人工智慧文本分析
(Artificial Intelligence for Text Analytics)

文本摘要和主題模型
(Text Summarization and Topic Models)

1082AITA07
MBA, IMTKU (M2455) (8410) (Spring 2020)
Wed 8, 9 (15:10-17:00) (B605)

Min-Yuh Day
Associate Professor
Dept. of Information Management, Tamkang University

http://mail.tku.edu.tw/myday/
2020-04-22
<table>
<thead>
<tr>
<th>週次 (Week)</th>
<th>日期 (Date)</th>
<th>內容 (Subject/Topics)</th>
</tr>
</thead>
</table>
| 1 2020/03/04 |  | 人工智慧文本分析課程介紹  
(Course Orientation on Artificial Intelligence for Text Analytics) |
| 2 2020/03/11 |  | 文本分析的基礎：自然語言處理  
(Foundations of Text Analytics: Natural Language Processing; NLP) |
| 3 2020/03/18 |  | Python自然語言處理  
(Python for Natural Language Processing) |
| 4 2020/03/25 |  | 處理和理解文本  
(Processing and Understanding Text) |
| 5 2020/04/01 |  | 文本表達特徵工程  
(Feature Engineering for Text Representation) |
| 6 2020/04/08 |  | 人工智慧文本分析個案研究 I  
(Case Study on Artificial Intelligence for Text Analytics I) |
課程大綱 (Syllabus)

週次 (Week) 日期 (Date) 內容 (Subject/Topics)
7 2020/04/15 文本分類 (Text Classification)
8 2020/04/22 文本摘要和主題模型 (Text Summarization and Topic Models)
9 2020/04/29 期中報告 (Midterm Project Report)
10 2020/05/06 文本相似度和分群 (Text Similarity and Clustering)
11 2020/05/13 語意分析和命名實體識別 (Semantic Analysis and Named Entity Recognition; NER)
12 2020/05/20 情感分析 (Sentiment Analysis)
課程大綱 (Syllabus)

週次 (Week)  日期 (Date)  內容 (Subject/Topics)
13  2020/05/27  人工智慧文本分析個案研究Ⅱ  
     (Case Study on Artificial Intelligence for Text Analytics Ⅱ)
14  2020/06/03  深度學習和通用句子嵌入模型  
     (Deep Learning and Universal Sentence-Embedding Models)
15  2020/06/10  問答系統與對話系統  
     (Question Answering and Dialogue Systems)
16  2020/06/17  期末報告 I (Final Project Presentation I)
17  2020/06/24  期末報告 II (Final Project Presentation II)
18  2020/07/01  教師彈性補充教學
Outline

• Text Summarization
• Topic Models
  • Topic Modeling
  • Latent Dirichlet Allocation (LDA)
Text Summarization and Topic Models
NLP

Classical NLP

Documents

Language Detection

- English
- Spanish
- Arabic

Pre-processing

Toxination (English)
POS Tagging (English)
Stopword Removal (EN)
...

Toxination (Spanish)
POS Tagging (Spanish)
Stopword Removal (ES)
...

Toxination (Arabic)
POS Tagging (Arabic)
Stopword Removal (AR)
...

Modeling

- Feature Extraction (EN)
- Modeling (English)
- Inference (English)

- Feature Extraction (ES)
- Modeling (Spanish)
- Inference (Spanish)

- Feature Extraction (AR)
- Modeling (Arabic)
- Inference (Arabic)

Output

- Sentiment
- Classification
- Entity Extraction
- Translation
- Topic Modelling
...

Deep Learning-based NLP

Documents

Preprocessing

Dense Embeddings
obtained via word2vec, doc2vec, GloVe, etc.

Hidden Layers

Output Units

Output

- Sentiment
- Classification
- Entity Extraction
- Translation
- Topic Modelling
...

Source: http://blog.aylien.com/leveraging-deep-learning-for-multilingual/
Modern NLP Pipeline

Modern NLP Pipeline

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

Documents → Preprocessing → Dense Word Embeddings → Deep Neural Network

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

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

Source: http://mattfortier.me/2017/01/31/nlp-intro-pt-1-overview/
Natural Language Processing (NLP) and Text Mining

1. Raw text
2. Sentence Segmentation
3. Tokenization
4. Part-of-Speech (POS)
5. Stop word removal
6. Stemming / Lemmatization
7. Dependency Parser
8. String Metrics & Matching

Word’s stem:
- am → am
- having → hav

Word’s lemma:
- am → be
- having → have

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

Topic Modeling

Seeking Life’s Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life.

One research team, using computer analyses to compare known genomes, concluded that today’s organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn’t be enough.

Although the numbers don’t match precisely, those predictions “are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Stiv Anderson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. “It may be a way of organizing any newly sequenced genome,” explains Aracdy Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an...

Text Summarization and Information Extraction

• Key-phrase extraction
  – extracting key influential phrases from the documents.

• Topic modeling
  – Extract various diverse concepts or topics present in the documents, retaining the major themes.

• Document summarization
  – Summarize entire text documents to provide a gist that retains the important parts of the whole corpus.

Topic Model in Bioinformatics

Topic Modeling

Topic Modeling (Unsupervised Learning) vs. Text Classification (Supervised Learning)

Topic Modeling

Term Document Matrix to Topic Distribution

Term Document Matrix

Word Assignment to Topics

Topic Importance

Topic Distribution Across Documents

m*m Matrix  m*n Singular Matrix  n*n Diagonal Matrix  n*m Singular Matrix

Topic Modeling
Latent Dirichlet Allocation (LDA)

\[ D \text{ documents} \]
\[ N \text{ words} \]
\[ K \text{ topics} \]

Latent Dirichlet Allocation (Blei et al., 2003)

Latent Dirichlet Allocation

David M. Blei
Computer Science Division
University of California
Berkeley, CA 94720, USA

Andrew Y. Ng
Computer Science Department
Stanford University
Stanford, CA 94305, USA

Michael I. Jordan
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA

Editor: John Lafferty

Abstract

We describe latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. We present efficient approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation. We report results in document modeling, text classification, and collaborative filtering, comparing to a mixture of unigrams model and the probabilistic LSI model.
gensim

topic modelling for humans

Gensim is a FREE Python library

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

https://radimrehurek.com/gensim/
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/
Text Summarization with Gensim Summarization

- Source: Text Summarization with Gensim Summarization:
  https://radimrehurek.com/gensim/auto_examples/tutorials/run_summarization.html

```python
1 from pprint import pprint as print
2 from gensim.summarization import summarize

1 text = {
2     "Thomas A. Anderson is a man living two lives. By day he is an "
3     "average computer programmer and by night a hacker known as "
4     "Neo. Neo has always questioned his reality, but the truth is "
5     "far beyond his imagination. Neo finds himself targeted by the "
6     "police when he is contacted by Morpheus, a legendary computer "
7     "hacker branded a terrorist by the government. Morpheus awakens "
8     "Neo to the real world, a ravaged wasteland where most of "
9     "humanity have been captured by a race of machines that live "
10    "off of the humans' body heat and electrochemical energy and "
11    "who imprison their minds within an artificial reality known as "
12    "the Matrix. As a rebel against the machines, Neo must return to "
13    "the Matrix and confront the agents: super-powerful computer "
14    "programs devoted to snuffing out Neo and the entire human "
15    "rebellion."
16 }
17 print(text)
```

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

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

https://tinyurl.com/imtkupython101
# 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>Text Summarization</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>Reading Comprehension</td>
<td>ARC</td>
<td><a href="http://data.allenai.org/arc/">http://data.allenai.org/arc/</a></td>
</tr>
<tr>
<td>Question Answering</td>
<td>CliCR</td>
<td><a href="http://aclweb.org/anthology/N18-1140">http://aclweb.org/anthology/N18-1140</a></td>
</tr>
<tr>
<td>Question Generation</td>
<td>CNN/DM</td>
<td><a href="https://cs.nyu.edu/~kcho/DMQA/">https://cs.nyu.edu/~kcho/DMQA/</a></td>
</tr>
<tr>
<td></td>
<td>NewsQA</td>
<td><a href="https://datasets.maluuba.com/NewsQA">https://datasets.maluuba.com/NewsQA</a></td>
</tr>
<tr>
<td></td>
<td>RACE</td>
<td><a href="http://www.qizheixie.com/data/RACE_leaderboard">http://www.qizheixie.com/data/RACE_leaderboard</a></td>
</tr>
<tr>
<td></td>
<td>SQuAD</td>
<td><a href="https://rajpurkar.github.io/SQuAD_leaderboard/">https://rajpurkar.github.io/SQuAD_leaderboard/</a></td>
</tr>
<tr>
<td></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/bdhenra/quasar">https://github.com/bdhenra/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>Semantic Parsing</td>
<td>AMR parsing</td>
<td><a href="http://ai.stanford.edu/index.html">http://ai.stanford.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>Sentiment Analysis</td>
<td>IMDB Reviews</td>
<td><a href="http://www.di.unipi.it/~gulli/AG_corpus_of_news_articles.html">http://www.di.unipi.it/~gulli/AG_corpus_of_news_articles.html</a></td>
</tr>
<tr>
<td></td>
<td>SST</td>
<td><a href="https://wiki.dbpedia.org/Datasets">https://wiki.dbpedia.org/Datasets</a></td>
</tr>
<tr>
<td></td>
<td>Yelp Reviews</td>
<td><a href="https://trec.nist.gov/data.html">https://trec.nist.gov/data.html</a></td>
</tr>
<tr>
<td></td>
<td>Subjectivity Dataset</td>
<td><a href="http://qwone.com/~jason/20Newsgroups/">http://qwone.com/~jason/20Newsgroups/</a></td>
</tr>
<tr>
<td>Text Classification</td>
<td>AG News</td>
<td><a href="http://www-lium.univ-lemans.fr/~schwenk/cslm_joint_paper/">http://www-lium.univ-lemans.fr/~schwenk/cslm_joint_paper/</a></td>
</tr>
<tr>
<td></td>
<td>DBpedia</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TREC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20 NewsGroup</td>
<td></td>
</tr>
<tr>
<td>Natural Language Inference</td>
<td>SNLI Corpus</td>
<td><a href="https://nlp.stanford.edu/projects/snlip/scitail/">https://nlp.stanford.edu/projects/snlip/scitail/</a></td>
</tr>
<tr>
<td></td>
<td>SciTail</td>
<td><a href="http://data.allenai.org/scitail/">http://data.allenai.org/scitail/</a></td>
</tr>
<tr>
<td>Semantic Role Labeling</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 Summarization

• Topic Models
  • Topic Modeling
  • Latent Dirichlet Allocation (LDA)
References


• Selva Prabhakaran (2020), Topic modeling visualization – How to present the results of LDA models?, https://www.machinelearningplus.com/nlp/topic-modeling-visualization-how-to-present-results-lda-models/