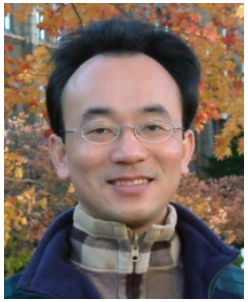
Tamkang University 淡江大學

人工智慧文本分析
(AI for Text Analytics)

NTPU 國立臺北大學 National Taipei University

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

1091AITA07
MBA, IMTKU (M2455) (8418) (Fall 2020)
Thu 3, 4 (10:10-12:00) (B206)



Min-Yuh Day
戴敏育
Associate Professor
副教授
Institute of Information Management, National Taipei University
國立臺北大學 資訊管理研究所
https://web.ntpu.edu.tw/~myday

2020-11-12

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
1	2020/09/17	人工智慧文本分析課程介紹 (Course Orientation on Artificial Intelligence for Text Analytics)
2	2020/09/24	文本分析的基礎：自然語言處理 (Foundations of Text Analytics: Natural Language Processing; NLP)
3	2020/10/01	中秋節 (Mid-Autumn Festival) 放假一天 (Day off)
4	2020/10/08	Python自然語言處理 (Python for Natural Language Processing)
5	2020/10/15	處理和理解文本 (Processing and Understanding Text)
6	2020/10/22	文本表達特徵工程 (Feature Engineering for Text Representation)

課程大綱 (Syllabus)

週次 (Week)　日期 (Date)　內容 (Subject/Topics)

7 2020/10/29 人工智慧文本分析個案研究 I
(Case Study on Artificial Intelligence for Text Analytics I)

8 2020/11/05 文本分類
(Text Classification)

9 2020/11/12 文本摘要和主題模型
(Text Summarization and Topic Models)

10 2020/11/19 期中報告 (Midterm Project Report)

11 2020/11/26 文本相似度和分群
(Text Similarity and Clustering)

12 2020/12/03 語意分析和命名實體識別
(Semantic Analysis and Named Entity Recognition; NER)

課程大綱 (Syllabus)

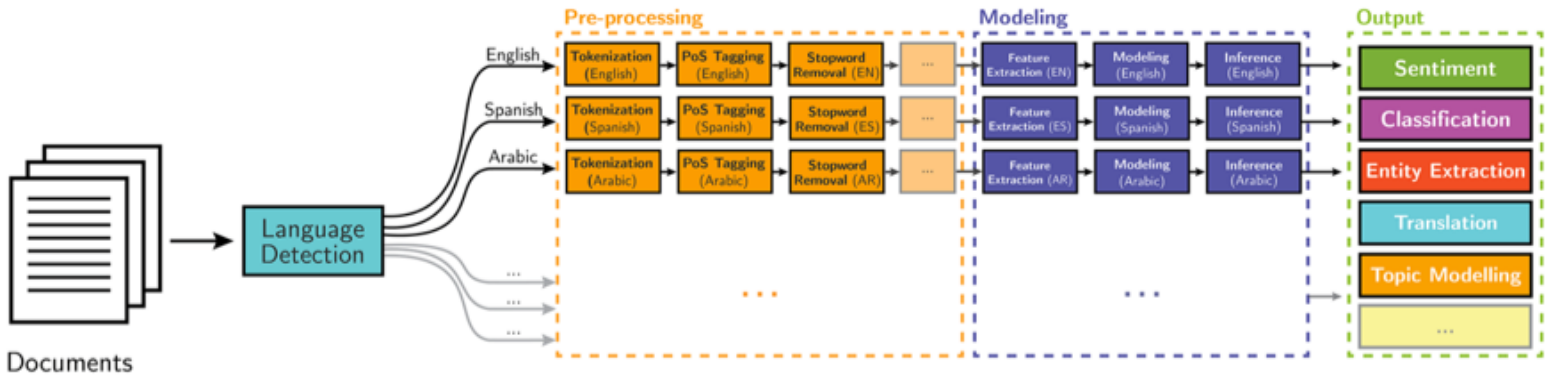
週次 (Week)	日期 (Date)	內容 (Subject/Topics)
13	2020/12/10	情感分析 (Sentiment Analysis)
14	2020/12/17	人工智慧文本分析個案研究 II (Case Study on Artificial Intelligence for Text Analytics II)
15	2020/12/24	深度學習和通用句子嵌入模型 (Deep Learning and Universal Sentence-Embedding Models)
16	2020/12/31	問答系統與對話系統 (Question Answering and Dialogue Systems)
17	2021/01/07	期末報告 I (Final Project Presentation I)
18	2021/01/14	期末報告 II (Final Project Presentation II)

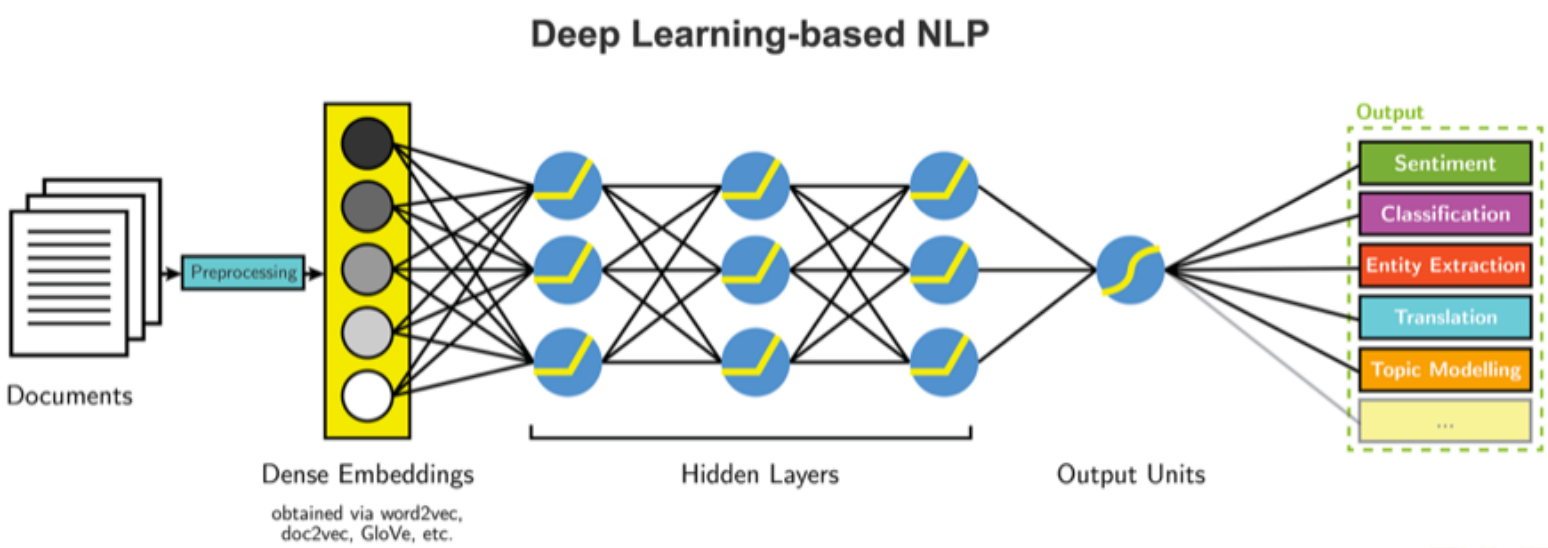
Outline

- Text Summarization
- Topic Models
  - Topic Modeling
  - Latent Dirichlet Allocation (LDA)

Text Summarization and Topic Models

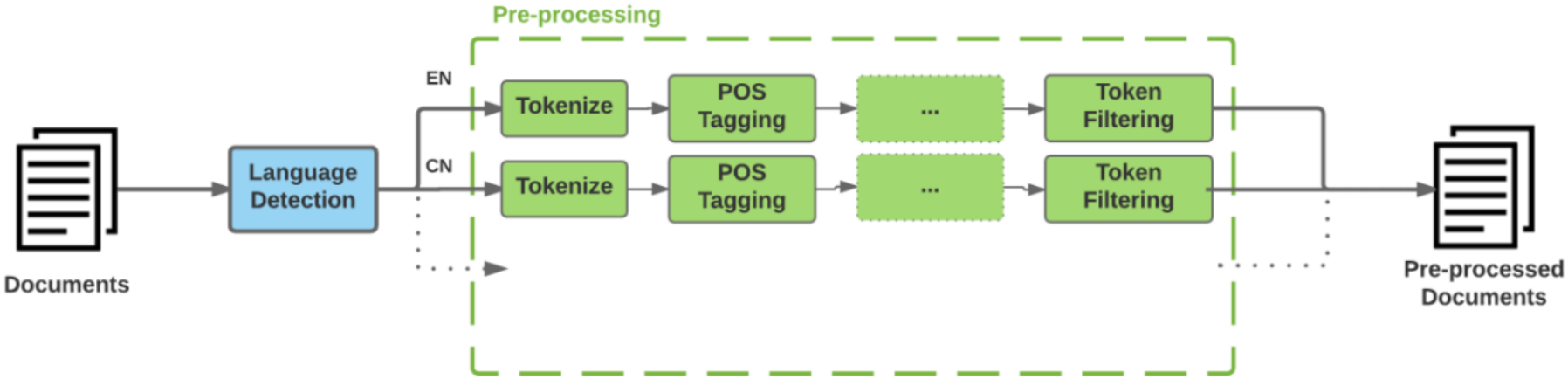
NLP

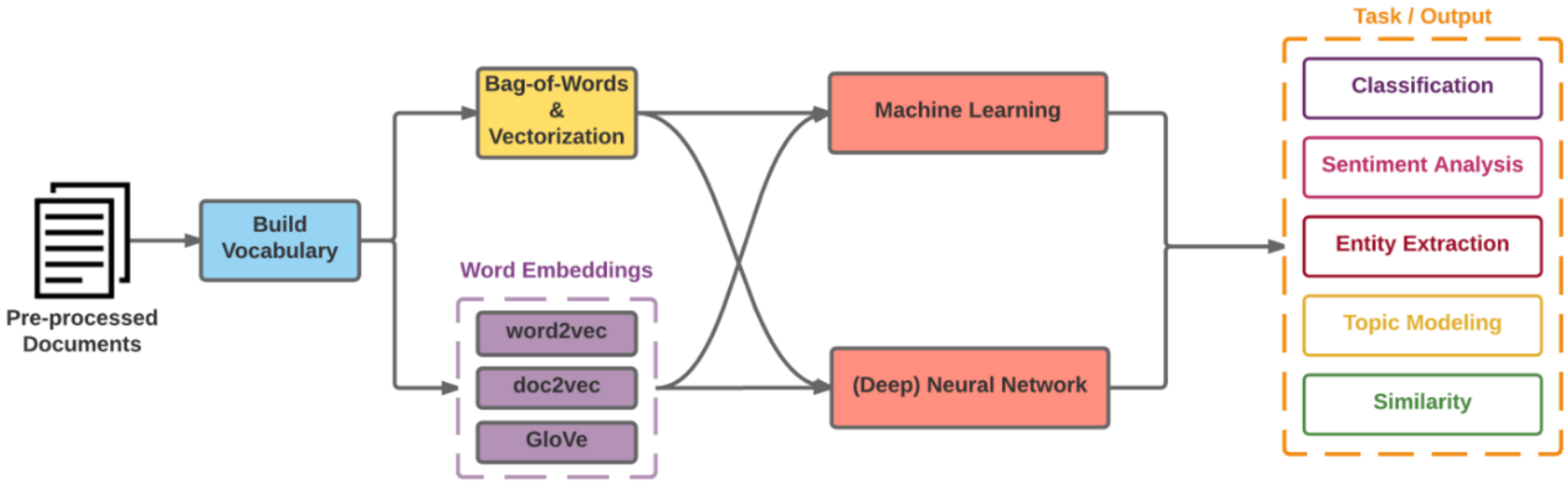




Source: http://blog.aylien.com/leveraging-deep-learning-for-multilingual/

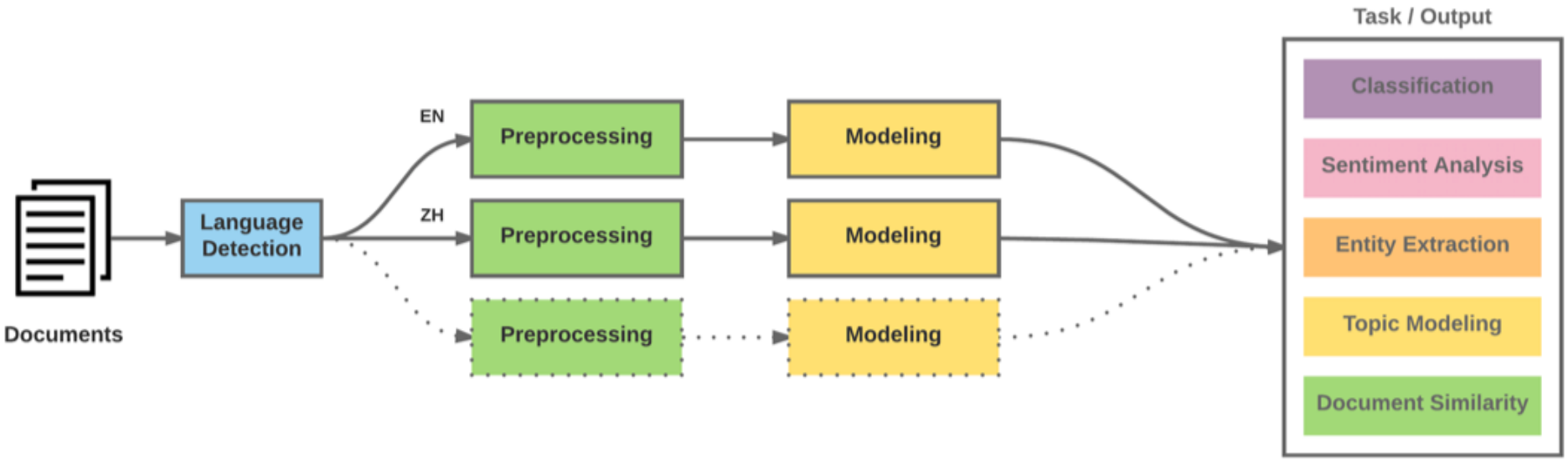
Modern NLP Pipeline





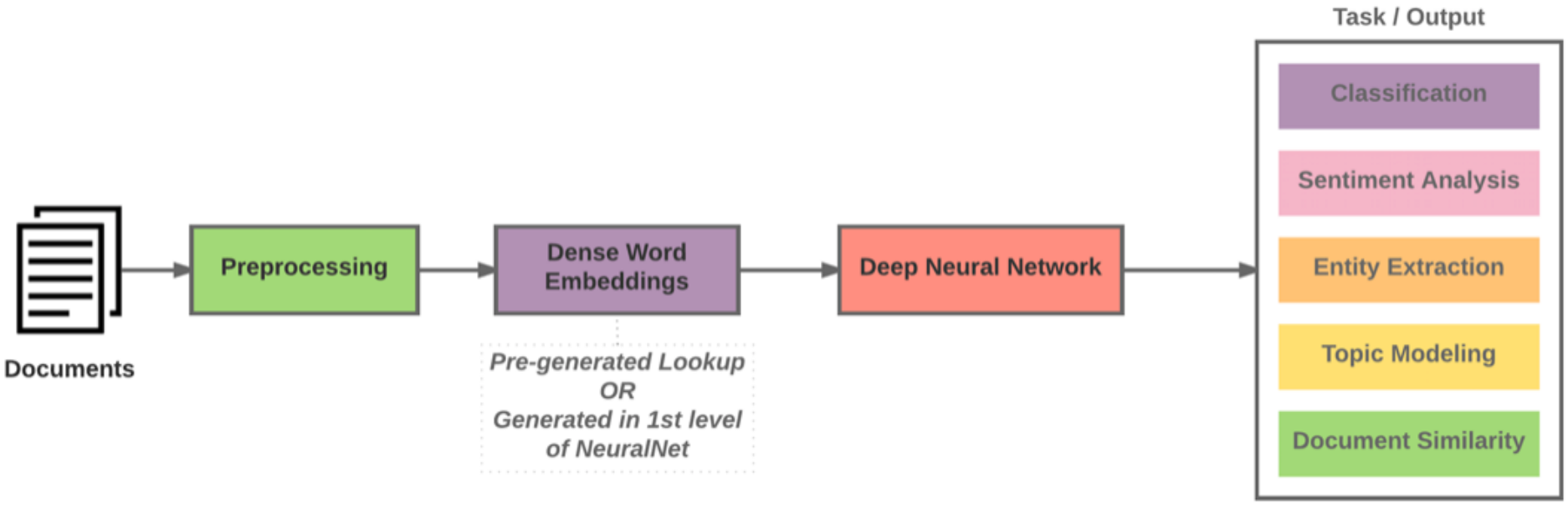
Source: https://github.com/fortiema/talks/blob/master/opendata2016sh/pragmatic-nlp-opendata2016sh.pdf

Modern NLP Pipeline



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

Deep Learning NLP



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

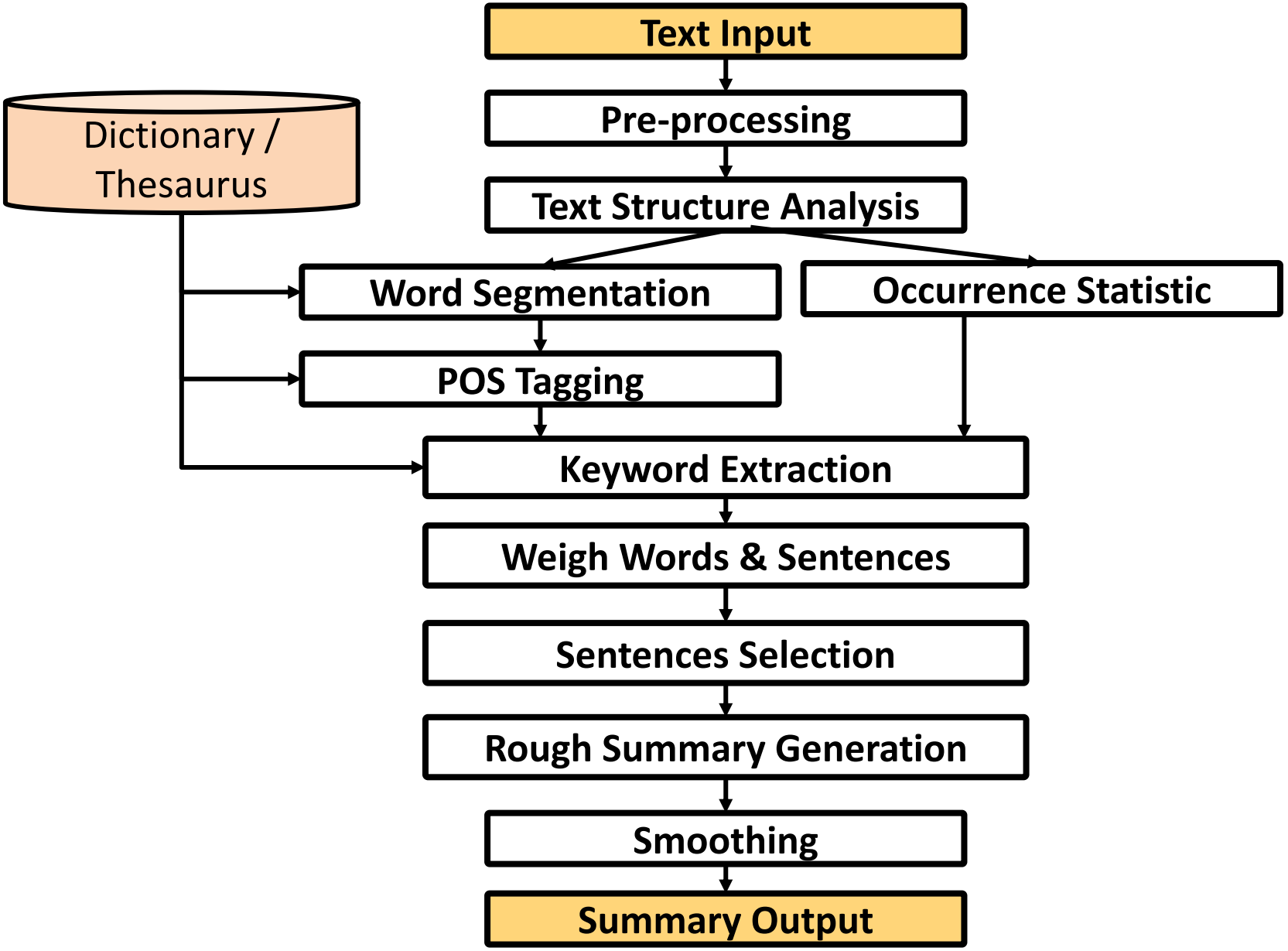
Natural Language Processing (NLP) and Text Mining

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

word's stem	word's lemma
am → am	am → be
having → hav	having → have

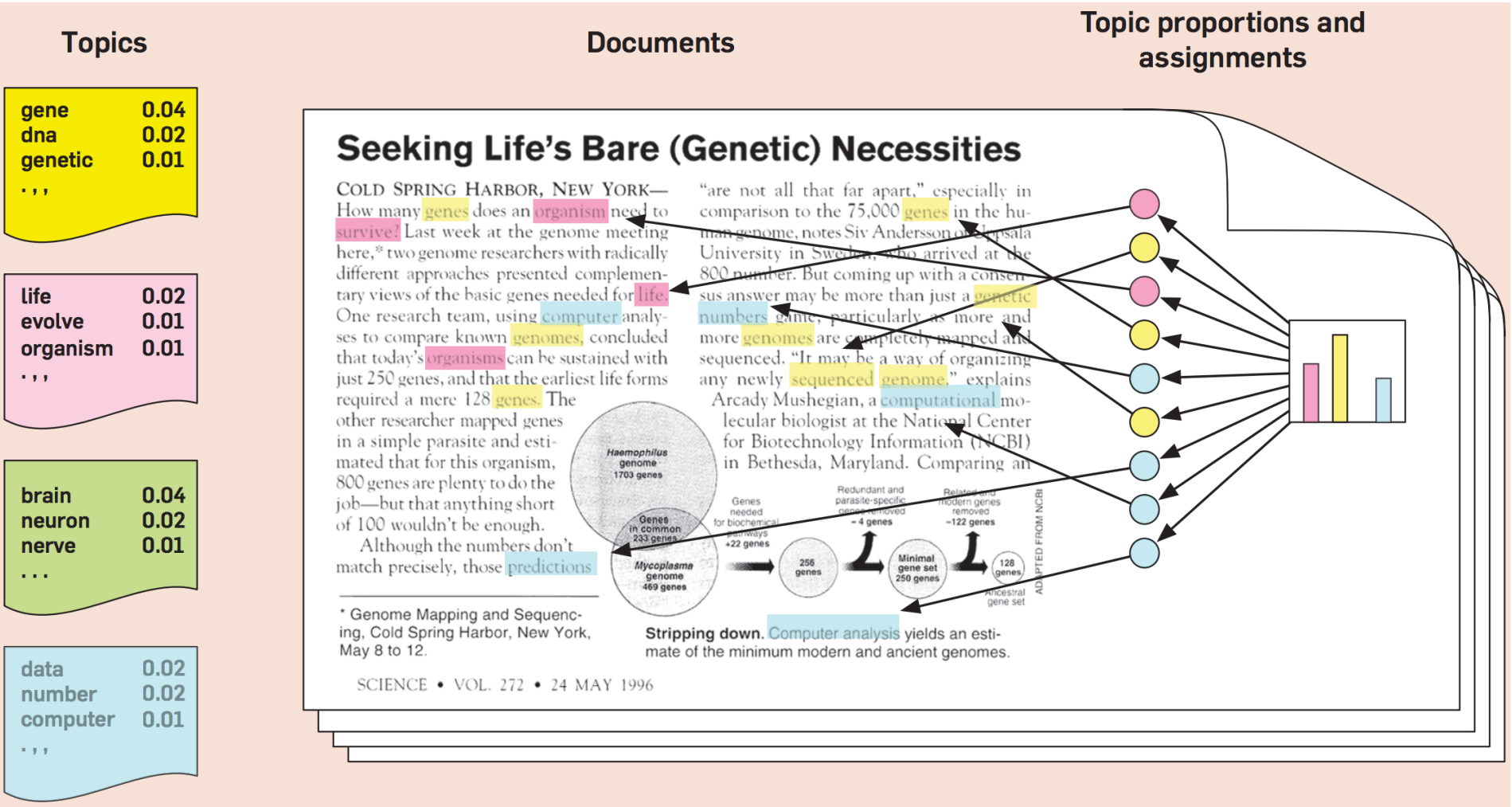
Source: Nitin Hardeniya (2015), NLTK Essentials, Packt Publishing; Florian Leitner (2015), Text mining - from Bayes rule to dependency parsing

Text Summarization



Source: Vishal Gupta and Gurpreet S. Lehal (2009), "A survey of text mining techniques and applications," Journal of emerging technologies in web intelligence, vol. 1, no. 1, pp. 60-76.

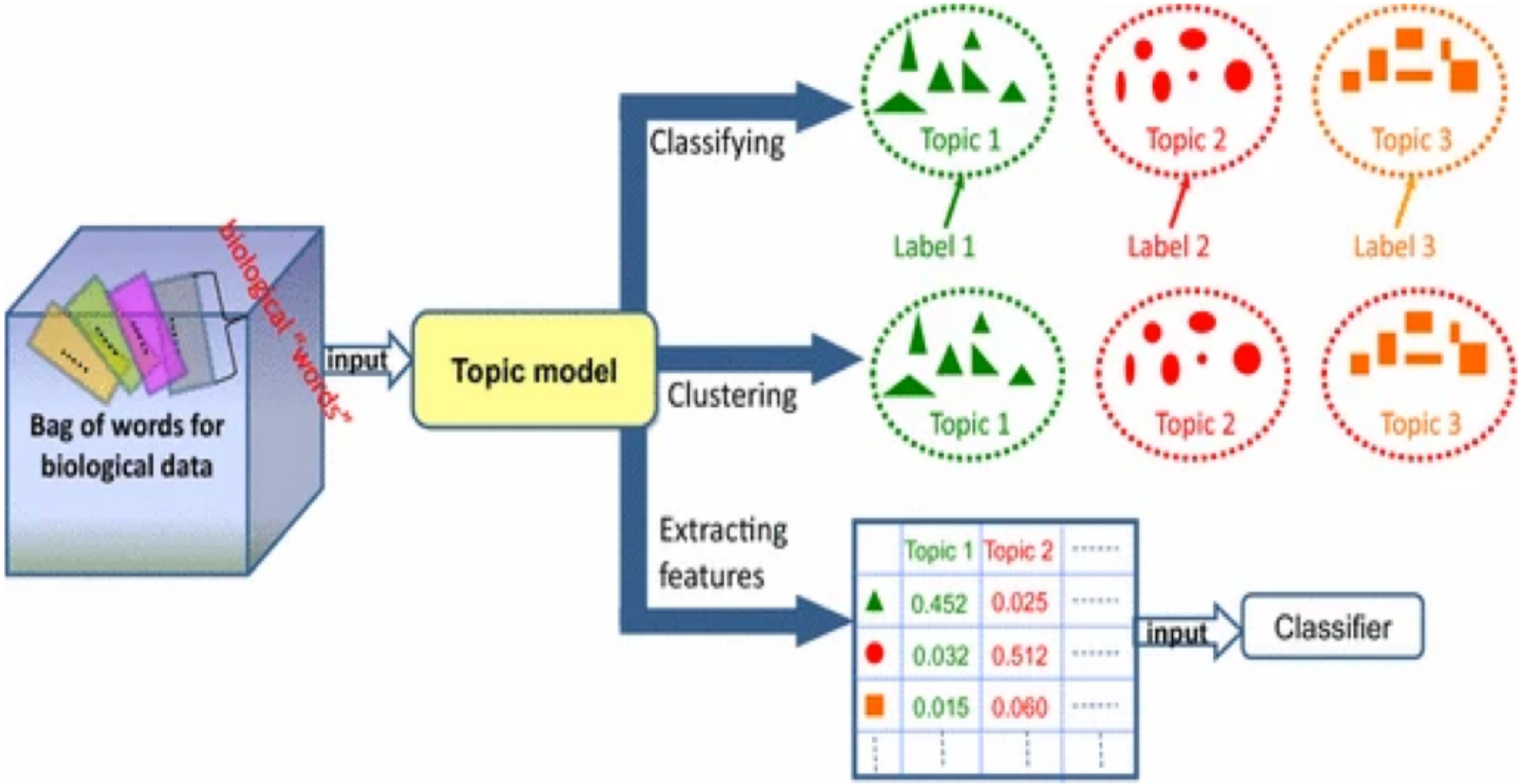
Topic Modeling



Text Summarization and Information Extraction

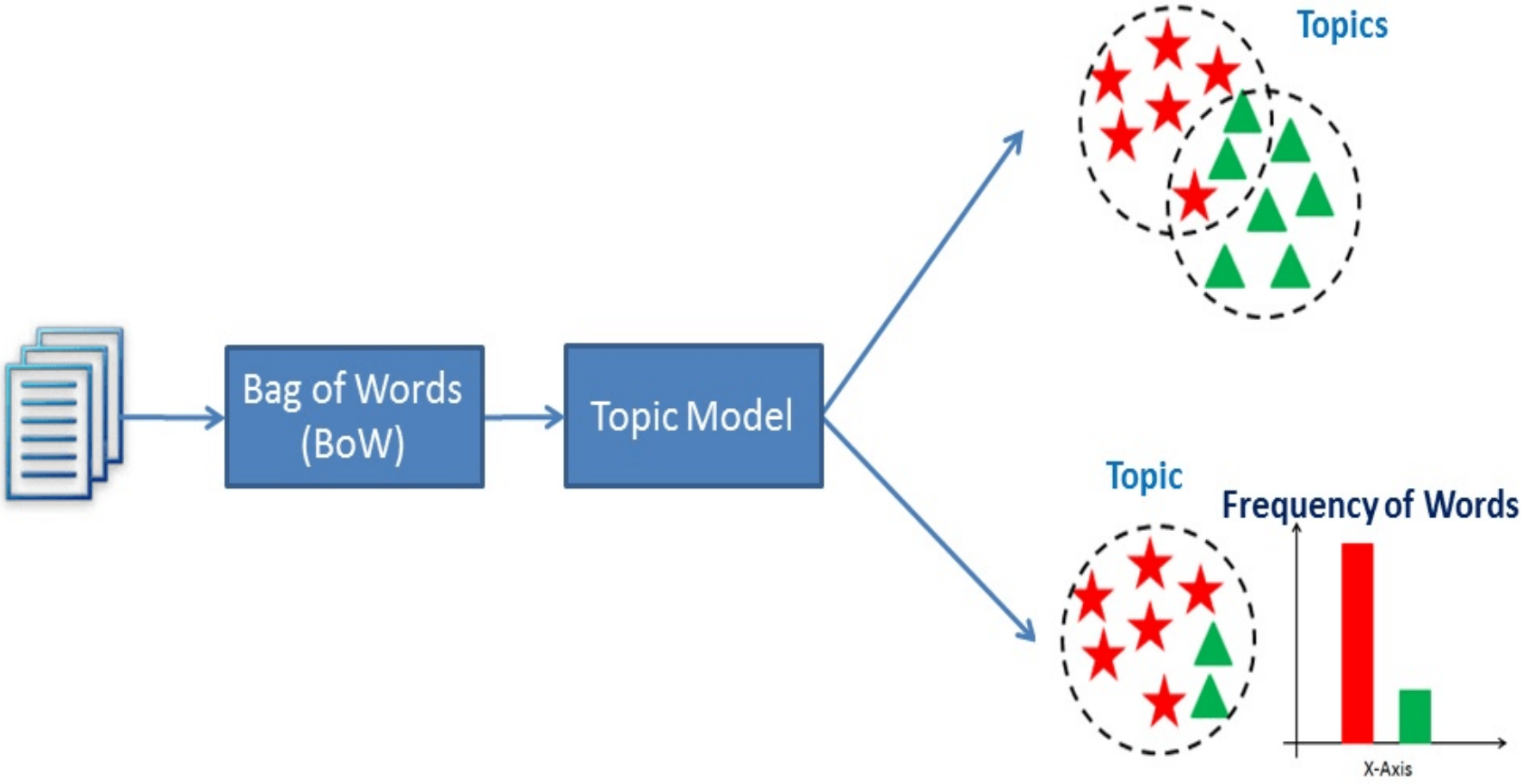
- Key-phrase extraction
  - extracting key influential phrases from the documents.
- Topic modeling
  - Extract various diverse concepts or topics present in the documents, retaining the major themes.
- Document summarization
  - Summarize entire text documents to provide a gist that retains the important parts of the whole corpus.

Topic Model in Bioinformatics



Source: Liu, Lin, Lin Tang, Wen Dong, Shaowen Yao, and Wei Zhou. "An overview of topic modeling and its current applications in bioinformatics." SpringerPlus 5, no. 1 (2016): 1608.

Topic Modeling

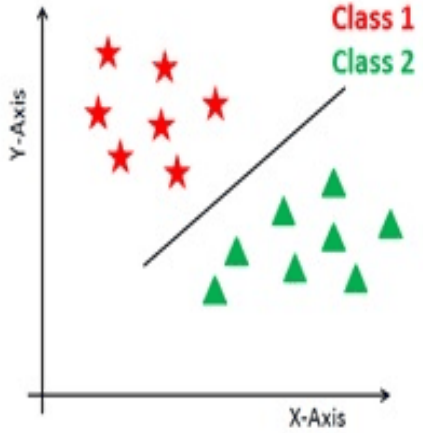


Source: Avinash Navlani (2018), Latent Semantic Analysis using Python,
https://www.datacamp.com/community/tutorials/discovering-hidden-topics-python

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

Bag of Words (BoW)

Text Classification

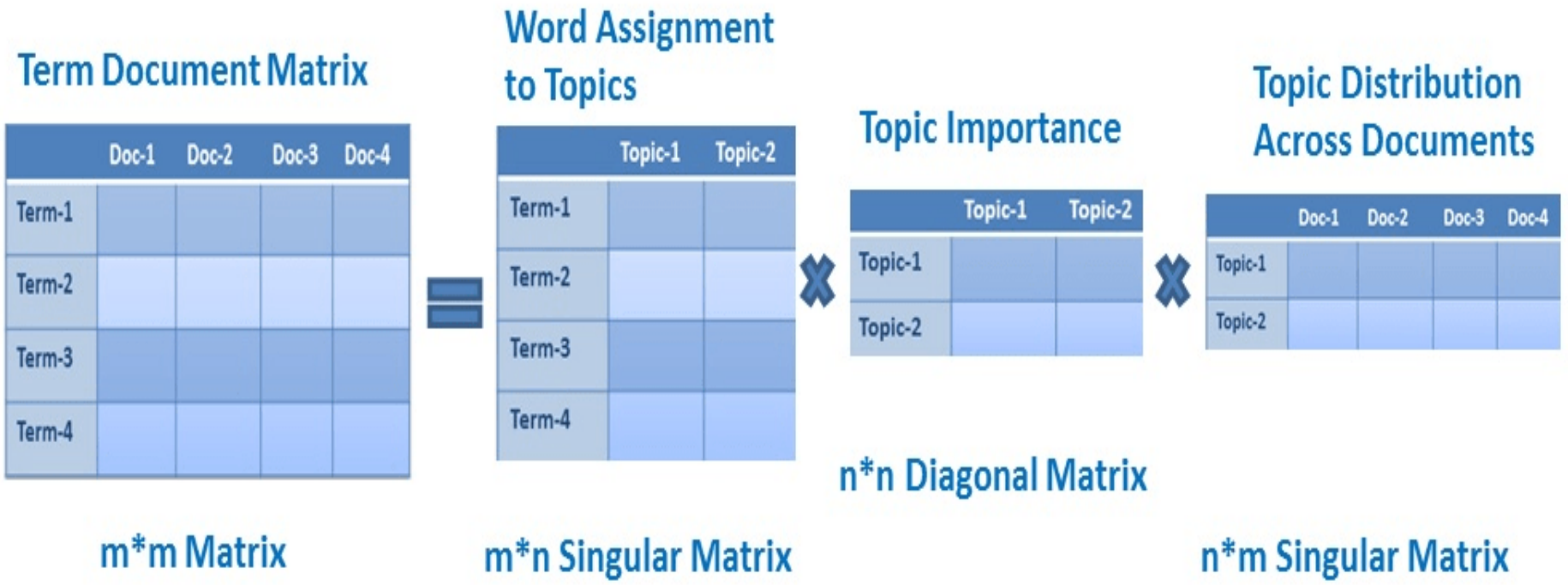


Topic Modeling

	Doc-1	Doc-2	Doc-3	Doc-4
Topic-1	0.425	0.032	0.015	0.553
Topic-2	0.052	0.512	0.623	0.121

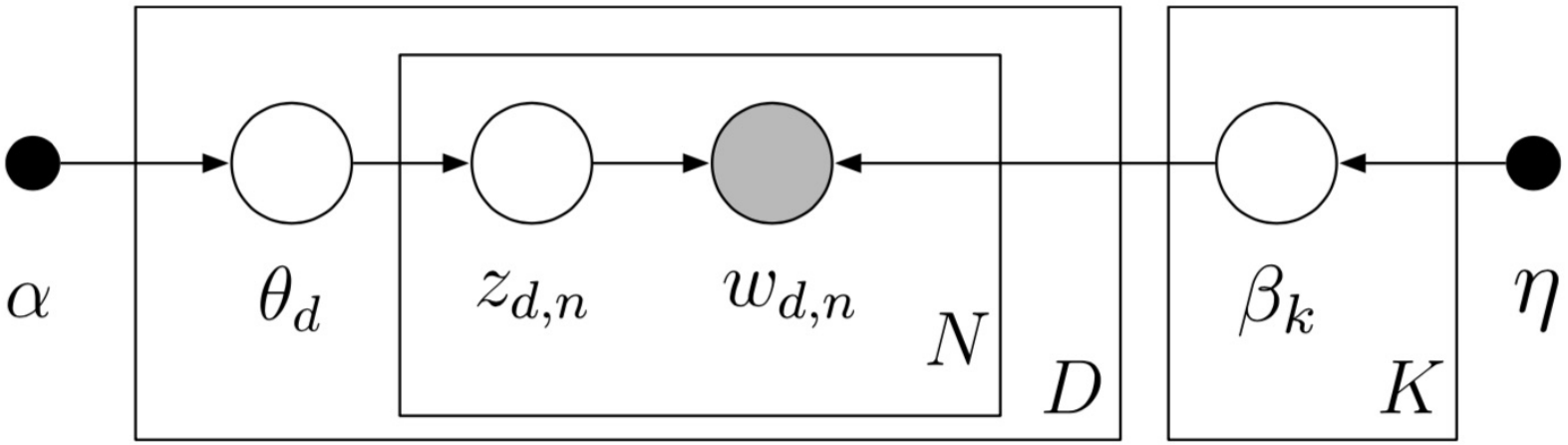
Source: Avinash Navlani (2018), Latent Semantic Analysis using Python, https://www.datacamp.com/community/tutorials/discovering-hidden-topics-python

Topic Modeling
Term Document Matrix to Topic Distribution



Source: Avinash Navlani (2018), Latent Semantic Analysis using Python, https://www.datacamp.com/community/tutorials/discovering-hidden-topics-python

Topic Modeling
Latent Dirichlet Allocation (LDA)



***D* documents**
***N* words**
***K* topics**

Source: Hoffman, Matthew D., David M. Blei, Chong Wang, and John Paisley. "Stochastic variational inference." The Journal of Machine Learning Research 14, no. 1 (2013): 1303-1347.

Latent Dirichlet Allocation (Blei et al., 2003)

Journal of Machine Learning Research 3 (2003) 993-1022 Submitted 2/02; Published 1/03

Latent Dirichlet Allocation

David M. Blei BLEI@CS.BERKELEY.EDU
Computer Science Division
University of California
Berkeley, CA 94720, USA

Andrew Y. Ng ANG@CS.STANFORD.EDU
Computer Science Department
Stanford University
Stanford, CA 94305, USA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
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.

Source: Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3, no. Jan (2003): 993-1022.

gensim

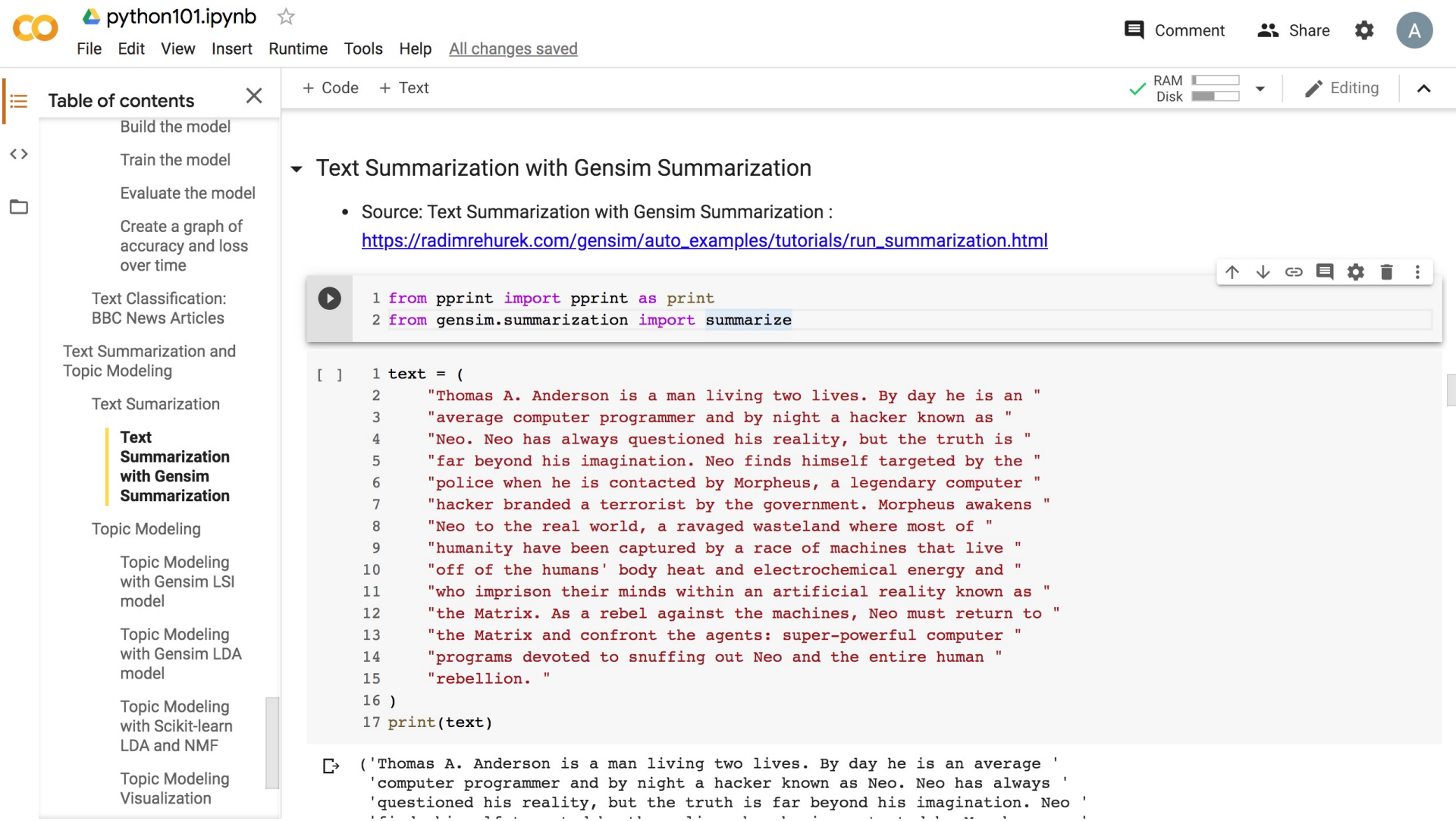
https://radimrehurek.com/gensim/

spaCy

https://spacy.io/

Python in Google Colab (Python101)

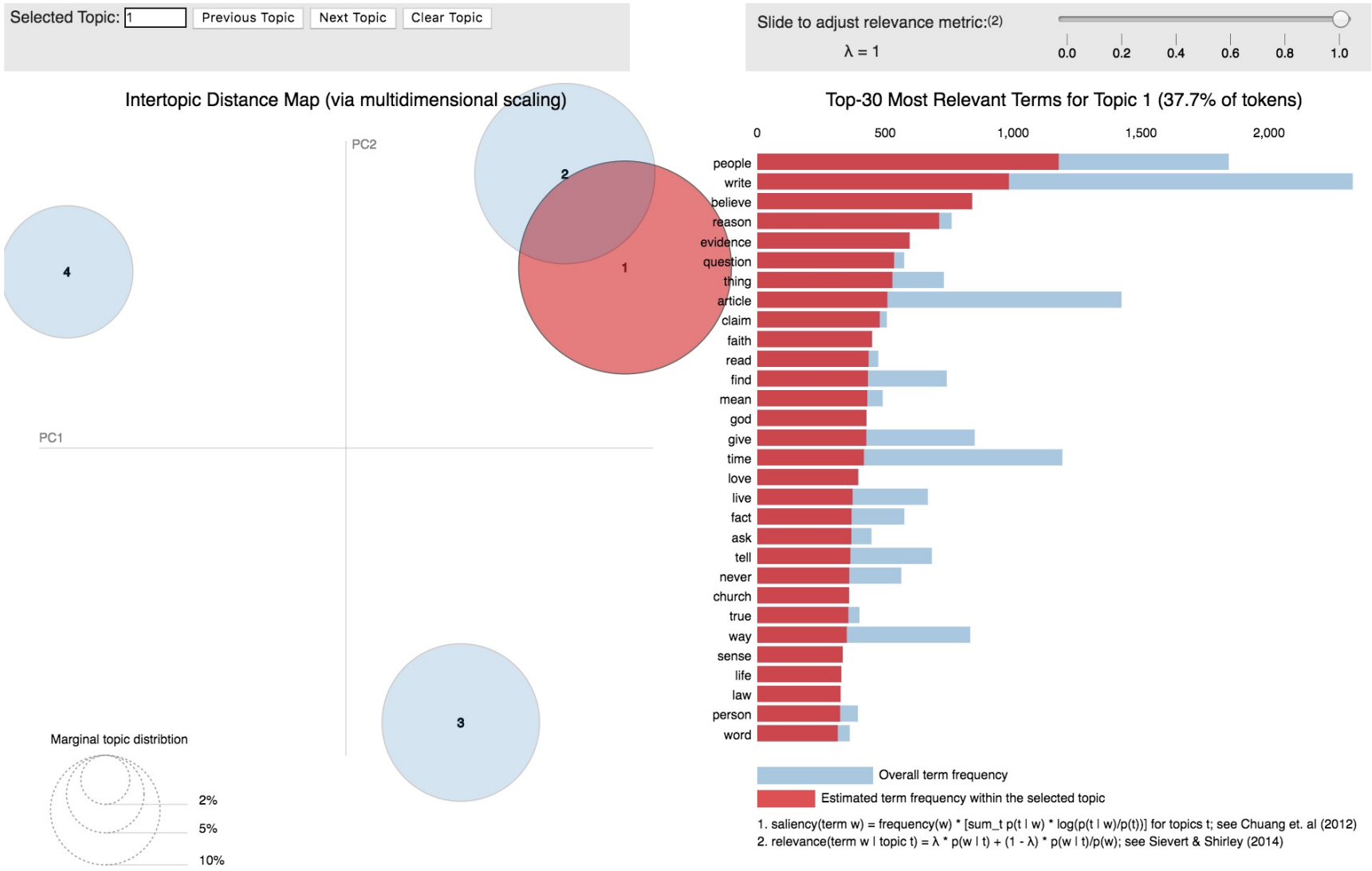
https://colab.research.google.com/drive/1FEG6DnGvwfUbeo4zJ1zTunjMqf2RkCrT



https://tinyurl.com/aintpupython101

Python in Google Colab (Python101)

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



https://tinyurl.com/aintpupython101

NLP Benchmark Datasets

Task	Dataset	Link
Machine Translation	WMT 2014 EN-DE WMT 2014 EN-FR	http://www-lium.univ-lemans.fr/~schwenk/cslm_joint_paper/
Text Summarization	CNN/DM Newsroom DUC Gigaword	https://cs.nyu.edu/~kcho/DMQA/ https://summari.es/ https://www-nlpir.nist.gov/projects/duc/data.html https://catalog.ldc.upenn.edu/LDC2012T21
Reading Comprehension Question Answering Question Generation	ARC CliCR CNN/DM NewsQA RACE SQuAD Story Cloze Test NarativeQA Quasar SearchQA	http://data.allenai.org/arc/ http://aclweb.org/anthology/N18-1140 https://cs.nyu.edu/~kcho/DMQA/ https://datasets.maluuba.com/NewsQA http://www.qizhexie.com/data/RACE_leaderboard https://rajpurkar.github.io/SQuAD-explorer/ http://aclweb.org/anthology/W17-0906.pdf https://github.com/deepmind/narrativeqa https://github.com/bdhingra/quasar https://github.com/nyu-dl/SearchQA
Semantic Parsing	AMR parsing ATIS (SQL Parsing) WikiSQL (SQL Parsing)	https://amr.isi.edu/index.html https://github.com/jkkummerfeld/text2sql-data/tree/master/data https://github.com/salesforce/WikiSQL
Sentiment Analysis	IMDB Reviews SST Yelp Reviews Subjectivity Dataset	http://ai.stanford.edu/~amaas/data/sentiment/ https://nlp.stanford.edu/sentiment/index.html https://www.yelp.com/dataset/challenge http://www.cs.cornell.edu/people/pabo/movie-review-data/
Text Classification	AG News DBpedia TREC 20 NewsGroup	http://www.di.unipi.it/~gulli/AG_corpus_of_news_articles.html https://wiki.dbpedia.org/Datasets https://trec.nist.gov/data.html http://qwone.com/~jason/20Newsgroups/
Natural Language Inference	SNLI Corpus MultiNLI SciTail	https://nlp.stanford.edu/projects/snli/ https://www.nyu.edu/projects/bowman/multinli/ http://data.allenai.org/scitail/
Semantic Role Labeling	Proposition Bank OneNotes	http://propbank.github.io/ https://catalog.ldc.upenn.edu/LDC2013T19

Source: Amirsina Torfi, Rouzbeh A. Shirvani, Yaser Keneshloo, Nader Tavvaf, and Edward A. Fox (2020). "Natural Language Processing Advancements By Deep Learning: A Survey." arXiv preprint arXiv:2003.01200.

Summary

- Text Summarization
- Topic Models
  - Topic Modeling
  - Latent Dirichlet Allocation (LDA)

References

- Dipanjan Sarkar (2019),
  Text Analytics with Python: A Practitioner's Guide to Natural Language Processing, Second Edition. APress. https://github.com/Apress/text-analytics-w-python-2e
- Benjamin Bengfort, Rebecca Bilbro, and Tony Ojeda (2018), Applied Text Analysis with Python,
  O'Reilly Media.
  https://www.oreilly.com/library/view/applied-text-analysis/9781491963036/
- 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/
- Min-Yuh Day (2020), Python 101, https://tinyurl.com/aintpupython101