Natural Language Processing
SoSe 2014

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

# Exercise 2: Text Classification and Sentiment Analysis

*Dr. Mariana Neves*

*June 18th, 2014*

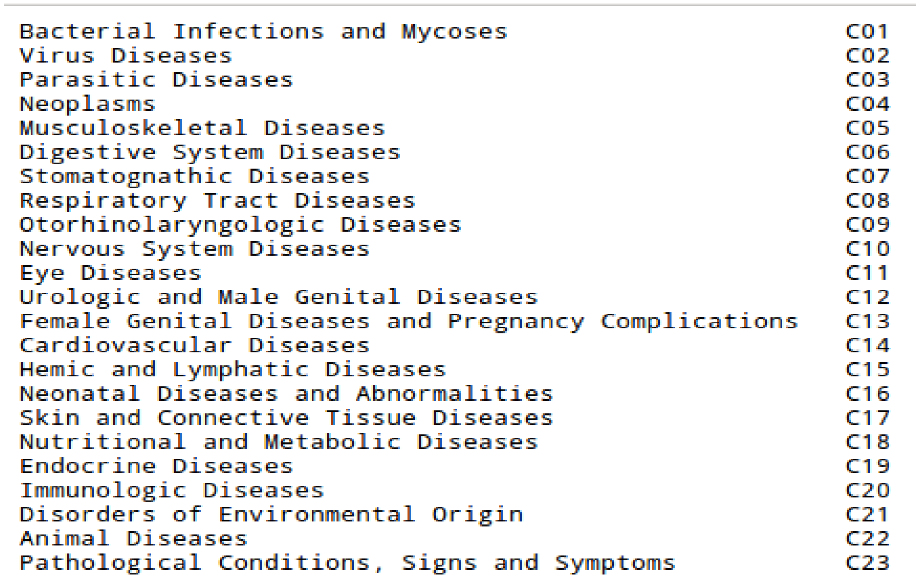
# Tasks

- Text Classification

- Sentiment Analysis

# Task 1: Text Classification

- Ohsumed corpus (20,000 documents)
  - http://disi.unitn.it/moschitti/corpora.htm
  - Split in training and test datasets
  - 23 diseases/categories



| | |
|---|---|
| Bacterial Infections and Mycoses | C01 |
| Virus Diseases | C02 |
| Parasitic Diseases | C03 |
| Neoplasms | C04 |
| Musculoskeletal Diseases | C05 |
| Digestive System Diseases | C06 |
| Stomatognathic Diseases | C07 |
| Respiratory Tract Diseases | C08 |
| Otorhinolaryngologic Diseases | C09 |
| Nervous System Diseases | C10 |
| Eye Diseases | C11 |
| Urologic and Male Genital Diseases | C12 |
| Female Genital Diseases and Pregnancy Complications | C13 |
| Cardiovascular Diseases | C14 |
| Hemic and Lymphatic Diseases | C15 |
| Neonatal Diseases and Abnormalities | C16 |
| Skin and Connective Tissue Diseases | C17 |
| Nutritional and Metabolic Diseases | C18 |
| Endocrine Diseases | C19 |
| Immunologic Diseases | C20 |
| Disorders of Environmental Origin | C21 |
| Animal Diseases | C22 |
| Pathological Conditions, Signs and Symptoms | C23 |

# Task 1: Text Classification

> Laser photodynamic therapy for papilloma viral lesions.
>
> Photodynamic therapy was tested for its therapeutic efficacy in eradicating rabbit papilloma warts.
>
> The wild-type viral warts suspension was used to induce treatable papilloma warts in the cutaneous tissue of Dutch Belted rabbits.
>
> The photosensitizing agents used intravenously were Photofrin II at 10 mg/kg of body weight and Chlorin e6 monoethylene diamine monohydrochloric acid (Chlorin e6 med HCl) at 1 mg/kg of body weight.
>
> The lasers used were an argon-dye laser at 628 and 655 nm and a gold vapor laser at 628 nm.
>
> The irradiances of 25 to 180 mW/cm2 were applied topically with an end-on lens optical fiber with total radiant doses of 7.5 to 54 J/cm2.
>
> Photofrin II and the argon-dye laser at the highest light dosage (54 J/cm2) and Chlorin e6 monoethylene diamine monohydrochloride administered 2 hours before argon-dye laser irradiation at 655 nm at the highest light dosage (54 J/cm2) produced wart regression.
>
> Total wart regression without recurrence was achieved with Photofrin II and the gold vapor laser at all light dosages.
>
> The difference observed between the argon-dye laser and the gold vapor laser might be explained by the pulsed nature of the gold vapor laser, with its high-peak powers, some 5000 x the average measured light dose.
>
> In this model, the smaller, less cornified lesions were more effectively treated with photodynamic therapy.

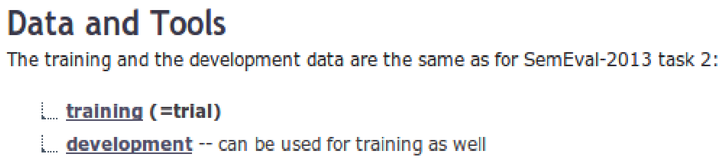
C02 - Virus Diseases

# Task 1: Text Classification

- Multi-class classification
- Features:
  - Bag of words (unigram), stopwords removal
  - And any other extra feature (optional)
- Classification: any classifier
  - SVM, Naïve Bayes, KNN, etc.
- External libraries/resources:
  - Tokenization
  - Machine learning algorthms (Weka, etc.)
  - Stopwords list: http://norm.al/2009/04/14/list-of-english-stop-words/

# Task 2: Sentiment Analysis

- Sentiment Analysis in Twitter
  - http://alt.qcri.org/semeval2014/task9/
  - Task: polarity classification
    - positive, negative, neutral
  - Training and development (Data and tools tab)

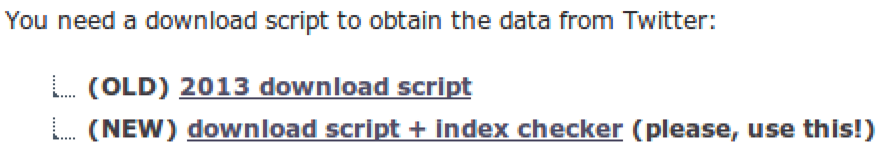


## Data and Tools

The training and the development data are the same as for SemEval-2013 task 2:

- **training (=trial)**
- **development** -- can be used for training as well

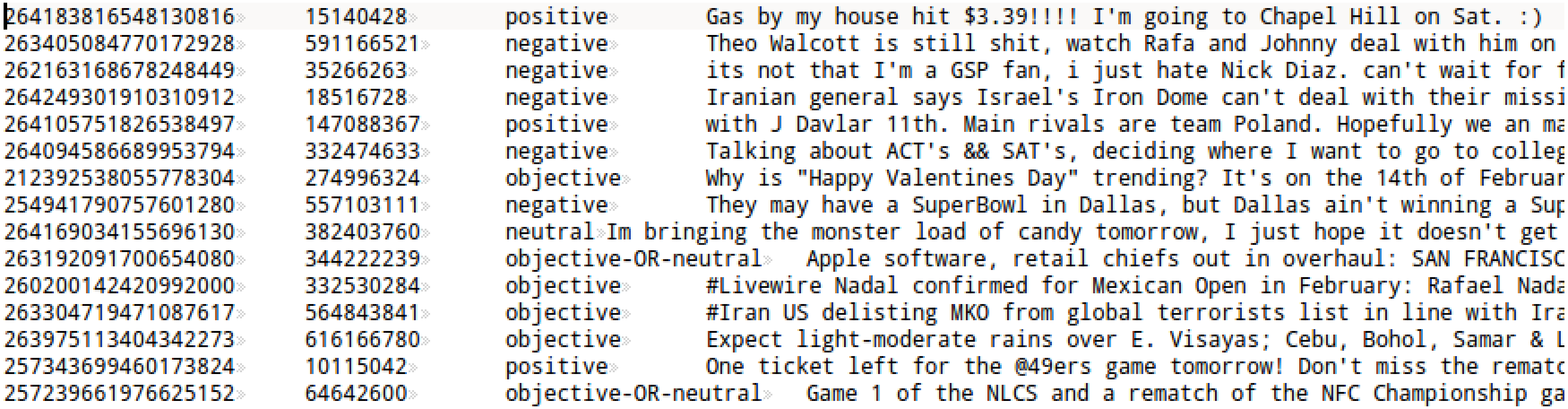
  - Download script



You need a download script to obtain the data from Twitter:

- **(OLD) 2013 download script**
- **(NEW) download script + index checker (please, use this!)**

# Task 2: Sentiment Analysis



```
264183816548130816	15140428	positive	Gas by my house hit $3.39!!!! I'm going to Chapel Hill on Sat. :)
263405084770172928	591166521	negative	Theo Walcott is still shit, watch Rafa and Johnny deal with him on
262163168678248449	35266263	negative	its not that I'm a GSP fan, i just hate Nick Diaz. can't wait for f
264249301910310912	18516728	negative	Iranian general says Israel's Iron Dome can't deal with their missi
264105751826538497	147088367	positive	with J Davlar 11th. Main rivals are team Poland. Hopefully we an ma
264094586689953794	332474633	negative	Talking about ACT's && SAT's, deciding where I want to go to colleg
212392538055778304	274996324	objective	Why is "Happy Valentines Day" trending? It's on the 14th of Februar
254941790757601280	557103111	negative	They may have a SuperBowl in Dallas, but Dallas ain't winning a Sup
264169034155696130	382403760	neutral	Im bringing the monster load of candy tomorrow, I just hope it doesn't get
263192091700654080	344222239	objective-OR-neutral	Apple software, retail chiefs out in overhaul: SAN FRANCISC
260200142420992000	332530284	objective	#Livewire Nadal confirmed for Mexican Open in February: Rafael Nada
263304719471087617	564843841	objective	#Iran US delisting MKO from global terrorists list in line with Ira
263975113404342273	616166780	objective	Expect light-moderate rains over E. Visayas; Cebu, Bohol, Samar & L
257343699460173824	10115042	positive	One ticket left for the @49ers game tomorrow! Don't miss the rematc
257239661976625152	64642600	objective-OR-neutral	Game 1 of the NLCS and a rematch of the NFC Championship ga
```

# Task 2: Sentiment Analysis

- Multi-class (multi-label) classification
- Bag of words, stopwords removal
- Lexicon words (register)
  - Subjectivity Lexicon: http://mpqa.cs.pitt.edu/
  - SentiStrength: http://sentistrength.wlv.ac.uk/
  - ...
- Classification: any classifier
  - SVM, Naive Bayes, KNN, etc.
- External libraries/resources:
  - Tokenization, Machine learning algorthms (Weka, etc.), Stopwords list, lexicons, etc.

# Exercise 2

- Collaborative work
  - Downloading corpus and resources
- Deadline: July 16th (no extensions)
  - Mail with results (P/R/FM per class/sentiment) and source code
- 20% final grade
  - 30% for both tasks