COMP 472/6721 **Assignment 4** **Fall 2012**

Due date:	December 3, 2012.
Late submission:	20% per day.
Teams:	You can do the assignment individually or in teams of 2. Teams must submit only 1 copy of the assignment.
Purpose:	The purpose of this assignment is to make you practice language modeling.

Language Modeling for Automatic Language Identification

In this assignment you will build a probabilistic language identification system to identify the language of an input sentence. You will first build a character-based language model for 3 languages: English, French and a language of your choice that uses the same character set. To build the English and French language models, you will use different translations of the first chapters of *Le Petit Prince* by Antoine de Saint-Exupery. On the course Moodle page, you will find the first 2 chapters of the English and the French versions. For the other language, you must find an online version of a text yourself (it doesn't have to be the same book). The Web is a great place to find electronic corpora. Look at the Web page of Project Gutenberg[1] for a good starting point.

Once you have your 3 training corpora, you must:

- Build a bigram character-based language model for each language.
- Use your language models to identify the most probable language of a sentence given as input to your program.

For example, if your smoothed-probability language models are the following (note that these values are made-up):

```
Language model for French :
AA: 7.9834e-005  AB: 1.3572e-003  AC: 3.5925e-003  AD: 2.3950e-004
AE: 7.9834e-005  AF: 2.3950e-004  AG: 3.2732e-003  AH: 7.9834e-005
AI: 1.3971e-002  AJ: 7.9834e-005  AK: 7.9834e-005  AL: 3.9119e-003
AM: 2.3152e-003  AN: 6.6262e-003  AO: 2.3950e-004  AP: 2.7942e-003
AQ: 3.9917e-004  AR: 9.6599e-003  AS: 3.4329e-003  AT: 6.9456e-003
AU: 4.5505e-003  AV: 3.1135e-003  AW: 7.9834e-005  AX: 7.9834e-005
AY: 2.3950e-004  AZ: 7.9834e-005
…

Language model for English :
AA: 7.9834e-005  AB: 1.3572e-003  AC: 3.5925e-003  AD: 2.3950e-004
AE: 7.9834e-005  AF: 2.3950e-004  AG: 3.2732e-003  AH: 7.9834e-005
AI: 1.3971e-002  AJ: 7.9834e-005  AK: 7.9834e-005  AL: 3.9119e-003
AM: 2.3152e-003  AN: 6.6262e-003  AO: 2.3950e-004  AP: 2.7942e-003
AQ: 3.9917e-004  AR: 8.5592e-003  AS: 3.4329e-003  AT: 6.9456e-003
AU: 4.5505e-003  AV: 3.1135e-003  AW: 7.9834e-005  AX: 7.9834e-005
AY: 2.3950e-004  AZ: 7.9834e-005
…

Language model for Italian :
AA: 7.9834e-005  AB: 1.3572e-003  AC: 3.5925e-003  AD: 2.3950e-004
AE: 7.9834e-005  AF: 2.3950e-004  AG: 3.2732e-003  AH: 7.9834e-005
AI: 1.3971e-002  AJ: 7.9834e-005  AK: 7.9834e-005  AL: 3.9119e-003
AM: 2.3152e-003  AN: 6.6262e-003  AO: 2.3950e-004  AP: 2.7942e-003
AQ: 3.9917e-004  AR: 6.0509e-003  AS: 3.4329e-003  AT: 6.9456e-003
AU: 4.5505e-003  AV: 3.1135e-003  AW: 7.9834e-005  AX: 7.9834e-005
AY: 2.3950e-004  AZ: 7.9834e-005
…
```

[1] http://www.gutenberg.org/

Your program should behave this way:

```
Enter a sentence:  Mary is eating an apple.

BIGRAM: MA
FRENCH:  P(M,A) = 2.6345e-003 ==> log prob of sequence so far: -2.5793
ENGLISH: P(M,A) = 2.3552e-004 ==> log prob of sequence so far: -3.6280
ITALIAN: P(M,A) = 1.8693e-003 ==> log prob of sequence so far: -2.7283

BIGRAM: AR
FRENCH:  P(A,R) = 9.6599e-003 ==> log prob of sequence so far: -4.5943
ENGLISH: P(A,R) = 8.5592e-003 ==> log prob of sequence so far: -5.6956
ITALIAN: P(A,R) = 6.0509e-003 ==> log prob of sequence so far: -4.9465

BIGRAM: RY
FRENCH:  P(R,Y) = 2.6666e-003 ==> log prob of sequence so far: -7.1683
ENGLISH: P(R,Y) = 5.5555e-002 ==> log prob of sequence so far: -6.9509
ITALIAN: P(R,Y) = 3.3333e-004 ==> log prob of sequence so far: -8.4236

…

The sentence is in English
```

Notes:

1. Make sure that all 3 languages use the same character set. In particular, you should not take diacritics (accents, cedillas...) into account. The French version of *Le Petit Prince* available on Moodle has been cleaned of diacritics. In fact, as long as you explain them in your report, you can make a series of assumptions about the character set you are using. For example, you can reduce all uppercases to lowercases, group punctuation together, ignore non-alphabetic characters…

2. In this assignment, you will smooth your probabilities with the add-λ method with λ=0.5. Recall that:

$$Prob(bigram) = \frac{Freq(bigram) + \lambda}{N + \lambda B}$$

where:
- λ=0.5
- N is the number of bigrams instances in the training text (the total number of bigrams in each training text)
- B is the number of possible bigrams types (e.g. 28x28 if you have 28 characters in your character set)

3. In order to avoid arithmetic underflow, remember to work in log space (add the log of the probabilities instead of multiplying the actual probabilities). Any logarithmic base would work; but use base 10 for this assignment.

One-page Report:

Your program must be accompanied by a written report (~1 page) that includes:

- Instructions on how to run your program
- The assumptions that you made (e.g. what character set you used)
- The language you chose and why

In addition to the report, include a trace of your output for 10 sentences:

- show 5 sentences that your system classifies correctly;
- and 5 sentences that your system gets wrong.

Two of these 10 sentences must be the following:

1. *Where are Jim and Julia?*
2. *Ils reviendront au final, peut-etre.*

Programming Language:

You are not required to use a specific programming language; Java, C++ or C are appropriate for this task If you wish to use another programming language, you need to check with me first.

Divertitevi! (Have fun!)

Submission:

The assignment must be handed electronically by midnight on the due date.

1. Make sure that you have signed the expectation of originality form (available on the Web page; or at: http://www.encs.concordia.ca/documents/expectations.pdf) and given it to me.

2. In addition, write one of the following statements on your assignment:
   - For individual work: *"I certify that this submission is my original work and meets the Faculty's Expectations of Originality***",** with your signature, I.D. #, and the date.
   - For group work*: "We certify that this submission is the original work of members of the group and meets the Faculty's Expectations of Originality"*, with the signatures and I.D. #s of all the team members and the date.

3. To hand in your assignment electronically:
   - Create one zip file, containing all files for your assignment.
   - Name your zip file this way:
     - For individual work: name the zip file: *a2_studentID*, where *studentID* is your ID number.
     - For group work: name the zip file: *a2_studentID1_studentID2*, where *studentID1* and *studentID2* are the ID numbers of each student.
   - Upload your zip file at: https://fis.encs.concordia.ca/eas/