the process of extracting interesting and non-trivial information and knowledge from unstructured text.

Text Mining: discovery by computer of new, previously unknown information, by automatically extracting information from different written resources.

An example of Text Mining

Analyze Text
- Information Extraction
- Classification
- Summarization
- Clustering

Retrieve and preprocess document

Document Collection

Knowledge

Management Information System

Overview of Information Extraction based Text Mining Framework

Text Data Mining

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
- Dependency Parser
- String Metrics & Matching

Word's stem: am → am
Word's lemma: am → be
Having: hav
Hav: have

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

**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 as few as 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 Stu Anderson, a genetics professor at the University of Texas in Austin. 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...
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

Documents

Language Detection

Preprocessing

Preprocessing

Modeling

Preprocessing

Modeling

Task / Output

Classification

Sentiment Analysis

Entity Extraction

Topic Modeling

Document Similarity

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

Pre-generated Lookup
OR
Generated in 1st level of NeuralNet

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

Prepare data

N-gram

N-gram range

unigram

bigram

trigram

Count mode

binary

tf-idf

count

Scoring method

none

f_classif

chi2

Select top_k features [score]

min(top: 1K, 2K, 15K, 20K, 25K, 90K, all)

Normalization mode

samplewise

None

featurewise

Build model

SVM

MLP

GBDT

Yes S/W < 1500

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

Yes
- S/W < 15K
  - Fine-tuned pre-trained embedding
  - Frozen pre-trained embedding

No
- 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' =
\[1, 2, 3, 6\]

Source: https://developers.google.com/machine-learning/guides/text-classification/step-3
One-hot encoding

'The mouse ran up the clock' =

\[
\begin{align*}
\text{The} & \quad 1 & \quad [0, 1, 0, 0, 0, 0, 0, 0], \\
\text{mouse} & \quad 2 & \quad [0, 0, 1, 0, 0, 0, 0, 0], \\
\text{ran} & \quad 3 & \quad [0, 0, 0, 1, 0, 0, 0, 0], \\
\text{up} & \quad 4 & \quad [0, 0, 0, 0, 1, 0, 0, 0], \\
\text{the} & \quad 1 & \quad [0, 1, 0, 0, 0, 0, 0, 0], \\
\text{clock} & \quad 5 & \quad [0, 0, 0, 0, 0, 1, 0, 0] \quad ] \\
\end{align*}
\]

\[ [0, 1, 2, 3, 4, 5, 6] \]
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

The mouse ran down

Source: https://developers.google.com/machine-learning/guides/text-classification/step-3
Sequence to Sequence (Seq2Seq)

Encoder:
- $e_0$ → $e_1$ → $e_2$ → $e_3$ → $e_4$ → $e_5$ → $e_6$

Attention:

Decoder:
- $d_0$ → $d_1$ → $d_2$ → $d_3$

Knowledge is power

Source: https://google.github.io/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 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.

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

BERT (Bidirectional Encoder Representations from Transformers)

**BERT input representation**

<table>
<thead>
<tr>
<th>Input</th>
<th>[CLS]</th>
<th>my</th>
<th>dog</th>
<th>is</th>
<th>cute</th>
<th>[SEP]</th>
<th>he</th>
<th>likes</th>
<th>play</th>
<th># # ing</th>
<th>[SEP]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Token Embeddings</strong></td>
<td>$E_{[CLS]}$</td>
<td>$E_{my}$</td>
<td>$E_{dog}$</td>
<td>$E_{is}$</td>
<td>$E_{cute}$</td>
<td>$E_{[SEP]}$</td>
<td>$E_{he}$</td>
<td>$E_{likes}$</td>
<td>$E_{play}$</td>
<td>$E_{# # ing}$</td>
<td>$E_{[SEP]}$</td>
</tr>
<tr>
<td><strong>Segment Embeddings</strong></td>
<td>$E_A$</td>
<td>$E_A$</td>
<td>$E_A$</td>
<td>$E_A$</td>
<td>$E_A$</td>
<td>$E_A$</td>
<td>$E_B$</td>
<td>$E_B$</td>
<td>$E_B$</td>
<td>$E_B$</td>
<td>$E_B$</td>
</tr>
<tr>
<td><strong>Position Embeddings</strong></td>
<td>$E_0$</td>
<td>$E_1$</td>
<td>$E_2$</td>
<td>$E_3$</td>
<td>$E_4$</td>
<td>$E_5$</td>
<td>$E_6$</td>
<td>$E_7$</td>
<td>$E_8$</td>
<td>$E_9$</td>
<td>$E_{10}$</td>
</tr>
</tbody>
</table>

The input embeddings is the sum of the token embeddings, the segmentation embeddings and the position embeddings.

Fine-tuning BERT on NLP 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

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

<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>392k</td>
<td>363k</td>
<td>108k</td>
<td>67k</td>
<td>8.5k</td>
<td>5.7k</td>
<td>3.5k</td>
<td>2.5k</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>BERT_{BASE}</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>BERT_{LARGE}</td>
<td>86.7/85.9</td>
<td>72.1</td>
<td>91.1</td>
<td>94.9</td>
<td>60.5</td>
<td>86.5</td>
<td>89.3</td>
<td>70.1</td>
<td>81.9</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
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

A High-Level Depiction of DeepQA Architecture

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
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: 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:
    print(sentence.sentiment.polarity)
# 0.060
```

Visit [https://textblob.readthedocs.io](https://textblob.readthedocs.io) for more information.
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 Natural Language Processing Group

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 CoreNLP

Output format: Visualise

Please enter your text here:

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

Submit  Clear

Part-of-Speech:

```
1 Stanford University is located in California.
  PRP VBZ DT JJ IN NN
2 It is a great university.
  NNP NNP VBZ JJ IN NNP
```

Named Entity Recognition:

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

Coreference:

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

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

### Stanford CoreNLP XML Output

#### Document

#### Document Info

#### Sentences

**Sentence #1**

<table>
<thead>
<tr>
<th>Tokens</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>Word</td>
<td>Lemma</td>
<td>Char begin</td>
<td>Char end</td>
<td>POS</td>
<td>NER</td>
<td>Normalized NER</td>
</tr>
<tr>
<td>1</td>
<td>Stanford</td>
<td>Stanford</td>
<td>0</td>
<td>8</td>
<td>NNP</td>
<td>ORGANIZATION</td>
<td></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>
</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>
</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>
</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>
</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>
</tr>
<tr>
<td>7</td>
<td>.</td>
<td>.</td>
<td>44</td>
<td>45</td>
<td>.</td>
<td>.</td>
<td></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

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

Output format: XML

Please enter your text here:

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

Submit  Clear

<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet href="CoreNLP-to-HTML.xsl" type="text/xsl"?>
<root>
  <document>
    <sentences>
      <sentence id="1">
        <tokens>
          <token id="1">
            <word>Stanford</word>
            <lemma>Stanford</lemma>
            <CharacterOffsetBegin>0</CharacterOffsetBegin>
            <CharacterOffsetEnd>8</CharacterOffsetEnd>
            <POS>NNP</POS>
            <NER>ORGANIZATION</NER>
            <Speaker>PER0</Speaker>
          </token>
          <token id="2">
            <word>University</word>
            <lemma>University</lemma>
            <CharacterOffsetBegin>9</CharacterOffsetBegin>
            <CharacterOffsetEnd>19</CharacterOffsetEnd>
            <POS>NNP</POS>
            <NER>ORGANIZATION</NER>
            <Speaker>PER0</Speaker>
          </token>
          ...
        </tokens>
      </sentence>
      ...
    </sentences>
  </document>
</root>
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 (NER)

http://nlp.stanford.edu:8080/ner/process
Stanford Named Entity Tagger (NER)

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

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.
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.
Stanford Named Entity Tagger (NER)

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

Stanford Named Entity Tagger

Classifier: english.conll.4class.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.

Submit  Clear

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. Bill 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
MISC

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
ORGANIZATION
PERSON
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

Classifier: english.muc.7class.distsim.crf.ser.gz

Classifier: english.all.3class.distsim.crf.ser.gz
Bill Gates no longer <ORGANIZATION>Microsoft</ORGANIZATION>'s biggest shareholder By <PERSON>Patrick M. Sheridan</PERSON> @CNNTech <DATE>May 2, 2014</DATE>: 5:46 PM ET Bill Gates sold nearly 8 million shares of <ORGANIZATION>Microsoft</ORGANIZATION> over the past two days. <LOCATION>NEW YORK</LOCATION> (CNNMoney) For the first time in <ORGANIZATION>Microsoft</ORGANIZATION>'s history, founder <PERSON>Bill Gates</PERSON> is no longer its largest individual shareholder. In the <DATE>past two days</DATE>, Gates has sold nearly 8 million shares of <ORGANIZATION>Microsoft</ORGANIZATION> (<ORGANIZATION>MSFT</ORGANIZATION>, Fortune 500), bringing down his total to roughly 330 million. That puts him behind <ORGANIZATION>Microsoft</ORGANIZATION>'s former CEO <PERSON>Steve Ballmer</PERSON> who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire <PERSON>Ballmer</PERSON>, who was <ORGANIZATION>Microsoft</ORGANIZATION>'s CEO until <DATE>earlier this year</DATE>, 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 <ORGANIZATION>Bill & Melinda Gates</ORGANIZATION> foundation. The foundation has spent <MONEY>$28.3 billion</MONEY> fighting hunger and poverty since its inception back in <DATE>1997</DATE>. 
Bill Gates no longer Microsoft's biggest shareholder

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

NEW YORK - LRB - CNNMoney - RRB - 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 - LRB - OMSFT, Fortune 500 - RRB -, 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. O 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.
線上展示使用簡化詞類進行斷詞標記，僅供參考並且系統不再進行更新。線上服務斷詞和授權mirror site僅提供精簡詞類，結果也與舊版的展示系統不同。

自 2014/01/06 起，本斷詞系統已經處理過 28270134 篇文章

歐巴馬是美國的一位總統

文章的文字檔
擷取未知詞過程
包含未知詞的斷詞標記結果
未知詞列表

歐巴馬(Nb) 是(SHI) 美國(Nc) 的(DE) 一(Neu) 位(Nf) 總統(Na)
莎士比亞在淡江 遇見賽萬提斯
2016-04-26 02:27 聯合報 記者徐葳倫／淡水報導

分享4月23日是「世界閱讀日」，也是英國大文豪莎士比亞的生日與忌日，及「唐吉訶德」作者賽萬提斯逝世之日。英專起家的淡江大學舉辦「當莎士比亞遇見賽萬提斯」活動，規畫主題書展、彩繪活動，並添購新書，拉近學生與經典文學的距離。

首波登場的「主題書展」，展出2大文豪經典作品的原著、各種譯本以及DVD、電子書等數位化資料，校方也添購許多新書，吸引學生「搶鮮」閱讀經典名作。現場還規畫「彩繪大師」，讓學生發揮創意，畫出五彩繽紛的莎士比亞和賽萬提斯人像。

英語系四年級學生陳彥伶說，讀英語系接觸莎士比亞作品，但過去沒有舉辦書展時，這些作品都放在圖書館8樓，現在搬到1樓大廳陳列，不僅有很多莎士比亞、賽萬提斯的經典新書，還可藉由電子書、電影理解兩位作家，是以前沒有過的體驗。

英語系四年級學生鄭少淮表示，莎士比亞的「馬克白」、「羅密歐與茱麗葉」都已經讀過很多次，從經典文學中理解不同城市、國家的文化。

日文系學生賴喻郁說，原本只是喜歡塗鴉才來參加活動，後來才知道畫的是2個大文豪，接觸他們的作品，文學經典「原來離我這麼近」。

淡江大學外語學院院長陳小雀表示，莎士比亞的「to be, or not to be; that is the question」，賽萬提斯的「看得越多，行得越遠；書讀得越多，知識就越廣博」，都是來自文學的名言，校方希望用最簡單的方式，讓學生知道「文學不難」，就在你我身邊。

http://udn.com/news/story/7323/1653437-%E8%8E%8E%E5%A3%AB%E6%AF%94%E4%BA%9E%E5%9C%A8%E6%B7%A1%E6%B1%9F-%E9%81%87%E8%A6%8B%E8%B3%BD%E8%90%AC%E6%8F%90%E6%96%AF
自 2014/01/06 起，本斷詞系統已經處理過 28270134 篇文章

莎士比亞在淡江 遇見賽萬提斯
2016-04-26 02:27 聯合報 記者徐載倫 / 淡化報導

分享4月23日是「世界閱讀日」，也是英國大文豪莎士比亞的生日與忌日，及「唐吉訶德」作者塞萬提斯逝世之日，英專起家的淡江大學舉辦「當莎士比亞遇見賽萬提斯」活動，規劃主題書展、彩繪活動，並設色新書，拉近學生與經典文學的距離。

首波登場的「主題書展」，展出2大文豪經典作品的原著、各種譯本以及DVD、電子書等數位化資料，校方也添購許多新書，吸引學生「搶鮮」閱讀經典名作。現場還規畫「彩繪大師」，讓學生發揮創意，畫出五彩繽紛的莎士比亞和賽萬提斯人像。英語系四年級學生陳靜伶說，讀英語系接觸莎士比亞作品，但過去沒有舉辦書展時，這些作品都放在圖書館3樓，現在搬到1樓大廳陳列，不僅有很多莎士比亞、賽萬提斯的經典新書，還可藉由電子書、電影理解兩位作家，是以前沒有過的體驗。

英語系四年級學生鄭少華表示，莎士比亞的「馬克白」、「羅密歐與茱麗葉」都已經讀過很多次，從經典文學中理解不同城市、國家的文化。

日文系學生鄭紅郁說，原本只是喜歡塗鴉才來參加活動，後來才知道畫的是2個大文豪，接觸他們的作品，文學經典「原來離我這麼近」。

淡江大學外語學院院長陳小雀表示，莎士比亞的「to be, or not to be; that is the question」，賽萬提斯的「看得越多，行得越遠；書讀得越多，知識就越廣博」，都是來自文學的名言，校方希望用最簡單的方式，讓學生知道「文學不難」，就在你我身邊。
CKIP 中研院中文斷詞系統
http://ckipsvr.iis.sinica.edu.tw/
分享4月23日是「世界閱讀日」，也是英國大文豪莎士比亞的生日與忌日，及「唐吉訶德」作者塞萬提斯逝世之日。英專起家的淡江大學舉辦「當莎士比亞遇見賽萬提斯」活動，規畫主題書展、彩繪活動，並添購新書，拉近學生與經典文學的距離。

莎士比亞(Nb) 在(P) 淡江(Nb) 遇見(VC) 賽萬提(Nb) 斯(Nep) 2016(Neu) -(FW) 04(Neu) -(FW) 2 6 0 2(Neu) :(COLONCATEGORY) 2 7(Neu) 聯合報(Nb) 記者(Na) 徐葳倫(Nb) 淡水(Nc) 報導(Na) 分享(VJ) 4月(Nd) 23日(Nd) 是(SHI) 「(PARENTTHESISCATEGORY) 世界(Nc) 閱讀日(Na) 」(PARENTTHESISCATEGORY) ，(COMMACATEGORY) 也(D) 是(SHI) 英國(Nc) 大(VH) 文豪(Na) 莎士比亞(Nb) 的(DE) 生日(Na) 與(Caa) 忌日(Na) ，(COMMACATEGORY) 及(Caa) 「(PARENTTHESISCATEGORY) 唐吉訶德(Nb) 」(PARENTTHESISCATEGORY) 作者(Na) 賽萬提(Nb) 斯(Nep) 逝世(VH) 之(DE) 日(Na) 。(PERIODCATEGORY) 英(Nc) 專(D) 起家(VA) 的(DE) 淡江(Nb) 大學(Nc) 舉辦(VC) 「(PARENTTHESISCATEGORY) 當(P) 莎士比亞(Nb) 遇見(VC) 賽萬提(Nb) 斯(Nep) 」(PARENTTHESISCATEGORY) 活動(Na) ，(COMMACATEGORY) 規畫(VC) 主題(Na) 書展(Na) 、(PAUSECATEGORY) 彩繪(VC) 活動(Na) ，(COMMACATEGORY) 並(Cbb) 添購(VC) 新書(Na) ，(COMMACATEGORY) 拉近(VC) 學生(Na) 與(Caa) 經典(Na) 文學(Na) 的(DE) 距離(Na) 。(PERIODCATEGORY)
Vector Representations of Words

Word Embeddings

Word2Vec

GloVe
Modern NLP Pipeline
Facebook Research FastText

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 <u>skip-gram model</u> 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>TOKENIZE (ABSOLUTE)</th>
<th>TAG (ABSOLUTE)</th>
<th>PARSE (ABSOLUTE)</th>
<th>TOKENIZE (RELATIVE)</th>
<th>TAG (RELATIVE)</th>
<th>PARSE (RELATIVE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>spaCy</td>
<td>0.2ms</td>
<td>1ms</td>
<td>19ms</td>
<td>1x</td>
<td>1x</td>
<td>1x</td>
</tr>
<tr>
<td>CoreNLP</td>
<td>2ms</td>
<td>10ms</td>
<td>49ms</td>
<td>10x</td>
<td>10x</td>
<td>2.6x</td>
</tr>
<tr>
<td>ZPar</td>
<td>1ms</td>
<td>8ms</td>
<td>850ms</td>
<td>5x</td>
<td>8x</td>
<td>44.7x</td>
</tr>
<tr>
<td>NLTK</td>
<td>4ms</td>
<td>443ms</td>
<td>n/a</td>
<td>20x</td>
<td>443x</td>
<td>n/a</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><strong>Parsey McParseface</strong></td>
<td>94.15</td>
<td>89.08</td>
<td>94.77</td>
</tr>
<tr>
<td><strong>Martins et al. (2013)</strong></td>
<td>93.10</td>
<td>88.23</td>
<td>94.21</td>
</tr>
<tr>
<td><strong>Zhang and McDonald (2014)</strong></td>
<td>93.32</td>
<td>88.65</td>
<td>93.37</td>
</tr>
<tr>
<td><strong>Weiss et al. (2015)</strong></td>
<td>93.91</td>
<td>89.29</td>
<td>94.17</td>
</tr>
<tr>
<td><strong>Andor et al. (2016)</strong></td>
<td><strong>94.44</strong></td>
<td><strong>90.17</strong></td>
<td><strong>95.40</strong></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 Analytics with Python
spaCy: Natural Language Processing

Industrial-Strength Natural Language Processing

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. We like to think of spaCy as the Ruby on Rails of Natural Language Processing.

Blazing fast

spaCy excels at large-scale information extraction tasks. It’s written from the ground up in carefully memory-managed Cython. Independent research in 2015 found spaCy to be the fastest in the world. If your application needs to process entire web dumps, spaCy is the library you want to be using.

Deep learning

spaCy is the best way to prepare text for deep learning. It interoperates seamlessly with TensorFlow, PyTorch, scikit-learn, Gensim and the rest of Python’s awesome AI ecosystem. With spaCy, you can easily construct linguistically sophisticated statistical models for a variety of NLP problems.

https://spacy.io/
Python in Google Colab (Python101)

https://colab.research.google.com/drive/1FEG6DnGvwfUbeo4zJ1zTunjMqf2RkCrT

https://tinyurl.com/imtkupython101
Python in Google Colab (Python101)

https://colab.research.google.com/drive/1FEG6DnGvwfUbeo4zJ1zTunjMqf2RkCrT

Text Analytics and Natural Language Processing (NLP)

Python for Natural Language Processing

spaCy Chinese Model
Open Chinese Convert (OpenCC, 開放中文轉換)
Jieba 結巴中文分詞
Natural Language Toolkit (NLTK)
Stanza: A Python NLP Library for Many Human Languages

Text Processing and Understanding

NLTK (Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit)
NLP Zero to Hero

Natural Language Processing - Tokenization (NLP Zero to Hero, part 1)
Natural Language Processing - Sequencing - Turning sentence into data (NLP Zero to Hero, part 2)
Natural Language Processing - Training a model to recognize sentiment in text (NLP Zero to Hero, part 3)

Python for Natural Language Processing

spaCy

- spaCy: Industrial-Strength Natural Language Processing in Python
- Source: https://spacy.io/usage/spacy-101

```python
[1] 1 !python -m spacy download en_core_web_sm

[3] 1 import spacy
  2 nlp = spacy.load("en_core_web_sm")
  3 doc = nlp("Apple is looking at buying U.K. startup for $1 billion")
  4 for token in doc:
  5     print(token.text, token.pos_, token.dep_)
```

Apple PROPN nsubj
  is AUX aux
  looking VERB ROOT
  at ADP prep
  buying VERB pcomp
  U.K. PROPN compound
  startup NOUN dobj
  for ADP prep
  $ SYM quantmod
  1 NUM compound
  billion NUM pobj

https://tinyurl.com/imtkupython101
Python in Google Colab (Python101)

https://colab.research.google.com/drive/1FEG6DnGvwfUbeo4zJ1zTunjMqf2RkCrT

---

import spacy
nlp = spacy.load("en_core_web_sm")
doc = nlp("Apple is looking at buying U.K. startup for $1 billion")
import pandas as pd
cols = ("text", "lemma", "POS", "explain", "stopword")
rows = []
for t in doc:
    row = [t.text, t.lemma_, t.pos_, spacy.explain(t.pos_), t.is_stop]
    rows.append(row)
df = pd.DataFrame(rows, columns=cols)
df

---

https://tinyurl.com/imtkuppython101
Python in Google Colab (Python101)

https://colab.research.google.com/drive/1FEG6DnGvwfUbeo4zJ1zTunjMqf2RkCrT

https://tinyurl.com/imtkupython101
```python
# Importing Spacy
import spacy
nlp = spacy.load("en_core_web_sm")
text = "Stanford University is located in California. It is a great university."
doc = nlp(text)
for ent in doc.ents:
    print(ent.text, ent.label_)
```

- Stanford University ORG
- California GPE

```python
# Displacy for Visualizing spaCy Annotations
from spacy import displacy

# Sentence for Displacy
from spacy import displacy
text = "Stanford University is located in California. It is a great university."
doc = nlp(text)
displacy.render(doc, style="ent", jupyter=True)
```

- Stanford University ORG
- California GPE

https://tinyurl.com/imtkupython101
```python
1 from spacy import displacy
2 text = "Stanford University is located in California. It is a great university."
3 doc = nlp(text)
4 displacy.render(doc, style="ent", jupyter=True)
5 displacy.render(doc, style="dep", jupyter=True)
```

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

https://tinyurl.com/imtkupython101
Python in Google Colab (Python101)

https://colab.research.google.com/drive/1FEG6DnGvwfUbeo4zJ1zTunjMqf2RkCrT

https://tinyurl.com/imtkupyteron101
MONPA 囧拍：
正體中文斷詞、詞性標註以及命名實體辨識的多任務模型

```python
# MONPA 囧拍：正體中文斷詞、詞性標註以及命名實體辨識的多任務模型
# Source: https://github.com/monpa-team/monpa
!pip install monpa

import monpa
sentence = "銀行產業正在改變，金融機構欲挖角科技人才"
words = monpa.cut(sentence)
print(sentence)
print(" ".join(words))
result_pseg = monpa.pseg(sentence)
for item in result_pseg:
    print(item)
```

銀行產業正在改變，金融機構欲挖角科技人才
銀行 產業 正在 改變 ， 金融 機構 欲 挖角 科技 人才
('銀行', 'ORG')
('產業', 'Na')
('正在', 'D')
('改變', 'VC')
('，', 'COMMACATEGORY')
('金融', 'Na')
('機構', 'Nc')
('欲', 'VK')
('挖角', 'VA')
('科技', 'Na')
('人才', 'Na')

https://tinyurl.com/imtkupyterthon101
jieba

words = jieba.cut(sentence)

```python
import jieba
import jieba.posseg as pseg
sentence = "銀行產業正在改變，金融機構欲挖角科技人才"
words = jieba.cut(sentence)
print(sentence)
print(" ".join(words))
wordspos = pseg.cut(sentence)
result = ''
for word, pos in wordspos:
    print(word + ' (' + pos + ')')
    result = result + ' ' + word + ' (' + pos + ')
print(result.strip())
```

銀行產業正在改變，金融機構欲挖角科技人才
銀行 產業 正在 改變 ， 金融 機構 欲 挖角 科技人才
銀行 (n)
產業 (n)
正在 (t)
改變 (v)
， (x)
金融 (n)
機構 (n)
欲 (d)
挖角 (n)
科技人才 (n)
銀行 (n) 產業 (n) 正在 (t) 改變 (v) ， (x) 金融 (n) 機構 (n) 欲 (d) 挖角 (n) 科技人才 (n)

https://tinyurl.com/imtkupyter101
# 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="https://summari.es/">https://summari.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><strong>Question Answering</strong></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>SQuAD</td>
<td><a href="https://rajpurkar.github.io/SQuAD-explorer/">https://rajpurkar.github.io/SQuAD-explorer/</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/bdhwgra/quasar">https://github.com/bdhwgra/quasar</a></td>
</tr>
<tr>
<td></td>
<td>SearchQA</td>
<td><a href="https://github.com/nyu-dl/SearchQA">https://github.com/nyu-dl/SearchQA</a></td>
</tr>
<tr>
<td><strong>Semantic Parsing</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>Sentiment Analysis</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>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>Natural Language Inference</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><strong>Semantic Role Labeling</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>

Summary

• Text Analytics and Text Mining
• Natural Language Processing (NLP)
• Text Analytics with Python
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.
- Rajesh Arumugam (2018), Hands-On Natural Language Processing with Python: A practical guide to applying deep learning architectures to your NLP applications, Packt.
- Christopher D. Manning and Hinrich Schütze (1999), Foundations of Statistical Natural Language Processing, The MIT Press.
自然語言處理核心技術
與文字探勘
(Core Technologies of Natural Language Processing and Text Mining)

Time: 2020/05/15 (Fri) (9:10 -12:00)
Place: 國立臺北護理健康大學 (台北市明德路365號) G210
Host: 祝國忠 院長 (健康科技學院院長)

Min-Yuh Day
戴敏育
Associate Professor
副教授

Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系

http://mail.tku.edu.tw/myday/
2020-05-15