www.zhuanzhi.ai

自然语言处理（NLP）——专知荟萃

入门学习

1. 《数学之美》吴军

   这个书写得特别生动形象，没有太多公式，科普性质。看完对于 nlp 的许多技术原理都有了一点初步认识。可以说是自然语言处理最好的入门读物。

   https://book.douban.com/subject/10750155/

2. 如何在 NLP 领域第一次做成一件事 by 周明 微软亚洲研究院首席研究员、自然语言处理顶会 ACL 候任主席

   http://www.msra.cn/zh-cn/news/features/nlp-20161124

3. 深度学习基础 by 邱锡鹏 邱锡鹏 复旦大学 2017 年 8 月 17 日

206 页 PPT 带你全面梳理深度学习要点。

http://nlp.fudan.edu.cn/xpqiu/slides/20170817-CIPS-ATT-DL.pdf

https://nndl.github.io/

4. Deep learning for natural language processing 自然语言处理中的深度学习 by 邱锡鹏

   主要讨论了深度学习在自然语言处理中的应用。其中涉及的模型主要有卷积神经网络，递归神经网络，循环神经网络网络等，应用领域主要包括了文本生成，问答系统，机器翻译以及文本匹配等。

   http://nlp.fudan.edu.cn/xpqiu/slides/20160618*DL4NLP@CityU.pdf*

5. Deep Learning, NLP, and Representations （深度学习，自然语言处理及其表达)

   来自著名的 colah's blog，简要概述了 DL 应用于 NLP 的研究，重点介绍了 Word Embeddings。

   http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/

   翻译： http://blog.csdn.net/ycheng_sjtu/article/details/48520293

6. 《中文信息发展报告》 by 中国中文信息学会 2016 年 12 月

   是一份非常好的中文 NLP 总览性质的文档，通过这份报告可以了解中文和英文 NLP 主要的技术方向。

   http://cips-upload.bj.bcebos.com/cips2016.pdf

7. Deep Learning in NLP （一）词向量和语言模型 by Lai Siwei(来斯惟) 中科院自动化所 2013

   比较详细的介绍了 DL 在 NLP 领域的研究成果，系统地梳理了各种神经网络语言模型

   http://licstar.net/archives/328

8. 语义分析的一些方法(一，二，三) by 火光摇曳 腾讯广点通

   http://www.flickering.cn/ads/2015/02/

9. 我们是这样理解语言的-3 神经网络语言模型 by 火光摇曳 腾讯广点通

   总结了词向量和常见的几种神经网络语言模型

   http://www.flickering.cn/nlp/2015/03/

10. 深度学习 word2vec 笔记之基础篇 by falao*beiliu*

    http://blog.csdn.net/mytestmy/article/details/26961315

11. Understanding Convolutional Neural Networks for NLP 卷积神经网络在自然语言处理的应用 by WILDML

    http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp

    翻译：http://www.csdn.net/article/2015-11-11/2826192

12. The Unreasonable Effectiveness of Recurrent Neural Networks. 循环神经网络惊人的有效性 by Andrej Karpathy

    http://karpathy.github.io/2015/05/21/rnn-effectiveness/

    翻译： https://zhuanlan.zhihu.com/p/22107715

13. Understanding LSTM Networks 理解长短期记忆网络（LSTM NetWorks） by colah

    http://colah.github.io/posts/2015-08-Understanding-LSTMs/

    翻译：http://www.csdn.net/article/2015-11-25/2826323?ref=myread

14. 注意力机制（Attention Mechanism）在自然语言处理中的应用 by robert*ai*

    http://www.cnblogs.com/robert-dlut/p/5952032.html

15. 初学者如何查阅自然语言处理（NLP）领域学术资料 刘知远

    http://blog.sina.com.cn/s/blog_574a437f01019poo.html

进阶论文

Word Vectors

1. Word2vec Efficient Estimation of Word Representations in Vector Space

   http://arxiv.org/pdf/1301.3781v3.pdf

2. Doc2vec Distributed Representations of Words and Phrases and their Compositional-ity

http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

3. Word2Vec tutorialhttp://tensorflow.org/tutorials/word2vec/index.html in TensorFlow-http://tensorflow.org/
4. GloVe : Global vectors for word representation http://nlp.stanford.edu/projects/glove/glove.pdf
5. How to Generate a Good Word Embedding? 怎样生成一个好的词向量？ Siwei Lai, Kang Liu, Liheng Xu, Jun Zhao https://arxiv.org/abs/1507.05523
   code： https://github.com/licstar/compare
   note：http://licstar.net/archives/620
6. tweet2vec http://arxiv.org/abs/1605.03481
7. tweet2vec https://arxiv.org/abs/1607.07514
8. author2vec http://dl.acm.org/citation.cfm?id=2889382
9. item2vec http://arxiv.org/abs/1603.04259
10. lda2vec https://arxiv.org/abs/1605.02019
11. illustration2vec http://dl.acm.org/citation.cfm?id=2820907
12. tag2vec http://ktsaurabh.weebly.com/uploads/3/1/7/8/31783965/distributed*representations*for*contentbased*and*personalized*tag*recommendation.pdf*
13. category2vec http://www.anlp.jp/proceedings/annual*meeting/2015/pdf*dir/C43.pdf]
14. topic2vec http://arxiv.org/abs/1506.08422
15. image2vec http://arxiv.org/abs/1507.08818
16. app2vec http://paul.rutgers.edu/
17. prod2vec http://dl.acm.org/citation.cfm?id=2788627
18. metaprod2vec http://arxiv.org/abs/1607.07326
19. sense2vec http://arxiv.org/abs/1511.06388
20. node2vec http://www.kdd.org/kdd2016/papers/files/Paper*218.pdf*
21. subgraph2vec http://arxiv.org/abs/1606.08928

22. wordnet2vec http://arxiv.org/abs/1606.03335
23. doc2sent2vec http://research.microsoft.com/apps/pubs/default.aspx?id=264430
24. context2vec http://u.cs.biu.ac.il/
25. rdf2vec http://iswc2016.semanticweb.org/pages/program/acceptedpapers.html#re-search*ristoski*32
26. hash2vec http://arxiv.org/abs/1608.08940
27. query2vec http://www.cs.cmu.edu/
28. gov2vec http://arxiv.org/abs/1609.06616
29. novel2vec http://aics2016.ucd.ie/papers/full/AICS*2016*paper*48.pdf*
30. emoji2vec http://arxiv.org/abs/1609.08359
31. video2vec https://staff.fnwi.uva.nl/t.e.j.mensink/publications/habibian16pami.pdf
32. video2vec http://www.public.asu.edu/
33. sen2vec https://arxiv.org/abs/1610.08078
34. content2vec http://104.155.136.4:3000/forum?id=ryTYxh5ll
35. cat2vec http://104.155.136.4:3000/forum?id=HyNxRZ9xg
36. diet2vec https://arxiv.org/abs/1612.00388
37. mention2vec https://arxiv.org/abs/1612.02706
38. POI2vec http://www.ntu.edu.sg/home/boan/papers/AAAI17*Visitor.pdf*
39. wang2vec http://www.cs.cmu.edu/
40. dna2vec https://arxiv.org/abs/1701.06279
41. pin2vec https://labs.pinterest.com/assets/paper/p2pwww17.pdf, (cited blog(https://medium.com/thegraph/applyingdeeplearningtorelat-edpinsa6fee3c92f5e#.erb1i5mze))
42. paper2vec https://arxiv.org/abs/1703.06587
43. struc2vec https://arxiv.org/abs/1704.03165
44. med2vec http://www.kdd.org/kdd2016/papers/files/rpp0303choiA.pdf
45. net2vec https://arxiv.org/abs/1705.03881
46. sub2vec https://arxiv.org/abs/1702.06921

47. metapath2vec https://ericdongyx.github.io/papers/KDD17dongchawlaswa-mimetapath2vec.pdf
48. concept2vec http://knoesis.cs.wright.edu/sites/default/files/Concept2vecEvaluat-ingQualityofEmbeddingsforOntologicalConcepts%20%284%29.pdf
49. graph2vec http://arxiv.org/abs/1707.05005
50. doctag2vec https://arxiv.org/abs/1707.04596
51. skill2vec https://arxiv.org/abs/1707.09751
52. style2vec https://arxiv.org/abs/1708.04014
53. ngram2vec http://www.aclweb.org/anthology/D171023

Machine Translation

1. Neural Machine Translation by jointly learning to align and translate
   http://arxiv.org/pdf/1409.0473v6.pdf
2. Sequence to Sequence Learning with Neural Networks
   http://arxiv.org/pdf/1409.3215v3.pdf
   PPT: nips presentation[http://research.microsoft.com/apps/video/?id=239083
   seq2seq tutorialhttp://tensorflow.org/tutorials/seq2seq/index.html
3. Cross-lingual Pseudo-Projected Expectation Regularization for Weakly Supervised Learning
   http://arxiv.org/pdf/1310.1597v1.pdf
4. Generating Chinese Named Entity Data from a Parallel Corpus
   http://www.mt-archive.info/IJCNLP-2011-Fu.pdf
5. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools
   http://www.lrec-conf.org/proceedings/lrec2014/pdf/775*Paper.pdf*

Summarization

1. Extraction of Salient Sentences from Labelled Documents

arxiv: http://arxiv.org/abs/1412.6815

github: https://github.com/mdenil/txtnets

2. A Neural Attention Model for Abstractive Sentence Summarization. EMNLP 2015. Facebook AI Research

   arxiv: http://arxiv.org/abs/1509.00685

   github: https://github.com/facebook/NAMAS

   github(TensorFlow): https://github.com/carpedm20/neuralsummarytensorflow

3. A Convolutional Attention Network for Extreme Summarization of Source Code

   homepage: http://groups.inf.ed.ac.uk/cup/codeattention/

   arxiv: http://arxiv.org/abs/1602.03001

   github: https://github.com/jxieeducation/DIYDataScience/blob/master/paper-notes/2016/02/convattentionnetworksourcecodesummarization.md

4. Abstractive Text Summarization Using SequencetoSequence RNNs and Beyond. BM Watson & Université de Montréal

   arxiv: http://arxiv.org/abs/1602.06023

5. textsum: Text summarization with TensorFlow

   blog: https://research.googleblog.com/2016/08/textsummarizationwithtensor-flow.html

   github: https://github.com/tensorflow/models/tree/master/textsum

6. How to Run Text Summarization with TensorFlow

   blog: https://medium.com/@surmenok/howtoruntextsummarizationwithtensor-flowd4472587602d#.mll1rqgjg

   github: https://github.com/surmenok/TextSum

Text Classification

1. Convolutional Neural Networks for Sentence Classification

   arxiv: http://arxiv.org/abs/1408.5882

   github: https://github.com/yoonkim/CNN*sentence*

github: https://github.com/harvardnlp/sentconvtorch

github: https://github.com/alexanderrakhlin/CNNforSentenceClassificationinKeras

github: https://github.com/abhaikollara/CNNSentenceClassification

2. Recurrent Convolutional Neural Networks for Text Classification

   paper: http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9745/9552

   github: https://github.com/knok/rcnntextclassification

3. Characterlevel Convolutional Networks for Text Classification.NIPS 2015. "Text Understanding from Scratch"

   arxiv: http://arxiv.org/abs/1509.01626

   github: https://github.com/zhangxiangxiao/Crepe

   datasets: http://goo.gl/JyCnZq

   github: https://github.com/mhjabreel/CharCNN

4. A CLSTM Neural Network for Text Classification

   arxiv: http://arxiv.org/abs/1511.08630

   RationaleAugmented Convolutional Neural Networks for Text Classification

   arxiv: http://arxiv.org/abs/1605.04469

5. Text classification using DIGITS and Torch7

   github: https://github.com/NVIDIA/DIGITS/tree/master/examples/textclassification

6. Recurrent Neural Network for Text Classification with MultiTask Learning

   arxiv: http://arxiv.org/abs/1605.05101

7. Deep MultiTask Learning with Shared Memory. EMNLP 2016

   arxiv: https://arxiv.org/abs/1609.07222

8. Virtual Adversarial Training for SemiSupervised Text

   arxiv: http://arxiv.org/abs/1605.07725

   notes: https://github.com/dennybritz/deeplearning-papernotes/blob/master/notes/adversarial-text-classification.md

9. Bag of Tricks for Efficient Text Classification. Facebook AI Research

   arxiv: http://arxiv.org/abs/1607.01759

github: https://github.com/kemaswill/fasttext*torch*

github: https://github.com/facebookresearch/fastText

10. Actionable and Political Text Classification using Word Embeddings and LSTM

arxiv: http://arxiv.org/abs/1607.02501

Implementing a CNN for Text Classification in TensorFlow

blog: http://www.wildml.com/2015/12/implementingacnnfortextclassificationinten-sorflow/

11. fancycnn: Multiparadigm Sequential Convolutional Neural Networks for text classification

github: https://github.com/textclf/fancycnn

12. Convolutional Neural Networks for Text Categorization: Shallow Wordlevel vs. Deep Characterlevel

arxiv: http://arxiv.org/abs/1609.00718

Tweet Classification using RNN and CNN

github: https://github.com/ganeshjawahar/tweetclassify

13. Hierarchical Attention Networks for Document Classification. NAACL 2016

paper: https://www.cs.cmu.edu/

github: https://github.com/raviqqe/tensorflowfont2char2word2sent2doc

github: https://github.com/ematvey/deeptextclassifier

14. ACBLSTM: Asymmetric Convolutional Bidirectional LSTM Networks for Text Classification

arxiv: https://arxiv.org/abs/1611.01884

github: https://github.com/Ldpe2G/ACBLSTM

15. Generative and Discriminative Text Classification with Recurrent Neural Networks. DeepMind

arxiv: https://arxiv.org/abs/1703.01898

16. Adversarial Multitask Learning for Text Classification. ACL 2017

arxiv: https://arxiv.org/abs/1704.05742

data: http://nlp.fudan.edu.cn/data/

17. Deep Text Classification Can be Fooled. Renmin University of China

    arxiv: https://arxiv.org/abs/1704.08006

18. Deep neural network framework for multilabel text classification

    github: https://github.com/inspirehep/magpie

19. MultiTask Label Embedding for Text Classification

    arxiv: https://arxiv.org/abs/1710.07210

Dialogs

1. A Neural Network Approach toContext-Sensitive Generation of Conversational Responses. by Sordoni 2015. Generates responses to tweets.

   http://arxiv.org/pdf/1506.06714v1.pdf

2. Neural Responding Machine for Short-Text Conversation

   使用微博数据单轮对话正确率达到 75%

   http://arxiv.org/pdf/1503.02364v2.pdf

3. A Neural Conversation Model

   http://arxiv.org/pdf/1506.05869v3.pdf

4. Visual Dialog

   webiste: http://visualdialog.org/

   arxiv: https://arxiv.org/abs/1611.08669

   github: https://github.com/batra-mlp-lab/visdial-amt-chat

   github(Torch): https://github.com/batra-mlp-lab/visdial

   github(PyTorch): https://github.com/Cloud-CV/visual-chatbot

   demo: http://visualchatbot.cloudcv.org/

5. Papers, code and data from FAIR for various memory-augmented nets with application to text understanding and dialogue.

   post: https://www.facebook.com/yann.lecun/posts/10154070851697143

6. Neural Emoji Recommendation in Dialogue Systems

arxiv: https://arxiv.org/abs/1612.04609

Reading Comprehension

1. Text Understanding with the Attention Sum Reader Network. ACL 2016

   arxiv: https://arxiv.org/abs/1603.01547

   github: https://github.com/rkadlec/asreader

2. A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task

   arxiv: http://arxiv.org/abs/1606.02858

   github: https://github.com/danqi/rccnndailymail

3. Consensus Attentionbased Neural Networks for Chinese Reading Comprehension

   arxiv: http://arxiv.org/abs/1607.02250

   dataset: http://hfl.iflytek.com/chineserc/

4. Separating Answers from Queries for Neural Reading Comprehension

   arxiv: http://arxiv.org/abs/1607.03316

   github: https://github.com/dirkweissenborn/qa*network*

5. AttentionoverAttention Neural Networks for Reading Comprehension

   arxiv: http://arxiv.org/abs/1607.04423

   github: https://github.com/OlavHN/attentionoverattention

6. Teaching Machines to Read and Comprehend CNN News and Children Books using Torch

   github: https://github.com/ganeshjawahar/torchteacher

7. Reasoning with Memory Augmented Neural Networks for Language Comprehension

   arxiv: https://arxiv.org/abs/1610.06454

8. Bidirectional Attention Flow: Bidirectional Attention Flow for Machine Comprehension

   project page: https://allenai.github.io/biattflow/

   github: https://github.com/allenai/biattflow

9. NewsQA: A Machine Comprehension Dataset

arxiv: https://arxiv.org/abs/1611.09830

dataset: http://datasets.maluuba.com/NewsQA

github: https://github.com/Maluuba/newsqa

10. GatedAttention Readers for Text Comprehension

    arxiv: https://arxiv.org/abs/1606.01549

    github: https://github.com/bdhingra/gareader

11. Get To The Point: Summarization with PointerGenerator Networks. ACL 2017. Stanford University & Google Brain

    arxiv: https://arxiv.org/abs/1704.04368

    github: https://github.com/abisee/pointergenerator

Memory and Attention Models

1. Reasoning, Attention and Memory RAM workshop at NIPS 2015.
   http://www.thespermwhale.com/jaseweston/ram/
2. Memory Networks. Weston et. al 2014
   http://arxiv.org/pdf/1410.3916v10.pdf
3. End-To-End Memory Networks
   http://arxiv.org/pdf/1503.08895v4.pdf
4. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Taskshttp://arxiv.org/pdf/1502.05698v7.pdf
5. Evaluating prerequisite qualities for learning end to end dialog systemshttp://arxiv.org/pdf/1511.06931.pdf
6. Neural Turing Machines
   http://arxiv.org/pdf/1410.5401v2.pdf
7. Inferring Algorithmic Patterns with Stack-Augmented Recurrent Netshttp://arxiv.org/pdf/1503.01007v4.pdf
8. Reasoning about Neural Attention
   https://arxiv.org/pdf/1509.06664v1.pdf

9. A Neural Attention Model for Abstractive Sentence Summarization
   https://arxiv.org/pdf/1509.00685.pdf
10. Neural Machine Translation by Jointly Learning to Align and Translate
    https://arxiv.org/pdf/1409.0473v6.pdf
11. Recurrent Continuous Translation Models
    https://www.nal.ai/papers/KalchbrennerBlunsom_EMNLP13
12. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation
    https://arxiv.org/pdf/1406.1078v3.pdf
13. Teaching Machines to Read and Comprehend
    https://arxiv.org/pdf/1506.03340.pdf

reinforcement learning in nlp

1. Generating Text with Deep Reinforcement Learning
   https://arxiv.org/abs/1510.09202
2. Improving Information Extraction by Acquiring External Evidence with Reinforcement Learning
   https://arxiv.org/abs/1603.07954
3. Language Understanding for Text-based Games using Deep Reinforcement Learning
   http://people.csail.mit.edu/karthikn/pdfs/mud-play15.pdf
4. On-line Active Reward Learning for Policy Optimisation in Spoken Dialogue Systems
   https://arxiv.org/pdf/1605.07669v2.pdf
5. Deep Reinforcement Learning with a Natural Language Action Space
   https://arxiv.org/pdf/1511.04636v5.pdf
6. 基于 DQN 的开放域多轮对话策略学习 宋皓宇, 张伟男 and 刘挺 2017

GAN for NLP

1. Generating Text via Adversarial Training
   https://web.stanford.edu/class/cs224n/reports/2761133.pdf
2. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient
   https://arxiv.org/pdf/1609.05473.pdf
3. Adversarial Learning for Neural Dialogue Generation
   https://arxiv.org/pdf/1701.06547.pdf
4. GANs for sequence of discrete elements with the Gumbel-softmax distribution
   https://arxiv.org/pdf/1611.04051.pdf
5. Connecting generative adversarial network and actor-critic methods
   https://arxiv.org/pdf/1610.01945.pdf

综述

1. A Primer on Neural Network Models for Natural Language Processing Yoav Goldberg. October 2015. No new info, 75 page summary of state of the art.
   http://u.cs.biu.ac.il/~yogo/nnlp.pdf
2. Deep Learning for Web Search and Natural Language Processinghttps://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/wsdm2015.v3.pdf
3. Probabilistic topic models
   https://www.cs.princeton.edu/
4. Natural language processing: an introduction
   http://jamia.oxfordjournals.org/content/18/5/544.short
5. A unified architecture for natural language processing: Deep neural networks with multitask learning
   http://arxiv.org/pdf/1201.0490.pdf
6. A Critical Review of Recurrent Neural Networksfor Sequence Learningttp://arxiv.org/pdf/1506.00019v1.pdf

7. Deep parsing in Watsonhttp://nlp.cs.rpi.edu/course/spring14/deepparsing.pdf
8. Online named entity recognition method for microtexts in social networking services: A case study of twitterhttp://arxiv.org/pdf/1301.2857.pdf
9. 《基于神经网络的词和文档语义向量表示方法研究》 by Lai Siwei(来斯惟) 中科院自动化所 2016

   来斯惟的博士论文基于神经网络的词和文档语义向量表示方法研究，全面了解词向量、神经网络语言模型相关的内容。

   https://arxiv.org/pdf/1611.05962.pdf

视频课程

1. Introduction to Natural Language Processing（自然语言处理导论） 密歇根大学

   https://www.coursera.org/learn/natural-language-processing
2. 斯坦福 cs224d 2015 年课程 Deep Learning for Natural Language Processing by Richard Socher

   2015 classeshttps://www.youtube.com/playlist?list=PLmImxx8Char8dxWB9LRqdpCTmewaml96q
3. 斯坦福 cs224d 2016 年课程 Deep Learning for Natural Language Processing by Richard Socher. Updated to make use of Tensorflow.

   https://www.youtube.com/playlist?list=PLmImxx8Char9Ig0ZHSyTqGsdhb9weEGam
4. 斯坦福 cs224n 2017 年课程 Deep Learning for Natural Language Processing by Chris Manning Richard Socher

   http://web.stanford.edu/class/cs224n/
5. Natural Language Processing - by 哥伦比亚大学 Mike Collins

   https://www.coursera.org/learn/nlangp
6. NLTK with Python 3 for Natural Language Processing by Harrison Kinsley. Good tutorials with NLTK code implementation.

https://www.youtube.com/playlist?list=PLQVvvaa0QuDf2JswnfiGkliBInZnIC4HL

7. Computational Linguistics by Jordan Boyd-Graber . Lectures from University of Maryland. https://www.youtube.com/playlist?list=PLQVvvaa0QuDf2JswnfiGkliBInZnIC4HL
8. Natural Language Processing - Stanford by Dan Jurafsky & Chris Manning. https://www.youtube.com/playlist?list=PL6397E4B26D00A269Previously on coursera. Lecture Noteshttp://www.mohamedaly.info/teaching/cmp-462-spring-2013

Tutorial

1. Deep Learning for Natural Language Processing without Magichttp://www.socher.org/index.php/DeepLearningTutorial/DeepLearningTutorial
2. A Primer on Neural Network Models for Natural Language Processinghttps://arxiv.org/abs/1510.00726
3. Deep Learning for Natural Language Processing: Theory and Practice Tutorialhttps://www.microsoft.com/en-us/research/publication/deep-learning-for-natural-language-processing-theory-and-practice-tutorial/
4. Recurrent Neural Networks with Word Embeddingshttp://deeplearning.net/tutorial/rnnslu.html
5. LSTM Networks for Sentiment Analysishttp://deeplearning.net/tutorial/lstm.html
6. Semantic Representations of Word Senses and Concepts 语义表示 ACL 2016 Tutorial by José Camacho-Collados, Ignacio Iacobacci, Roberto Navigli and Mohammad Taher Pilehvar http://acl2016.org/index.php?article_id=58 http://wwwusers.di.uniroma1.it/~collados/Slides_ACL16Tutorial_SemanticRepresentation.pdf
7. ACL 2016 Tutorial: Understanding Short Texts 短文本理解 http://www.wangzhongyuan.com/tutorial/ACL2016/Understanding-Short-Texts/

8. Practical Neural Networks for NLP EMNLP 2016
   https://github.com/clab/dynet_tutorial_examples
9. Structured Neural Networks for NLP: From Idea to Code
   https://github.com/neubig/yrsnlp-2016/blob/master/neubig16yrsnlp.pdf
10. Understanding Deep Learning Models in NLP
    http://nlp.yvespeirsman.be/blog/understanding-deeplearning-models-nlp/
11. Deep learning for natural language processing, Part 1
    https://softwaremill.com/deep-learning-for-nlp/
12. TensorFlow Tutorial on Seq2Seq Models
    https://www.tensorflow.org/tutorials/seq2seq/index.html
13. Natural Language Understanding with Distributed Representation Lecture Note by Cho
    https://github.com/nyu-dl/NLP*DL*Lecture*Note*
14. Michael Collinshttp://www.cs.columbia.edu/ - one of the best NLP teachers. Check out the material on the courses he is teaching.
15. Several tutorials by Radim Řehůřekhttps://radimrehurek.com/gensim/tutorial.html on using Python and gensimhttps://radimrehurek.com/gensim/index.html to process corpora and conduct Latent Semantic Analysis and Latent Dirichlet Allocation experiments.
16. Natural Language Processing in Actionhttps://www.manning.com/books/natural-language-processing-in-action - A guide to creating machines that understand human language.

图书

1. 《数学之美》（吴军）
   科普性质,看完对于 nlp 的许多技术原理都会有初步认识
2. 《自然语言处理综论》（Daniel Jurafsky）

这本书是冯志伟老师翻译的 作者是 Daniel Jurafsky，在 coursera 上面有他的课程。

本书第三版正尚未出版，但是英文版已经全部公开。

Speech and Language Processing (3rd ed. draft) by Dan Jurafsky and James H. Martin

https://web.stanford.edu/~jurafsky/slp3/

3. 《自然语言处理简明教程》（冯志伟）
4. 《统计自然语言处理(第 2 版)》（宗成庆）
5. 清华大学刘知远老师等合著的《互联网时代的机器学习和自然语言处理技术大数据智能》，科普性质。

领域专家

国内

1. 清华大学

   NLP 研究：孙茂松主要从事一些中文文本处理工作，比如中文文本分类，中文分词。刘知远从事关键词抽取，表示学习，知识图谱以及社会计算。刘洋从事数据驱动的机器学习。

   情感分析：黄民烈

   信息检索：刘奕群、马少平

   语音识别——王东

   社会计算：唐杰
2. 哈尔滨工业大学

   社会媒体处理：刘挺、丁效

   情感分析：秦兵 车万翔
3. 中科院

   语言认知模型：王少楠，宗成庆

   信息抽取：孙乐、韩先培

信息推荐与过滤：王斌（中科院信工所）、鲁骁（国家计算机网络应急中心）

自动问答：赵军、刘康，何世柱（中科院自动化研究所）

机器翻译：张家俊、宗成庆（中科院自动化研究所）

语音 合成——陶建华（中科院自动化研究所）

文字识别：刘成林（中科院自动化研究所）

文本匹配：郭嘉丰

4. 北京大学

篇章分析：王厚峰、李素建

自动文摘，情感分析：万小军、姚金戈

语音技术：说话人识别——郑方

多模态信息处理：陈晓鸥

冯岩松

5. 复旦大学

语言表示与深度学习：黄萱菁、邱锡鹏

6. 苏州大学

词法与句法分析：李正华、陈文亮、张民

语义分析：周国栋、李军

机器翻译：熊德意

7. 中国人民大学

表示学习，推荐系统：赵鑫

8. 微软亚洲研究院自然语言计算组

周明 刘铁岩 谢幸

9. 头条人工智能实验室

李航

10. 华为诺亚

前任 李航 吕正东

国际

1. 斯坦福大学

   知名的 NLP 学者：Daniel Jurafsky, Christopher Manning, Percy Liang 和 Chris Potts, Richard Socher

   NLP 研究：Jurafsky 和科罗拉多大学波尔得分校的 James Martin 合著自然语言处理方面的教材。这个 NLP 研究组从事几乎所有能够想象到的研究方向。今天 NLP 领域最被广泛使用的句法分析器和词性标注工具可能都是他们负责开发的。

   http://nlp.stanford.edu/

2. 加州大学圣巴巴拉分校

   知名 NLP 学者：William Wang(王威廉), Fermin Moscoso del Prado Martin

   NLP 研究：William 研究方向为信息抽取和机器学习，Fermin 研究方向为心理语言学和计量语言学。

   http://www.cs.ucsb.edu/~william William Wang(王威廉)经常在微博分享关于 NLP 的最近进展和趣事，几乎每条都提供高质量的信息。

   微博：https://www.weibo.com/u/1657470871

3. 加州大学圣迭戈分校

   知名的 NLP 学者：Lawrence Saul(Roger Levy 今年加入 MIT)

   NLP 研究：主要研究方向是机器学习，NLP 相关的工作不是很多，但是在计算心理语言学有些比较有趣的工作。

   http://grammar.ucsd.edu/cpl/

4. 加州大学圣克鲁兹分校

   知名 NLP 学者：Pranav Anand, Marilyn Walker 和 LiseGetoor

   NLP 研究：Marilyn Walker 主要研究方向为对话系统。

   http://people.ucsc.edu/~panand/

   http://users.soe.ucsc.edu/~maw/

5. 卡内基梅隆大学

   知名 NLP 学者：Jaime Carbonell,Alon Lavie, Carolyn Rosé, Lori Levin, Roni Rosenfeld, Chris Dyer (休假中), Alan Black, Tom Mitchell 以及 Ed Hovy

NLP 研究：在多个 NLP 领域做了大量工作，包括机器翻译、文摘、交互式对话系统、语音、信息检索以及工作最为突出的机器学习领域。Chris 主要方向为机器学习和机器翻译交叉研究，做了一些非常出色的工作。虽然 Tom Mitchell 属于机器学习系而不是语言技术研究所，但是由于他在 CMU 的“永不停息的语言学习者”项目中的重要贡献，我们必须在这里提到他。

http://www.cs.cmu.edu/~nasmith/nlp-cl.html

http://www.lti.cs.cmu.edu/

6. 芝加哥大学(以及芝加哥丰田科技学院 TTIC)

   知名 NLP 学者：John Lafferty, John Goldsmith, Karen Livescu, MichelGalley (兼职)和 Kevin Gimpel.

   NLP 研究：芝加哥大学以及丰田科技学院有许多机器学习、语音以及 NLP 方向的研究人员。John Lafferty 是一个传奇性人物，其参与原始 IBM MT 模型研发，同时也是 CRF 模型的发明人之一。Goldsmith 的团队是无监督的形态归纳法(unsupervised morphology induction)的先驱。Karen 主要研究方向为语音，特别是对发音方式的建模。Michel 主要研究结构化预测问题，特别是统计机器翻译。Kevin 在许多结构化预测问题上都做出出色工作。

   http://ai.cs.uchicago.edu/faculty/

   http://www.ttic.edu/faculty.php

7. 科罗拉多大学博尔德分校

   知名 NLP 学者：Jordan Boyd-Graber, Martha Palmer, James Martin,Mans Hulden 以及 Michael Paul

   NLP 研究：Martha Palmer 主要研究资源标注和创建，其中代表性有 FrameNet, VerbNet, OntoNotes 等，此外其也在词汇语义学（Lexical semantics）做了一些工作。Jim Martin 主要研究语言的向量空间模型，此外与 Dan Jurafsky(以前在科罗拉多大学博尔德分校，之后去了斯坦福)合作编写语音和语言处理的著作。Hulden, Boyd-Graber 和 Paul 最近加入科罗拉多大学博尔德分校。Hulden 主要使用有穷状态机相关技术，做一些音位学(phonology)和形态学(morphology)相关工作，

Boyd-Graber 主要研究主题模型和机器学习在问答、机器翻译上的应用。Michael Paul 主要研究机器学习在社交媒体监控(social media monitoring)上的应用。

http://clear.colorado.edu/start/index.php

8. 哥伦比亚大学

知名的 NLP 学者：有多位 NLP 领域顶级学者，Kathy McKeown, Julia Hirschberg, Michael Collins(休假中), Owen Rambow, Dave Blei, Daniel Hsu 和 Becky Passonneau

NLP 研究:在文摘、信息抽取以及机器翻译上面做了大量的研究。Julia 团队主要在语音领域做一些研究。Michael Collins 是从 MIT 离职后加入哥伦比亚 NLP 团队的，其主要研究内容为机器翻译和 parsing。DaveBlei 和 Daniel Hsu 是机器学习领域翘楚，偶尔也会做一些语言相关的工作。

http://www1.cs.columbia.edu/nlp/index.cgi

9. 康纳尔大学

NLP 知名学者：Lillian Lee, Thorsten Joachims, Claire Cardie, Yoav Artzi, John Hale,David Mimno, Cristian Danescu-Niculescu-Mizil 以及 Mats Rooth

NLP 研究：在机器学习驱动 NLP 方面有许多有趣的研究。Lillian 与其学生做了许多独辟蹊径的研究，如电影评论分类，情感分析等。Thorsten，支持向量机的先驱之一，SVMlight 的作者。John 研究内容包括计算心理语言学和认知科学。Mats 研究领域包括语义学和音位学。Claire Cardie 在欺诈性评论方面的研究室非常有影响的。Yoav Artzi 在语义分析和情景化语言理解方面有许多重要的工作。David Mimno 在机器学习和数位人文学（digital humanities）交叉研究的顶级学者。

http://nlp.cornell.edu/

10. 佐治亚理工学院

知名 NLP 学者：Jacob Eisenstein 和 Eric Gilbert

NLP 研究：Jacob 在机器学习和 NLP 交叉领域做了一些突出性的工作，特别是无监督学习以及社交媒体领域。在 MIT,他是 Regina Barzilay 的学生，在 CMU 和 UIUC 分别与 Noah Smith、Dan Roth 做博士后研究。此外，Eric Gilbert 在计

算社会学(computationalsocial science)上做了许多研究。这些研究经常与 NLP 进行交叉。

http://www.cc.gatech.edu/~jeisenst/

http://smlv.cc.gatech.edu/

http://comp.social.gatech.edu/

11. 伊利诺伊大学厄巴纳-香槟分校

知名的 NLP 学者：Dan Roth, Julia Hockenmaier, ChengXiang Zhai, Roxana Girju 和 Mark Hasegawa-Johnson

NLP 研究：机器学习在 NLP 应用，NLP 在生物学上应用（BioNLP），多语言信息检索，计算社会学，语音识别

http://nlp.cs.illinois.edu/

12. 约翰·霍普金斯大学(JHU)

知名 NLP 学者：Jason Eisner, Sanjeev Khudanpur, David Yarowsky,Mark Dredze, Philipp Koehn 以及 Ben van Durme，详细情况参考链接（http://web.jhu.edu/HLTCOE/People.html）

NLP 研究：约翰·霍普金斯有两个做 NLP 的研究中心，即 the Center for Language and Speech Processing (CLSP) 和 the Human Language Technology Center of Excellence(HLTCOE)。他们的研究几乎涵盖所有 NLP 领域，其中机器学习、机器翻译、parsing 和语音领域尤为突出。Fred Jelinek,语音识别领域的先驱，其于 2010 年 9 月去世，但是语音识别研究一直存在至今。在过去十年内，JHU 的 NLP summer research workshop 产生出许多开创性的研究和工具。

http://web.jhu.edu/HLTCOE/People.html

http://clsp.jhu.edu/

13. 马里兰大学学院市分校

知名的 NLP 学者：Philip Resnik, Hal Daumé, Marine Carpuat, Naomi Feldman

NLP 研究：和 JHU 一样，其 NLP 研究比较全面。比较大的领域包括机器翻译，机器学习，信息检索以及计算社会学。此外，还有一些团队在计算心理语言学上做一些研究工作。

https://wiki.umiacs.umd.edu/clip/index.php/Main_Page

14. 马萨诸塞大学阿默斯特分校

知名的 NLP 学者：Andrew McCallum, James Allan (不是罗彻斯特大学的 James Allan), Brendan O'Connor 和 W. Bruce Croft

NLP 研究：机器学习和信息检索方向顶尖研究机构之一。Andrew 的团队在机器学习在 NLP 应用方面做出许多重要性的工作，例如 CRF 和无监督的主题模型。其与 Mark Dredze 写了一篇指导性文章关于“如何成为一名成功 NLP/ML Phd”。Bruce 编写了搜索引擎相关著作“搜索引擎：实践中的信息检索”。James Allan 是现代实用信息检索的奠基人之一。IESL 实验室在信息抽取领域做了大量的研究工作。另外，其开发的 MalletToolkit，是 NLP 领域非常有用工具包之一。

http://ciir.cs.umass.edu/personnel/index.html

http://www.iesl.cs.umass.edu/

http://people.cs.umass.edu/~brenocon/complang_at_umass/

http://mallet.cs.umass.edu/

15. 麻省理工学院

知名的 NLP 学者：Regina Barzilay, Roger Levy (2016 年加入)以及 Jim Glass

NLP 研究：Regina 与 ISI 的 Kevin Knight 合作在文摘、语义、篇章关系以及古代文献解读做出过极其出色的工作。此外，开展许多机器学习相关的工作。另外，有一个比较大团队在语音领域做一些研究工作，Jim Glass 是其中一员。

http://people.csail.mit.edu/regina/

http://groups.csail.mit.edu/sls//sls-blue-noflash.shtml

16. 纽约大学

知名 NLP 学者：Sam Bowman, Kyunghyun Cho, Ralph Grishman

NLP 研究：Kyunghyun and Sam 刚刚加入 NLP 团队，主要研究包括机器学习/深度学习在 NLP 以及计算语言学应用。与 CILVR machine learning group、Facebook AI Research 以及 Google NYC 有紧密联系。

https://wp.nyu.edu/ml2/

17. 北卡罗来纳大学教堂山分校

知名的 NLP 学者：Mohit Bansal, Tamara Berg, Alex Berg, Jaime Arguello

NLP 研究：Mohit 于 2016 年加入该团队，主要研究内容包括 parsing、共指消解、分类法(taxonomies)以及世界知识。其最近的工作包括多模态语义、类人语言理解(human-like language understanding)以及生成/对话。Tamara 和 Alex Berg 在语言和视觉领域发了许多有影响力的论文，现在研究工作主要围绕 visual referring expressions 和 visual madlibs。Jaime 主要研究对话模型、web 搜索以及信息检索。UNC 语言学系还有 CL 方面一些研究学者，例如 Katya Pertsova（计算形态学(computational morphology)）以及 Misha Becker(computational language acquisition)

http://www.cs.unc.edu/~mbansal/

http://www.tamaraberg.com/

http://acberg.com/

https://ils.unc.edu/~jarguell/

18. 北德克萨斯大学

知名的 NLP 学者：Rodney Nielsen

NLP 研究：Rodney 主要研究 NLP 在教育中的应用，包括自动评分、智能教学系统

http://www.rodneynielsen.com/

19. 东北大学

知名 NLP 学者：David A. Smith, Lu Wang, Byron Wallace

NLP 研究：David 在数位人文学（digital humanities）特别是语法方面做了许多重要的工作。另外，其受 google 资助做一些语法分析工作，调研结构化语言(structural language)的变化。Lu Wang 主要在文摘、生成以及论元挖掘(argumentation mining)、对话、计算社会学的应用以及其他交叉领域。Byron Wallace 的工作包括文本挖掘、机器学习，以及它们在健康信息学上的应用。

http://www.northeastern.edu/nulab/

20. 纽约市立学院（CUNY）

知名 NLP 学者：Martin Chodorow 和 WilliamSakas

NLP 研究：Martin Chodorow，ETS 顾问，设计 Leacock-Chodorow WordNet 相似度指标计算公式，在语料库语言学、心理语言学有一些有意义的工作。此外 NLP@CUNY 每个月组织一次讨论，有很多高水平的讲者。

http://nlpatcuny.cs.qc.cuny.edu/

21. 俄亥俄州立大学（OSU）

知名的 NLP 学者：Eric Fosler-Lussier, Michael White, William Schuler,Micha Elsner, Marie-Catherine de Marneffe, Simon Dennis, 以及 Alan Ritter, Wei Xu

NLP 研究：Eric 的团队研究覆盖从语音到语言模型到对话系统的各个领域。Michael 主要研究内容包括自然语言生成和语音合成。William 团队研究内容主要有 parsing、翻译以及认知科学。Micha 在 Edinburgh 做完博士后工作，刚刚加入 OSU，主要研究内容包括 parsing、篇章关系、narrative generation 以及 language acquisition。Simon 主要做一些语言认知方面的工作。Alan 主要研究 NLP 在社交媒体中应用和弱监督学习。Wei 主要做一些社交媒体、机器学习以及自然语言生成的交叉研究。

http://cllt.osu.edu/

22. 宾夕法尼亚大学

知名的 NLP 学者：Arvind Joshi, Ani Nenkova, Mitch Marcus, Mark Liberman 和 Chris Callison-Burch

NLP 研究：这里是 LTAG(Lexicalized Tree Adjoining Grammar)、Penn Treebank 的起源地，他们做了大量 parsing 的工作。Ani 从事多文档摘要的工作。同时，他们也有很多机器学习方面的工作。Joshi 教授获得 ACL 终身成就奖。

http://nlp.cis.upenn.edu/

23. 匹兹堡大学

知名的 NLP 学者：Rebecca Hwa, Diane Litman 和 Janyce Wiebe

NLP 研究：Diane Litman 从事对话系统和评价学生表现方面的研究工作。Janyce Wiebe 在情感 / 主观分析任务上有一定的影响力。

http://www.isp.pitt.edu/research/nlp-info-retrieval-group

24. 罗切斯特大学

知名的 NLP 学者：Len Schubert, James Allen 和 Dan Gildea

NLP 研究：James Allen 是篇章关系和对话任务上最重要的学者之一，他的许多学生在这些领域都很成功，如在 AT&T 实验室工作的 Amanda Stent，在南加州大学资讯科学研究院 USC/ISI 的 David Traum。Len Schubert 是计算语义学领域的重要学者，他的许多学生是自然语言处理领域内的重要人物，如在 Hopkins（约翰•霍普金斯大学）的 Ben Van Durme。Dan 在机器学习、机器翻译和 parsing 的交叉研究上有一些有趣的工作。

http://www.cs.rochester.edu/~james/

http://www.cs.rochester.edu/~gildea/

http://www.cs.rochester.edu/~schubert/

25. 罗格斯大学

知名的 NLP 学者：Nina Wacholder 和 Matthew Stone

NLP 研究：Smaranda 和 Nina 隶属通讯与信息学院(School of Communication and Information)的 SALTS(Laboratory for the Study of Applied Language Technology and Society)实验室。他们不属于计算机专业。Smaranda 主要做自然语言处理方面的工作，包括机器翻译、信息抽取和语义学。Nina 虽然之前从事计算语义学研究，但是目前更专注于认知方向的研究。Matt Stone 是计算机专业的，从事形式语义（formal semantics）和多模态交流（multimodal communication）的研究。

http://salts.rutgers.edu/

http://www.cs.rutgers.edu/~mdstone/

26. 南加州大学

知名的 NLP 学者：信息科学学院有许多优秀的自然语言处理专家，如 Kevin Knight, Daniel Marcu, Jerry Hobbs 和 Zornitsa Kozareva

NLP 研究：他们从事几乎所有可能的自然语言处理研究方向。其中主要的领域包括机器翻译、文本解密（decipherment）和信息抽取。Jerry 主要从事篇章关系和对话任务的研究工作。Zornitsa 从事关系挖掘和信息抽取的研究工作。

http://nlg.isi.edu/

27. 加州大学伯克利分校

知名的 NLP 学者：Dan Klein, Marti Hearst, David Bamman

NLP 研究：可能是做 NLP 和机器学习交叉研究的最好研究机构之一。Dan 培养了许多优秀学生，如 Aria Haghighi, John DeNero 和 Percy Liang。

http://nlp.cs.berkeley.edu/Members.shtml

28. 德克萨斯大学奥斯汀分校

知名的 NLP 学者：Ray Mooney, Katrin Erk, Jason Baldridge 和 Matt Lease

NLP 研究：Ray 是自然语言处理与人工智能领域公认的资深教授。他广泛的研究方向包括但不限于机器学习、认知科学、信息抽取和逻辑。他仍然活跃于研究领域并且指导很多学生在非常好的期刊或者会议上发表文章。Katrin 专注于计算语言学的研究并且也是该领域著名研究者之一。Jason 从事非常酷的研究，和半监督学习、parsing 和篇章关系的交叉领域相关。Matt 研究信息检索的多个方面，最近主要发表了许多在信息检索任务上使用众包技术的论文。

http://www.utcompling.com/

http://www.cs.utexas.edu/~ml/

29. 华盛顿大学

知名的 NLP 学者：Mari Ostendorf, Jeff Bilmes, Katrin Kirchoff, Luke Zettlemoyer, Gina Ann Levow, Emily Bender, Noah Smith, Yejin Choi 和 Fei Xia

NLP 研究：他们的研究主要偏向于语音和 parsing，但是他们也有通用机器学习的相关工作。他们最近开始研究机器翻译。Fei 从事机器翻译、parsing、语言学和 bio-NLP 这些广泛的研究工作。Emily 从事语言学和自然语言处理的交叉研究工作，并且负责著名的计算语言学相关的专业硕士项目。Gina 从事对话、语音和信息检索方向的工作。学院正在扩大规模，引入了曾在卡内基梅隆大学担任教职的 Noah 和曾在纽约州立大学石溪分校担任教职的 Yejin。

https://www.cs.washington.edu/research/nlp

https://ssli.ee.washington.edu/

http://turing.cs.washington.edu/

http://depts.washington.edu/lingweb/

30. 威斯康辛大学麦迪逊分校

知名的 NLP 学者：Jerry Zhu

NLP 研究：Jerry 更加偏向机器学习方面的研究，他主要从事半监督学习的研究工作。但是，最近也在社交媒体分析方向发表论文。

http://pages.cs.wisc.edu/~jerryzhu/publications.html

31. 剑桥大学

知名的 NLP 学者：Stephen Clark, Simone Teufel, Bill Byrne 和 Anna Korhonen

NLP 研究：有很多基于 parsing 和信息检索的工作。最近，也在其他领域发表了一些论文。Bill 是语音和机器翻译领域非常知名的学者。

http://www.cl.cam.ac.uk/research/nl/

32. 爱丁堡大学

知名的 NLP 学者：Mirella Lapata, Mark Steedman, Miles Osborne, Steve Renals, Bonnie Webber, Ewan Klein, Charles Sutton, Adam Lopez 和 Shay Cohen

NLP 研究：他们在几乎所有的领域都有研究，但我最熟悉的工作是他们在统计机器翻译和基于机器学习方法的篇章连贯性方面的研究。

http://www.ilcc.inf.ed.ac.uk/

33. 新加坡国立大学

知名的 NLP 学者：Hwee Tou Ng

NLP 研究：Hwee Tou 的组主要从事机器翻译（自动评价翻译质量是焦点之一）和语法纠错(grammatical error correction)方面的研究。他们也发表了一些词义消歧和自然语言生成方面的工作。Preslav Nakov 曾是这里的博士后，但现在去了卡塔尔。

http://www.comp.nus.edu.sg/~nlp/home.html

34. 牛津大学

知名的 NLP 学者：Stephen Pulman 和 Phil Blunsom

NLP 研究：Stephen 在第二语言学习(second language learning)和语用学方面做了许多工作。Phil 很可能是机器学习和机器翻译交叉研究领域的领导者之一。

http://www.clg.ox.ac.uk/people.html

35. 亚琛工业大学

知名的 NLP 学者：Hermann Ney

NLP 研究：Aachen 是世界上研究语音识别和机器翻译最好的地方之一。任何时候，都有 10-15 名博士生在 Hermann Ney 的指导下工作。一些统计机器翻译最厉害的人来自 Aachen，如 Franz Och（Google Translate 负责人），Richard Zens（目前在 Google）和 Nicola Ueffing（目前在 NRC 国家研究委员会，加拿大）。除了通常的语音和机器翻译的研究，他们同时在翻译和识别手语(sign language)方面有一些有趣的工作。但是，在其他 NLP 领域没有许多相关的研究。

http://www-i6.informatik.rwth-aachen.de/web/Homepage/index.html

36. 谢菲尔德大学

知名的 NLP 学者：Trevor Cohn, Lucia Specia, Mark Stevenson 和 Yorick Wilks

NLP 研究：Trevor 从事机器学习与自然语言处理交叉领域的研究工作，主要关注图模型和贝叶斯推理(Bayesian inference)。Lucia 是机器翻译领域的知名学者并在这个领域组织（或共同组织）了多个 shared tasks 和 workshops。Mark 的组从事计算语义学和信息抽取与检索的研究工作。Yorick 获得 ACL 终身成就奖，并在大量的领域从事研究工作。最近，他研究语用学和信息抽取。

http://nlp.shef.ac.uk/

37. 达姆施塔特工业大学, The Ubiquitous Knowledge Processing 实验室

知名的 NLP 学者：Irena Gurevych, Chris Biemann 和 Torsten Zesch

NLP 研究：这个实验室进行许多领域的研究工作：计算词汇语义学（computational lexical semantics）、利用和理解维基百科以及其他形式的 wikis、情感分析、面向教育的 NLP 以及数位人文学（digital humanities）。Irena 是计算语言学（CL）和自然语言处理（NLP）领域的著名学者。Chris 曾在 Powerset 工作，现在在语义学领域有一些有趣的项目。Torsten 有许多学生从事不同领域的研究。UKP 实验室为（NLP）社区提供了许多有用的软件，JWPL（Java Wikipedia Library）就是其中之一。

http://www.ukp.tu-darmstadt.de/

38. 多伦多大学

知名的 NLP 学者：Graeme Hirst, Gerald Penn 和 Suzanne Stevenson

NLP 研究：他们有许多词汇语义学（lexical semantics）的研究以及一些 parsing 方面的研究。Gerald 从事语音方面的研究工作。

http://www.cs.utoronto.ca/compling/

39. 伦敦大学学院

知名的 NLP 学者：Sebastian Riedel

NLP 研究：Sebastian 主要从事自然语言理解方面的研究工作，大部分是知识库和语义学相关的工作。

http://mr.cs.ucl.ac.uk/

会议

自然语言处理国际会议

1. Association for Computational Linguistics (ACL)
2. Empirical Methods in Natural Language Processing (EMNLP)
3. North American Chapter of the Association for Computational Linguistics
4. International Conference on Computational Linguistics (COLING)
5. Conference of the European Chapter of the Association for Computational Linguistics (EACL)

相关包含 NLP 内容的其他会议

1. SIGIR: Special Interest Group on Information Retrieval
2. AAAI: Association for the Advancement of Artificial Intelligence
3. ICML: International Conference on Machine Learning
4. KDD: Association for Knowledge Discovery and Data Mining
5. ICDM: International Conference on Data Mining

期刊

1. Journal of Computational Linguistics
2. Transactions of the Association for Computational Linguistics
3. Journal of Information Retrieval
4. Journal of Machine Learning

国内会议 通常都包含丰富的讲习班和 Tutorial 公开的 PPT 都是很好的学习资源

1. CCKS 全国知识图谱与语义计算大会
   http://www.ccks2017.com/index.php/att/ 成都 8 月 26-8 月 29
2. SMP 全国社会媒体处理大会
   http://www.cips-smp.org/smp2017/ 北京 9.14-9.17
3. CCL 全国计算语言学学术会议
   http://www.cips-cl.org:8080/CCL2017/home.html 南京 10.13-10.15
4. NLPCC Natural Language Processing and Chinese Computing
   http://tcci.ccf.org.cn/conference/2017/ 大连 11.8-11.12
5. NCMMSC 全国人机语音通讯学术会议
   http://www.ncmmsc2017.org/index.html 连云港 11.11 － 11.13

Toolkit Library

Python Libraries

1. fastText by Facebookhttps://github.com/facebookresearch/fastText - for efficient learning of word representations and sentence classification
2. Scikit-learn: Machine learning in Pythonhttp://arxiv.org/pdf/1201.0490.pdf
3. Natural Language Toolkit NLTKhttp://www.nltk.org/

4. Patternhttp://www.clips.ua.ac.be/pattern - A web mining module for the Python programming language. It has tools for natural language processing, machine learning, among others.
5. TextBlobhttp://textblob.readthedocs.org/ - Providing a consistent API for diving into common natural language processing NLP tasks. Stands on the giant shoulders of NLTK and Pattern, and plays nicely with both.
6. YAlignhttps://github.com/machinalis/yalign - A sentence aligner, a friendly tool for extracting parallel sentences from comparable corpora.
7. jiebahttps://github.com/fxsjy/jieba#jieba-1 - Chinese Words Segmentation Utilities.
8. SnowNLPhttps://github.com/isnowfy/snownlp - A library for processing Chinese text.
9. KoNLPyhttp://konlpy.org - A Python package for Korean natural language processing.
10. Rosettahttps://github.com/columbia-applied-data-science/rosetta - Text processing tools and wrappers e.g. Vowpal Wabbit
11. BLLIP Parserhttps://pypi.python.org/pypi/bllipparser/ - Python bindings for the BLLIP Natural Language Parser also known as the Charniak-Johnson parser
12. PyNLPlhttps://github.com/proycon/pynlpl - Python Natural Language Processing Library. General purpose NLP library for Python. Also contains some specific modules for parsing common NLP formats, most notably for FoLiAhttp://proycon.github.io/folia/, but also ARPA language models, Moses phrasetables, GIZA
13. python-ucto https://github.com/proycon/python-ucto - Python binding to ucto a unicode-aware rule-based tokenizer for various languages
14. Parseratorhttps://github.com/datamade/parserator - A toolkit for making domain-specific probabilistic parsers
15. python-frog https://github.com/proycon/python-frog - Python binding to Frog, an NLP suite for Dutch. pos tagging, lemmatisation, dependency parsing, NER
16. python-zparhttps://github.com/EducationalTestingService/python-zpar - Python bindings for ZParhttps://github.com/frcchang/zpar, a statistical part-of-speech-tagger, constiuency parser, and dependency parser for English.

17. colibri-corehttps://github.com/proycon/colibri-core - Python binding to C
18. spaCyhttps://github.com/spacy-io/spaCy - Industrial strength NLP with Python and Cython.
19. textacyhttps://github.com/chartbeat-labs/textacy - Higher level NLP built on spaCy
20. PyStanfordDependencieshttps://github.com/dmcc/PyStanfordDependencies - Python interface for converting Penn Treebank trees to Stanford Dependencies.
21. gensimhttps://radimrehurek.com/gensim/index.html - Python library to conduct unsupervised semantic modelling from plain text
22. scattertexthttps://github.com/JasonKessler/scattertext - Python library to produce d3 visualizations of how language differs between corpora.
23. CogComp-NlPyhttps://github.com/CogComp/cogcomp-nlpy - Light-weight Python NLP annotators.
24. PyThaiNLPhttps://github.com/wannaphongcom/pythainlp - Thai NLP in Python Package.
25. jPTDPhttps://github.com/datquocnguyen/jPTDP - A toolkit for joint part-of-speech POS tagging and dependency parsing. jPTDP provides pre-trained models for 40+ languages.
26. CLTKhttps://github.com/cltk/cltk: The Classical Language Toolkit is a Python library and collection of texts for doing NLP in ancient languages.
27. pymorphy2https://github.com/kmike/pymorphy2 - a good pos-tagger for Russian
28. BigARTMhttps://github.com/bigartm/bigartm - a fast library for topic modelling
29. AllenNLPhttps://github.com/allenai/allennlp - An NLP research library, built on PyTorch, for developing state-of-the-art deep learning models on a wide variety of linguistic tasks.

C++ Libraries

1. MIT Information Extraction Toolkithttps://github.com/mit-nlp/MITIE - C, C++, and Python tools for named entity recognition and relation extraction

2. CRF++https://taku910.github.io/crfpp/ - Open source implementation of Conditional Random Fields CRFs for segmenting/labeling sequential data & other Natural Language Processing tasks.
3. CRFsuitehttp://www.chokkan.org/software/crfsuite/ - CRFsuite is an implementation of Conditional Random Fields CRFs for labeling sequential data.
4. BLLIP Parserhttps://github.com/BLLIP/bllip-parser - BLLIP Natural Language Parser also known as the Charniak-Johnson parser
5. colibri-corehttps://github.com/proycon/colibri-core - C++ library, command line tools, and Python binding for extracting and working with basic linguistic constructions such as n-grams and skipgrams in a quick and memory-efficient way.
6. uctohttps://github.com/LanguageMachines/ucto - Unicode-aware regular-expression based tokenizer for various languages. Tool and C++ library. Supports FoLiA format.
7. libfoliahttps://github.com/LanguageMachines/libfolia - C++ library for the FoLiA formathttp://proycon.github.io/folia/
8. froghttps://github.com/LanguageMachines/frog - Memory-based NLP suite developed for Dutch: PoS tagger, lemmatiser, dependency parser, NER, shallow parser, morphological analyzer.
9. MeTAhttps://github.com/meta-toolkit/meta - MeTA : ModErn Text Analysishttps://meta-toolkit.org/ is a C++ Data Sciences Toolkit that facilitates mining big text data.
10. StarSpacehttps://github.com/facebookresearch/StarSpace - a library from Facebook for creating embeddings of word-level, paragraph-level, document-level and for text classification

Java Libraries

1. Stanford NLPhttp://nlp.stanford.edu/software/index.shtml
2. OpenNLPhttp://opennlp.apache.org/
3. ClearNLPhttps://github.com/clir/clearnlp

4. Word2vec in Javahttp://deeplearning4j.org/word2vec.html
5. ReVerbhttps://github.com/knowitall/reverb/ Web-Scale Open Information Extraction
6. OpenRegexhttps://github.com/knowitall/openregex An efficient and flexible token-based regular expression language and engine.
7. CogcompNLPhttps://github.com/CogComp/cogcomp-nlp - Core libraries developed in the U of Illinois' Cognitive Computation Group.
8. MALLEThttp://mallet.cs.umass.edu/ - MAchine Learning for LanguagE Toolkit - package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text.
9. RDRPOSTaggerhttps://github.com/datquocnguyen/RDRPOSTagger - A robust POS tagging toolkit available in both Java & Python together with pre-trained models for 40+ languages.

中文

1. THULAC 中文词法分析工具包 http://thulac.thunlp.org/ by 清华 C++/Java/Python
2. NLPIRhttps://github.com/NLPIR-team/NLPIR by 中科院 Java
3. LTP 语言技术平台 https://github.com/HIT-SCIR/ltp by 哈工大 C++
4. FudanNLPhttps://github.com/FudanNLP/fnlp by 复旦 Java
5. HanNLPhttps://github.com/hankcs/HanLP Java
6. SnowNLPhttps://github.com/isnowfy/snownlp Python Python library for processing Chinese text
7. YaYaNLPhttps://github.com/Tony-Wang/YaYaNLP 纯 python 编写的中文自然语言处理包，取名于“牙牙学语”
8. DeepNLPhttps://github.com/rockingdingo/deepnlp Deep Learning NLP Pipeline implemented on Tensorflow with pretrained Chinese models.
9. chinese*nlphttps://github.com/taozhijiang/chinese*nlp] C++ & Python Chinese Natural Language Processing tools and examples
10. Jieba 结巴中文分词 https://github.com/fxsjy/jieba 做最好的 Python 中文分词组件

11. kcws 深度学习中文分词 https://github.com/koth/kcws BiLSTM+CRF 与 IDCNN+CRF
12. Genius 中文分词 https://github.com/duanhongyi/genius Genius 是一个开源的 python 中文分词组件，采用 CRFConditional Random Field 条件随机场算法。
13. loso 中文分词 https://github.com/fangpenlin/loso
14. Information-Extraction-Chinesehttps://github.com/crownpku/Information-Extraction-Chinese Chinese Named Entity Recognition with IDCNN/biLSTM+CRF, and Relation Extraction with biGRU+2ATT 中文实体识别与关系提取

datasets

1. *Apache Software Foundation Public Mail Archives*http://aws.amazon.com/de/datasets/apache-software-foundation-public-mail-archives/
2. Blog Authorship Corpushttp://u.cs.biu.ac.il/: consists of the collected posts of 19,320 bloggers gathered from blogger.com in August 2004. 681,288 posts and over 140 million words.
3. Amazon Fine Food Reviews Kagglehttps://www.kaggle.com/snap/amazon-fine-food-reviews: consists of 568,454 food reviews Amazon users left up to October 2012. Paperhttp://i.stanford.edu/. 240 MB
4. Amazon Reviewshttps://snap.stanford.edu/data/web-Amazon.html: Stanford collection of 35 million amazon reviews. 11 GB
5. ArXivhttp://arxiv.org/help/bulk*data*s3: All the Papers on archive as fulltext 270 GB + sourcefiles 190 GB
6. ASAP Automated Essay Scoring Kagglehttps://www.kaggle.com/c/asap-aes/data: For this competition, there are eight essay sets. Each of the sets of essays was generated from a single prompt. Selected essays range from an average length of 150 to 550 words per response. Some of the essays are dependent upon source information and others are

not. All responses were written by students ranging in grade levels from Grade 7 to Grade 10. All essays were hand graded and were double-scored. 100 MB

7. ASAP Short Answer Scoring Kagglehttps://www.kaggle.com/c/asap-sas/data: Each of the data sets was generated from a single prompt. Selected responses have an average length of 50 words per response. Some of the essays are dependent upon source information and others are not. All responses were written by students primarily in Grade 10. All responses were hand graded and were double-scored. 35 MB
8. Classification of political social mediahttps://www.crowdflower.com/data-for-everyone/: Social media messages from politicians classified by content. 4 MB
9. CLiPS Stylometry Investigation CSI Corpushttp://www.clips.uantwerpen.be/datasets/csi-corpus: a yearly expanded corpus of student texts in two genres: essays and reviews. The purpose of this corpus lies primarily in stylometric research, but other applications are possible.
10. ClueWeb09 FACChttp://lemurproject.org/clueweb09/FACC1/: ClueWeb09http://lemurproject.org/clueweb09/ with Freebase annotations 72 GB
11. ClueWeb11 FACChttp://lemurproject.org/clueweb12/FACC1/: ClueWeb11http://lemurproject.org/clueweb12/ with Freebase annotations 92 GB
12. Common Crawl Corpushttp://aws.amazon.com/de/datasets/common-crawl-corpus/: web crawl data composed of over 5 billion web pages 541 TB
13. Cornell Movie Dialog Corpushttp://www.cs.cornell.edu/: contains a large metadata-rich collection of fictional conversations extracted from raw movie scripts: 220,579 conversational exchanges between 10,292 pairs of movie characters, 617 movies 9.5 MB
14. DBpediahttp://aws.amazon.com/de/datasets/dbpedia-3-5-1/?tag=datasets%23keywords%23encyclopedic: a community effort to extract structured information from Wikipedia and to make this information available on the Web 17 GB
15. Del.icio.ushttp://arvindn.livejournal.com/116137.html: 1.25 million bookmarks on delicious.com

16. Disasters on social mediahttps://www.crowdflower.com/data-for-everyone/: 10,000 tweets with annotations whether the tweet referred to a disaster event 2 MB
17. Economic News Article Tone and Relevancehttps://www.crowdflower.com/data-for-everyone/: News articles judged if relevant to the US economy and, if so, what the tone of the article was. Dates range from 1951 to 2014. 12 MB
18. Enron Email Datahttp://aws.amazon.com/de/datasets/enron-email-data/: consists of 1,227,255 emails with 493,384 attachments covering 151 custodians 210 GB
19. Event Registryhttp://eventregistry.org/: Free tool that gives real time access to news articles by 100.000 news publishers worldwide. Has APIhttps://github.com/gregorleban/EventRegistry/.
20. Federal Contracts from the Federal Procurement Data Center USASpending.govhttp://aws.amazon.com/de/datasets/federal-contracts-from-the-federal-procurement-data-center-usaspending-gov/: data dump of all federal contracts from the Federal Procurement Data Center found at USASpending.gov 180 GB
21. Flickr Personal Taxonomieshttp://www.isi.edu/: Tree dataset of personal tags 40 MB
22. Freebase Data Dumphttp://aws.amazon.com/de/datasets/freebase-data-dump/: data dump of all the current facts and assertions in Freebase 26 GB
23. Google Books Ngramshttp://storage.googleapis.com/books/ngrams/books/datasetsv2.html: available also in hadoop format on amazon s3 2.2 TB
24. Google Web 5gramhttps://catalog.ldc.upenn.edu/LDC2006T13: contains English word n-grams and their observed frequency counts 24 GB
25. Gutenberg Ebook Listhttp://www.gutenberg.org/wiki/Gutenberg:Offline*Catalogs*: annotated list of ebooks 2 MB
26. Harvard Libraryhttp://library.harvard.edu/open-metadata#Harvard-Library-Bibliographic-Dataset: over 12 million bibliographic records for materials held by the Harvard Library, including books, journals, electronic resources, manuscripts, archival materials, scores, audio, video and other materials. 4 GB

27. Hillary Clinton Emails Kagglehttps://www.kaggle.com/kaggle/hillary-clinton-emails: nearly 7,000 pages of Clinton's heavily redacted emails 12 MB
28. Machine Translation of European Languageshttp://statmt.org/wmt11/translation-task.html#download: 612 MB
29. News article / Wikipedia page pairingshttps://www.crowdflower.com/data-for-every-one/: Contributors read a short article and were asked which of two Wikipedia articles it matched most closely. 6 MB
30. NIPS2015 Papers version 2 Kagglehttps://www.kaggle.com/benhamner/nips-2015-papers/version/2: full text of all NIPS2015 papers 335 MB
31. NYTimes Facebook Datahttp://minimaxir.com/2015/07/facebook-scraper/: all the NYTimes facebook posts 5 MB
32. Open Library Data Dumpshttps://openlibrary.org/developers/dumps: dump of all revisions of all the records in Open Library. 16 GB
33. Personae Corpushttp://www.clips.uantwerpen.be/datasets/personae-corpus: collected for experiments in Authorship Attribution and Personality Prediction. It consists of 145 Dutch-language essays by 145 different students.
34. Reddit Commentshttps://www.reddit.com/r/datasets/comments/3bxlg7/*ihavee*very*publicly*available*reddit*comment/: every publicly available reddit comment as of july 2015. 1.7 billion comments 250 GB
35. Reddit Comments May '15Kagglehttps://www.kaggle.com/reddit/reddit-comments-may-2015: subset of above dataset 8 GB
36. Reddit Submission Corpushttps://www.reddit.com/r/datasets/comments/3mg812/full*reddit*submission*corpus*now*available*2006/: all publicly available Reddit submissions from January 2006 - August 31, 2015]. 42 GB
37. Reuters Corpushttp://trec.nist.gov/data/reuters/reuters.html: a large collection of Reuters News stories for use in research and development of natural language processing, information retrieval, and machine learning systems. This corpus, known as "Reuters

Corpus, Volume 1" or RCV1, is significantly larger than the older, well-known Reuters-21578 collection heavily used in the text classification community. Need to sign agreement and sent per post to obtain. 2.5 GB

38. SMS Spam Collectionhttp://www.dt.fee.unicamp.br/: 5,574 English, real and non-enconded SMS messages, tagged according being legitimate ham or spam. 200 KB
39. Stackoverflowhttp://data.stackexchange.com/: 7.3 million stackoverflow questions + other stackexchanges
40. Twitter Cheng-Caverlee-Lee Scrapehttps://archive.org/details/twitter*cikm*2010: Tweets from September 2009 - January 2010, geolocated. 400 MB
41. Twitter New England Patriots Deflategate sentimenthttps://www.crowdflower.com/data-for-everyone/: Before the 2015 Super Bowl, there was a great deal of chatter around deflated footballs and whether the Patriots cheated. This data set looks at Twitter sentiment on important days during the scandal to gauge public sentiment about the whole ordeal. 2 MB
42. Twitter sentiment analysis: Self-driving carshttps://www.crowdflower.com/data-for-everyone/: contributors read tweets and classified them as very positive, slightly positive, neutral, slightly negative, or very negative. They were also prompted asked to mark if the tweet was not relevant to self-driving cars. 1 MB
43. Twitter Tokyo Geolocated Tweetshttp://followthehashtag.com/datasets/200000-tokyo-geolocated-tweets-free-twitter-dataset/: 200K tweets from Tokyo. 47 MB
44. Twitter US Airline Sentiment Kagglehttps://www.kaggle.com/crowdflower/twitter-airline-sentiment: A sentiment analysis job about the problems of each major U.S. airline. Twitter data was scraped from February of 2015 and contributors were asked to first classify positive, negative, and neutral tweets, followed by categorizing negative reasons such as "late flight" or "rude service". 2.5 MB
45. Wikipedia Extraction http://aws.amazon.com/de/datasets/wikipedia-extraction-wex/: a processed dump of english language wikipedia

46. Wikipedia XML Datahttp://aws.amazon.com/de/datasets/wikipedia-xml-data/: complete copy of all Wikimedia wikis, in the form of wikitext source and metadata embedded in XML.
47. Yahoo! Answers Comprehensive Questions and Answershttp://webscope.sandbox.yahoo.com/catalog.php?datatype=l: Yahoo! Answers corpus as of 10/25/2007. Contains 4,483,032 questions and their answers.
48. Yahoo! Answers Manner Questionshttp://webscope.sandbox.yahoo.com/catalog.php?datatype=l: subset of the Yahoo! Answers corpus from a 10/25/2007 dump, selected for their linguistic properties. Contains 142,627 questions and their answers.
49. Yahoo! N-Grams, version 2.0http://webscope.sandbox.yahoo.com/catalog.php?datatype=l: n-grams n = 1 to 5, extracted from a corpus of 14.6 million documents 126 million unique sentences, 3.4 billion running words crawled from over 12000 news-oriented sites
50. Yahoo! Search Logs with Relevance Judgmentshttp://webscope.sandbox.yahoo.com/catalog.php?datatype=l: Annonymized Yahoo! Search Logs with Relevance Judgments
51. Yelphttps://www.yelp.com/academic*dataset*: including restaurant rankings and 2.2M reviews
52. Youtubehttps://www.reddit.com/r/datasets/comments/3gegdz/17*millions*youtube*videos*description/: 1.7 million youtube videos descriptions
53. 开放知识图谱 OpenKG.cnhttp://openkg.cn
54. CLDC 中文语言资源联盟 http://www.chineseldc.org/
55. 用于训练中英文对话系统的语料库 https://github.com/candlewill/Dialog*Corpus* Datasets for Training Chatbot System
56. 中文 Wikipedia Dumphttps://dumps.wikimedia.org/zhwiki/
57. 98 年人民日报词性标注库@百度盘 https://pan.baidu.com/s/1gd6mslt
58. 百度百科 100gb 语料@百度盘 http://pan.baidu.com/s/1i3wvfil 密码 neqs 出处应该是梁斌 penny 大神

59. 搜狗 20061127 新闻语料包含分类@百度盘 https://pan.baidu.com/s/1bnhXX6Z
60. UDChinesehttps://github.com/UniversalDependencies/UD*Chinese* for training spaCy POS
61. 八卦版問答中文語料 https://github.com/zake7749/Gossiping-Chinese-Corpus
62. 中文 word2vec 模型 https://github.com/to-shimo/chinese-word2vec
63. 中文word2vec模型之维基百科中文https://github.com/Samurais/wikidata-corpus 使用 2017 年 6 月 20 日中文维基百科语料训练的脚本和模型文件。
64. Synonyms:中文近义词工具包 https://github.com/huyingxi/Synonyms/ 基于维基百科中文和 word2vec 训练的近义词库，封装为 python 包文件。
65. 中文突发事件语料库 https://github.com/shijiebei2009/CEC-Corpus Chinese Emergency Corpus
66. dgk*lost*conv 中文对白语料 https://github.com/rustch3n/dgk*lost*conv chinese conversation corpus
67. 漢語拆字字典 https://github.com/kfcd/chaizi
68. 中国股市公告信息爬取 https://github.com/startprogress/China*stock*announcement 通过 python 脚本从巨潮网络的服务器获取中国股市（sz,sh）的公告上市公司和监管机构
69. tushare 财经数据接口 http://tushare.org/ TuShare 是一个免费、开源的 python 财经数据接口包。
70. 保险行业语料库 https://github.com/Samurais/insuranceqa-corpus-zh 52nlp 介绍 Blog http://www.52nlp.cn/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0%E4%BF%9D%E9%99%A9%E8%A1%8C%E4%B8%9A%E9%97%AE%E7%AD%94%E5%BC%80%E6%94%BE%E6%95%B0%E6%8D%AE%E9%9B%86 OpenData in insurance area for Machine Learning Tasks
71. 最全中华古诗词数据库 https://github.com/chinese-poetry/chinese-poetry 唐宋两朝近一万四千古诗人, 接近 5.5 万首唐诗加 26 万宋诗. 两宋时期 1564 位词人，21050 首词。

72. 中文语料小数据 https://github.com/crownpku/Small-Chinese-Corpus 包含了中文命名实体识别、中文关系识别、中文阅读理解等一些小量数据