COMP 6781 **Assignment 2** **Winter 2016**

Due date: ~~February 11, 2016~~ **March 17, 2016**
Late submission: 25% per day.
Teams: No teams allowed. The assignment must be done individually.

Language Identification

In this assignment you will build a probabilistic language identification system to identify the language of tweets. This will be done using a simplification of the dataset used in the TweetLID 2014 shared task [1].

For this assignment, you are required to do the programming yourself. You cannot use any 3rd party tool, unless you specifically have my permission to do so. However, you are not required to use a specific programming language; C++, C, C#, Python or Java are appropriate for this task. If you wish to use any other programming language, please check with me first.

The description of the TweetLID 2014 shared task is available at the URL: http://komunitatea.elhuyar.eus/tweetlid. In this assignment, you will use a simplified version of the dataset used in this shared task. It contains tweets for 6 related languages: Basque (eu), Catalan (ca), Galician (gl), Spanish (es), English (en) and Portuguese (pt). The simplified dataset is available on the Moodle page.

The assignment consists in 3 parts:

1- Building a character-based unigram & bigram language models from the training set provided (the file `simple-training-tweets`). For each language, you will build:
   a) A unigram model, smoothed with add-delta smoothing with δ=0.1.
   b) A bigram model, smoothed with add-delta smoothing with δ =0.1.

   As long as you explain them, you can make a series of assumptions about the character set you are using. For example, you can reduce diacritics to non-diacritic forms, ignore case distinction, use only letters, group punctuations together, …

   Dump your LMs in 2 files called `unigramLM.txt` and `bigramLM.txt`. These files must include the first 50 "grams" (i.e. unigram or bigram) along with their unsmoothed probability and their smoothed probability.

2- Test your language models with the test set provided (the file `simple-testing-tweets`)
   For each tweet in the test file, use both your LMs above and classify the tweet with its most likely language. Dump your results in 2 files called `results-unigram.txt` and `results-bigram.txt` which must include the tweet number and its most likely language.

3- Evaluate & analyse your results using the file `simple-testing-tweets`
   Create 2 text files called `analysis-unigram.txt` and `analysis-bigram.txt` which includes the following information:
   a) The overall accuracy of the LMs
   b) The accuracy for each language
   c) A confusion matrix indicating the correct classification versus the classification of your system

The report:
Write a report (~4 pages) to describe your code and your results. Your report must describe:

1. The program:
   a) Describe your code itself (choice of language, data structures, …)
   b) Indicate the instructions necessary to run your code (files, commands, …)
2. The analysis of your results
   a) Which method/language seems to give the best results?
   b) Why?
   c) …

Reference:

[1] Zubiaga, A., San Vicente, I., Gamallo, P., Pichel, J. R., Alegria, I., Aranberri, N., Ezeiza A., Fresno, V. (2014). Overview of TweetLID: Tweet language identification at SEPLN 2014. Proceedings of the TweetLID Worshop at SEPLN2014. Girona. pp. 1-11.

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: https://www.concordia.ca/content/dam/encs/docs/Expectations-of-Originality-Feb14-2012.pdf) and given it to me.

2. In addition, write one of the following statements on your assignment:

   "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.

3. Upload your files:
   - Create one zip file, containing all files of your assignment.
   - Name your zip file *a1_studentID*, where *studentID* is your ID number.
   - Upload your zip file at: https://fis.encs.concordia.ca/eas/