Introduction to Machine Learning Applications

Spring 2021

Lecture-18

Lydia Manikonda

manikl@rpi.edu

Rensselaer

Agenda for today

- Text and NLP
- Class exercises

Announcements

- Project – Individual
  - First 3 Sections Due 04/08/2021 11:59 pm via LMS.
  - Final presentations in-class -- 04/26/2021; 04/29/2021; 05/03/2021
  - Final report due – 04/29/2021 11:59 pm via LMS
- HW7
- Modified Syllabus

Natural Language Processing

- Language is complex and challenging

"The Scarecrow and the Tin Woodman stood up in a corner and kept quiet all night, although of course they could not sleep." - The Wonderful Wizard of Oz, L. Frank Baum

Language is challenging

- I looked straight at an eclipse of the sun.
- I looked after my little sister.
- I looked at the catalogue and saw the dress that I wanted to order.

Language is challenging

- River bank/ financial bank
- Mountain bike/ tandem bike
- Wine glass/ sea glass
- …

Subfields of Linguistics

Subfield	Description
Phonetics	The study of the sounds of human language
Phonology	The study of sound systems in human language
Morphology	The study of formation and internal structure of words
Syntax	The study of formation and internal structure of sentences
Semantics	The study of the meaning of sentences
Pragmatics	The study of the way sentences with their semantic meanings are used for particular communicative goals.

Table from: Linguistic Fundamentals for Natural Language Processing, June 2013, Emily M. Bender

Different Components

- Language models
- Part Of Speech tagging
- Parsing
- Named Entity Recognition
- Co-reference
- Automatic Speech Recognition
- Text To Speech
- Semantic roles

- Weather is gloomy. = high probability = p(is/weather)*p(gloomy/is) =
- Gloomy weather is. = low probability = p(weather/gloomy)*p(is/weather) =
- A b c d = p(b/a)*p(c/b)*p(d/c)

Important questions to ask

- What level of **representation**?
  - Words, sequences, trees
- What are the assumptions of a **model**?
  - Grammar, Hidden Markov model
- Efficient algorithms for **learning**?
  - Viterbi, Forward-Backward

Applications

- Text classification
  - Spam detection
- Question answering systems
  - Watson
- Machine translation
  - English to Korean
- Spoken dialogue systems
  - Siri, Alexa

In this class..

- Sentiment analysis
- Topic modeling
- Bag of words approach
- Word embeddings

Types of text

- Subjective text:
  - Blogs, Online product reviews, Movie reviews, Social media posts such as Tweets, etc.
- Objective text:
  - Wikipedia articles, NYT news articles, WSJ articles, etc.

Sentiment Analysis

- Computational inference of opinion, sentiment subjectivity in text

- Businesses are interested in customer opinions
- Individuals looking at the existing product reviews before purchasing a product
- Tweets during a presidential debate to predict who is the most favorable candidate

Sentiment Analysis

- There are practical issues
  - Cleaning the text
  - Tokenization
  - POS tagging
  - Semantics
  - Named entity recognition
  - Coreference resolution
  - …

Sentiment Analysis

Positive	Negative
My experience so far has been fantastic!	Your support team is useless
Package is nice.	Absolutely horrible!
This product is second to none.	This is better than nothing.
I will absolutely recommend it.	Yeah sure. So smooth!

Sentiment Analysis

- Unsupervised classification
  - Using POS tagging, bigrams and calculating Semantic Orientation (SO)
- Supervised classification
  - Naïve Bayes
  - Maximum entropy
  - Support vector machines
- Large area of research – Issues with fake text (such as fake reviews), spam filters,..

Exercise – 1 – Fake?

1. I was kinda doubtful about the "electroluminescent technology" of this USB cable. That it actually would work. But it actually did. Not only did the blue light function as they should, they were clear and bright, plus they turned off when the charge was finished. It's rad. They should make them in other colors too. Let's just say we're really impressed and are going to order a few more...
2. Such a cool product. I was so happy with how bright the lights on the cable are. It shipped super fast. The light shuts off when the charging is complete, so that’s super helpful. I don’t have to keep checking.

Bag of words model

S1 = “weather is nice”

S2 = “nice weather”

S3 = “it is nice and sunny”

Vocabulary = {nice, it, and, sunny, weather, is} = 6-D

S1 = [1, 0, 0, 0, 1, 1]

S2 = [1, 0, 0, 0, 1, 0]

S3 = [1, 1, 1, 1, 0, 1]

Bag of words model

S1 = “weather is nice weather weather”

S2 = “nice sunny weather”

S3 = “it is nice and sunny nice”

Vocabulary = [weather, is, nice, sunny, it, and] = length of 6

S1 = [3, 1, 1, 0, 0, 0]

S2 = [1, 0, 1, 1, 0, 0]

S3 = [0, 1, 2, 1, 1, 1]

What is the similarity between S2 and S3?

Cosine similarity

S2.S3/|S2|*|S3| = (1*0 + 0*1 + 1*2 + 1*1 + 0*1 + 0*1)/..

|S2| = sqrt(1*1 + 0*0 + 1*1 + 1*1 + 0*0 + 0*0) = sqrt(3)

|S3| = sqrt(0+1+4+1+1+1) = sqrt(8)

Similarity between S2 and S3 = 3/(sqrt(3)*sqrt(8))