



The Handbook of NLP with Gensim : Leverage Topic Modeling to Uncover Hidden Patterns, Themes, and Valuable Insights Within Textual Data

Kuo, Chris

Monografia

Elevate your natural language processing skills with Gensim and become proficient in handling a wide range of NLP tasks and projects

Key Features

- Advance your NLP skills with this comprehensive guide covering detailed explanations and code practices
- Build real-world topical modeling pipelines and fine-tune hyperparameters to deliver optimal results
- Adhere to the real-world industrial applications of topic modeling in medical, legal, and other fields

Purchase of the print or Kindle book includes a free PDF eBook

Book Description

Navigating the terrain of NLP research and applying it practically can be a formidable task made easy with *The Handbook of NLP with Gensim*. This book demystifies NLP and equips you with hands-on strategies spanning healthcare, e-commerce, finance, and more to enable you to leverage Gensim in real-world scenarios. You'll begin by exploring motives and techniques for extracting text information like bag-of-words, TF-IDF, and word embeddings. This book will then guide you on topic modeling using methods such as Latent Semantic Analysis (LSA) for dimensionality reduction and discovering latent semantic relationships in text data, Latent Dirichlet Allocation (LDA) for probabilistic topic modeling, and Ensemble LDA to enhance topic modeling stability and accuracy. Next, you'll learn text summarization techniques with Word2Vec and Doc2Vec to build the modeling pipeline and optimize models using hyperparameters. As you get acquainted with practical applications in various industries, this book will inspire you to design innovative projects. Alongside topic modeling, you'll also explore named entity handling and NER tools, modeling procedures, and tools for effective topic modeling applications. By the end of this book, you'll have mastered the techniques essential to create applications with Gensim and integrate NLP into your business processes.

What you will learn

- Convert text into numerical values such as bag-of-word, TF-IDF, and word embedding
- Use various NLP techniques with Gensim, including Word2Vec, Doc2Vec, LSA, FastText, LDA, and Ensemble LDA
- Build topical modeling pipelines and visualize the results of topic models
- Implement text summarization for legal, clinical, or other documents
- Apply core NLP techniques in healthcare, finance, and e-commerce
- Create efficient chatbots by harnessing Gensim's NLP capabilities

Who this book is for

This book is for data scientists and professionals who want to become proficient in topic modeling with Gensim. NLP practitioners can use this book as a code reference, while students or those considering a career transition will find this a valuable resource for advancing in the field of NLP. This book contains real-world applications for biomedical, healthcare, legal, and operations, making it a helpful guide for project managers designing their own topic modeling applications

Título: The Handbook of NLP with Gensim Leverage Topic Modeling to Uncover Hidden Patterns, Themes, and Valuable Insights Within Textual Data

Edición: 1st ed

Editorial: Birmingham Packt Publishing, Limited 2023 2023

Descripción física: 1 online resource (310 pages)

Contenido: Cover -- Title Page -- Copyright and Credits -- Contributors -- Table of Contents -- Preface -- Part 1: NLP Basics -- Chapter 1: Introduction to NLP -- Introduction to natural language processing -- NLU + NLG = NLP -- NLU -- NLG -- Gensim and its NLP modeling techniques -- BoW and TF-IDF -- LSA/LSI -- Word2Vec -- Doc2Vec -- LDA -- Ensemble LDA -- Topic modeling with BERTopic -- Common NLP Python modules included in this book -- spaCy -- NLTK -- Summary -- Questions -- References -- Chapter 2: Text Representation -- Technical requirements -- What word embedding is -- Simple encoding methods -- One-hot encoding -- BoW -- Bag-of-N-grams -- What TF-IDF is -- Shining applications of BoW and TF-IDF -- Coding - BoW -- Gensim for BoW -- scikit-learn for BoW (CountVectorizer) -- Coding - Bag-of-N-grams -- Gensim for N-grams -- scikit-learn for N-grams -- NLTK for N-grams -- Coding - TF-IDF -- Gensim for TF-IDF -- scikit-learn for TF-IDF -- Summary -- Questions -- References -- Chapter 3: Text Wrangling and Preprocessing -- Technical requirements -- Key steps in NLP preprocessing -- Tokenization -- Lowercase conversion -- Stop word removal -- Punctuation removal -- Stemming -- Lemmatization -- Coding with spaCy -- spaCy for lemmatization -- spaCy for PoS -- Coding with NLTK -- NLTK for tokenization -- NLTK for stop-word removal -- NLTK for lemmatization -- Coding with Gensim -- Gensim for preprocessing -- Gensim for stop-word removal -- Gensim for stemming -- Building a pipeline with spaCy -- Summary -- Questions -- References -- Part 2: Latent Semantic Analysis/Latent Semantic Indexing -- Chapter 4: Latent Semantic Analysis with scikit-learn -- Technical requirements -- Understanding matrix operations -- An orthogonal matrix -- The determinant of a matrix -- Understanding a transformation matrix -- A transformation matrix in daily life examples Understanding eigenvectors and eigenvalues -- An introduction to SVD -- Truncated SVD -- Truncated SVD for LSI -- Coding truncatedSVD with scikit-learn -- Using TruncatedSVD -- randomized_SVD -- Using TruncatedSVD for LSI with real data -- Loading the data -- Creating TF-IDF -- Using TruncatedSVD to build a model -- Interpreting the outcome -- Summary -- Questions -- Chapter 5: Cosine Similarity -- Technical requirements -- What is cosine similarity? -- How cosine similarity is used in images -- How to compute cosine similarity with scikit-learn -- Summary -- Questions -- References -- Chapter 6: Latent Semantic Indexing with Gensim -- Technical requirements -- Performing text preprocessing -- Performing word embedding with BoW and TF-IDF -- BoW -- TF-IDF -- Modeling with Gensim -- BoW -- TF-IDF -- Using the coherence score to find the optimal number of topics -- Saving the model for production -- Using the model as an information retrieval tool -- Loading the dictionary list -- Preprocessing the new document -- Scoring the document to get the latent topic scores -- Calculating the similarity scores with the new document -- Finding documents with high similarity scores -- Summary -- Questions -- References -- Part 3: Word2Vec and Doc2Vec -- Chapter 7: Using Word2Vec -- Technical requirements -- Introduction to Word2Vec -- Advantages of Word2Vec -- Reviewing the real-world applications of Word2Vec -- Introduction to Skip-Gram (SG) -- Data preparation -- The input and output layers -- The hidden layer -- Should I remove stop words for training Word2Vec? -- Model computation -- Introduction to CBOW -- Using a pretrained model for semantic search -- Adding and subtracting words/concepts -- Example 1 -- Example 2 -- Visualizing Word2Vec with TensorBoard -- Training your own Word2Vec model in CBOW and Skip-Gram -- Load the data -- Text preprocessing Training your own Word2Vec model in CBOW -- Training your own Word2Vec model in Skip-Gram -- Visualizing your Word2Vec model with t-SNE -- Comparing Word2Vec with Doc2Vec, GloVe, and fastText -- Word2Vec versus Doc2Vec -- Word2Vec versus GloVe -- Word2Vec versus FastText -- Summary -- Questions -- References -- Chapter 8: Doc2Vec with Gensim -- Technical requirements -- From Word2Vec to Doc2Vec -- PV-DBOW -- The input layer -- The hidden layer -- The output layer -- Model optimization -- PV-DM -- The real-world applications of Doc2Vec -- Doc2Vec modeling with Gensim -- Text preprocessing for Doc2Vec -- Modeling -- Saving the model -- Saving the training data -- Putting the model into production -- Loading the

model -- Loading the training data -- Use case 1 - find similar articles -- Use case 2 - find relevant documents based on keywords -- Tips on building a good Doc2Vec model -- Summary -- Questions -- References -- Part 4: Topic Modeling with Latent Dirichlet Allocation -- Chapter 9: Understanding Discrete Distributions -- Technical requirements -- The basics of discrete probability distributions -- Bernoulli distributions -- The formal definition of a Bernoulli distribution -- What does it look like? -- Fun facts -- Binomial distributions -- The real-world examples -- The formal definition of a binomial distribution -- What does it look like? -- Plotting it with Python -- Fun facts -- Multinomial distributions -- The real-world examples -- The formal definition of a multinomial distribution -- What does it look like? -- Fun facts -- Beta distributions -- The real-world examples -- The formal definition of a beta distribution -- What does it look like? -- The beta distribution in Bayesian inference -- Fun fact -- Dirichlet distributions -- Real-world examples -- The formal definition of a Dirichlet distribution -- What is a simplex? What does the Dirichlet distribution look like? -- The Dirichlet distribution in Bayesian inference -- Fun fact -- Summary -- Questions -- References -- Chapter 10: Latent Dirichlet Allocation -- What is generative modeling? -- Discriminative modeling -- Generative modeling -- Bayes' theorem -- Expectation-Maximization (EM) -- Understanding the idea behind LDA -- Dirichlet distribution of topics -- Understanding the structure of LDA -- Variational inference -- Variational E-M -- Gibbs sampling in LDA -- Variational E-M versus Gibbs sampling -- Summary -- Questions -- References -- Chapter 11: LDA Modeling -- Technical requirements -- Text preprocessing -- Preprocessing -- Experimenting with LDA modeling -- A model built on BoW data -- A model built on TF-IDF data -- Building LDA models with a different number of topics -- Models built on BoW data -- Models built on TF-IDF data -- Determining the optimal number of topics -- Using the model to score new documents -- Text preprocessing -- Scoring new texts -- Outcome -- Summary -- Questions -- References -- Chapter 12: LDA Visualization -- Technical requirements -- Designing an infographic -- Data visualization with pyLDAvis -- The interactive graph -- Summary -- Questions -- References -- Chapter 13: The Ensemble LDA for Model Stability -- Technical requirements -- From LDA to Ensemble LDA -- The process of Ensemble LDA -- Understanding DBSCAN and CBDBSCAN -- DBSCAN -- CBDBSCAN (Checkback DBSCAN) -- Building an Ensemble LDA model with Gensim -- Preprocessing the training data -- Creating text representation with BOW and TF-IDF -- Saving the dictionary -- Building the Ensemble LDA model -- Scoring new documents -- Summary -- Questions -- References -- Part 5: Comparison and Applications -- Chapter 14: LDA and BERTopic -- Technical requirements -- Understanding the Transformer model Understanding BERT -- Describing how BERTopic works -- BERT - word embeddings -- UMAP - reduce the dimensionality of embeddings -- HDBSCAN - cluster documents -- c-TFIDF - create a topic representation -- Maximal Marginal Relevance -- Building a BERTopic model -- Loading the data - no text preprocessing -- Modeling -- Reviewing the results of BERTopic -- Getting the topic information -- Inspecting the keywords of a single topic -- Getting document information -- Getting representative documents -- Visualizing the BERTopic model -- Visualizing topics -- Visualizing the hierarchy of topics -- Visualizing the top words of topics -- Visualizing on a heatmap -- Predicting new documents -- Using the modular property of BERTopic -- Word embeddings -- Dimensionality reduction -- Clustering -- Comparing BERTopic with LDA -- Approach -- Word embeddings -- Text preprocessing -- Language understanding -- Topic clarity -- Determination of the number of topics -- Determination of word significance in a topic -- Summary -- Questions -- References -- Chapter 15: Real-World Use Cases -- Word2Vec for medical fraud detection -- Background -- Questions -- NLP solution -- Takeaways -- Background -- Questions -- NLP solution -- Takeaways -- Background -- Questions -- NLP solution -- Takeaways -- Comparing LDA/NMF/BERTopic on Twitter/X posts -- Background -- Questions -- NLP solution -- Takeaways -- Interpretable text classification from electronic health records -- Background -- Questions -- NLP solution -- Takeaways -- BERTopic for legal documents -- Background -- Questions -- NLP solution -- Takeaways -- Word2Vec for 10-K financial documents to the SEC -- Background -- Questions -- NLP solution -- Takeaways -- Summary -- References -- Assessments -- Chapter 1 - Introduction to NLP -- Chapter 2 - Text Representation -- Chapter 3 - Text Wrangling and Preprocessing Chapter 4 - Latent Semantic Analysis with scikit-learn

ISBN: 1-80324-550-6

Materia: Natural language processing (Computer science)- Software Python (Computer program language) Open source software

Enlace a formato físico adicional: 9781803244945

Baratz Innovación Documental

- Gran Vía, 59 28013 Madrid
- (+34) 91 456 03 60
- informa@baratz.es