Middle East Technical University | Department of Computer Engineering

CENG 463

Introduction to Natural Language Processing

Fall '2012-2013

Term Project

Due date: 17 January 2013, Thursday, 23:55

1 Objectives

In this assignment you are expected to build a sentiment analysis system for Turkish. You will work on movie reviews and try to classify them as positive or negative.
You will work on the project as a group. You are also expected to write a report on your contribution to the project.

Keywords: *sentiment analysis, opinion mining, Turkish, movie reviews*

2 Sentiment Analysis

Sentiment analysis aims to extract opinions, emotions and attitude from documents. It is closely related to computational linguistics, natural language processing, text mining and also known as subjectivity analysis, opinion mining, and appraisal extraction. The idea is to find out what other people think.

General task of sentiment analysis is to find opinions on a subject. It could be a product, brand, person or a company. Basic opinions categories are positive, negative or neutral. This is also called polarity.

Sentiment analysis gained a lot of importance with the vast amount of information directly available via internet. Companies using sentiment analysis tools to check how well their name is doing out there or how public sees them, how one of their products is received by people. Sentiment analysis is a fast and efficient way to get feedback from customers. Shopping online gained a lot of popularity in recent years and a lot of internet shoppers read product reviews and search the general attitude towards the brand and product they are considering. Political opinions are also another major interest as seen in recent American presidential elections. A leaders recent speech or a recent development about a candidate has major reflections through internet community that is directly observable.

3 Tools

This is a classification task, therefore you will need tools for feature extraction and classification. NLTK has several classifiers and tools available for you. Please study relevant chapters of your book if you will use NLTK in your project. You can also use any of the several other tools available; there is no limitation. Consider using support vector machines which are highly powerful and popular.

3.1 Morphological Analysis

You may want to use morphological analysis of words while selecting your features. Xerox has openly available tools for NLP tasks (Xerox Finite State Technology Tools for Natural Language Processing) and one of them is a finite state transducer based morphological analyzer.
Assume you are in directory a called "test", Xerox tools are in a directory called "software" and you would like to see the morphological analysis of the word "beğenmedim"

```
>> echo beğenmedim | ../software/linux/bin/lookup ../software/tr-tagger/↩
    tlexmorph.fst
beğenmedim  (beGen)beğen+Verb(+mA)+Neg(+dH)+Past(+m)+A1sg
```

3.2 LIBSVM

LIBSVM is a support vector machine implementation by Chang and Lin including multiple classification and regression models and multiple kernel functions. Both Java and C++ implementations are available and it has interfaces for multiple programming languages including Python. LIBSVM has a simple data format. Each instance of your data occupies a single line in a text file. First token of the line is the label of the data. After the label, each feature of your data follows in index value pairs separated by a colon and represented in numeric form. The format of training and testing data file is:

```
<label> <index1>:<value1> <index2>:<value2> ...
.
.
```

Or you can build an "svm_problem" instance as shown below in an example usage

```
from libsvm import *
# x has the data
x=[ [1,0,1],[-1,0,-1],[-1,0,-1], [9,10,222] ]
# y has the corresponding class labels of each item in x
y=[1,1,-1,1]
prob = svm_problem(y, x)
param = svm_parameter(kernel_type = LINEAR, C = 10)
# training the model
m = svm_model(prob, param)
#testing the model
m.predict([1, 1, 1])
```

All you need to do is build an instance of svm_problem with your data and set parameters for training with svm_parameter class. It is quite easy and you can refer to README file or project website "**http://www.csie.ntu.edu.tw/~cjlin/libsvm/**". Note that SVM problems usually do binary classification ie between two classes. You may want to train two classifiers: one for opinionated vs neutral and one for positive vs negative.

3.3 Scikit-Learn

Scikit-Learn is a Python module for using machine learning algorithms including supervised and unsupervised methods and useful tools such as model selection and feature extraction. It has an svm implementation based on LIBSVM. It is a comprehensive package.

4 Data

You will use movie reviews as subjects in this assignment. You will work with reviews from the website Beyazperde (`http://www.beyazperde.com`) which is a popular Turkish website focused on movies. You will use reviews at "`http://www.beyazperde.com/filmler/elestiriler-beyazperde/`" . The review with four and five stars will be considered as positive, three stars as neutral and ratings below three stars are considered negative. You may implement a web crawler to automatically extract labeled data.

```
1 *     NEGATIVE
2 **    NEGATIVE
3 ***   NEUTRAL
4 ****  POSITIVE
5 ***** POSITIVE
```

Examples:

```
1  (htr2b : Dönüşüm *)
2  \url{http://www.beyazperde.com/filmler/film-214686/elestiriler-beyazperde/}
3  NEGATIVE
4  Yazıya yine, yeni, yeniden 'Türk Korku Sineması' tanımını kullanarak başlamak en ↩
       kolay çıkış yolu olarak gözük ...
5
6  (Anna Karenina ****)
7  \url{http://www.beyazperde.com/filmler/film-191856/elestiriler-beyazperde/}
8  POSITIVE
9  önetmenlik kariyerine kısa filmler ve televizyon dizilerinde bölüm yönetmeni ↩
       olarak başlayan ...
10
11 (Kıyamet Günü ***)
12 \url{http://www.beyazperde.com/filmler/film-146630/elestiriler-beyazperde/}
13 NEUTRAL
14 Juan Antonio Bayona'nın ismini ilk duyduğumuz çalışması, kendisinin de ilk uzun ↩
       met ...
```

5 Specifications

1. **Y**ou are expected to implement a sentiment analysis tool for movie reviews in Turkish.
2. **Y**ou will work on the same project as a group.
3. **E**ach of you will implement a part of the project and will write a report about it.
4. **B**egin working on the project as early as possible since there is no late submission for this assignment.
5. **T**here will a demonstration at the end of the semester.
6. **D**ecide on the parts of the project and divide them among yourselves as early as possible.

6 Regulations

1. **Programming Language:** You are free to choose any implementation method.
2. **Late Submission:** Late submission is not allowed for this assignment.
3. **Cheating:** We have zero tolerance policy for cheating. In case of cheating, all parts involved (source(s) and receiver(s)) get zero. People involved in cheating will be punished according to the university regulations.
4. **R**emember that students of this course are bounded to code of honor and its violation is subject to severe punishment.
5. **Newsgroup:** You must follow the newsgroup (news.ceng.metu.edu.tr) for discussions and possible updates on a daily basis.
6. **Evaluation:** Your grade will be harmonic mean of your precision and recall in other words F1 score.

$$grade = \frac{2 * precision * recall}{precision + recall} \tag{1}$$

$$precision; = \frac{true\ positives}{true\ positives + false\ positives} \quad recall; = \frac{true\ positives}{true\ positives + false\ negatives} \tag{2}$$

7 Submission

- Submission will be done via COW.
  Create a tar.gz file named `project.tar.gz` that contains your source files and a copy of your report in pdf format.

8 References

- Speech and Language Processing, Daniel Jurafsky and James H. Martin
- Foundations of Statistical Natural Language Processing, Christopher D. Manning and Hinrich Schütze
- Natural Language Processing with Python, Steven Bird, Ewan Klein and Edward Loper
- Sentiment Analysis in Turkish, Umut Erogul, 2009
- Xerox Finite State Technology Tools for Natural Language Processing, `http://open.xerox.com/Services/fst-nlp-tools/Pages/morphology`
- C.-C. Chang and C.-J. Lin. LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011
- Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.