Introduction to Machine Learning: Autumn Semester 2013
Instructor: Prof. Andreas Krause

Course Project
Part 3

Text classification

1 Introduction

A long time ago, in a galaxy far, far away... there was a planet, called "Htrae". We only have some small fragments of information about this planet: we know there were countries and cities and that all the cities had different names (consisting usually of more then one word) and numeric codes. We also know that their language structure was somewhat similar to most European languages.

Scientists at ETH Zurich are currently exploring the history of Htrae and its geopolitical structure. In order to do that they collected information from different sources and detected phrases that can correspond to the names of the cities. Sometimes along with the phrases the numeric code of the city and country are obtained, but not always.

One of the major problems that arise in the analysis is that in different sources the name of the city can appear differently: Some words can be missing, some can be misspelled. There are also typos in the spelling of the cities names. And one has to agree that this is not surprising when you have a city name consisting of up to 25 words.

For these reasons, scientists face a challenging task of finding the numeric codes for the name of every city. And that is why they decided to ask you for help.

Your task is to design a multiclass-classifier which, given a name of the city, will return the corresponding city and country code. Of course the city code is more important, so if you guess the city code – great, otherwise the scientists ask you to guess at least the country code.

2 Dataset Description

2.1 Input

Each sample in the dataset is a city name, consisting usually of more than one word. Words are separated one from another by a space character. Each word is a sequence of characters in Htrae alphabet that luckily appears to be very similar to the English alphabet with symbols like numbers, hyphen, and slash.

The most common misspellings in city names are

- Different capitalization: *yfirjhnjcnfy* can become *Yfirjhnjcnfy* or *YFIRJHNJCNFY*,
- Missing words: for example, name *eas cjdtncrbv u hy hedl* becomes *eas cjdtncrbv u hedl*,
- Missing letters in a word: *yfirjhnjcnfy* becomes *firjhnjcnfy*,
- Wrong letters in a word: *yfirjhnjcnfy* becomes *ybirjhnjcnfy*,
- Different forms of the word (e.g. nominative and accusative cases): *yfirjhnjcnfy* becomes *yfirjhnjcndi*.

2.2 Output

There are two numeric output codes. The first one is a 6-digit city code. The second one is a 3-digit country code. In addition to the misspellings in the city names, also the city codes are sometimes assigned to the wrong city name. But the scientists are pretty sure that they only messed this up for cities within the same country.

2.3 Training Set

The training data set consists of one comma separated file. The first column of the file is a phrase representing the name of one city. The second column is the code of the city, and the third column is the code of the country which this city belongs to. The training set is in the file "training.csv".

The number of lines in the training set is 6108 and therefore for training you are given 6108 city names. There are 122 possible city codes and 15 possible country codes.

2.4 Validation and Test Sets

Both validation and test set contain city names that have not been classified yet. Your task is to reconstruct a city and a country code for each of the given city names. The data sets are given in the files "validation.csv" and "testing.csv". Each line in both files contains only the name of the city (that corresponds to the first column of the training set file).

Required output: A file that contains one line for each city with the predicted city code and country code separated by a comma. The "i^{th}" line of the output file should contain both codes prediction for the instance in the "i^{th}" row of the input file.

There are two submission pages: One for the validation set and one for the test set. You will receive feedback for each submission on the validation set and you can use it as a way to compare the prediction performance of your algorithm with other submissions. **You will not receive immediate feedback for the submissions on the test set: it will be used to calculate your grade.**

3 Evaluation and Grading

Each submission (upload of a prediction file for a given data set) will be ranked according to the following error metric:

- You are penalized by 1 for every city code that is missclassified
- You are penalized by 0.25 for every country code that is missclassified

So for example if you predict both city and country code wrong you suffer a penalty 1.25. If you missclassify the city code, but correctly predict the country code, you are penalized by only 1.

Of course, the less penalty points you get, the better your solution is.

Let y_i^{city} and y_i^{country} be respectively the ground truth city and country codes of the object i, let $\hat{y}_i^{\text{city}}$ and $\hat{y}_i^{\text{country}}$ be your prediction and let $I[x]$ be the indicator function that equals 1 if x is true and 0 otherwise. Then the cost of your solution is computed as follows:

$$\text{CE} = \sum_i \left(I[y_i^{\text{city}} \neq \hat{y}_i^{\text{city}}] + 0.25 I[y_i^{\text{country}} \neq \hat{y}_i^{\text{country}}] \right)$$

We compare the cost of the submission to two baseline predictions: a weak one (called "baseline easy") and a strong one (called "baseline hard"). These will have costs of $\mathrm{CE}_{\mathrm{BE}}$ and $\mathrm{CE}_{\mathrm{BH}}$ respectively, calculated as described above. Both baselines will appear in the rankings together with the error measure of your submitted predictions.

Performing better than the weak baseline on the **validation set** will give you 50% of the grade, and matching or exceeding the strong baseline on the **test set** will give you 100% of the grade. This allows you to check if you are getting at least 50% of the grade by looking at the ranking. If your prediction performance on the **test set** ($\mathrm{CE}_{\mathrm{test}}$) is in between the baselines, the grade is computed as:

$$\text{Grade} = \left(1 - \left(\frac{\mathrm{CE}_{\mathrm{test}} - \mathrm{CE}_{\mathrm{BH,test}}}{\mathrm{CE}_{\mathrm{BE,test}} - \mathrm{CE}_{\mathrm{BH,test}}}\right)\right) \times 50\% + 50\%$$

Your last submission on the test set will be used for grading.

3.1 Submission

In addition to your predictions on the test set you need to provide a brief report that explains how you obtained your results. We include a template for LaTeX in the file "report.tex". If you do not want to use LaTeX, please use the same sections as shown in "report.pdf".

Upload a zip file with the report (as a PDF file) along with your code or parameters/screenshots of the tools you used. For further instructions refer to the report template.

We might ask you to show us what you did, so please keep the necessary files until the end of the semester.

3.2 Deadline

You will be able to submit predictions starting from **Monday, 25.11.2013, 17:00** until **Friday, 20.12.2013, 23:59:59**.