Natural Language Processing
SoSe 2015

Exercise 3: Text Classification and Sentiment Analysis

Dr. Mariana Neves

June 15th, 2015

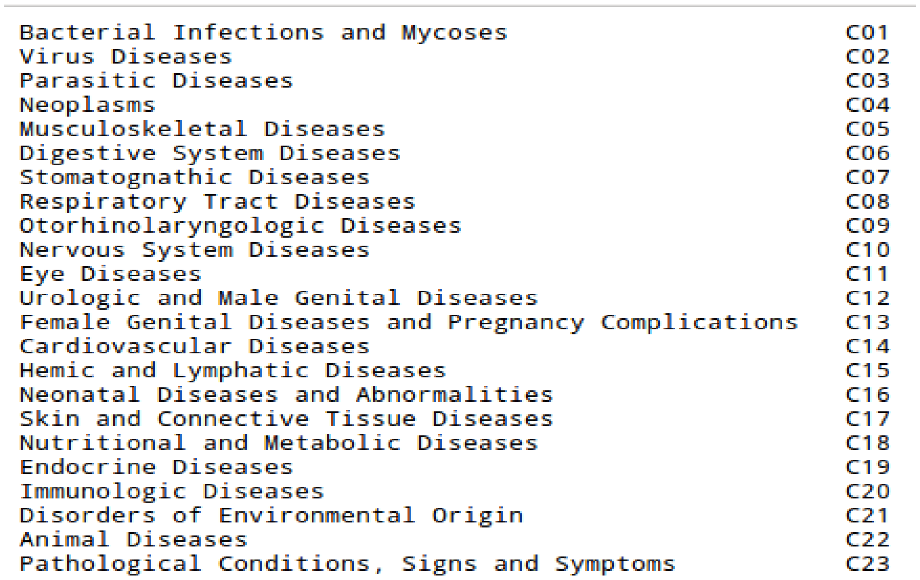
Tasks

- Text Classification

- Sentiment Analysis

Task 1: Text Classification

- Ohsumed corpus (20,000 documents)
  - http://disi.unitn.it/moschitti/corpora.htm
  - Already 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

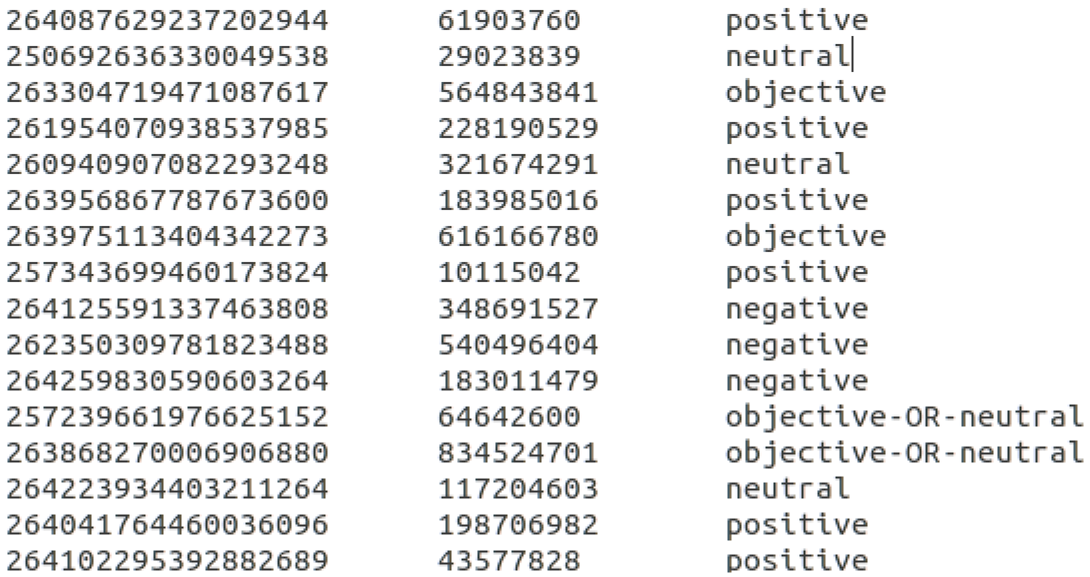
Task 1: Text Classification

- Multi-class, multi-label classification
- Features:
  - Bag of words (unigram), stopwords removal
- Classification: any classifier
  - SVM, Naïve Bayes, KNN, etc.
- External libraries/resources:
  - Tokenization
  - Machine learning algorthms (Weka, etc.)
  - Stopwords list:
    - http://xpo6.com/list-of-english-stop-words/

Task 2: Sentiment Analysis

- Sentiment Analysis in Twitter
  - SemEval'2015
  - Subtask B – Message polarity classification
  - Polarity classification: positive, negative, neutral
  - http://alt.qcri.org/semeval2015/task10/index.php?id=data-and-tools

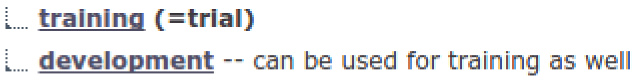
Task 2: Sentiment Analysis



```
264087629237202944      61903760        positive
250692636330049538      29023839        neutral
263304719471087617      564843841       objective
261954070938537985      228190529       positive
260940907082293248      321674291       neutral
263956867787673600      183985016       positive
263975113404342273      616166780       objective
257343699460173824      10115042        positive
264125591337463808      348691527       negative
262350309781823488      540496404       negative
264259830590603264      183011479       negative
257239661976625152      64642600        objective-OR-neutral
263868270006906880      834524701       objective-OR-neutral
264223934403211264      117204603       neutral
264041764460036096      198706982       positive
264102295392882689      43577828        positive
```

Task 2: Sentiment Analysis

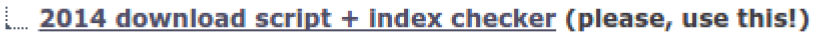
- Sentiment Analysis in Twitter
  - Training and development



training (=trial)

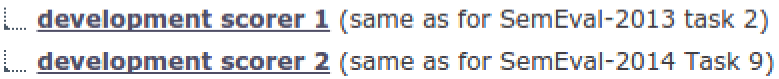
development -- can be used for training as well

  - Download script



2014 download script + index checker (please, use this!)

  - Development scorer



development scorer 1 (same as for SemEval-2013 task 2)

development scorer 2 (same as for SemEval-2014 Task 9)

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 3

- Collaborative work (sentiment analysis)
  - Download corpora and resources
- Deadline on July 6th, upload to HPI owncloud
  - README file with comments, instructions and results (precision)
  - Zipped source code
- Late submissions will lose 0.5/20 point in the exam