ASSIGNMENT 1

COMP 599, Fall 2015

Due: September 29^{th}, 2015 in class. No late assignments accepted.

You must do this assignment individually. You may consult with other students orally, but may not take notes or share code, and you must complete the final submission on your own.

Question 1:	30 points
Question 2:	20 points
Question 3:	10 points
Question 4:	40 points
	100 points total

Assignment

Question 1: Identify the Ambiguity (30 points)

Analyze the following passages by identifying the most salient instances of linguistic ambiguity that they exhibit. Write a *short* paragraph for each that answers the following questions. What is the ambiguity, and what are the different possible interpretations? What in the passage specifically causes this ambiguity? What domain does this ambiguity operate over (phonological, lexical, syntactic, orthographic, etc.)? What sort of knowledge is needed for a natural language understanding system to disambiguate the passage, whether the system is human or machine?

1. *All that glitters is not gold.*
2. *Put down the dog. It doesn't want to be held.*
3. *My computer's got a Miley Virus. It's stopped twerking.* (Source: The 2014 UK Pun Championship)
4. *The reporters want to meet the oldest living person.*
5. *Stolen jewels found by tree.*

Question 2: FST for French Verbal Conjugation (20 points)

Develop a FST to perform morphological analysis for the following French verbal conjugation table, which shows verbs conjugated in the present tense:

Infinitive	**1 Sg**	**2 Sg**	**3 Sg**	**1 Pl**	**2 Pl**	**3 Pl**
-er verbs						
parler (to speak)	parle	parles	parle	parlons	parlez	parlent
jouer (to play)	joue	joues	joue	jouons	jouez	jouent
regarder (to watch)	regarde	regardes	regarde	regardons	regardez	regardent
-re verbs						
attendre (to wait)	attends	attends	attend	attendons	attendez	attendent
perdre (to lose)	perds	perds	perd	perdons	perdez	perdent
Irregular verbs						
être (to be)	suis	es	est	sommes	êtes	sont
avoir (to have)	ai	as	a	avons	avez	ont

The morphological analyzer should provide the infinitive form of the verb, which we will take to be its lemma, along with its POS, person and number agreement. For example, feeding "*ont#*" as input to the final FST should result in the output "*avoir* +V +3 +Pl".

Your response should include three components:

- A schematic transducer in the style of Figure 3.13 in J&M (page 61)
- A lexicon table as in the top half of Figure 3.14 in J&M (page 62)
- A "fleshed-out" FST in the format of the bottom half of Figure 3.14 for the lexical items presented above

Question 3: Good-Turing Smoothing (10 points)

Prove that the simple version of Good-Turing discounting given in the lecture notes leads to a proper probability distribution. That is, the version in which:

$$P(UNK) = f_1/N,$$
$$P(w_c) = \frac{(c+1)f_{c+1}}{Nf_c},$$

where *UNK* represents the set of unknown words, w_c represents a word seen during training with a frequency of c, and N represents the total number of tokens in the training corpus.

Question 4: Document Classification (40 points)

The goal of this question is to give you experience in using existing tools for machine learning and natural language processing to solve a document classification task. Before you attempt this question, you will need to install Python 2 on the machine you plan to work on, as well as the following Python packages and their dependencies:

- NLTK: `http://www.nltk.org/`
- NumPy: `http://www.numpy.org/`
- scikit-learn: `http://scikit-learn.org/stable/`

Download the corpus of text available on the course website. This corpus is a collection of news articles that were used in the TAC 2010 and 2011 summarization data sets, and contains articles separated into five categories:

1. Accidents and Natural Disasters
2. Attacks
3. Health and Safety
4. Endangered Resources
5. Investigations and Trials

Your task is to train a document classification system to distinguish documents from these five topic categories.

Data storage and format

The raw text files are stored in the subfolders */tac2010* and */tac2011*. Included in each data set is also a file that stores the topic category (i.e., the label to be predicted), for example */tac2010/tac2010.labels*, where each line is in the format of "document_id *tab* category". The category will be an integer from 1 to 5, corresponding to the five categories listed above.

Preprocessing and feature extraction

Preprocess the input documents to extract feature vector representations of them. Your features should be N-gram counts, for $N \leq 2$. Use NLTK's tokenizer to help you in this task. You should experiment with the complexity of the N-gram features (i.e., unigrams, or unigrams and bigrams), whether to distinguish upper and lower case, whether to remove stop words, etc. NLTK contains a list of stop words in English. Also, remove infrequently occurring words and bigrams as features. You may tune the threshold at which to remove infrequent words and bigrams. You may choose to experiment with the amount of smoothing/regularization in training the models to achieve better results, though you can also just leave these at the default settings. Read scikit-learn's documentation for more information on how to do this.

Model selection and tuning by cross-validation

Use TAC 2010 as your training set and TAC 2011 as your testing set. There will not be a separate development set for model selection. Instead, you will compare the following two methods:

1. Select the best model based on results on the training set
2. Do two-fold cross-validation on the training set to choose the best model.

Compare the logistic regression, support vector machine (with a linear kernel), and Naive Bayes algorithms for each of the feature sets that you plan to explore. For the cross-validation part, scikit-learn provides a package to help you with this. Please make use of existing resources! Evaluate the six models (two for each classifier) from the model selection step on the final test set.

Report

Write a *short* report on your method and results, carefully document the range of parameter settings that you tried and your experimental procedure. It should be no more than one page long. Report on the performance in terms of accuracy, and speculate on the successes and failures of the models. What were the results of the two methods of model selection on the training set vs. the test set? For the overall best performing model, include a confusion matrix as a form of error analysis. Also, explain the role of cross-validation in the above experiment.

What To Submit

On paper: Submit a hard copy of your solutions to Questions 1 to 3, as well as the report part of Question 4 in class.

For the programming part of Question 4, you should submit one zip file with your source code to MyCourses under Assignment 1.