

Homework #6 (due in 2 weeks – 4/16/09; accepted late with 15 points off on 4/23/09)

Multilingual Issues (100=15+15+70 points)

This assignment will give you a chance to consider multilingual issues in text retrieval. Specifically you will build a look at an integrated web-based translation and search system and build a simple approach to language identification.

Google Translate (15 points)

Investigate the translate and search utility at http://translate.google.com/translate_s. Try the following English queries (optionally others) to search pages written in (1) Spanish, (2) Thai, and (3) Ukrainian: “NATO troops in Kosovo”, “1000 places to see before you die”, “Guantanamo Bay”. Make observations about the quality of the results and consider issues such as: are the translated (to-English) pages useful, what types of errors occur, what do you find that is impressive/substandard, and the ease of use of the interface.

Dictionary Translation (15 points)

Describe three significant problems with using dictionaries to translate queries in cross-language information retrieval.

Language Identification (70 points)

There are several approaches to language identification. Typical methods include: (1) using common words (stopwords) as features; (2) using character-based language modeling techniques; (3) vector space comparison between a training document and each test document (the training 'document' may be a large sample of text); and, (4) compression techniques (i.e., train a compression model for each language and see how each 'test' document compresses using each model). On the course web page I have included samples of English, French, and Spanish text, along with test files for each. Each test file contains 1000 sentences with one sentence on each line. The files are in the ISO-8859-1 (Latin-1) encoding. Your task is to build a classifier to predict language and evaluate its results on the test documents. Describe your methods and results.

Evaluation. Assess the performance of your classifier by calculating precision, recall, and F-scores for each language; you will obtain three metrics for each language. Precision(*Lang*) = percentage of time that you predict language=*Lang* and you are correct. Recall(*Lang*) = percentage of cases where the true language is *Lang* and your prediction is correct. Both precision and recall are values between 0.0 and 1.0. F-scores can be computed as $2 * P * R / (P + R)$. You should try to obtain 90% accuracy on the test sets.

You are not required to use the *training* data that I provided. And you may use other sources if you like. My texts are works of fiction/literature, from Project Gutenberg. You may use other approaches to those mentioned above, and you may use publicly available tools (e.g., *gzip*, language modeling toolkits, SVM_light, decision trees); however, you should not use software intended to solve the entire identification problem (i.e., you should not rely on products such as *Rosette* (by BASIS Technology) or demos such as <http://odur.let.rug.nl/~vannoord/TextCat/Demo/>).

The first two lines of each test file are shown below:

English

Resumption of the session

I declare resumed the session of the European Parliament adjourned on Friday 17 December 1999 , and I would like once again to wish you a happy new year in the hope that you enjoyed a pleasant festive period .

Spanish

Reanudación del período de sesiones

Declaro reanudado el período de sesiones del Parlamento Europeo , interrumpido el viernes 17 de diciembre pasado , y reitero a Sus Señorías mi deseo de que hayan tenido unas buenas vacaciones .

French

Reprise de la session

Je déclare reprise la session du Parlement européen qui avait été interrompue le vendredi 17 décembre dernier et je vous renouvelle tous mes vux en espérant que vous avez passé de bonnes vacances .