Homework #6 (due in 2 weeks – 4/17/17)

Multilingual Issues (100=10+10+10+70 points)

This assignment will give you a chance to consider multilingual issues in text retrieval. Specifically you will take a look at a web-based CLIR system and build a simple approach to language identification. I assume you have read the assigned paper by Kishida.

Dictionary Translation (10 points)
Describe three significant problems that arise when using dictionaries to translate queries in cross-language information retrieval.

Translation ambiguity (10 points)
Briefly explain how Pirkola has attempted to cope with translation ambiguity in CLIR.

Language Models (10 points)
Briefly explain how statistical language models can be modified to support cross-language information retrieval.

Language Identification (70 points)
There are several approaches to written language identification. Typical methods include: (1) using common words (stopwords) as indicators; (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) standard machine learning approaches (naive Bayes, SVMs). 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 encoded in UTF-8. Your task is to build a classifier to predict language using any method[1], and to 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). Show your work for calculating precision and recall (i.e., show numerators and denominators) and report scores with at least four digits of precision. You should try to obtain 90% accuracy on the test sets.[2]

You are not required to use the *training* data that I provided. And you may use other sources of training data 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., 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. Note how punctuation is separated by space, however, this is not the case in the provided training data.

[1] See below. You should not use an out of the box tool designed specifically for language identification.
[2] I obtained 90% accuracy in each language by simply using small stopword lists.

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 .

Extra Credit (up to 4 points)
I'll give 4 points extra-credit on the assignment to the student with the highest accuracy (F-score) on the Spanish data. And 2 points to the student with the second highest accuracy.