**CROSS LANGUAGE INFORMATION RETRIEVAL**

**What is CLIR?**

CLIR (cross-language information retrieval) is a subsection of information retrieval that deals with retrieving data written in a language other than the user's query language.

**About our project:**

The goal of this project is to create a cross-language information retrieval system (CLIR) that can search text documents published in English and show the results in German, given a query in German. In a nutshell we'll utilize machine translation, a vector space model for information retrieval, and IR assessment techniques to evaluate the system's performance.

**Why this project?**

CLIR systems have progressed a lot. So, we want to contribute to this field by developing an accurate multilingual and cross-lingual adhoc information retrieval system which is as successful as other monolingual systems.

**Approaching the problem statement:**

* Tokenizing the input and loading the data files
* The lexicon is preprocessed by stemming and deleting stopwords.
* Calculating the TF/IDF representation for all of the documents in the Wikipedia corpus.
* Using an inverted index to quickly find pages based on a query word.
* Using BM25 to implement querying.
* Tests are carried out.

**Data to be used:**

* bilingualTraslationText.(en,de): A sentence aligned, parallel German-English corpus from the Europarl corpus (which is a collection of debates held in the EU parliament over a number of years). This information will be used to create word-alignment tools and a translation likelihood table.
* crossval.(en,de): A smaller parallel corpus used to evaluate the translation system.
* sample.(docs,queries,qrel): A collection of English documents (from Wikipedia), German queries, and relevance judgment scores for each query-document pair.

**Evaluation:**

We'll employ the Mean Average Precision to assess our CLIR engine's performance. In IR, MAP is a typical assessment metric.

**Project by:**

Vijaya Teja Rayavarapu Ankur Yadav

Keerthi Gosika Shreeyash Amit Yende