

Assignment 7 + Text Classification Basics

(SNLP Tutorial 7)

Vilém Zouhar

15th, 17th June

Assignment 7

- Exercise 1: Count Tree
- Exercise 2: Kneser-Ney Smoothing
- Bonus: Smoothing Techniques

Text Classification

Fill in the classes:

- $f : \text{Text} \rightarrow C$ (classes/categories)
- Topic detection: Document \rightarrow
 {politics, NLP, healthcare, sport, ...}
- Spam detection: Document \rightarrow
 {SPAM, BENIGN, MARKETING}
- Author identification/profiling: Document(s) \rightarrow
 {F. Bacon, W. Shakespeare, ...}
- Native language identification: Document \rightarrow
 {German, Polish, ...}
- POS Tagging: Sentence \rightarrow
 {*NN*, *VERB*, *PART.*, ...}^{|S|}
- Sense Disambiguation: Word+sentence \rightarrow
 Senses of Word

Issues with this?

Classification vs. Clustering

	Classification	Clustering
Method	???	???
Classes	???	???
# Classes	???	???

	Classification	Clustering
Method	Supervised	Unsupervised
Classes	Given	Unknown
# Classes	Given	(Mostly) unknown

Binary vs. Multi-Class Classification

Multi-Class

- $f : D \rightarrow \{\text{politics, NLP, healthcare, sport, ...}\}$

How to turn this into a binary classification?

Binary

- $f_1 : D \rightarrow \{\text{politics, not politics}\}$
- $f_2 : D \rightarrow \{\text{NLP, not NLP}\}$
- $f_3 : D \rightarrow \{\text{healthcare, not healthcare}\}$
- ...

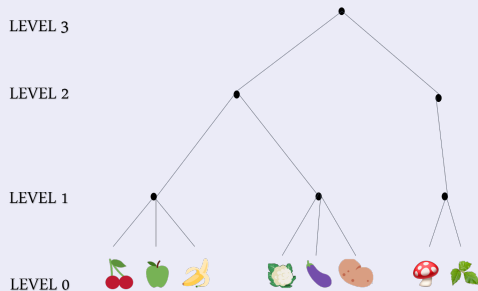
How to turn multiple multi-class into a single multi-class?

Flat vs. Hierarchical

Flat Classification

$f_1 : D \rightarrow$        

Hierarchical Classification



Single-Category vs Multi-Category

- $f : D \rightarrow 2^C$
- Topic detection: Document $\rightarrow \{\text{politics, NLP, healthcare, sport, ...}\}$
- Sentiment analysis: Document $\rightarrow \{\text{positive, negative, interested, ...}\}$
- Topic detection: Document $\rightarrow \{(\text{politics, news}), (\text{NLP, Machine Learning}), (\text{healthcare, nutrition}), (\text{sport, biography}), ... \}$
- Sentiment analysis:
Document $\rightarrow \{(\text{positive, happy}), (\text{negative, sad}), (\text{neutral, ambivalent}), ... \}$

Feature Extraction

- Move from text to more processable domain
- How? (at least three “approaches”)

Binary/indicator features

$$f_b(doc) = \begin{cases} 1 & \text{Contains string "Super free $$$ discount"} \\ 0 & \text{Otherwise} \end{cases}$$

Integer features

$$f_i(doc) = \text{Number of occurrences of "buy"}$$

Real-valued features

$$f_r(doc) = \frac{\text{Number of occurrences of "buy"}}{|doc|}$$

Feature Selection

TODO

Document Frequency

DF

$$df(term) = \frac{|\{doc | term \in doc, doc \in D\}|}{|D|}$$

- Remove rare items ($df \leq \frac{2}{|D|}$)
Won't occur in new documents anyway
- Remove frequent items ($df = 1$)
Usually stop words
No information
- Sometimes not a good idea (interaction with other terms, etc.)
- Stopword distribution gives information in author identification

Information Gain

- Information gained (reduction in entropy) by knowing whether a term is present

$$\begin{aligned} G(C, t) &= H(C) - H(C|t) \\ &= - \sum_i p(c_i) \log p(c_i) \\ &\quad + p(t) \sum_i p(c_i, t) \log p(c_i, t) \\ &\quad + p(\bar{t}) \sum_i p(c_i, \bar{t}) \log p(c_i, \bar{t}) \end{aligned}$$

Pointwise Mutual Information

- Difference between observed distribution and independent

$$\text{pmi}(c_i, t) = \log \frac{p(c_i, t)}{p(c_i) \cdot p(t)}$$

- TODO (relation to mutual information)

Chi Square χ^2

$$\chi^2(c_1, c_2) = \sum_{tt,tf,ft,ff} (O - E)^2$$

- TODO example
- TODO table

Term Strength

- Two documents: d_1, d_2
- Term t
- $p(t \in d_2 | t \in d_1)$
- *What is the probability that the term t will be in d_2 given that it is in d_1 ?*
- If two documents related \rightarrow high probability
- If two documents not related \rightarrow low probability
- “Constant” with stop words

Resources

- ① UdS SNLP Class, WSD: <https://teaching.lsv.uni-saarland.de/snlp/>