Lecture Note - 11: Document Classification, Support Vector Machine (SVM), Sentiment Analysis

Dihui Lai

November 15, 2020

Contents

1 Document Representation

1.1 Document Representations

A document can be represented by the words and the number of their occurrence using term-document matrix

Words	As You Like It	Twelfth Night	Ulius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	5
wit	20	15	2	3

1.2 tf-idf Weighted Measure

Certain words are more common in all documents e.g. the, it, they. The less frequent like "litigation" might be more important that the more frequent word "good". The tf-idf algorithms can be used to adjust the intuition.

$$w_{t,d} = tf_{t,d} \times idf_t$$

Here,

$$tf_{t,d} = 1 + log_{10}(count(t,d)) \text{ if } count(t,d) > 0, \text{ else } 0$$

With $count(t,d)$ the number of occurence of the term t in the document d

$$idf_t = log_{10}\left(\frac{N}{df_t}\right)$$

N is the total number of documents in the collection, and df_t is the number of documents in which term t occurs. If word good appears in all collected document, $idf_t = 0$

1.3 Document/Paragraph to Vector Representation

- Average of the word embedding: take the mean of the vector embedding of the words in the document.
- LDA (Latent Dirichlet Allocation) [David Blei, Andrew Ng, and Michael I. Jordan, 2003]
- Distributed Memory version of Paragraph Vector (PV-DM)
- ...

2 Document Classification, Sentiment Analysis

- Document classification: use machine learning to classify a document type, novel, news, entertainment etc.
- Sentiment Analysis: given a document/paragraph, infer the underlying sentiment (positive/negative, like/dislike)

3 Support Vector Machine (SVM)

- SVM is a classification/regression model that use hyperplane to categorize data points in high dimension space.
- Loss function: $L = \frac{1}{2}|w|^2$ when $y^i(\vec{w} \cdot \vec{x^i} + b) \geq 1$
- or equivalently $L' = \frac{1}{2}|w|^2 - \sum_{i=1}^{N} \alpha_i(y^i(\vec{w} \cdot \vec{x^i} + b) - 1)$
- Optimization: $\frac{\partial L'}{\partial w_i} = 0$

4 Other NLP Challenges and Transfer Learning

- Question and answer, e.g. SQuAD (the Stanford Question Answering Dataset)
- Machine translation
- Multi-lingual NLP
- Transfer learning e.g. BERT (`https://arxiv.org/abs/1810.04805`)

5 Regular Expression

- Algebraic notation for characterizing a set of strings.
- Useful for searching in corpus of texts
- Python example: str = "The rain in Spain"; x = re.findall("Spain", str)

- []: Used to indicate a set of characters e.g. [abc] will match 'a', 'b', or 'c'.
- '\d': matches any decimal digit; equivalent to the set [0-9].
- '+': match 1 or more repetitions of the preceding RE

- '\D': anything but a number (a non-digit)
- '\s': space (tab,space,newline etc.)
- '\S': anything but a space e.g. '\S+@\S+' represents email address
- '^': This expression matches the start of a string

5.1 Expression in Python

re.search(regex, text) returns a match object when the pattern is found or not match if the pattern is not found.

```
import re
m = re.search('hello', 'hello world, hello all, good afternoon')
print m.group(0)
#hello
```

re.findall(regex, text) will return a list of all the matches.

```
m = re.findall('hello', 'hello world, hello all, good afternoon')
print m
#['hello', 'hello']
```