Homework 2: Document Classification

Form:	Jupyter notebook file
Language:	English
Requirements:	The report should be clear, readable and include all code documented
Submission:	Ipynb document sent by e-mail to contact
Contact:	asnatm@post.bgu.ac.il
Deadline for submission:	**December 19, 2017**

Students will form teams of two people each, and submit a single homework for each team. The same score for the homework will be given to each member of the team.

Submit your solution in the form of an Jupyter notebook file (with extension ipynb). Images should be submitted as PNG or JPG files. The whole code should also be submitted as a separate folder with all necessary code to run the questions separated in clearly documented functions. Python 2.7 should be used.

The goal of this homework is to let you practice basic web scraping as well as data and text analysis with python.

Submission: Submission of the homework will be done by uploading the Jupyter notebook to Moodle. Please include also screen shots of the results and the saved models, so I'll not have to retrain the models. The homework needs to be entirely in English. The deadline for submission of Homework 2 is set to December 19, 2017 end of day Israel.

Task

The Ohsumed corpus (20,000 documents) includes medical abstracts assigned to 23 diseases/categories. This is a multi-class classification task http://disi.unitn.it/moschitti/corpora.htm Use the external libraries and resources presented in the class for task implementation. Please set a variable at the beginning of the exercise, with the dataset folder.

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

1. **Text pre-processing and exploration:**

- Download the corpus.
- Split to train and test
- Clean and normalize the text (e.g. tokenization, lower case, stop words removal, stemming)
- Explore the dataset (#of categories, #of docs from each category, terms distribution per category). Present a table of top 10 words per category.
- Explain the expected challenges (e.g. top words which are common to multiple categories)

2. **Document classification:**

Here, you should test combinations using 2 feature extraction methods and 3 machine learning models to train a classification model. Test the impact of changing at least one parameter per feature extraction and machine learning model on classification result.

- Implement feature extraction (Bag of words, n-grams, TF-IDF, any other feature - optional)
- Classify using machine learning methods (e.g. SVM, Naïve Bayes)
- Tune each model parameters, as well as pre-processing and parameters steps to optimize the results
- Use accuracy metrics to compare between the different models
- Use the best model selected in the previous steps for prediction on the test set. Present the accuracy of the model and the challenges.
- Describe the task challenges, and explain effective solutions

Good luck