Data Mining
Project #1 & #2

Contents of this project

- Know your data
  - Feature understanding (attribute types)
  - Mean (median), percentile
  - Biasness, Distribution
- Data pre-processing
- Problem definition
  - Task
  - Input, Output
  - Evaluation metric
- Method
- Result

Project Requirements

- Template
  - 3+ pages report
    - Introduction (1-2+ paragraphs)
    - Problem definition (1-2+ paragraphs)
    - Data understanding (3-5+ paragraphs)
    - Data preprocessing (3+ paragraphs)

Part #1: ~ 6/3(Mon)

    - Method (3+ paragraphs)
    - Experimental Result (3+ paragraphs)
    - Conclusion (1 paragraphs)

Part #2: ~ 6/17(Mon)

  - MS word format
  - Language: English or Korean

- Tasks:
  - Step 1: Know your data
  - Step 2: Data preprocessing
  - Step 3: Machine learning analysis

- WRITE YOUR REPORT WITH FULL SENTENCES

Task for Project #1

- You can choose one of the following tasks
  - Item recommendation
  - Classification (e.g. news category classification)
  - Regression (e.g. rating prediction)

Method for Project #1

- You can choose any method for your task
  - Item recommendation → e.g. matrix factorization
  - Classification → e.g. Decision tree, Naive Bayes
  - Regression → e.g. linear regression, logistic regression

Dataset

- Amazon dataset
- News dataset
- Movie dataset
- Any dataset you want (e.g. music dataset, sports dataset)

BASIC WRITING INFORMATION

- 1 PARAGRAPH = 1 IDEA
- 1 PARAGRAPH = 5-9 SENTENCES
- ALWAYS START YOUR PARAGRAPH WITH TOPIC SENTENCE

Example (News Dataset)

- News category classification
  - Input: News article
  - Output: News category

- News article


```
"root":{
        "category":"CRIME",
        "headline":"There Were 2 Mass Shootings
                In Texas Last Week, But Only 1 On TV",
        "authors":"Melissa Jeltsen",
        "link":"https://www.huffing....",
        "short_description":"She left her husband.
                He killed their children.
                Just another day in America.",
        "date":"2018-05-26"
}
```

Example (News Dataset)

- Introduction (1-2+ paragraphs) which contains
  - why this problem is important
  - why do you want to solve this task
  - what is the application scenario of the task
  - …

Example (News Dataset)

- Problem definition (1-2+ paragraphs) which contains
  - formal definition of the task with notations
    - Input / output format (train / test)
      - News document $d = \{w_1, w_2, \dots, w_m\}$, which is combined with headline and short description.
      - A set of category $C = \{c_1, c_2, \dots, c_n\}$
    - Type of attributes
  - Intuitive figure for the toy example
  - description of each attribute

"headline":"There Were 2 Mass Shootings
In Texas Last Week, But Only 1 On TV",
"authors":"Melissa Jeltsen",
"link":"https://www.huffing....",
"short_description":"She left her husband.
He killed their children.
Just another day in America.",
"date":"2018-05-26"

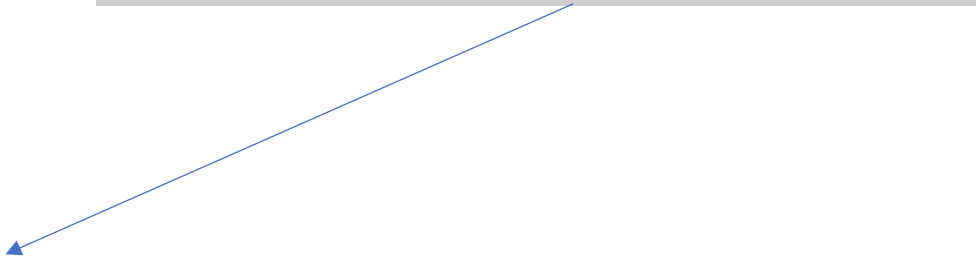
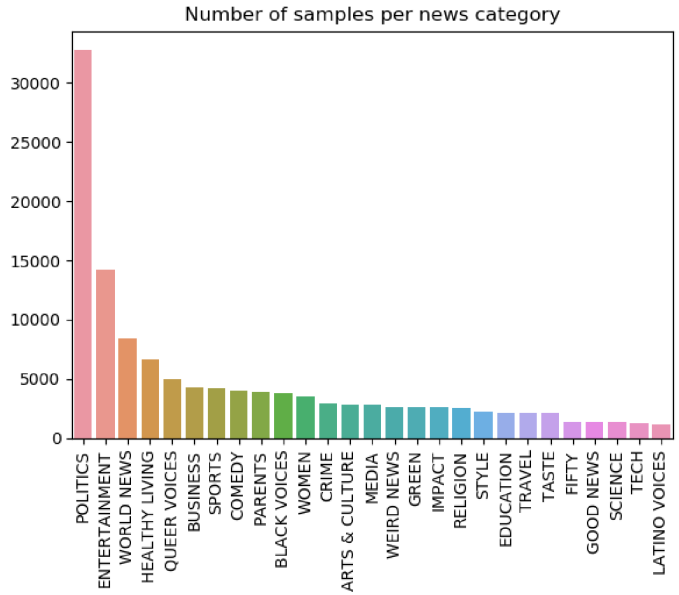
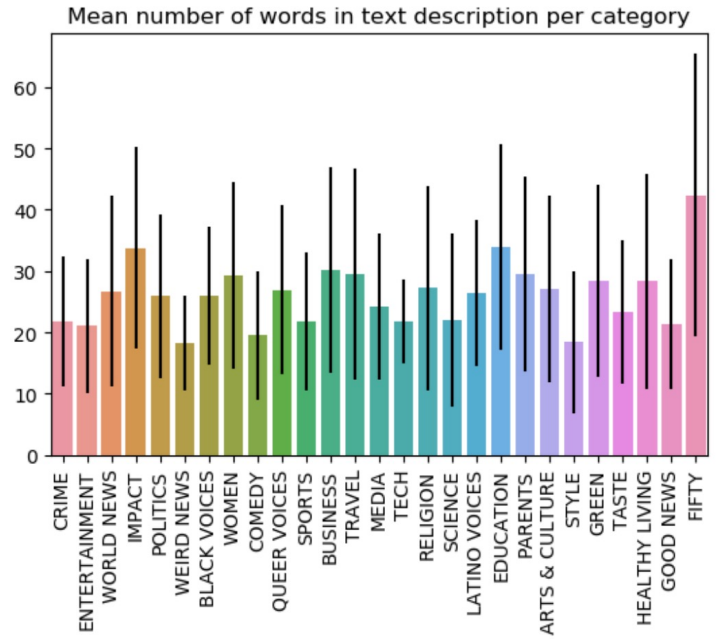


Figure 1. example of news classification

Example (News Dataset)

- Data understanding (3-5+ paragraphs)
  - 3-5 data (statistical) analysis
    - Including 2-3 figures for the analysis
      - Statistical observation
    - Distribution analysis



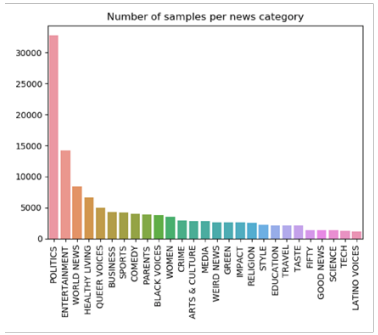


Example (News Dataset)

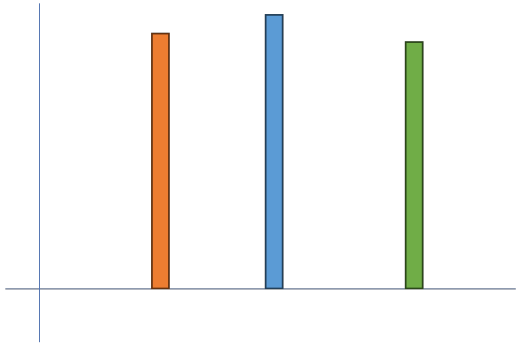
- Data preprocessing (3+ paragraphs) that contains
  - Why do we need this processing
    - Distribution analysis
  - How to handle it, how to solve it
  - Effectiveness of the preprocessing

Example (News Dataset)

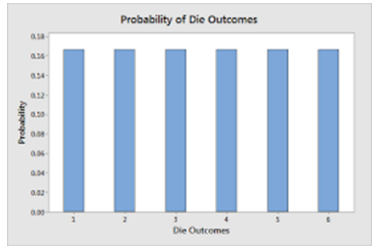
- Data imbalance problem



Semantic clustering for data imbalance



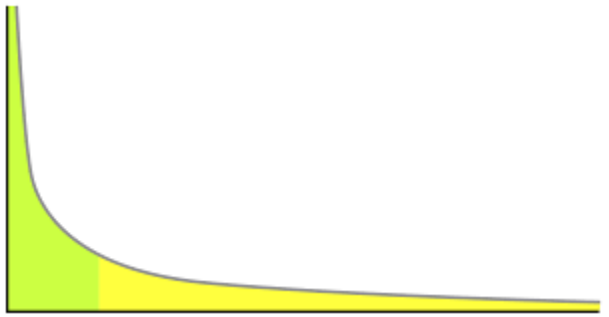
Uniform sampling



- Data preprocessing
  - Convert Unicode into its equivalent character.

  - Before: `Joe Biden Urges National Unity In Speech On Renewed \u2018Cancer Moonshot\u2019`

  - After: `Joe Biden Urges National Unity In Speech On Renewed ‘Cancer Moonshot’`
- Frequency handling with threshold
- Capitalization (lower case handling)

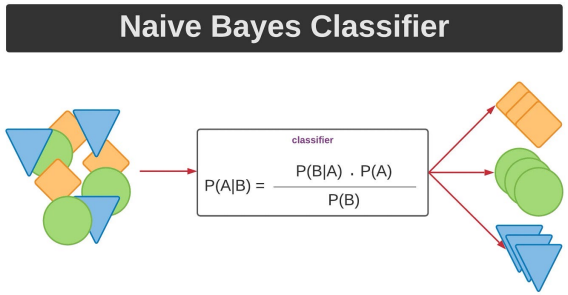


Example (News Dataset)

- Method (3+ paragraphs) that contains
  - Basic explanation of the method
    - Including framework or method figure (optional)
  - 1 or 2 more baselines for the method validation

  - e.g., **Naive Bayes Algorithm** for classification.

- The Naive Bayes algorithm is a probabilistic classifier based on Bayes' Theorem.
- It calculates the probability that a given document d belongs to a certain category c based on the features (words) it contains.
- Bayes' theorem: $P(c|d) = \frac{P(d|c)\times P(c)}{P(d)} \propto P(d|c)\times P(c)$
- It assumes that attributes are conditionally independent: $P(d|c)\times P(c) = \prod_{j=1}^{n} P(w_j|c)\times P(c)$



Example (News Dataset)

- Experimental results (3+ paragraphs) that contains
  - Experimental setting
  - Evaluation metrics
  - Overall performance with tables or figures
    - Quantitative analysis
    - Qualitative analysis
  - Error case analysis
    - How to solve the error case

category	Accuracy	F1	Recall	Precision
Total	0.49	0.30	0.23	0.45
Politics	0.974	0.73	0.67	0.80
⋮	⋮	⋮	⋮	⋮
Sports	0.39	0.40	0.43	0.38

Example (News Dataset)

- Conclusion (1 paragraph) that contains
  - Brief summary of the report
  - Observation and result