Introduction:

Google allows website publishers to earn money by displaying ads on their websites. Google automatically delivers ads that are targeted to a publisher’s content or audience. In particular, contextual targeting uses keyword analysis and word frequency to determine what a webpage is about and precisely match Google ads to each page. We refer to this task of categorizing a webpage as text classification, and this is done for webpages in many languages, classifying into hundreds of categories such as sports, travel, fitness etc. It is important for Google to build text classifiers that have high precision (predicted categories are accurate) and high recall (all relevant categories are predicted). Continually improving text classification is extremely important to Google because it has a large positive impact for publishers and advertisers, and it improves the user experience.

The performance of our text classifiers depends on the models used, and the quality and quantity of labeled training data. Collecting human-labeled training data is relatively expensive, especially when this needs to be done in many languages. Because categories are language independent (Tennis in English is the same category/concept as Tennis in Czech, even if the words used on a webpage about Tennis are different in each language), one approach for obtaining training data in other languages is to use Google’s machine translation to convert training data in one language, say, English, to training data in another language. The focus of this project is to investigate approaches, benefits and limitations of using machine translation to generate training data for foreign-language text classifiers.

Technical background:

The team will investigate various aspects of using machine translation to train taxonomic
and binary classifiers in multiple languages. Sample research questions include:

- What are the benefits and limitations of using machine translation for training foreign-language text classifiers?
- What are optimal source-destination language pairs? For example, if one needs training data in Korean, what is the best source language?
- Understanding datasize effects. Collecting labeled data to train classifiers is expensive, so we want to collect as much data as needed, but no more. What volume of translated training data is optimal?
- Improving foreign-language binary classifiers for rare events. Rare events can be particularly difficult to predict accurately. But when using machine translation, do we need to be concerned about rare events classes that are idiomatic?
- How does translating bag-of-words compare to translating phrases or sentences?

Possible project tasks: The team will need to:

- **Create datasets.** This project will require labeled data across multiple languages, so the team will use existing datasets (since labeling data is beyond the scope of this project). Possible sources include Wikipedia in multiple languages, and/or other parallel corpora such as the European Parliament Proceedings.
- **Translate data.** Develop tools that use Google’s translation service to translate text from one language to another.
- **Develop classification tools.** Use existing off-the-shelf classifiers to train and evaluate models. These may include classifier test-suites such as Weka, or stand-alone package such as SVM-Light.
- **Devise experiments.** With the guidance of the Google mentor, create experiments that explore questions of interest related to using machine translation for text classification.

**Special Requirements:**

Knowledge of one or more of C++, Java, Python and Matlab is a requirement. No datasets or computer resources will be provided by Google. The team will make use of the Google translate service at http://translate.google.com/.
Expectations:
By the end of the project, the team should have a good understanding of the benefits and limitations of using machine translation for training text classifiers in foreign languages.

Suggested Reading:
