NLP Assignment 1

Due: 09/27 11:59pm

Undergraduate students can choose to do Problem 1 and 2 (10 pts) or Problem 3 (10+5 extra pts). Graduate students should do Problem 3.

1. Use the Penn Treebank tagset to tag each word in the following sentences from different genres.

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form

28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

(1). News (1pt)
On the eve of President Obama's speech laying out his strategy to combat the terror group ISIS, national security hawks are demanding the United States authorize airstrikes against the group's leaders.

(2). Tweet (1pt)
What The Heck Is AdWords Quality Score And Why Does It Matter Anyway? http://bit.ly/12lHLap (@WordStream)

(3). Discussion Forum (1pt)
Actually, this isn't all that unusual. A second term drop or slump is practically par for the course. People are pretty much burned out on Obama. He's been President now for six years and he's basically in the news everyday, and we all know there's still two more years of the same old "ra-ra"s and jeers to go. Having said that ... Would Romney really beat Obama if the election was held this year? No. Elections are a lot more complicated than a single poll in an off election year. There would be very very very little voter turnover.

(4). Speech Conversation (1pt)
How has your life changed since The Big Bang Theory took off?
The Big Bang Theory has completely changed my life.
To be blunt, it's been several years since I've had to cash in any unemployment cheques - and that's really nice.
I know that sounds mundane, but it's very true.

(5). Shakespeare (1pt)
So shaken as we are, so wan with care,
Find we a time for frighted peace to pant,
And breathe short-winded accents of new broils
To be commenced in strands afar remote.

(6). (Extra 2pt) You could provide detailed comparison and analysis on the challenges across different genres, and/or translate the above sentences into another language and tag POS for the translations.

2. Grammar L1 is given as follows:

Grammar:

S → NP VP
S → Aux NP VP
S → VP
NP→ Pronoun
NP→ Proper-Noun
NP→ Det Nominal
Nominal → Noun
Nominal → Nominal Noun
Nominal → Nominal PP
VP → Verb
VP → Verb NP
VP → Verb NP PP
VP → Verb PP
VP → VP PP
PP → Preposition NP

Lexicon:

Det → that | this | a
Noun → book | flight | meal | money
Verb → saw | book
Pronoun → I | she | me
Proper-Noun → New York
Aux → does
Preposition → with | at

(1) Create a new grammar L2 based on L1 and expand the lexical entries so that you can parse the following sentence (2pt):

The spy saw the cop with the telescope.

(2) Show the parsing procedure for this procedure using bottom-up CKY parsing algorithm (3pt).

3. (J&M 5.8, 10pt; extra 5pt for undergraduate students) Build a bigram HMM tagger. Data is packed in wsj_pos.zip.
From the labeled training set, train the transition and emission probabilities of the HMM tagger directly on the hand-tagged data. Then implement the Viterbi algorithm so that you can decode (label) an arbitrary test sentence. Now run your algorithm on the test set. Report its error rate and compare its performance to the most frequent tag baseline. And also report the solutions for unknown word tagging.