Assignment 7,8 + Text Classification Basics

(SNLP Tutorial 7)

Vilém Zouhar

15th, 17th June

Overview

- Task, approaches
- Features
- ▶ Document Frequency
- ▶ Information Gain
- ▶ Pointwise Mutual Information
- ▶ χ^2
- ▶ Term Strength
- Homework

Text Classification

Fill in the classes:

- f : Text $\rightarrow C$ (classes/categories)

Issues with this?

Text Classification

Fill in the classes:

- f : Text $\rightarrow C$ (classes/categories)
- Topic detection: Document $\rightarrow$

Issues with this?

Text Classification

Fill in the classes:

- f : Text $\rightarrow C$ (classes/categories)
- Topic detection: Document $\rightarrow$
- $\{\text{politics}, \text{NLP}, \text{healthcare}, \text{sport}, \ldots\}$

Issues with this?

Text Classification

Fill in the classes:

- f : Text $\rightarrow C$ (classes/categories)
- Topic detection: Document $\rightarrow$
- $\{\text{politics}, \text{NLP}, \text{healthcare}, \text{sport}, \ldots\}$
- Spam detection: Document $\rightarrow$

Issues with this?

Text Classification

Fill in the classes:

- f : Text $\rightarrow C$ (classes/categories)
- Topic detection: Document $\rightarrow$
- $\{\text{politics}, \text{NLP}, \text{healthcare}, \text{sport}, \ldots\}$
- Spam detection: Document $\rightarrow$
- $\{\text{SPAM}, \text{BENIGN}, \text{MARKETING}\}$

Issues with this?

Text Classification

Fill in the classes:

- f : Text $\rightarrow C$ (classes/categories)
- Topic detection: Document $\rightarrow$
- $\quad\{\text{politics}, \text{NLP}, \text{healthcare}, \text{sport}, \ldots\}$
- Spam detection: Document $\rightarrow$
- $\quad\{\text{SPAM}, \text{BENIGN}, \text{MARKETING}\}$
- Author identification/profiling: Document(s) $\rightarrow$

Issues with this?

Text Classification

Fill in the classes:

- f : Text $\rightarrow$ C (classes/categories)
- Topic detection: Document $\rightarrow$
- $\quad\{\text{politics}, \text{NLP}, \text{healthcare}, \text{sport}, \ldots\}$
- Spam detection: Document $\rightarrow$
- $\quad\{\text{SPAM}, \text{BENIGN}, \text{MARKETING}\}$
- Author identification/profiling: Document(s) $\rightarrow$
- $\quad\{\text{F. Bacon}, \text{W. Shakespeare}, \ldots\}$

Issues with this?

Text Classification

Fill in the classes:

- f : Text $\rightarrow$ C (classes/categories)
- Topic detection: Document $\rightarrow$
- $\{\text{politics}, \text{NLP}, \text{healthcare}, \text{sport}, \ldots\}$
- Spam detection: Document $\rightarrow$
- $\{\text{SPAM}, \text{BENIGN}, \text{MARKETING}\}$
- Author identification/profiling: Document(s) $\rightarrow$
- $\{\text{F. Bacon}, \text{W. Shakespeare}, \ldots\}$
- Native language identification: Document $\rightarrow$

Issues with this?

Text Classification

Fill in the classes:

- f : Text $\rightarrow C$ (classes/categories)
- Topic detection: Document $\rightarrow$
- $\{$politics, NLP, healthcare, sport, $\ldots\}$
- Spam detection: Document $\rightarrow$
- $\{$SPAM, BENIGN, MARKETING$\}$
- Author identification/profiling: Document(s) $\rightarrow$
- $\{$F. Bacon, W. Shakespeare, $\ldots\}$
- Native language identification: Document $\rightarrow$
- $\{$German, Polish, $\ldots\}$

Issues with this?

Text Classification

Fill in the classes:

- f : Text $\rightarrow C$ (classes/categories)
- Topic detection: Document $\rightarrow$
- $\quad \{\text{politics}, \text{NLP}, \text{healthcare}, \text{sport}, \ldots\}$
- Spam detection: Document $\rightarrow$
- $\quad \{\text{SPAM}, \text{BENIGN}, \text{MARKETING}\}$
- Author identification/profiling: Document(s) $\rightarrow$
- $\quad \{\text{F. Bacon}, \text{W. Shakespeare}, \ldots\}$
- Native language identification: Document $\rightarrow$
- $\quad \{\text{German}, \text{Polish}, \ldots\}$
- POS Tagging: Sentence $\rightarrow$

Issues with this?

Text Classification

Fill in the classes:

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

Issues with this?

Text Classification

Fill in the classes:

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

Issues with this?

Text Classification

Fill in the classes:

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

Issues with this?

Classification vs. Clustering

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

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

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

How to turn this into a binary classification?

Binary vs. Multi-Class Classification

Multi-Class

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

How to turn this into a binary classification?

Binary

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

Binary vs. Multi-Class Classification

Multi-Class

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

How to turn this into a binary classification?

Binary

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

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

Flat vs. Hiearchical

TODO

Single-Category vs Multi-Category

- $f : D \rightarrow 2^C$

TODO

Single-Category vs Multi-Category

- $f : D \rightarrow 2^C$
- Topic detection: Document $\rightarrow 2^{\{\text{politics},\text{NLP},\text{healthcare},\text{sport},\ldots\}}$

TODO

Single-Category vs Multi-Category

- $f : D \to 2^C$
- Topic detection: Document $\to 2^{\{\text{politics},\text{NLP},\text{healthcare},\text{sport},\ldots\}}$
- Sentiment analysis: Document $\to 2^{\{\text{positive},\text{negative},\text{interested},\ldots\}}$

TODO

Feature Extraction

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

Feature Extraction

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

Binary/indicator features

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

Integer features

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

Real-valued features

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

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

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 term present or not

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

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 (expansion using Bayes)
- TODO (average, max)
- TODO (relation to mutual information)

χ^2

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

- TODO example
- TODO table
- χ^2 avg vs. χ^2 max (multiple categories)

Term Strength

- Two documents: d_1, d_2
- Term t
- $p(t \in d_2 | t \in d1)$
- *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/