Birla Institute of Technology and Science, Pilani

CS F469-Information Retrieval

Second Semester 2016-17

Due Date: 16/4/2017 **Total Marks: 30**

Assignment - 2

In this assignment you are required to implement a **Cross Lingual Document Translator** based on statistical machine translation model**.** It is recommended that each group maintains a **github repository** for this assignment. You should commit in small logical commits with proper commit messages. Here are the links to Git Tutorial and Git Documentation.

Cases of plagiarism would be penalized by awarding zero marks. This includes using any of the code submitted in any of the present/ previous offerings of this course.

Languages - C, C++, Java, Python.

Corpus - You are required to use the following corpus:
https://drive.google.com/open?id=0B6v0ngBbAZWvcUstV3Z4bWJLRHM

Submission: You are required to work in the same groups as assignment 1.

Here is an outline of the assignment in stages-

1) **Pre-Processing** (if required)
   Any required preprocessing should be carried out, and any steps taken or lack thereof, should be justified in the design document.

2) **Statistical Machine Translation**
   A statistical model has to be trained for alignment and translation using the indices built in the previous step. Kindly refer IBM Models and Expectation Maximization (EM) Algorithm covered during class. No external library should be used in this phase.

3) **Compute Similarity**
   Build a module which takes input two documents of the same language at a time and outputs the cosine similarity and Jaccard Coefficient for the two. No external libraries are allowed in this phase. Take the total number of documents to be translated as user input. This number could be large.

4) **Performance Metrics:**
   Your submission would be evaluated as follows-
   You will be given a few test cases at the time of your demo. Each test case would be a French/English document. The translator must generate the corresponding English/French translation of the same. (You can be asked to translate in either direction.) This generated translation would then be compared against an accurate translation (provided at the time of Demo) using Cosine Similarity and Jaccard Coefficient, which is supposed to be implemented inherently into your translator.
   You will be awarded marks accordingly.

Deliverables

1. **Code -** Should be well documented. The purpose and intent of each method, class and modules should be mentioned.
2. **Design Document -** Should justify each and every aspect of the implementation including the data structures that you used for storing the terms, alignments, etc. and all the algorithms used.
3. **README file -** List down all the steps for compiling your code and running the translator. Also list down all the assumptions that you have made while implementing this assignment (if any), along with the Github Repository link.

Interface

The translator can have a simple command line interface (GUI doesn't carry any marks). But the driver code, when run, should have the following interface implemented -

1. Compute the cosine similarity and Jaccard Coefficient between two documents specified by paths (show them for each doc pair)
2. Show the translated document.
3. Show the average cosine similarity and Jaccard Coefficient for all the test cases.

Innovation

Any kind of extra efforts made by the group to improve the performance would be rewarded suitably. Some of these could be improving upon IBM model 1 using any of the higher degree IBM Models or any other optimization/heuristics.

Submission Details

Make a Zip of all the above mentioned deliverables and mail it to lavika.goel@pilani.bits-pilani.ac.in before the deadline along with the subject line "IR Submission Group No. XX", and the zip file named as "Group no. XX" where XX should be substituted with the group number.
You will have to run the code on your own machine, at the time of the demo.

Evaluation Scheme

Task	**Marks**
Implementing Statistical Translational Model	7
Similarity Calculation	3
Innovation	3
Evaluation of Performance Metrics	3
Documentation	4
Viva	10