Assignment 1: Using the Stanford CoreNLP Toolkit

Due Thursday 1/22 @ 11:55PM
Submit via ecommons
30 point (+5 points extra credit)

1 Overview

In this assignment you will use the Stanford CoreNLP toolkit to analyze a handful of documents loosely related to amusement parks or proms in some way. You will be able to answer several of the questions with the same or very similar code, so you should **read the entire assignment before answering any of the questions**. **Each question will be worth 5 points.**

Collaboration: You may work in pairs for this assignment, but everyone should submit their assignment to ecommons. Please include the names of the team members at the top of your report.

Deliverables: Please turn in a PDF or .doc file with answers to the questions organized by each section. In addition you must submit 12 files named 01-12.chain and 1 file named 01.coreference (see below).

Late penalty: You may turn in the assignment late as long as it is turned in before the next class period when we will discuss it. However, **you will lose 6 points per day that it is late**.

2 Question 1: Document Classification

There are 4 genres of discourse in the included textual data. For example, personal stories.

- What are they? Give your best guess. There aren't specific words you will be graded against, but it should be relatively obvious what they are after examining them. **Hint:** There are 4 documents in genre 1, 3 documents in genre 2, 2 documents in genre 3 and 3 documents in genre 4.
- How can you tell? Give as many reasons as you can think of.
- Group the texts into the 4 genres and report them using their filenames.

3 Question 2: Part-Of-Speech

For each genre what are the most 10 most frequent part-of-speech tags and what is their relative frequency within that genre? Relative frequency is simply the count of the tag divided by the total number of tags.

$$\text{RelFreq}(x_{pos}) = \frac{\text{count}(x_{pos})}{\sum_y \text{count}(y_{pos})} \tag{1}$$

For example, if the tag **NN** was seen 10 times and there was a total of 100 words (the same as the number of part-of-speech tags) in the document, the relative frequency would be 0.1. This is the same as the unigram probability of the part-of-speech tag.

Please produce a table similar to Table 1, where the part-of-speech tags are sorted in descending order (largest-to-smallest) of their relative frequency.

Genre	Part-Of-Speech	Frequency	Relative Frequency
genre 1	DT	78	0.20
genre 1	NN	54	0.13
...	...	...	...
genre 2	PRP	66	0.18
genre 2	DT	50	0.15
...	...	...	...

Table 1: Example results table with made up numbers

4 Question 3: Named Entity Recognition

What are the relative frequencies for the 7 types of named entities recognized by the toolkit: Time, Location, Organization, Person, Money, Percent, Date. Please produce a table similar to the one for part-of-speech. Are 7 types enough? What are some other types you think would be useful?

5 Question 4: Dependency Relations

For each document, create a chain of verbs with their arguments (only consider *nsubj*, *dobj*, and *iobj*) extracted from the dependency parse. Verbs should be in the order in which you see them in the document. The output should be predicate argument structure, where the verb is the predicate and the arguments are the *nsubj*, *dobj*, and *iobj* in that order (when available). See Figure 1 for an example input and output format. Each chain should be saved to a separate text file named XX.chain, where XX is the original file name of the source document.

Extra credit (5 points): Many of the arguments are missing. For example, the Fox saw the crow fly off, but this is not reflected in the simple predicate argument structure. Modify your code to improve the predicate argument structure in some way that captures more of the information. Include the modified results using the file naming convention XX.chain.extra and explain what you did.

6 Question 5: Coreference

For each genre, what is:

- the longest, shortest, mean, and median coreference chain length
- the maximum, minimum, mean and median number of mentions

For document 01.txt, fix any errors in the coreference chains that you can find and turn in the corrected XML (just the <coreference> section) as a separate file named 01.coreference. How do the values change once you correct the errors?

7 Question 6: Overall

- What are your general impressions of the toolkit.
- Where did it have the most problems and what do you think the biggest reason is?

A Fox once saw a Crow fly off with a piece of cheese in its beak and settle on a branch of a tree. "That's for me, as I am a Fox," said Master Reynard, and he walked up to the foot of the tree. "Good day, Mistress Crow," he cried. "How well you are looking today: how glossy your feathers; how bright your eye. I feel sure your voice must surpass that of other birds, just as your figure does; let me hear but one song from you that I may greet you as the Queen of Birds." The Crow lifted up her head and began to caw her best, but the moment she opened her mouth the piece of cheese fell to the ground, only to be snapped up by Master Fox. "That will do," said he. "That was all I wanted. In exchange for your cheese I will give you a piece of advice for the future: "Do not trust flatterers."
saw(Fox) fly(Crow) settle(Crow) 's(That) said(Reynard) walked(he) cried(Crow) cried(he) looking(you, today) surpass(voice, that) does(figure) hear(me) great(I, you) lifted(Crow, head) began(Crow) caw(null, best) opened(she, mouth) fell(piece) do(that) said(he) wanted(I) give(I, piece, you) do(flatterers)

Figure 1: An example of a chain of verbs with their arguments from Aesop's fable *The Fox and the Crow.*