2024-2 Unstructured Data Analysis
2024.10.07

Text Representation

Week 6 – Code Practice

Professor: Misuk Kim
Teaching Assistant: Minjoo Son
minjoo77@hanyang.ac.kr

HANYANG UNIVERSITY

Text Representation
Week 6 – Code Practice

Contents

1. Review

2. Text Representation

3. Document Classification

4. Assignment

1. Review

❖ Text Preprocessing

- Purpose: Prepares text according to the problem's purpose

- A great website to understand the process of text preprocessing
  - English: https://chaelist.github.io/docs/text_analysis/english_text/
    - Text Cleaning: Remove unnecessary symbols/expressions (e.g., !, ., ", ;, etc.)
    - Case Conversion: Convert between uppercase and lowercase letters (lowercase ↔ uppercase)
    - Tokenization: Split text into words or tokens
    - POS Tagging: Identify the part of speech (POS) of each word
    - Select Desired POS: Choose only the words with specific parts of speech
    - Lemmatization (or Stemming): Find the base form (or root) of a word
    - Stopwords Removal: Eliminate common stopwords
  - Korean: https://chaelist.github.io/docs/text_analysis/korean_text/

1. Review

❖ Text Preprocessing

- In-Class Task Sample Code: https://colab.research.google.com/drive/1kdNC7tpiDnzTugJ8H2ELIpa2nUjEZirM?usp=drive_link

1. Reveiw

❖ Text Representation

- What we've learned in the past lectures?
  - Classic Methods
    - Bag of Words (BOW)
    - TF-IDF
    - N-Grams
  - Distributed Methods
    - Word2Vec (CBOW & Skip gram)
    - FastText
    - GloVe
    - Sentence/Paragraph/Document level

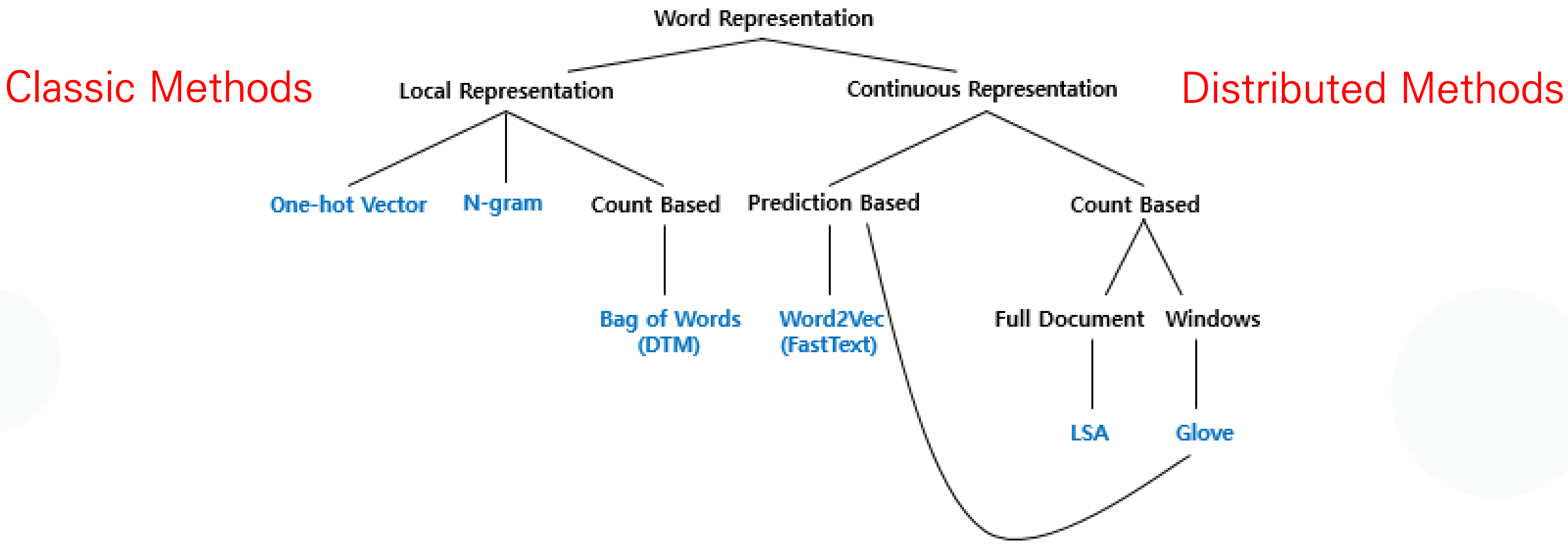
1. Review

❖ Text Representation

- What will we do in today's lectures?
  - – Practicing basic text representation methods in code
    - ▪ Classic methods
    - ▪ Distributed methods
  - – Document Classification
    - ▪ Classic methods
    - ▪ Distributed methods → Assignments!

❖ Text Representation

- For computers to understand and efficiently process text, it's essential to convert the text into numbers that the computers can interpret. (Feature vector extraction!)
- There are several methods for representing text in natural language processing.
  - Local Representation: This method maps a specific value to a word by looking only at the word itself.
  - Distributed Representation: This method represents a word by considering the surrounding words to capture its meaning.



Source: https://wikidocs.net/31767

2. Text Representation

❖ One-Hot Encoding

- It is one of the most basic methods for representing words among these techniques.
- It involves setting the dimension of the vector to the size of the vocabulary, assigning a value of 1 to the index of the word you want to represent, and assigning 0 to other indices.
- Vectors where most of the dimensions are 0 are called sparse vectors.

Red →

Red	Orange	Yellow	Green	...	Purple	Black
1	0	0	0	...	0	0

Yellow →

Red	Orange	Yellow	Green	...	Purple	Black
0	0	1	0	...	0	0

Green →

Red	Orange	Yellow	Green	...	Purple	Black
0	0	0	1	...	0	0

Limitations

- As the number of words increases, the space required to store vectors also continues to grow, which is a drawback.
- Additionally, it does not capture the similarity between words

2. Text Representation

❖ Distributed Representation

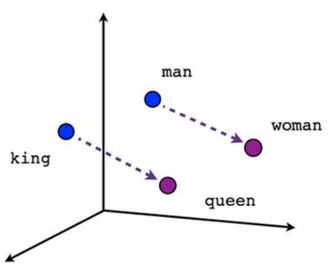
- Words are represented by distributing their meanings across multiple dimensions.
- Distributional Hypothesis: Words that appear in similar contexts tend to have similar meanings.
- Unlike one-hot vectors, where the dimension of the vector equals the size of the vocabulary, these methods allow the vector dimensions to be relatively lower.
- Learning methods include NNLM, RNNLM, Word2Vec, GloVe, FastText, etc.

Red	→	0.48	-0.3	1.8	1.1	-2.1	2.8	0.12
Yellow	→	0.17	2.13	-0.2	4.1	2.99	-3.02	0.04
Green	→	-0.27	4.36	-0.41	3.6	0.8	1.3	5.03

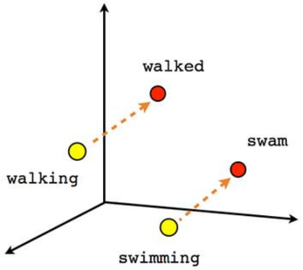
2. Text Representation

❖ Distributed Representation

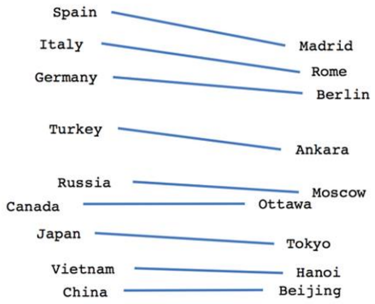
- Calculates similarity between words



Male-Female

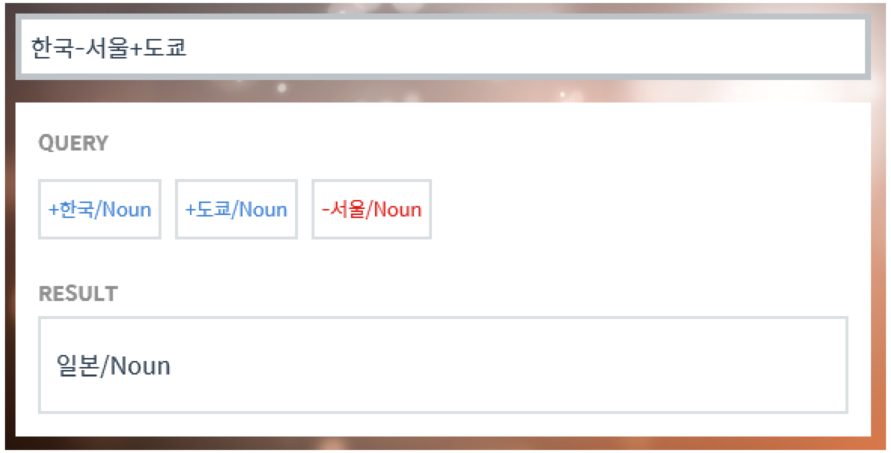


Verb tense



Country-Capital

- allows for algebraic operations



한국-서울+도쿄

QUERY

+한국/Noun +도쿄/Noun -서울/Noun

RESULT

일본/Noun

Source: https://glanceyes.com/entry/%EC%B6%94%EC%B2%9C-%EC%8B%9C%EC%8A%A4%ED%85%9C-Word2Vec%EC%9D%84-%EC%9D%91%EC%9A%A9%ED%95%9C-Item2Vec

2. Text Representation

❖ Code Practice

- Classic Methods
  - https://drive.google.com/file/d/1No4dnmXkVnvT5fJFx7O8gDEcOhA2Xc4Q/view?usp=drive_link (1. Classic Methods.ipynb)

- Distributed Methods
  - https://drive.google.com/file/d/1-ovCb-zFoQCfgz_w2s4QTTYVcRIwnspf/view?usp=drive_link (2. Distributed Methods.ipynb)

3. Document Classification

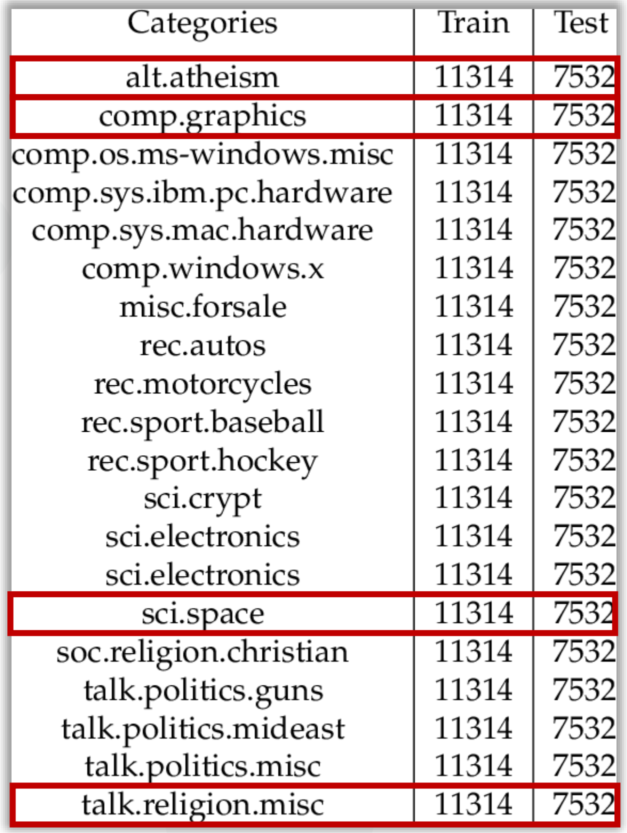
❖ Document Classification

- Document classification refers to the task of categorizing a given document into predefined classes.
  - For example, when we read a news article, we can usually tell if it belongs to categories like politics, economics, entertainment, or sports.
  - Using the text feature extraction and representation methods learned earlier, we will explore various approaches to text document classification.
  - Classification is a prominent field in machine learning.
    - When machine learning is broadly divided into supervised and unsupervised learning, classification falls under supervised learning. Various machine learning techniques exist for classification, including logistic regression, decision trees, Naive Bayes, and SVM.
    - Since document classification is a type of supervised learning, every document or text must have a label or a predefined class for the learning process.

3. Document Classification

❖ Document Classification

- Code Practice
  - 20 newsgroup dataset
    - In Scikit-learn, the 'sklearn.datasets.fetch_20newsgroups' module allows you to download over 18,000 documents belonging to 20 topics or categories.
    - You can use the 'categories' parameter to select specific topics from the 20 available topics.
    - The 'remove' parameter lets you delete unnecessary data.
    - Within each dataset, '.data' is used to retrieve the text content, and '.target' is used to get the labels (categories) represented as numbers.
    - In this class, we will retrieve documents from four topics: 'alt.atheism', 'talk.religion.misc', 'comp.graphics', and 'sci.space'. Since the topic names are sometimes included in the headers or footers, we use the 'remove' parameter to eliminate these hints.
    - The 'subset' parameter is used to distinguish between the training and evaluation datasets.



Categories	Train	Test
alt.atheism	11314	7532
comp.graphics	11314	7532
comp.os.ms-windows.misc	11314	7532
comp.sys.ibm.pc.hardware	11314	7532
comp.sys.mac.hardware	11314	7532
comp.windows.x	11314	7532
misc.forsale	11314	7532
rec.autos	11314	7532
rec.motorcycles	11314	7532
rec.sport.baseball	11314	7532
rec.sport.hockey	11314	7532
sci.crypt	11314	7532
sci.electronics	11314	7532
sci.electronics	11314	7532
sci.space	11314	7532
soc.religion.christian	11314	7532
talk.politics.guns	11314	7532
talk.politics.mideast	11314	7532
talk.politics.misc	11314	7532
talk.religion.misc	11314	7532

3. Document Classification

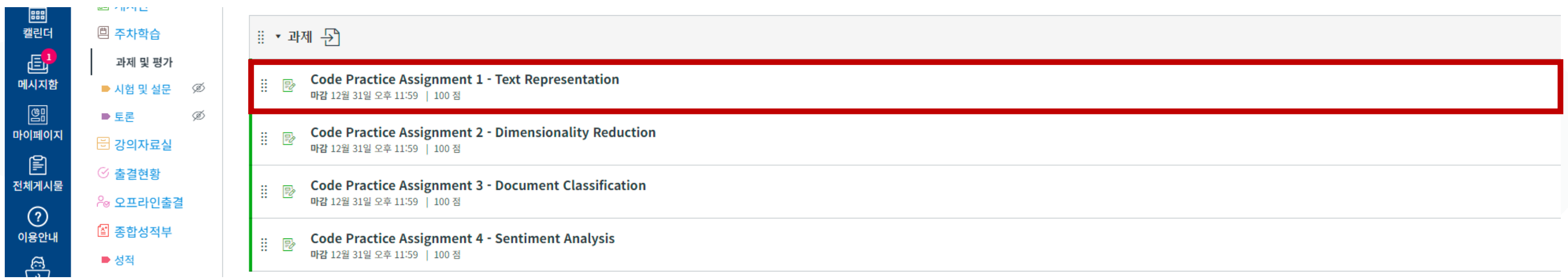
❖ Document Classification

- Code Practice
  - Classic Methods
    - https://drive.google.com/file/d/17pLXGRaTeLJDikHRhbaTVLivtsAlyP4W/view?usp=drive_link (3. Document Classification_Classic Methods Based.ipynb)
  - Distributed Methods
    - https://drive.google.com/file/d/1tQbavYxfWJ8EB1PVr2KaGsQq1cKTxqDL/view?usp=drive_link (Assignment.ipynb)

4. Assignment

❖ Document Classification

- Code Practice
  - – Distributed Methods
    - ▪ https://drive.google.com/file/d/1tQbavYxfWJ8EB1PVr2KaGsQq1cKTxqDL/view?usp=drive_link (Assignment.ipynb)
  - – Referencing the '1) Word2Vec' code from the provided 'Assignment.ipynb' file, create document representations using the '2) FastText' and '3) GloVe' models, and then train a classification model to record the 'Train Accuracy' and 'Test Accuracy’.
    - – (Optional) Try applying various parameters of the existing word embedding model and machine learning models to maximize performance.
  - – Save the file as 'Assignment_YourName_YourStudentID.ipynb' and submit it to the 'Code Practice Assignment 1 – Text Representation' section under 'Assignments’
  - – by ~~9:00 am on Tuesday, October 8 (tomorrow).~~ → 23:00 pm on Monday, October 14



Code Practice Assignment 1 - Text Representation
마감 12월 31일 오후 11:59 | 100 점

Code Practice Assignment 2 - Dimensionality Reduction
마감 12월 31일 오후 11:59 | 100 점

Code Practice Assignment 3 - Document Classification
마감 12월 31일 오후 11:59 | 100 점

Code Practice Assignment 4 - Sentiment Analysis
마감 12월 31일 오후 11:59 | 100 점

Text Representation
Week 6 – Code Practice

2024-2 Unstructured Data Analysis
2024.10.07

Q & A

Thank you for your attention. Any questions are welcome!

Minjoo Son

HANYANG UNIVERSITY