【导读】作者一年前整理了这份关于 NLP 与知识图谱的参考资源，涵盖内容与形式也是非常丰富，接下来人工智能头条还会继续努力，分享更多更好的新资源给大家，也期待能与大家多多交流，一起成长。

NLP 参考资源

自然语言处理（Natural Language Processing）是深度学习的主要应用领域之一。

教程

- CS224d: Deep Learning for Natural Language Processing
  http://cs224d.stanford.edu/

- CS224d 课程的课件
  http://web.stanford.edu/class/cs224n/syllabus.html

- CMU 的 NLP 教程。该网页下方还有美国其他高校的 NLP 课程的链接。
  http://demo.clab.cs.cmu.edu/NLP/

- 北京大学的 NLP 教程，特色：中文处理。缺点：传统方法居多，深度学习未涉及。
  http://ccl.pku.edu.cn/alcourse/nlp/

- COMS W4705: Natural Language Processing
  http://www.cs.columbia.edu/~cs4705/

- 初学者如何查阅自然语言处理（NLP）领域学术资料
  https://mp.weixin.qq.com/s/TSc4E8lKwgc-EvzP8OlJeg

- 揭开知识库问答 KB-QA 的面纱（知识图谱方面的系列专栏）
  https://zhuanlan.zhihu.com/kb-qa

- 《语音与语言处理》第三版，NLP 和语音合成方面的专著
  http://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

- CIPS ATT 2017 文本分析和自然语言课程 PPT
  https://mp.weixin.qq.com/s/5KhTWdOk-b84DXmoVr68-A

- CMU NN for NLP
  http://phontron.com/class/nn4nlp2017/assets/slides/

- CMU Machine Translation and Sequence to Sequence Models
  http://phontron.com/class/mtandseq2seq2017/

- Oxford Deep NLP 2017 course
  https://github.com/oxford-cs-deepnlp-2017/lectures

书籍

- 《Natural Language Processing with Python》，Steven Bird、Ewan Klein、Edward Loper 著。这本书的作者们创建了著名的 NLTK 工具库。
  http://ccl.pku.edu.cn/alcourse/nlp/LectureNotes/Natural%20Language%20Processing%20with%20Python.pdf

  注：Steven Bird，爱丁堡大学博士，墨尔本大学副教授。
  http://www.stevenbird.net/about.html
  Ewan Klein，苏格兰人，哥伦比亚大学博士（1978 年），爱丁堡大学教授。
  Edward Loper，宾夕法尼亚大学博士。

- 推荐 5 本经典自然语言处理书籍

https://mp.weixin.qq.com/s/0HmsMytif3INqAX1Si5ukA

网站

- 一个自然语言处理爱好者的群体博客。包括 52nlp、rickjin、liwei 等国内外华人大牛.
  http://www.52nlp.cn/

- 实战课程：自己动手做聊天机器人
  http://www.shareditor.com/bloglistbytag/?tagname=%E8%87%AA%E5%B7%B1%E5%8A%A8%E6%89%8B%E5%81%9A%E8%81%8A%E5%A4%A9%E6%9C%BA%E5%99%A8%E4%BA%BA

- 北京大学计算机科学技术研究所语言计算与互联网挖掘研究
  http://www.icst.pku.edu.cn/lcwm/

- NLP 深度学习方面的代码库
  https://github.com/rockingdingo/deepnlp

- NLP 专家李维的 blog
  https://liweinlp.com/

- 一个 NLP 方面的 blog
  http://www.shuang0420.com/

- 一个 DL+ML+NLP 的 blog
  http://www.cnblogs.com/Determined22/

- 一个 NLP 方面的 blog
  http://www.cnblogs.com/robert-dlut/

- 一个 NLP 方面的 blog
  https://blog.csdn.net/wangxinginnlp

工具

- Natural Language Toolkit(NLTK)
  官网：*http://www.nltk.org/*
  可使用 nltk.download()下载相关 nltk 官方提供的各种资源。

  参考：
  http://www.cnblogs.com/baiboy/p/nltk3.html

- OpenNLP
  http://opennlp.apache.org/

- FudanNLP
  https://github.com/FudanNLP/fnlp

- Stanford CoreNLP
  http://stanfordnlp.github.io/CoreNLP/

- THUCTC
  THUCTC(THU Chinese Text Classification)是由清华大学自然语言处理实验室推出的中文文本分类工具包。
  http://thuctc.thunlp.org/

- gensim
  gensim 是 Python 语言的计算文本相似度的程序包。
  http://radimrehurek.com/gensim/index.html

pip install --upgrade gensim
GitHub 地址：
https://github.com/RaRe-Technologies/gensim

参考学习：
情感分析的新方法——基于 Word2Vec/Doc2Vec/Python
http://www.open-open.com/lib/view/open1444351655682.html

Gensim Word2vec 使用教程
http://blog.csdn.net/Star_Bob/article/details/47808499

- GloVe
  GloVe:Global Vectors for Word Representation
  https://nlp.stanford.edu/projects/glove/

- textsum
  textsum 是一个基于深度学习的文本自动摘要工具。

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

参考：
http://www.jiqizhixin.com/article/1449

谷歌开源新的 TensorFlow 文本自动摘要代码：TensorFlow 文本摘要生成 - 基于注意力的序列到序列模型
http://blog.csdn.net/tensorflowshizhan/article/details/69230070

- jieba

https://github.com/fxsjy/jieba

- NLPIR
  NLPIR 汉语分词系统(又名 ICTCLAS2013)，是中科院张华平博士的作品。
  http://ictclas.nlpir.org/

  参考：
  这个网页对于 NLP 的大多数功能进行了可视化的展示。NLP 入门必看。
  http://ictclas.nlpir.org/nlpir/

- snownlp
  https://github.com/isnowfy/snownlp

- HanLP
  HanLP 是一个目前留学日本的中国学生的作品。
  http://hanlp.linrunsoft.com/

  作者 blog：
  http://www.hankcs.com/

  Github：
  https://github.com/hankcs/HanLP/

  从作者的名气来说，HanLP 无疑是最低的，性能也不见得有多好。然而对于初学者来说，这却是最适合的工具。这主要体现在以下几个方面：

1.中文处理能力。NLTK 和 OpenNLP 对中文支持非常差，这里不光是中文分词的问题，有些 NLP 算法需要一定的语言模型数据，但浏览 NLTK 官方的模型库，基本找不到中文模型数据。

2.jieba、IK 之类的功能太单一，多数局限在中文分词方面领域。gensim、THUCTC 专注于 NLP 的某一方面，也不是通用工具。

3.NLPIR 和 Stanford CoreNLP 算是功能最强的工具包了。前者的问题在于收费不开源，后者的问题在于缺少中文文档。FudanNLP 的相关文档较少，文档友好度不如 HanLP。

4.HanLP 在主页上提供了相关算法的 blog，便于初学者快速掌握相关概念。其词典是明文发布，便于用户修改。HanLP 执行时，会将明文词典以特定结构缓存，以提高执行效率。
注：不要以为中文有分词问题，就比别的语言复杂，英文还有词根问题呢。。。每种语言都不简单。

- AllenNLP
  AllenNLP 是 Allen AI 实验室的作品，采用深度学习技术，基于 PyTorch 开发。
  http://allennlp.org/

  Allen AI 实验室由微软联合创始人 Paul G. Allen 投资创立。
  http://allenai.org/

- 其他

  python 版的汉字转拼音软件
  https://github.com/mozillazg/python-pinyin

Java 分布式中文分词组件-word 分词

https://github.com/ysc/word

jena 是一个语义网络、知识图谱相关的软件。

http://jena.apache.org/

- NLPchina
  NLPchina(中国自然语言处理开源组织)旗下有许多好用的工具。
  http://www.nlpcn.org/

  Github：
  https://github.com/NLPchina

- Ansj
  Ansj 是一个 NLPchina 旗下的开源的 Java 中文分词工具，基于中科院的 ictclas 中文分词算法，比其他常用的开源分词工具（如 mmseg4j）的分词准确率更高。
  https://github.com/NLPchina/ansj_seg

- Word2VEC_java
  word2vec java 版本的一个实现。
  https://github.com/NLPchina/Word2VEC_java
  doc2vec java 版本的一个实现，基于 Word2VEC_java。
  https://github.com/yao8839836/doc2vec_java

- ansj_fast_lda
  LDA 算法的 Java 包。
  https://github.com/NLPchina/ansj_fast_lda

- nlp-lang
  这个项目是一个基本包.封装了大多数 nlp 项目中常用工具
  https://github.com/NLPchina/nlp-lang

- 词性标注
  ICTPOS3.0 汉语词性标记集
  http://jacoxu.com/ictpos3-0%E6%B1%89%E8%AF%AD%E8%AF%8D%E6%80%A7%E6%A0%87%E8%AE%B0%E9%9B%86/

- Word Hashing
  Word Hashing 是非常重要的一个 trick，以英文单词来说，比如 good，他可以写成#good#，然后按 tri-grams 来进行分解为#go goo ood od#，再将这个 tri-grams 灌入到 bag-of-word 中，这种方式可以非常有效的解决 vocabulary 太大的问题(因为在真实的 web search 中 vocabulary 就是异常的大)，另外也不会出现 oov 问题，因此英文单词才 26 个，3 个字母的组合都是有限的，很容易枚举光。

  那么问题就来了，这样两个不同的单词会不会产出相同的 tri-grams，paper 里面做了统计，说了这个冲突的概率非常的低，500K 个 word 可以降到 30k 维，冲突的概率为 0.0044%。

  但是在中文场景下，这个 Word Hashing 估计没有这么有效了。

- 词汇共现
  http://sewm.pku.edu.cn/TianwangLiterature/SEWM/2005(5)/%5b%b3%c2%c1%88,%20et%20al.,2005%5d/050929.pdf

词汇共现是指词汇在文档集中共同出现。以一个词为中心，可以找到一组经常与之搭配出现的词，作为它的共现词汇集。

词汇共现的其中一种用例：

有若干关键词，比如：水果、天气、风，有若干描述词，比如，很甜、晴朗、很大，然后现在要找出他们之间的搭配，在这个例子里，我们最终要找到：水果很甜、天气晴朗、风很大

- 关键词提取
  主要三种方法：1.基于统计特征，如 TF-IDF；2.基于词图模型，如 TextRank；3.基于主题模型，如 LDA。

- 自然语言理解
  Natural language understanding(NLU)属于 NLP 的一个分支，属于人工智能的一个部分，用来解决机器理解人类语言的问题，属于人工智能的核心难题。
  http://www.shuang0420.com/2017/04/27/NLP%E7%AC%94%E8%AE%B0%20-%20NLU%E4%B9%8B%E6%84%8F%E5%9B%BE%E5%88%86%E7%B1%BB/

- 论文
  《Distant Supervision for relation extraction without labeled data》
  《Using Recurrent Neural Networks for Slot Filling in Spoken Language Understanding》
  《Convolutional Neural Networks for Sentence Classification》：TextCNN 的开山之作

知识图谱参考资源

- 知识图谱构建技术综述
  https://wenku.baidu.com/view/38ad3ef7e109581b6bd97f19227916888586b959.html

- 知识图谱技术综述
  https://wenku.baidu.com/view/e69a3619fe00bed5b9f3f90f76c66137ee064f15.html

- 知识图谱技术原理介绍
  https://wenku.baidu.com/view/b3858227c5da50e2534d7f08.html

- 基于知识图谱的问答系统关键技术研究
  https://mp.weixin.qq.com/s/JLYegFP7kEg6n34crgP09g

- 什么是知识图谱？
  https://mp.weixin.qq.com/s/XgKvh63wgEe-CR9bchp03Q

- 当知识图谱遇上聊天机器人
  https://mp.weixin.qq.com/s/iqFXvhvYfOejaeNAhXxJEg

- 知识图谱前沿技术课程实录
  https://mp.weixin.qq.com/s/U-dlYhnaR8OQw2UKYKUWKQ

- 阿里知识图谱首次曝光：每天千万级拦截量，亿级别全量智能审核
  https://mp.weixin.qq.com/s/MZE_SXsNg6Yt4dz2fmB1sA

- 东南大学漆桂林：知识图谱的应用
  https://mp.weixin.qq.com/s/WIro7pk7kboMvdwpZOSdQA

- 东南大学高桓：知识图谱表示学习
  https://mp.weixin.qq.com/s/z1hhG4GaBQXPHHt9UGZPnA

- 复旦肖仰华：基于知识图谱的问答系统
  https://mp.weixin.qq.com/s/JZYH_m1eS93KRjkWA82GoA

- 多源信息表示学习在知识图谱中的应用
  https://mp.weixin.qq.com/s/cEmtOAtfP2gSBlaPfGXb3w

- 如何构建知识图谱
  https://mp.weixin.qq.com/s/cL1aKdu8ig8-ocOPirXk2w

- 中文通用百科知识图谱（CN-DBpedia）
  https://mp.weixin.qq.com/s/Nh7XJOLNBDdpibopVG4MrQ