

TEXT MINING

Lecture 03

TEXT MINING PRINCIPLES

KEUNGOU I KIM

awekim@handong.edu



Data Science

- Data
 - Record of history
 - Structured data → excel sheet
 - Unstructured data → text, audio, video
- Data Science
 - Mathematics / Statistics → probability, linear algebra, descriptive statistics, inferential statistics, Bayes' theorem
 - Computer science → R, Python, SQL
 - Domain Knowledge

- Data value
 - Either numerical or non-numerical value

2022

Mathematical computation
Statistical computation

“2022”

“two zero two two”

“two thousand and twenty two”

“twenty hundred and twenty two”

Mathematical computation
Statistical computation



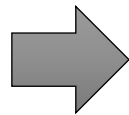
1. The meaning of the value that the computer understands is different.
2. The way computers handle values is different.

Data Science 2) Data Dimension

- In Data Science, data is recognized as a “dimension”
 - Big Data: Countless Variables and Countless Observations → Countless Dimensions and Countless Objects
 - Each dimension is "independent"

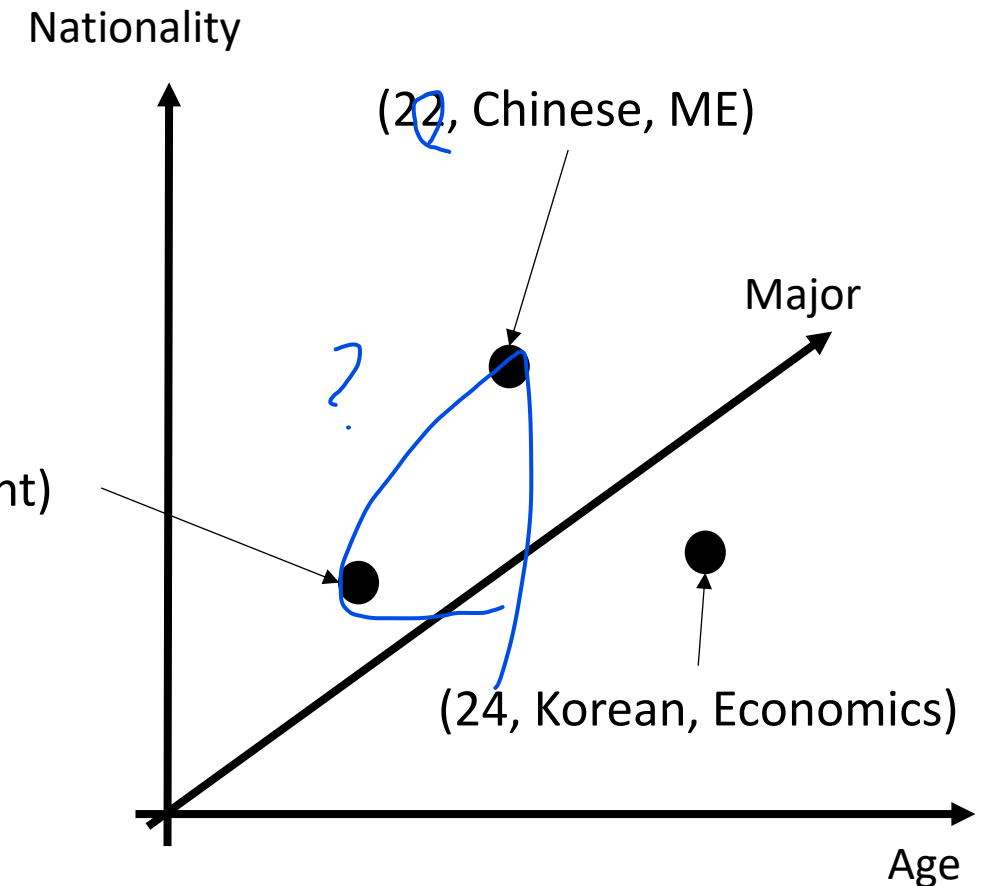
Name	Age	Nationality	Major
Kim	22	Korean	Management
Lee	24	Korean	Economics
Wang	22	Chinese	Mechanical Engineering

3 people
3 attributes
3 values



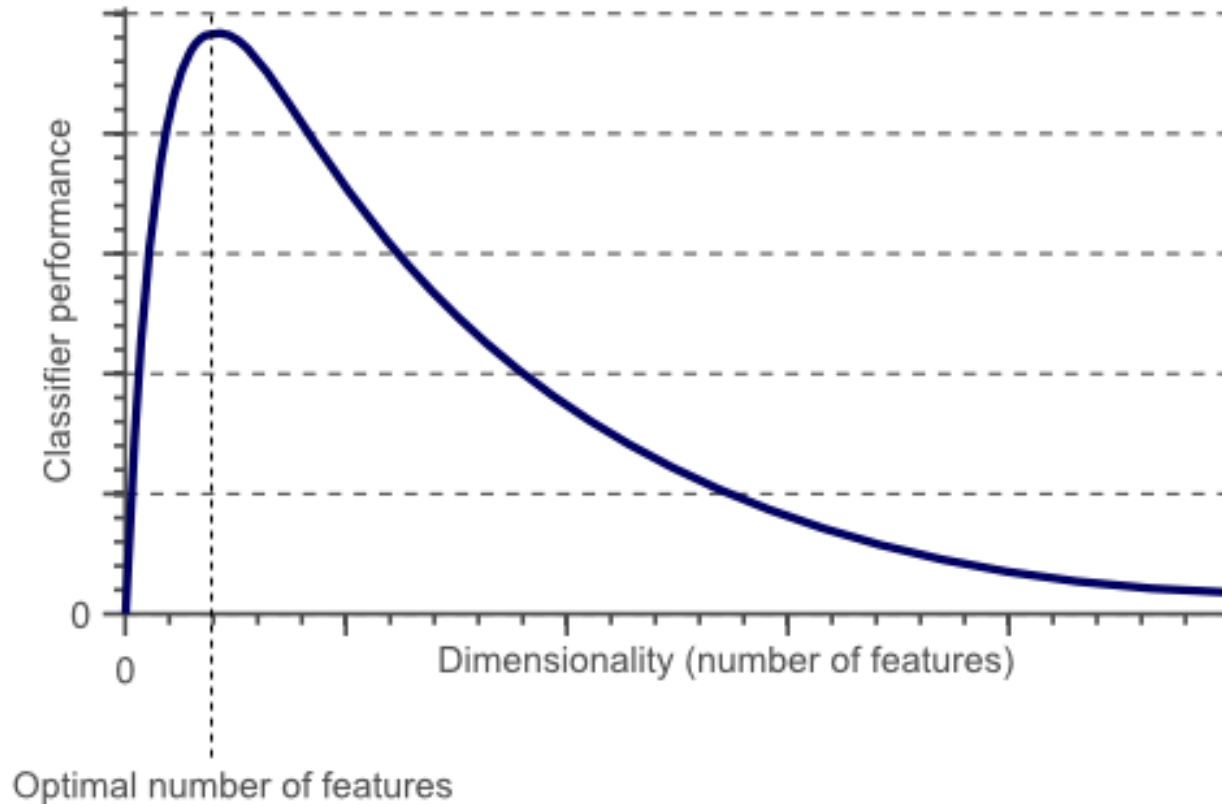
3 objects
3 axis
3 lengths

(22, Korean, Management)



- Dimensional Curse

- number of objects < number of dimensions
- Variables increase → More information is available.



Dimensional reduction: PCA (principal component analysis), LSA (Linear Discriminant Analysis), MDS (Multidimensional scaling), SVD (Singular Value Decomposition), LL (Locally-Linear Embedding), Kernel Principal Component Analysis, etc.

Data Science 3) Data Pre-processing

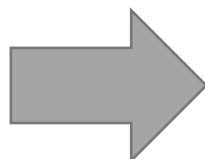
- Turning data into an “analyzable form”
 - Data processing work tailored to the “input value” of the data analysis methodology to be performed
 - Tidy data: A form in which one row has an independent value

Outliers

Missing values

Team	Game	W	L	Coach	Pt
ManC	38	29	3	Pep	93
Liv	38	28	-2	Klopp	
Chel	38	21	6	Tuchel	74
Tot	38	22	40	Konte	71
Ars	38	22	13	Arteta	69
ManU	38	16	12	Rangel	58
Man9	38	16	12	Rangel	58

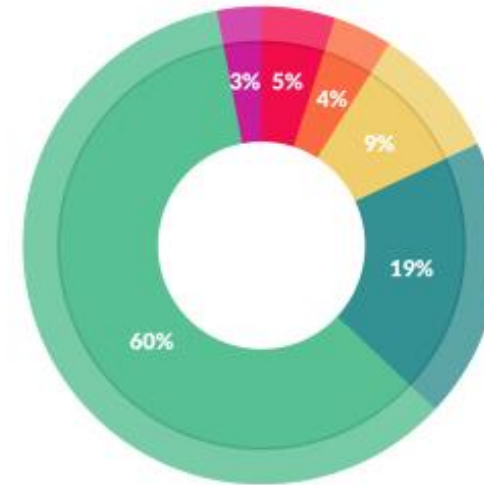
Duplicates



Team	Game	W	L	Coach	Pt
ManC	38	29	3	Pep	93
Liv	38	28	2	Klopp	92
Chel	38	21	6	Tuchel	74
Tot	38	22	11	Konte	71
Ars	38	22	13	Arteta	69
ManU	38	16	12	Rangel	58

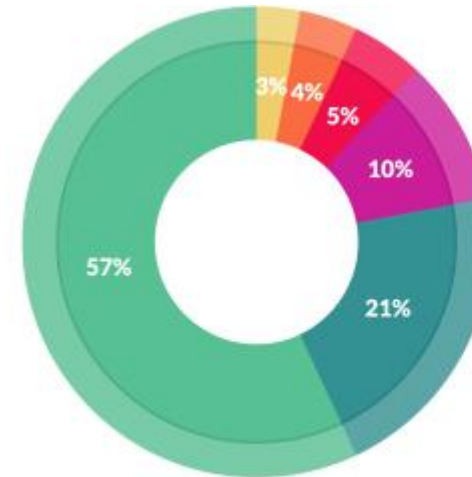
Data Science 3) Data Pre-processsing

- In most cases, data sets are imperfect for data analysis
 - Data cleansing decides the “quality” and “performance” of data analysis
 - It takes most times...
- Two purposes
 - Creating a “tidy” data
 - Cleansing a “messy” data



What data scientists spend the most time doing

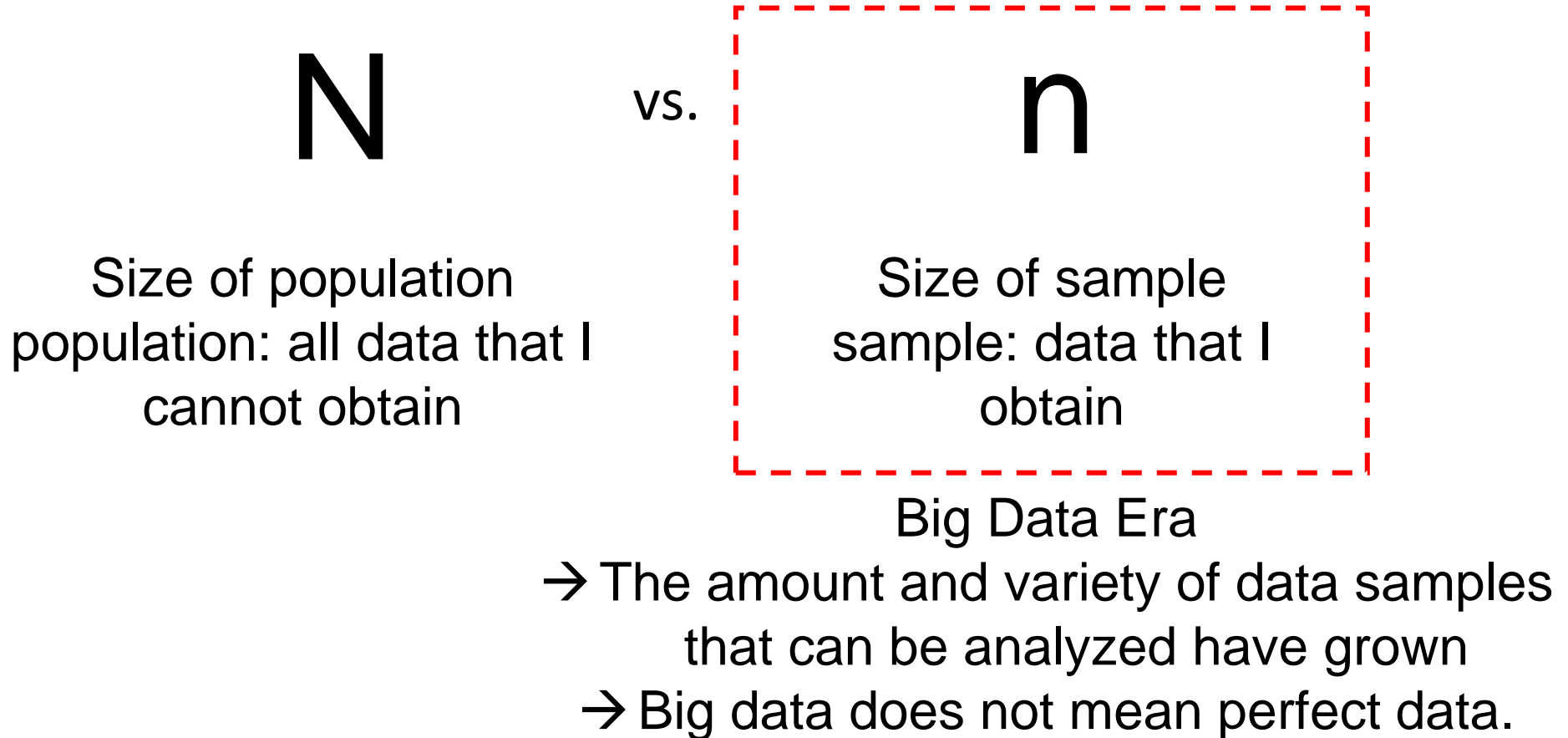
- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%



What's the least enjoyable part of data science?

- Building training sets: 10%
- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Other: 5%

- No data is perfect
 - Philosophy of Statistics: I cannot know everything.

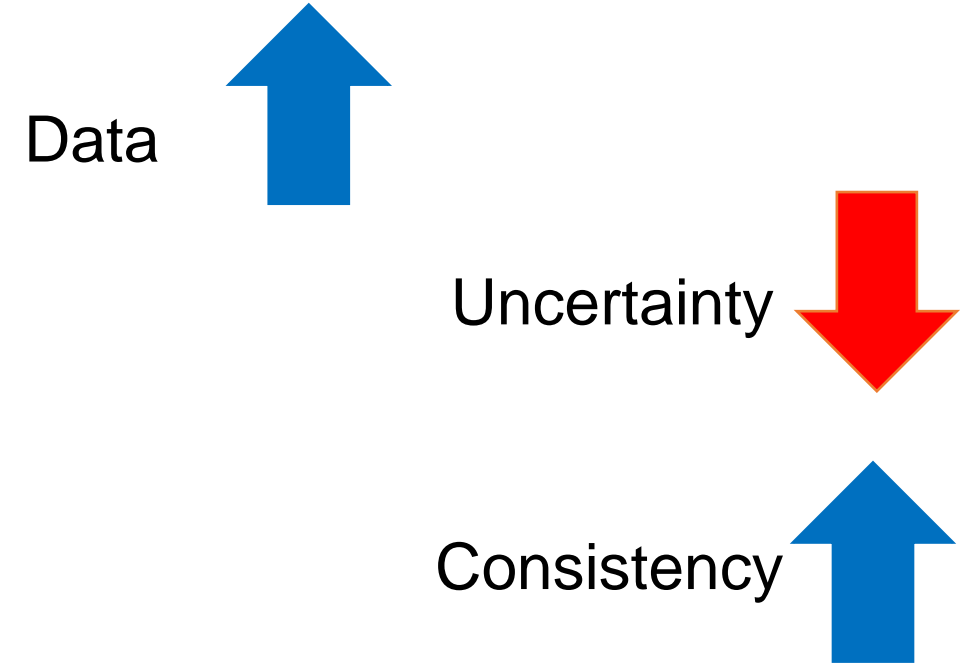


Data Science 4) Uncertainty

- No data is perfect



“Is it safe under extreme repetitive conditions?”
→ Removal of uncertainty about new cars



However, uncertainty still exists!!!

- No data is perfect



Winning move

=

Unknown information
that my algorithm has
not experienced yet

Text Data

Features of Text Data

- Value of text data is NOT numerical
 - Raw text data cannot be used for calculation, but it is highly compressed.

22

“22”

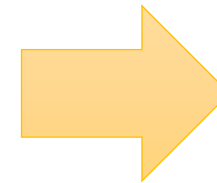
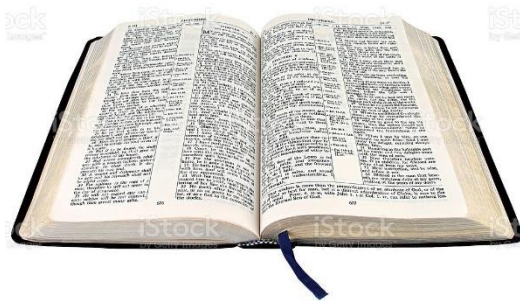
“Twenty-two”

“Two-two”

Text data



WIKIPEDIA
La enciclopedia libre



knowledge, opinion,
philosophy, belief + α

Features of Text Data

- Value of text
 - Value of bible – value of a book? Or value of content?
 - Value of bible as a book → 19,850 won
 - Value of bible as a content → ??



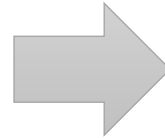
필사본 성경
500굴덴 (약 5억원)

루터 성경
1.5굴덴 (약 150만원)

- 1굴덴은 당시 일반 노동자들의 2~3주 급료
- 대학교수였던 루터의 월급은 8굴덴

- Unstructured format
 - No data structure = No columns and no rows
 - Can't conduct data analysis with raw text data

“Chelsea and Tottenham managers were both sent off after contentious 2-2 draw.”



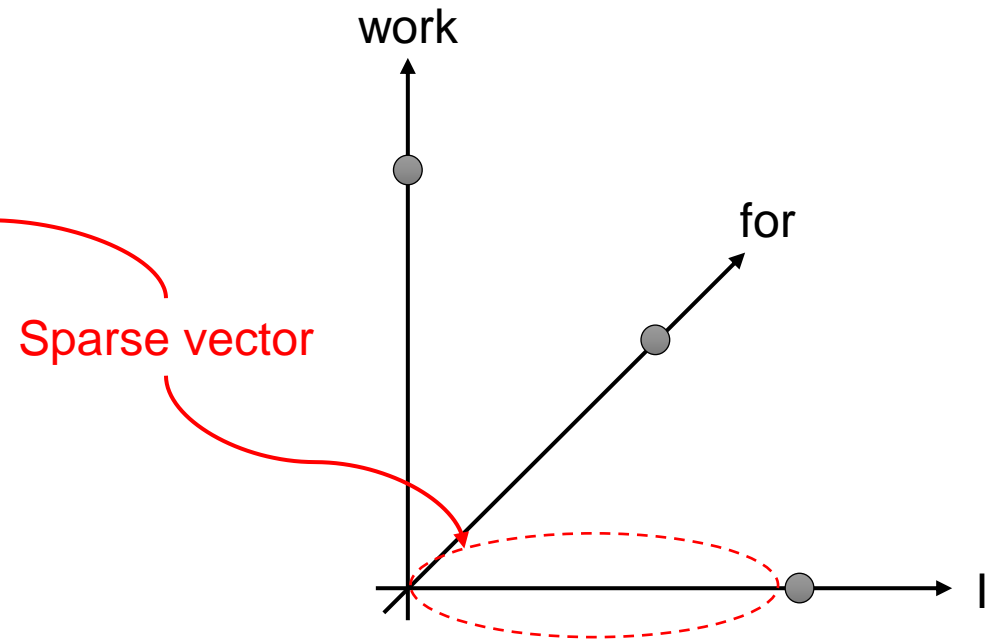
DataFrame

Features of Text Data

- Unstructured format
 - Cannot compute mathematical operations with text variables → text encoding?
 - Text encoding: converting text into a numerical value
 - The problem is that the encoded text does not have any meaning

	I	work	for	hyundai	motors
I	1	0	0	0	0
work	0	1	0	0	0
for	0	0	1	0	0
hyundai	0	0	0	1	0

I = [1, 0, 0, 0, 0]
work = [0, 1, 0, 0, 0]
for = [0, 0, 1, 0, 0]
hyundai = [0, 0, 0, 1, 0]



Features of Text Data

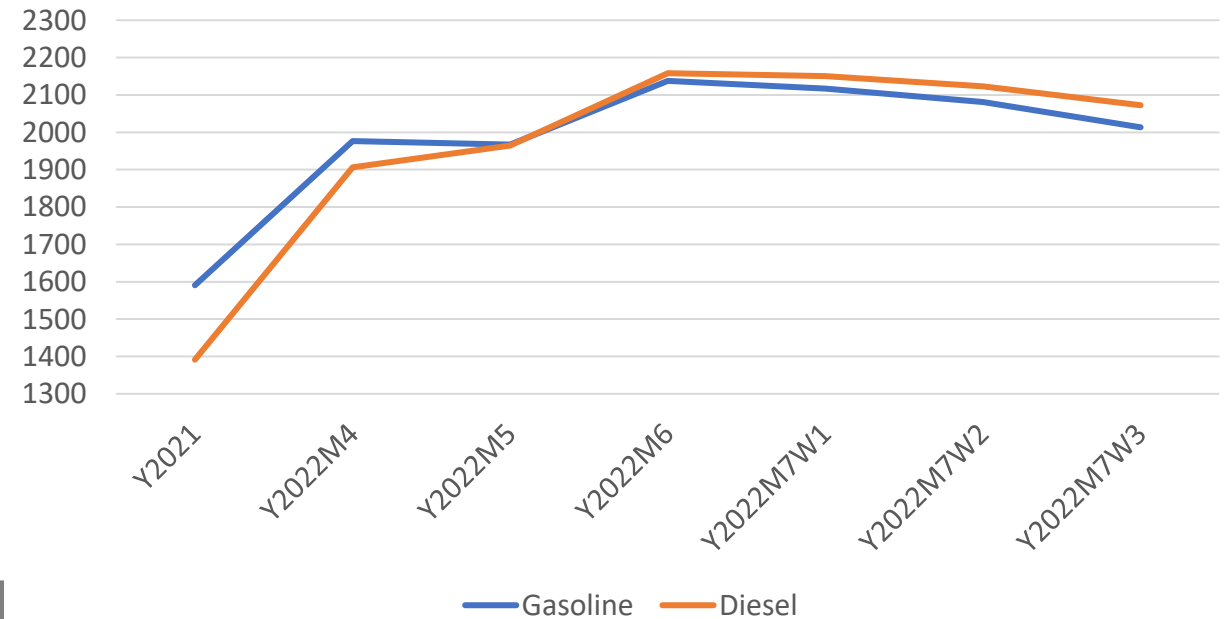
- Ambiguity 모호성

- Ambiguity means the capability of being understood in two or more possible senses or ways.
- Contextual background matters

Data without context 2013.1

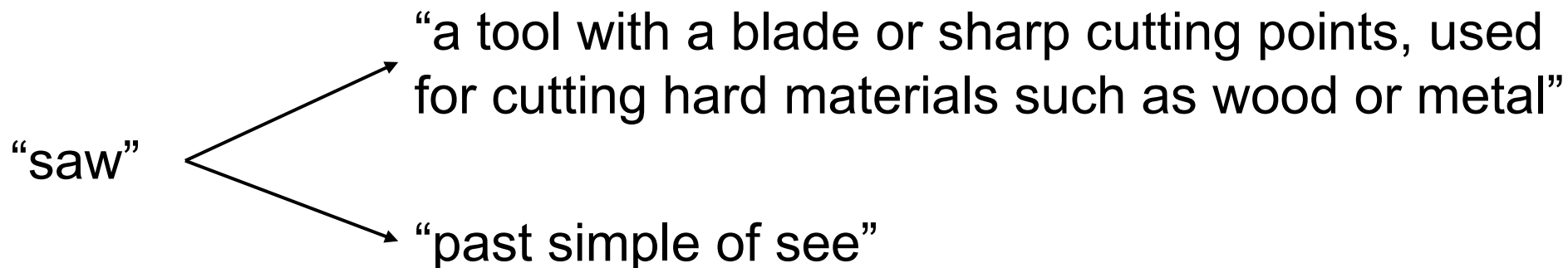
Data with context

	Gasoline	Diesel
Y2021	1590.6	1391.4
Y2022M4	1976.53	1906.42
Y2022M5	1967.07	1964.28
Y2022M6	2137.7	2158.2
Y2022M7W1	2116.8	2150.4
Y2022M7W2	2080.7	2123.3
Y2022M7W3	2013.1	2072.5

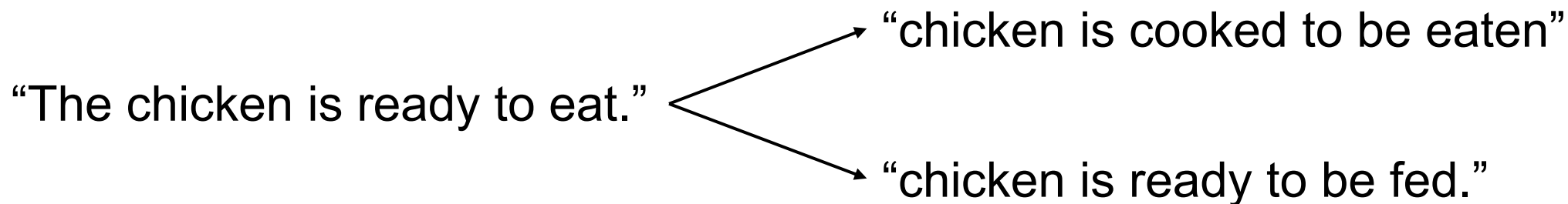


- Ambiguity 모호성

- Lexical ambiguity: a word with multiple meaning



- Syntactic ambiguity: a sentence with multiple meanings due to the relationship between words or sentences



- Ambiguity types

- Homonym 동형어: A word that has the same form but has a different meaning. different etymology. 형태는 같으나 뜻이 다른 단어. 어원이 다름

- pitcher: baseball pitcher or bowl

- 배: 과일 or 신체 부위

- polysemy 다의어: A word in which one form of a word has several related meanings. 한 형태의 단어가 여러 관련된 의미를 갖는 단어

- mouth: body mouth or entrance

- 저녁: 저녁 시간 or 저녁 밥

- Ambiguity types

- Synonym 동의어: A word in different forms but with the same meaning 다른 형태이지만 의미가 같은 단어

- ➔ tired or sleepy

- ➔ 아기, 유아

- Hypernym 상위어 Hyponym 하위어: A word of a higher concept or lower concept 상위 개념 혹은 하위 개념의 단어

- ➔ Bird → pigeon, eagle

- ➔ 새 → 비둘기, 독수리

- Ambiguity in our daily lives

"저녁에 짜장면
먹을까?"

"됐어"

짜장면 먹기 싫어

나 밀가루
안 먹는거 몰라?

지금 짜장면 먹을 때니?

Features of Text Data

- Paraphrase 의역

- Restatement or rewording of a paragraph or text
- Synonym restatement, structure change, etc.
- One meaning, numerous ways to express it

Modes: **Standard** Fluency Formal Simple Creative Expand Shorten Synonyms:

The reason I go so angry is because of the Korea national team played so bad last night.

18 Words

Rephrase

18 Words

The fact that the Korea national team performed so poorly last night is what makes me so irate.

The reason I go so angry is because of the Korea national team played so bad last night.

I became so enraged because of how poorly the Korea national team performed last night.

The fact that the Korea national team performed so poorly last night is what makes me so irate.

- High freedom of expression = Many possible answers

https://quillbot.com/?utm_medium=paid_search&utm_source=google&utm_campaign=sitelinks&campaign_type=extension

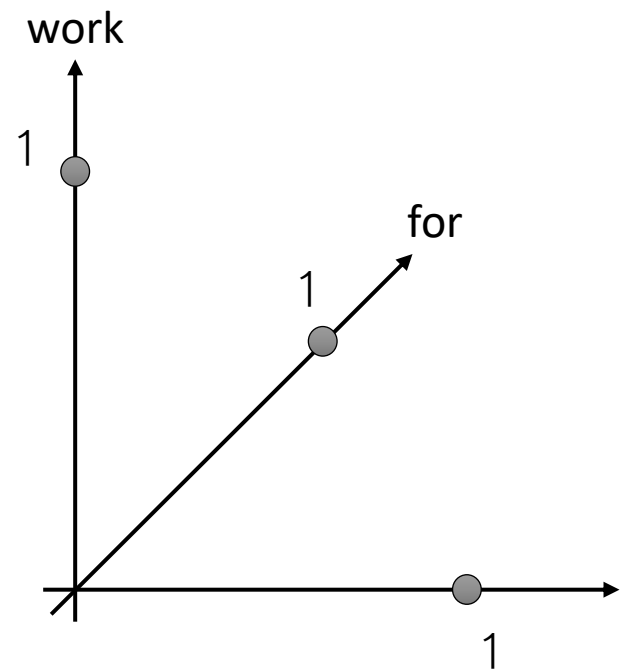
Text Mining Overview

- “A process where an interesting pattern or a useful value is discovered in unstructured Big Data.”
 - Collection
 - Preprocessing
 - Analysis
 - Summarization
- With text mining, we can find certain patterns from the unstructured text data.
- “text-as-data approach”

- Encoding (one-hot encoding vector)
 - Turning text data into “numbers”
 - Sparse vector (spare representation)
 - “Meaning” is not considered

	I	love	Handong
I	1	0	0
love	0	1	0
Handong	0	0	1

I = [1, 0, 0]
work = [0, 1, 0]
Handong = [0, 0, 1]



- Considering “text” as a data set
 - Matrix or Data.Frame or Numerical array (in encoded)

<Original text>

Ferguson) “Twitter is a waste of time.”

Kim) “Cyworld is a full of joy.”

Doc	Text
Ferguson	“Twitter is ...”
Kim	“Cyworld is ...”

Wide format

Doc	X1	X2	X3	X4	X5	...
Ferguson	Twitter	is	a	waste	of	...
Kim	Cyworld	is	a	full	of	...

Long format

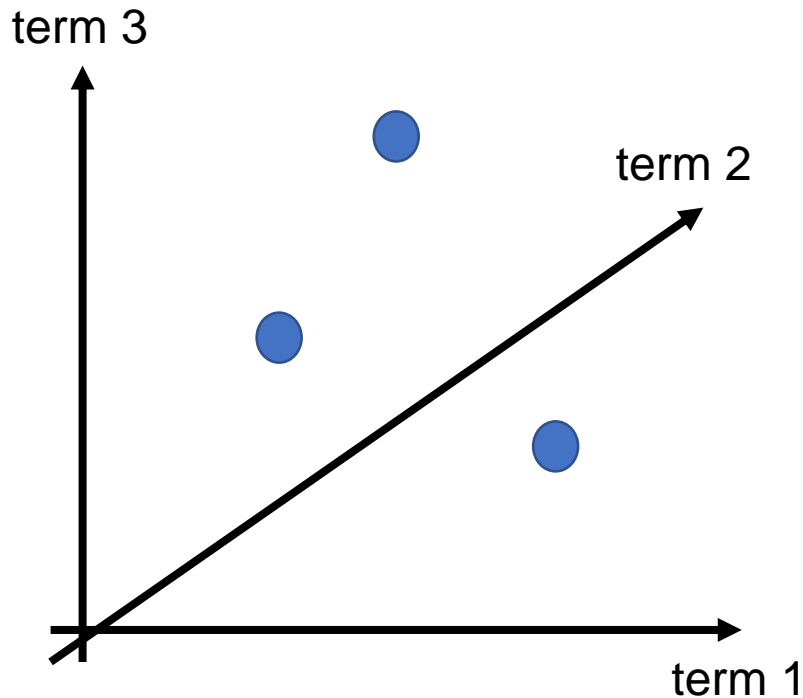
Doc	Text
Ferguson	Twitter
Ferguson	is
Ferguson	...

- Considering “text” as “data” means that text is non-numeric data with a hierarchical structure

Corpus > Document > Paragraph > Sentence > Word > Morpheme

- Corpus: a set of documents
- Morpheme: the smallest unit of language that has its own meaning (part of a word)
- In general, corpus > document > word is used. This is similar to sample > observation > variable in conventional data analysis.
- Depending on the “unit of text analysis”, the whole concept of text mining analysis changes.
- Remember that the unit of text analysis and the level of hierarchical structure depends on the researcher’s interest.

- Text dimension
 - Similar to the concept of data dimension, text data can also be described with an n-dimension
 - Text is ‘high-dimensionality’ data



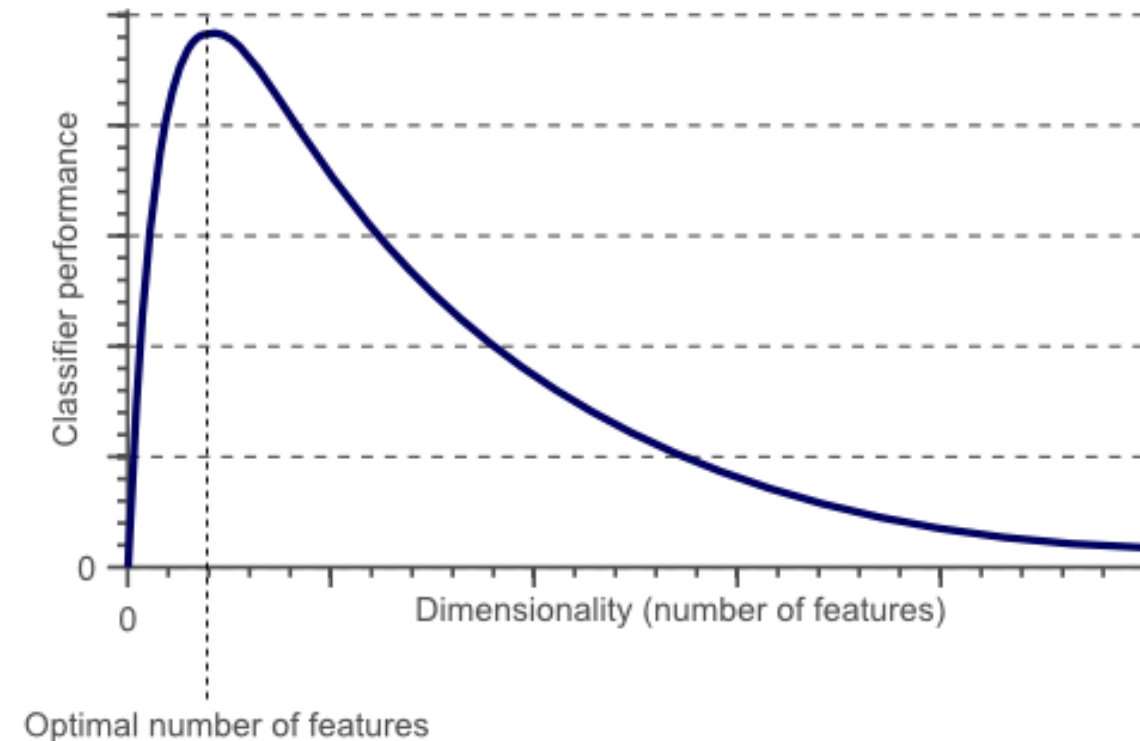
Number of dimension:

Number of unique terms:

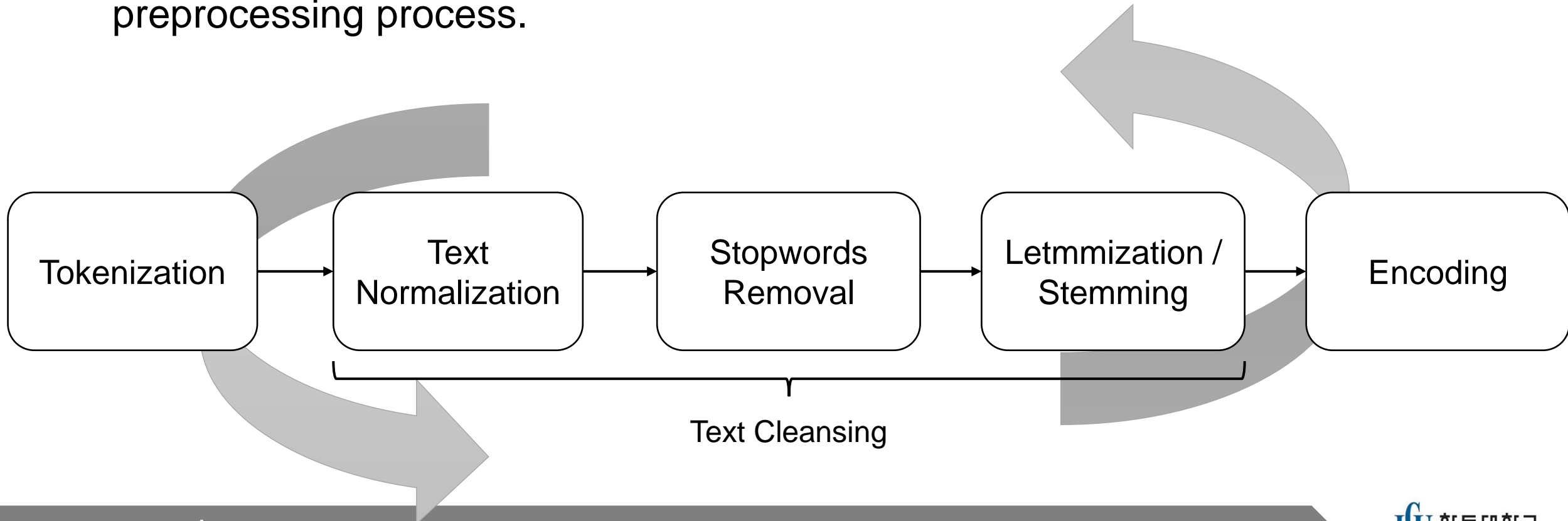
Number of objects:

Number of sentences (or documents):

- High Dimensionality
 - Assuming that each word is considered as a “dimension”, text data is a data set with very high dimension. Hundreds? Thousands?
 - Curse of dimensionality: Explosive nature of increasing data dimensions and its resulting exponential increase in computational efforts required for its processing and/or analysis



- Text data pre-processing process
 - Due to the nature of text data, text preprocessing is completed after several iterations rather than all at once.
 - The key to text data preprocessing is to 1) transform the text to fit the “researcher's purpose” and 2) minimize data distortion or loss during the preprocessing process.



- Text Mining analysis includes methods being used to “understand” text data and “extract” the hidden meaning from it
 - Always think : 1) “what sort of data am I dealing with? 2) “what do I want to find from the analysis?”
 - Selecting the proper analytic method is more important than anything
- Text Mining analysis
 - Word cloud
 - Cluster analysis
 - Network analysis
 - Sentiment analysis
 - Topic modelling

- Natural Language Processing (NLP)
 - Linguistic
 - Text Mining*
 - Artificial Intelligence
 - Domain knowledge
- Text Mining is often used as a prerequisite technique needed for proceeding to NLP
- While most Text Mining relies on “discrete representation,” NLP focuses on “distributional (or continuous) representation”

Text Mining Approaches

- Basic principals
 - “People express themselves through language and want to be approved”
→ There is 'something' in language that researchers need to focus on
- Three approaches
 - Conversation analysis
 - Discourse analysis
 - Critical discourse analysis

- Conversation analysis

- An approach to the study of social interaction, embracing both verbal and non-verbal conduct, in situations of everyday life [Wikipedia]
- Used to understand “how speakers work together to make the conversation happen” → “how speakers talk” & “how sentences are structured”

- Conversation

- Communication between two people including “turn-taking”, “backchanneling”, “pauses”, “interruptions” + “gesture”, “pitch”, etc.

You know what? I
went to
supermarket ~~~

~~~~ Can you  
believe it?

Uh-huh

wow

- Discourse analysis
  - An approach to the analysis of written, vocal, or sign language use, or any significant semiotic event [Wikipedia]
  - Interested in 'naturally occurring' language use
  - Used to understand “how speakers use language to construct meaning and identity”
- Tannen (2008)'s studies on women's discourse on their sisters
  - Small n-narratives: accounts of **specific events** or interactions that speakers said that had occurred with their sisters
  - Big n-narratives: **themes** speakers developed in telling the interviewee about their sisters, and in support of which they told the small-n narratives
  - Master narratives: **culture-wide ideologies** shaping the big-N Narratives.

Tannen, Deborah. 2008. We've never been close, we're very different: Three narrative types in sister discourse

- Critical perspective
  - Views social problems as institutionalized
  - Challenges unjust professional discourse
  - Seeks social change
- Critical discourse analysis: Critical perspective + Discourse analysis
  - 3 dimensions of text, discourse practice, and social practice
  - Text: The stage of describing the linguistic properties in the text, paying attention to vocabulary, tone, sentences, grammar, dialogue order, and newspaper article structure.
  - Discourse practice: The stage of interpretation involving various aspects of the process by which texts are produced and consumed.
  - Social practice: Analysis of social and cultural practices or relationships outside the text is an analysis of the relationship between texts' social practice practices and social structures.

- Content analysis
  - Method used to determine the presence of certain words, themes, or concepts within some given data, such as books, newspapers, magazines, speeches, interviews, web contents, etc.
  - Identifying “patterns” in recorded communication
  - Quantitative approach: counting and measuring
  - Qualitative approach: interpreting and understanding
  - Steps: (1) Categorize (or code) words, themes, and concepts within the text.  
(2) Then, analyze.

- Like a typical data analysis, text analysis can be divided into two types based on the learning type
- Supervised learning
  - If the classification, label, or meaning of the text is known
  - Sentimental analysis, document classification, etc.
- Unsupervised learning
  - If the meaning of the text is unknown
  - clustering, topic modeling, etc.

- Descriptive analysis
  - Text analysis with descriptive features (Text frequency, co-occurrence rate)
  - The result of descriptive analysis is intuitive and easy to understand
  - Using descriptive analysis, we can understand the characteristics of unstructured data (Ex. number of reviews, mentions, comments, etc.)
- Predictive analysis
  - Goal of machine learning is “prediction”
  - Pretrained text analysis model can be used for prediction
  - Text is sequential data → applicable for prediction?

# ***Word Representation***

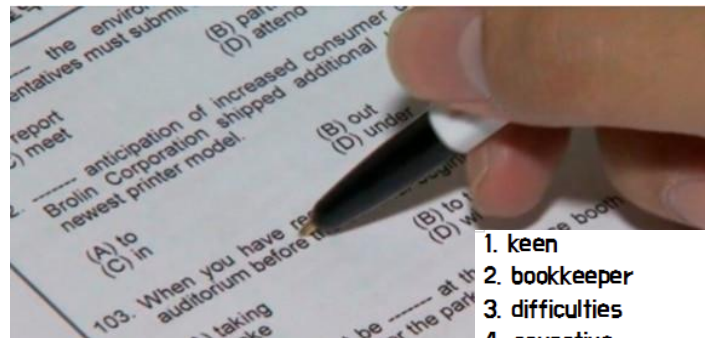


- From computer's perspective, text is just a “symbol”
- The goal of Text Mining is to convert a text into a form that a computer can understand (or CALCULATE)
- Two ways of word representation
  - Discrete representation
  - Distributed representation

- Discrete representation
  - Also known as local representation
  - Count based approach
  - One-hot encoding vector

- Related techniques

- Bag of Words
- Document-Term Matrix (DTM)
- Term-Document Matrix (TDM)
- Term Frequency-Inverse Document Frequency (TF-IDF)
- N-gram language model



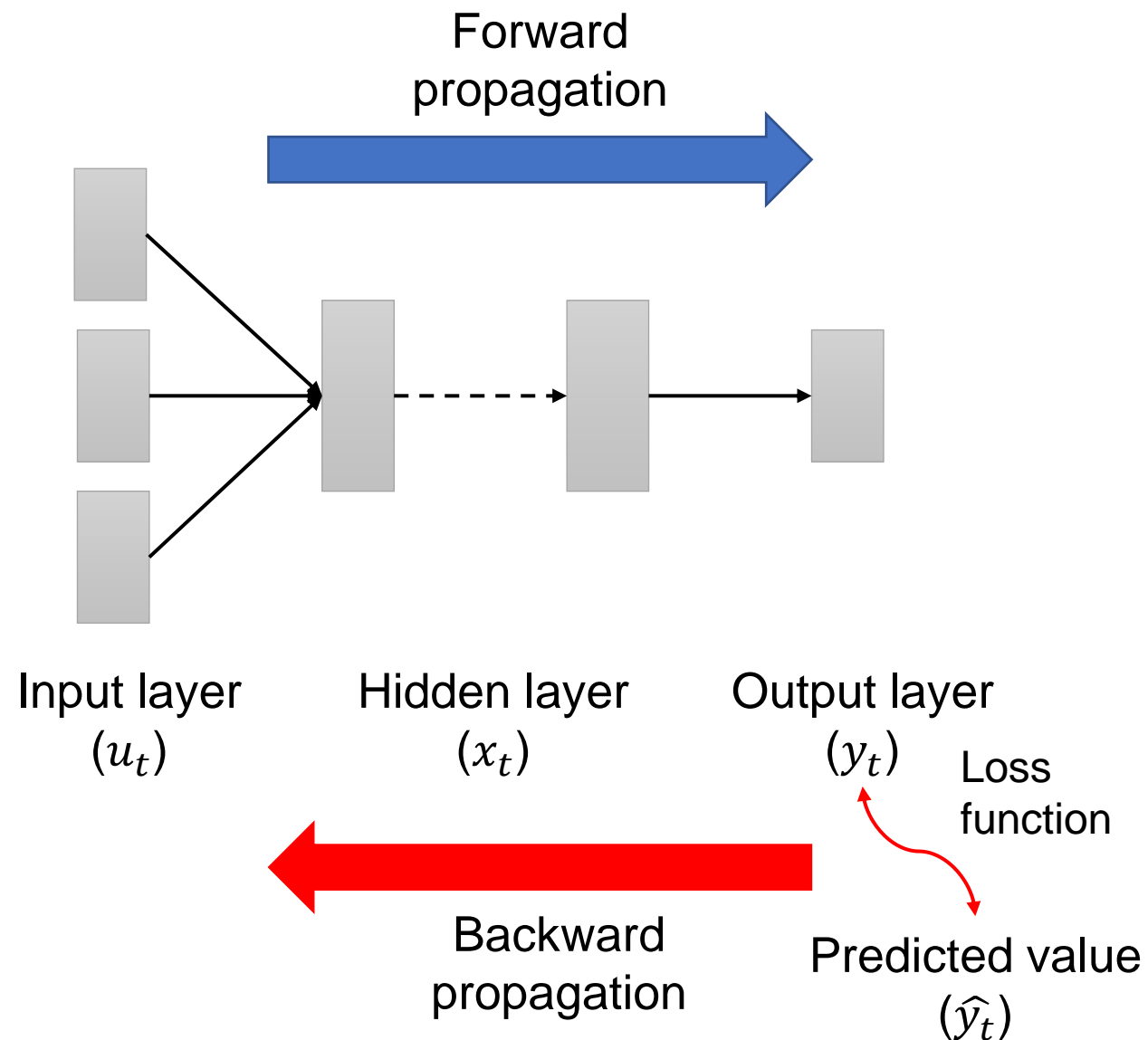
1. keen
2. bookkeeper
3. difficulties
4. causative
5. affected
6. athletics
7. affection
8. marginal
9. release
10. historic site
11. conduction
12. even-numbered
13. demonstration
14. undisclosed
15. undetermined
16. record
17. association
18. ballroom
19. witness
20. stringently
21. bring together
22. recognize
23. expedition
24. remote
25. serve as
26. isolated
27. vessel
28. therapeutic
29. contraction
30. metabolism

1. 예리한
2. 회계 장부 담당자
3. 말쑥, 곤란
4. 원인이 되는, 야기 시키는
5. 영향을 받는
6. 체육, 육상 경기
7. 감정, 감동
8. 최저의, 한계의
9. 양도하다, 놓아주다
10. 유적지
11. 전도
12. 짝수인
13. 입증, 설명, 시연
14. 나타내지 않은
15. 미확인의
16. 이력, 경력
17. 협회, 연계, 연관
18. 무도회장
19. 직접보다, 목격하다
20. 엄격하게, 용서 없이
21. 모으다, 합치다
22. 표창하다, 인정하다
23. 여행, 탐험
24. 외딴, 외진
25. ~의 역할을 하다
26. 외딴, 고립된
27. 선박, 배
28. 치료법의
29. 수축, 축소
30. 신진대사

- Limitations of Discrete Representation
  - Can't figure out the nuances of words
  - Subjective issue
  - Expensive labeling
  - Difficult to calculate the similarity between words
- It starts from the simple truth that “frequency matters,” but that is not the way we understand the text!

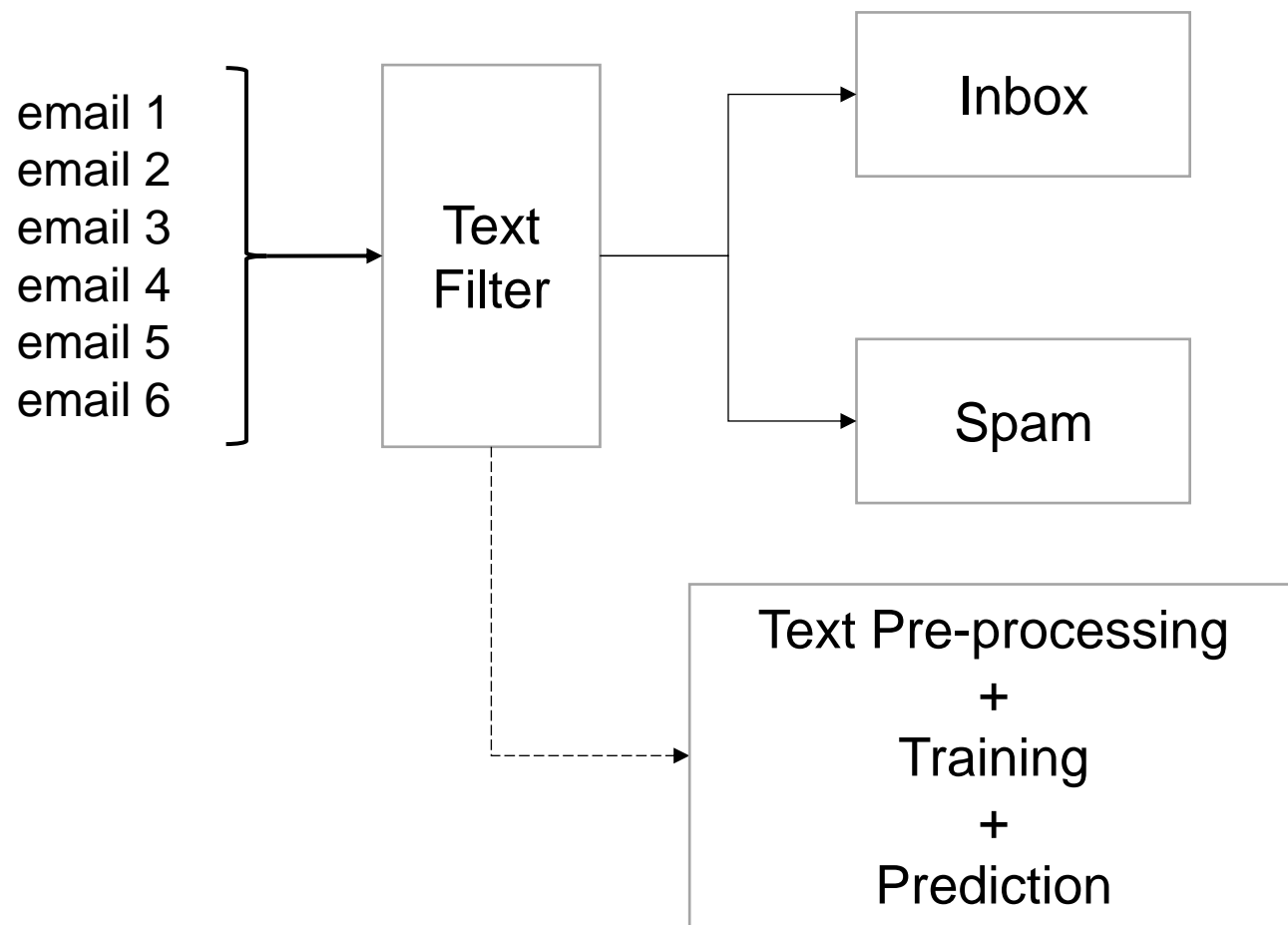
- Distributional (or continuous) representation
  - Vector expression that can understand the 'meaning of words' (ex. RGB)
- Distributional Hypothesis
  - The word itself has no meaning, and the context in which it is used forms the meaning
  - 'The meaning of a word is its conjugation in that language.'
  - Context of a word = words around it in a sentence
- Embedding vector
  - Word2Vec
  - GloVe
  - FastText
  - Deep Learning language model (BERT, GPT, LLaMA, etc.)

- Neural network language model (NNLM)
  - Language model: probability distribution over the sequence of words
- RNN
- LSTM
- BERT
- GPT



# ***Text Mining Applications***

- Spam email detection
  - Binary classification problem
  - Filter by word or word embedding
- Bad word detector
  - Classification problem
  - Filter by word or word embedding



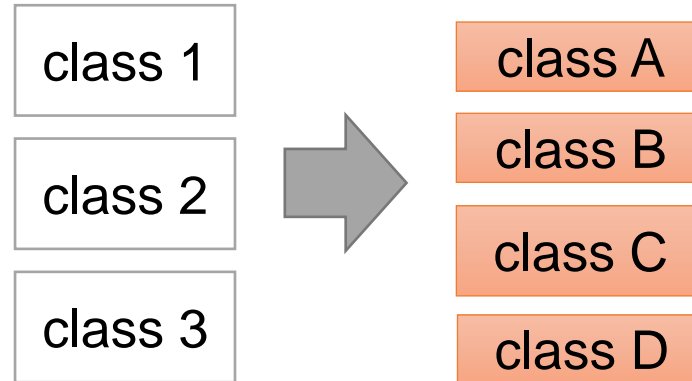
- Classification

- Used to manage and support diverse research or tasks
- Industry classification (Standard Industrial Classification, SIC; International Standard Industrial Classification of All Economic Activities, ISIC; Statistical Classification of Economic Activities in the European Community, NACE)
- Technology classification (International Patent Classification, IPC; Cooperative Patent Classification, CPC)

- Classification based on text analysis

## <Existing Classification>

- Based on the opinions of experts, classification systems such as industries and technologies are established and periodically revised
- Limitations in dealing with new or fusion taxonomy
- Intervention by subjective opinion



## <Text analysis-based Classification>

- Establish a new classification based on the text data
- Ability to reflect change faster and proactively



- Matching different data sets
  - To combine multiple data sets into one, a concordance table is often used.
  - Concordance table: a table combining two tables arranged in a sequence

## SAT | ACT Conversion Table

Based On 25th/75th Percentile Scores From  
Students Accepted to Top 100 Schools

|      |    |      |    |
|------|----|------|----|
| 1580 | 36 | 1330 | 30 |
| 1540 | 35 | 1300 | 29 |
| 1490 | 34 | 1280 | 28 |
| 1440 | 33 | 1240 | 27 |
| 1400 | 32 | 1210 | 26 |
| 1360 | 31 | 1170 | 25 |

### Analysis by mathchops

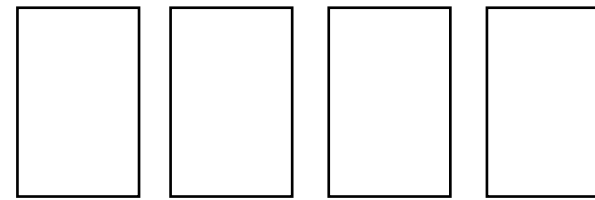
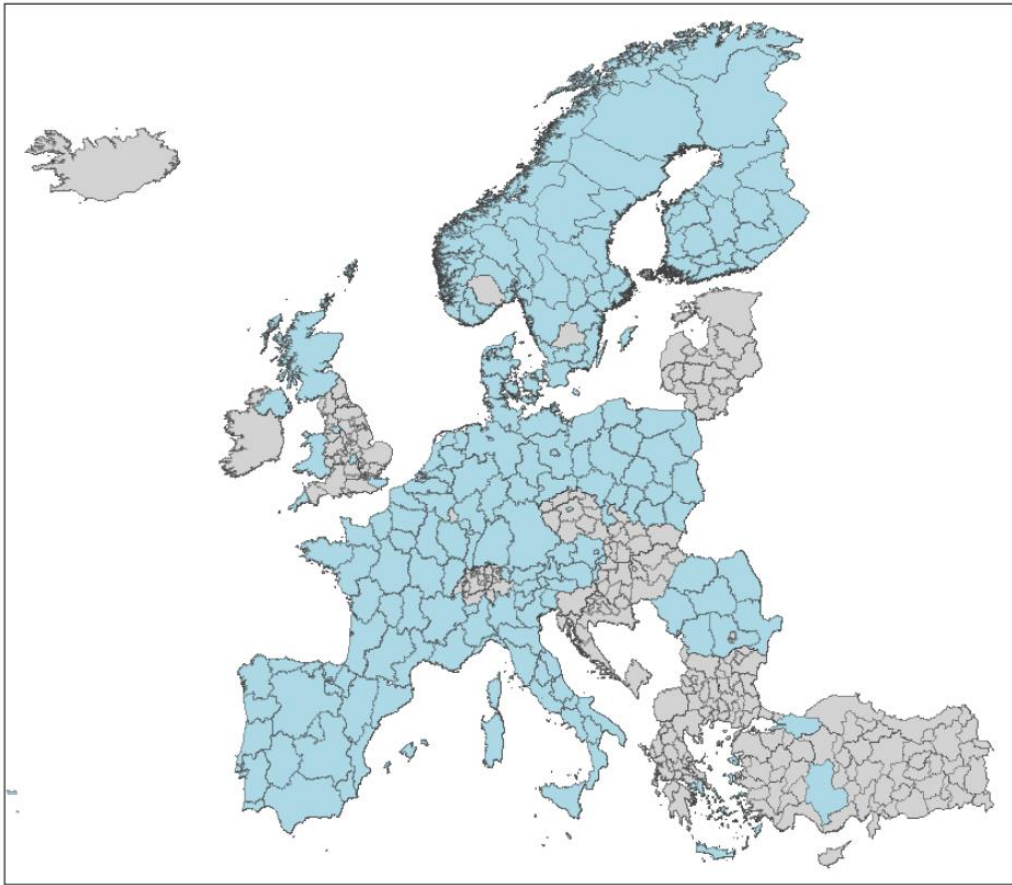
We analyzed the data from 40 schools listed in the US News and World Report's Top 100 Schools. This table differs substantially from the official concordance table, which inflates the value of ACT scores by an average of 35.6 SAT points over the 25 - 34 range. For more information, visit mathchops.com.

| IPC  | IOM  | SOU  | share of IPC<br>with this combo | # with this<br>combo | # with this IPC |
|------|------|------|---------------------------------|----------------------|-----------------|
| B60C | 3199 | 6351 | 1.71E-02                        | 15                   | 878             |
| B60C | 3250 | 3231 | 6.15E-02                        | 54                   | 878             |
| B60C | 3250 | 3250 | 5.69E-03                        | 5                    | 878             |
| B60C | 3251 | 3231 | 2.28E-03                        | 2                    | 878             |

<https://paranmir.wordpress.com/2010/11/08/%EA%B8%B0%EA%B3%A0-%EC%82%B0%EC%97%85%EB%B3%84%EB%A1%9C-%EC%82%B4%ED%8E%B4%EB%B3%B8-%ED%95%9C%EA%B5%AD%EA%B3%BC-%EB%AF%B8%EA%B5%AD%EC%9D%98-%ED%8A%B9%ED%97%88%EB%8F%99%ED%96%A5/>

- If there's no suitable concordance table, a text analysis method can be implemented
  - Matching with the name → (Company data) Hyundai Motors, Hyundai Motors Group, Hyundai LTD, etc.

- Smart specialization policy (SSP)
  - An approach in which a country or region identifies and invests in priorities based on “comparative advantage”
    - Are SSPs implemented in Europe based on regional “comparative advantage”?

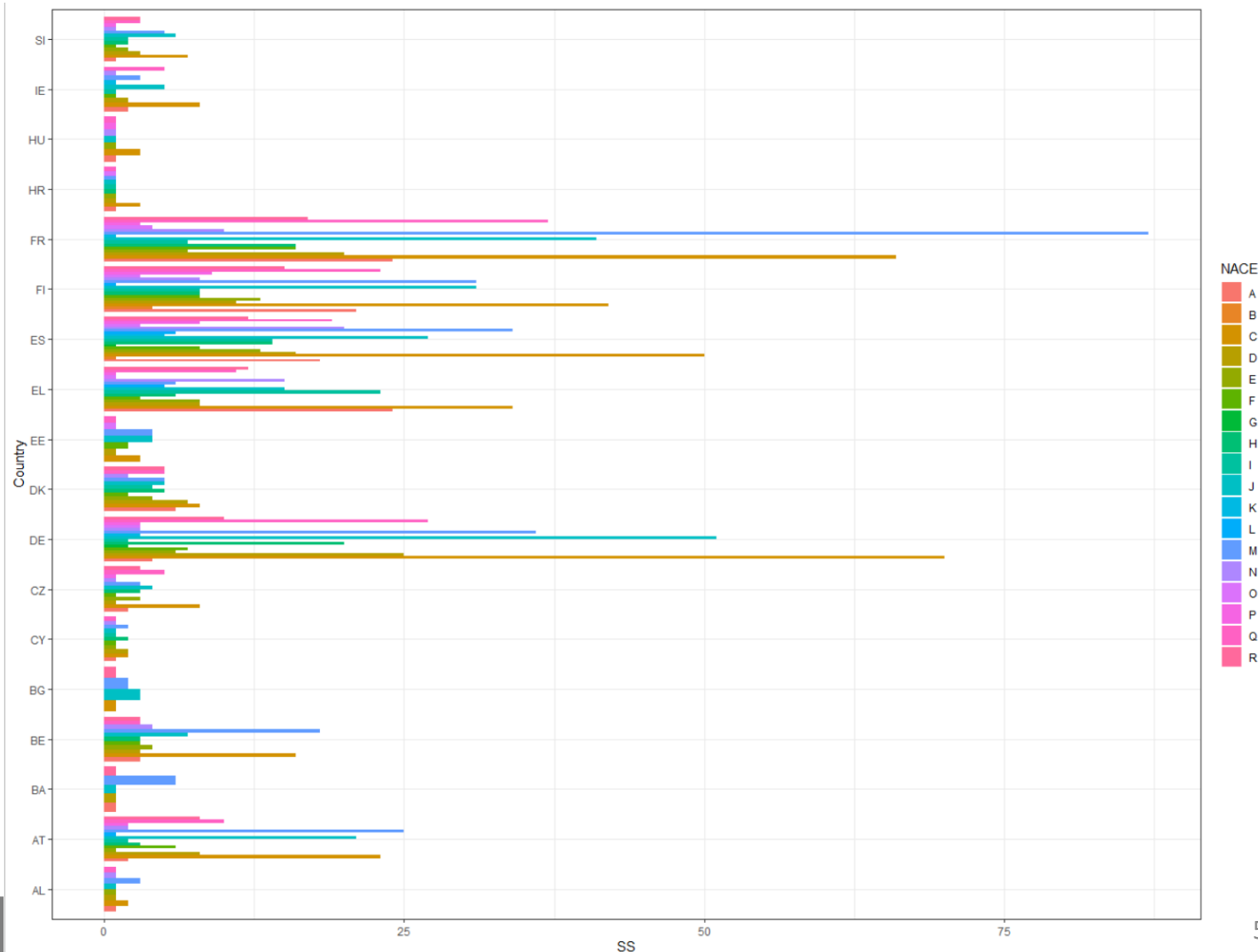


Policy Data

1. Extract keyword from regional policy
2. Match keywords to industry codes

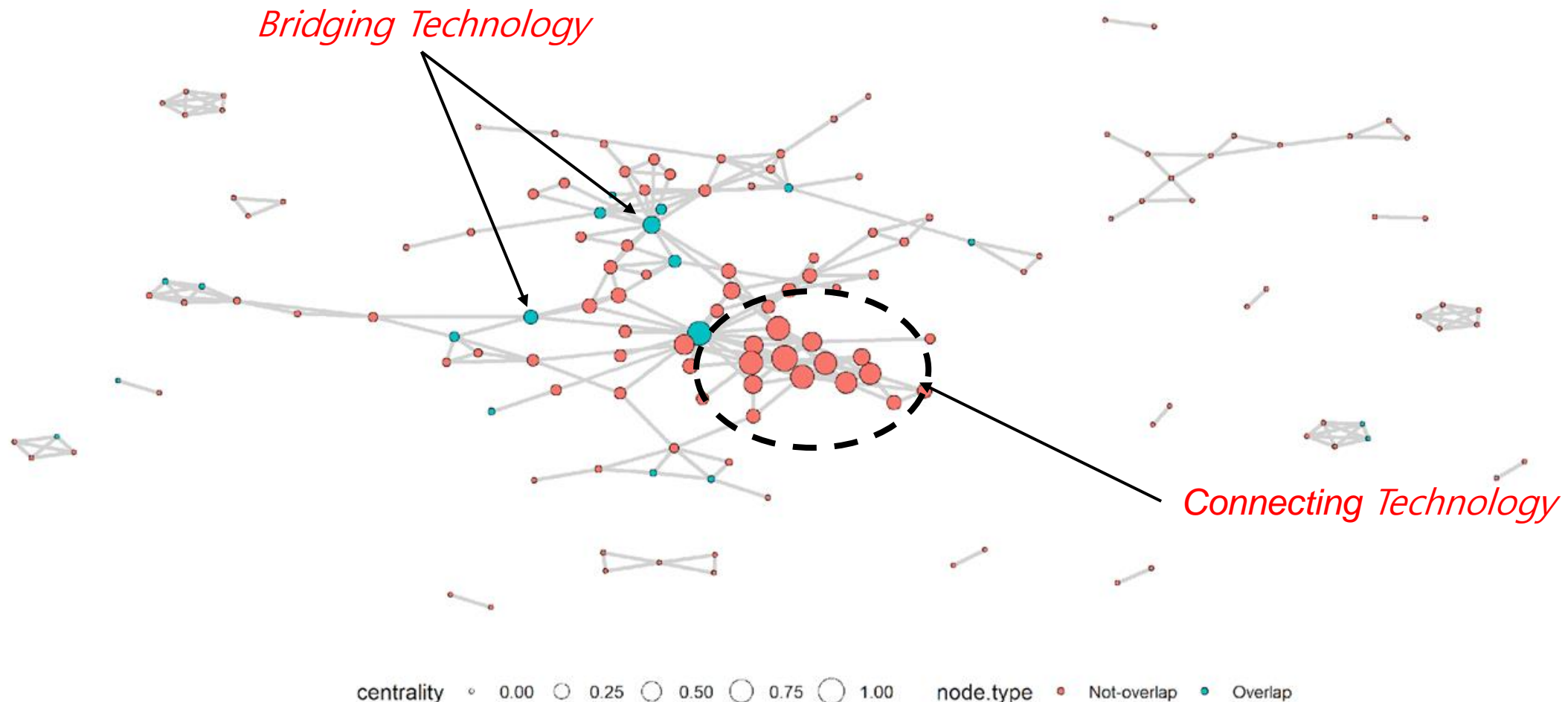
Region – Policy – Industry Code – Patent Code

- Finding the main keyword
- Evaluating the industry distribution of the country's SSP



- Analyzing region's knowledge structure and SSP status
  - Weighted degree centrality / betweenness centrality
  - Text Mining + Network Analysis

ITF1





- Evaluation of smart specialization policy with connectivity and brokerage

