CS-AD 220 – Spring 2016

Natural Language Processing

Session 6: 16-Feb-16

Prof. Nizar Habash

Some slides are adapted from Jurafsky and Martin's course slides on Speech and Language Processing

NYUAD Course CS-AD 220 – Spring 2016
Natural Language Processing

Assignment #1
Unix Tools and Regular Expressions
Assigned Feb 4, 2016

Due Feb 18, 2016 (11:59pm)

I. Grading & Submission

This assignment is about the use of regular expressions (regex) and a set of Unix tools for quick text processing. The assignment accounts for 10% of the full grade. Section III below has a set of questions. The student needs to answer them all. The specific number of points for each question is provided. The student should submit a **PDF** file containing the answers to each question and sub-question in order. The student should also include the commands and the result of applying the commands by copying and pasting from the terminal. Each student must work alone. This is not a group effort.

The assignment is due on Feb 18 before midnight (11:59pm). For late submissions, 10% will be deducted from the homework grade for any portion of each late day. The student should upload the answer to NYU Classes (Assignment #1).

Assignment #1 posted on NYU Classes

Moving Legislative Day Class

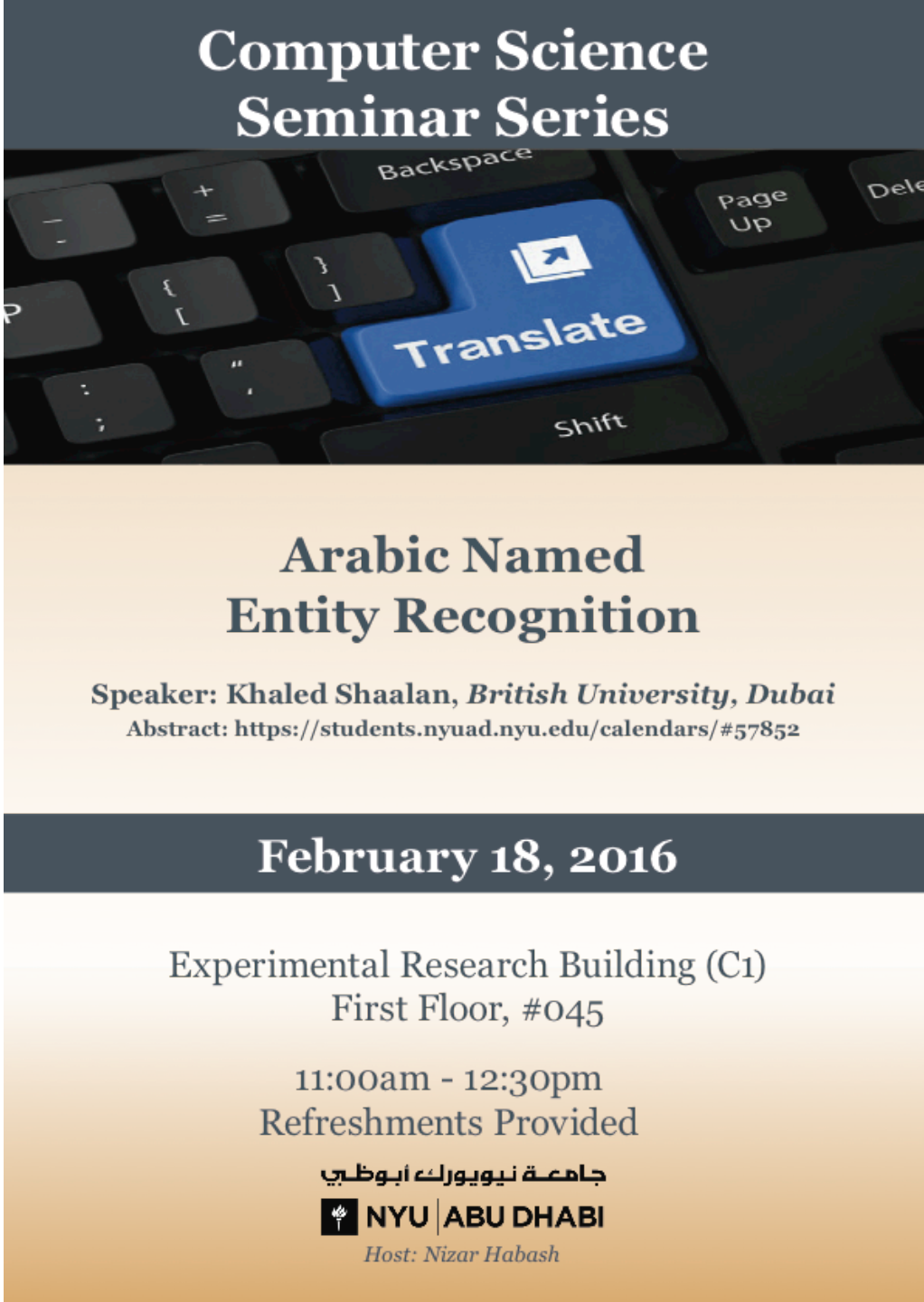
- Spring Break is March 18 – 25, 2016
- Sat March 26, 2016 is a Legislative *Thursday*
- Move to

Sat April 2, 2016 at 10am

Same Classroom C2-E049

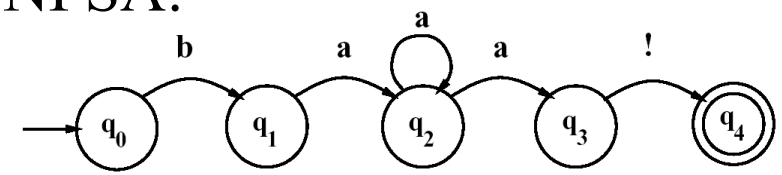
- Invited lecture >>>

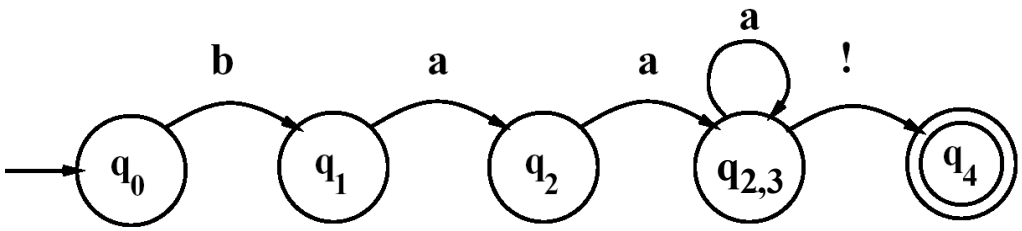
- Extra credit alert!
  - 1% of the whole grade.
  - Attend in person.
  - Submit a one page summary of the talk by February 25 in class.



Example of NFSA → DFSA

- Trace the states of the following NFSA:
  - q0 / b → q1
  - q1 /a → q2
  - q2 /a → q2,3 (an ambiguous state: q2 or q3)
  - q2,3 /a → q2,3 (here we trace the union of q2/a and q3/a)
  - q2,3 /! → q4 (again, trace the union of q2/! and q3/!)
- The DFSA states are q0, q1, q2, q2,3, q4
  - The DFSA looks like this





Morphology

- Morphology is the study of the ways that words are built up from smaller meaningful units called morphemes
  - Morpheme: the smallest unit of language that combines both a form (sound) and a meaning

- We can discuss morphology in terms of
  - Form
    - How to put words together
      - Affixational, templatic, reduplicative morphology
  - Function
    - The meaning of words from their parts
      - Derivational and Inflectional morphology
  - Form and Function operations are independent

Morphology

- Affixation
  - Prefix + Stem + Suffix
    - Un-interpret-able
    - Antidisestablishmentarianism
      - Anti-dis-establish-ment-ari-an-ism
      - Antidisestablishmentarian***istically***
  - Infix
    - tulong – t**um**ulong (Tagalog help – helped)
    - bili – b**um**ili (Tagalog buy– bought)
  - Circumfix
    - ge-kann-t (German known)
    - ge-zeig-t (German shown)

Morphology

- Agglutinative morphology
  - Stacking of morpheme
  - Highly semantically compositional
  - Turkish is the classic example

Turkish	English
ev	(the) house
evler	(the) houses
evin	your (sing.) house
eviniz	your (pl./formal) house
evim	my house
evimde	at my house
evlerinizin	of your houses
evlerinizden	from your houses
evlerinizdendi	(he/she/it) was from your houses
evlerinizdenmiş	(he/she/it) was (said to be) from your houses

Morphology

- Agglutinative morphology
  - Stacking of morpheme
  - Highly semantically compositional
  - Turkish is the classic example

muvaffakiyetsizleştiricileştiriveremeyebileceklerimizdenmişsinizcesine

(you are talking) as if you are one of those that we cannot easily convert into an unsuccessful-person-maker

Morphology

- Templatic morphology
  - Root + Template/Pattern
    - Arabic Root: KTB
    - CaCaC+nA katab+nA we wrote
    - na+CCuC na+ktub we write
    - maCCaC maktab office
    - CiCAC kitAb book
  - maCC[ai]C = location pattern
    - maktab /office; malEab / playground;
    - masraH / theater; madras+a / school;
    - masjid / mosque; majlis / assembly

Morphology

- Reduplication
  - Tagalog future tense
    - bili - bibili — buy – will buy
    - pasok – papasok — enter – will enter
    - lakad – lalakad — walk – will walk
- Ablaut and Suppletion
  - English irregular verbs
    - Fall – fell — ablaut
    - Think – thought — partial suppletion
    - Go – went — total suppletion

Morphology

- Allomorphy
  - Model variant forms of morphemes (allomorphs)
  - Morpheme +s (+PL) has three allomorphs
    - /s/, /z/, /iz/
    - s, s , es
    - cats, dogs, foxes
  - Arabic morpheme Al+ (Det) has 14 allomorphs…
    - aš+, as+, at+, ad+ …(preceding the so-called Sun Letters)
  - Turkish vowel harmony

Vowel Harmony in Turkish

adamlar	'men'	günler	'days'
anneler	'mothers'	ipler	'threads'
atlar	'horses'	yıllar	'years'
aylar	'months'	kalemler	'pencils'
bankalar	'banks'	kediler	'cats'
başlar	'heads'	kitaplar	'books'
camiler	'mosques'	kızlar	'girls'
çocuklar	'children'	masalar	'tables'
dersler	'lessons'	mevsimler	'seasons'
dişçiler	'dentists'	oteller	'hotels'

- Turkish has eight vowels
  - Front vowels: i ü e ö Back vowels: ı u a o
- The plural morpheme +ler has two allomorphs
  - +ler : following roots with front vowels
  - +lar : following roots ending with back vowels
  - Predictable complementary distribution that is phonologically conditioned

Morphology

- Phonotactics, Allophony
  - Model phonology independent of morphemes
  - The phones [b] and [p]
    - Different phonemes in English

      /bat/ and /pat/ ←minimal pair
    - Allophones of the same phoneme in Arabic

      /dibs/ [dips] not /dibs/ 'molasses'

Morphology

- Clitics
  - Clitics are affixational morphemes that phonologically dependent but syntactically independent
    - Proclitics (prefixing clitics)
    - Enclitics (suffixing clitics)
  - Clitics are often orthographically attached
  - Exmaples
    - Al+šams → /aš+šams/ 'the Sun' (Arabic)
    - ¡Ábrelo! 'Open it!' (Spanish)
- Contractions
  - A contraction is a shortned version of a word
    - Let's = let us
    - Won't = will not

Morphology

- In terms of functional operations, we can further divide morphology up into two broad types
  - Inflectional
  - Derivational

Word Classes

- By word class, we have in mind familiar notions like noun and verb
- We'll go into more details in Chapter 5
- Right now we're concerned with word classes because the way that stems and affixes combine is based to a large degree on the word class of the stem

Inflectional Morphology

- Inflectional morphology concerns the combination of morphemes where the resulting word:
  - Has the same word class as the original
    - Word classes are minimal distinctions based on part-of-speech (POS), but can include distinctions within the same POS (e.g., masculine/feminine/neuter nouns)
  - Serves a grammatical/semantic purpose that is
    - Different from the original
    - But is nevertheless transparently related to the original

Some Inflectional Features

- Person
  - 1st, 2nd, 3rd
- Gender
  - feminine, masculine, neutral, classes…
- Number
  - singular, dual, plural
- Case
  - nominative, accusative, dative, locative,…
- Tense
  - past, present, future
- Aspect
  - progressive, perfective, imperfective
- Mood
  - indicative, interrogative, imperative, conditional, subjunctive

Nouns and Verbs in English

- Nouns are simple
  - Markers for plural and possessive
- Verbs are only slightly more complex
  - Markers appropriate to the tense of the verb

Regulars and Irregulars

- It is a little complicated by the fact that some words misbehave (refuse to follow the rules)
  - Mouse/mice, goose/geese, ox/oxen
  - Go/went, fly/flew
- The terms regular and irregular are used to refer to words that follow the rules and those that don't

Regular and Irregular Verbs

- Regulars…
  - Walk, walks, walking, walked, walked
- Irregulars
  - Eat, eats, eating, ate, eaten
  - Catch, catches, catching, caught, caught
  - Cut, cuts, cutting, cut, cut
  - Go, goes, going, went, gone
  - Be, is, am, are, being, was, were, been

Derivational Morphology

- Derivational morphology is the messy stuff that no one ever taught you.
  - Quasi-systematicity
  - Irregular meaning change
  - Changes of word class

Derivational Examples

- Verbs and Adjectives to Nouns

-ation	computerize	computerization
-ee	appoint	appointee
-er	kill	killer
-ness	fuzzy	fuzziness

Derivational Examples

- Nouns and Verbs to Adjectives

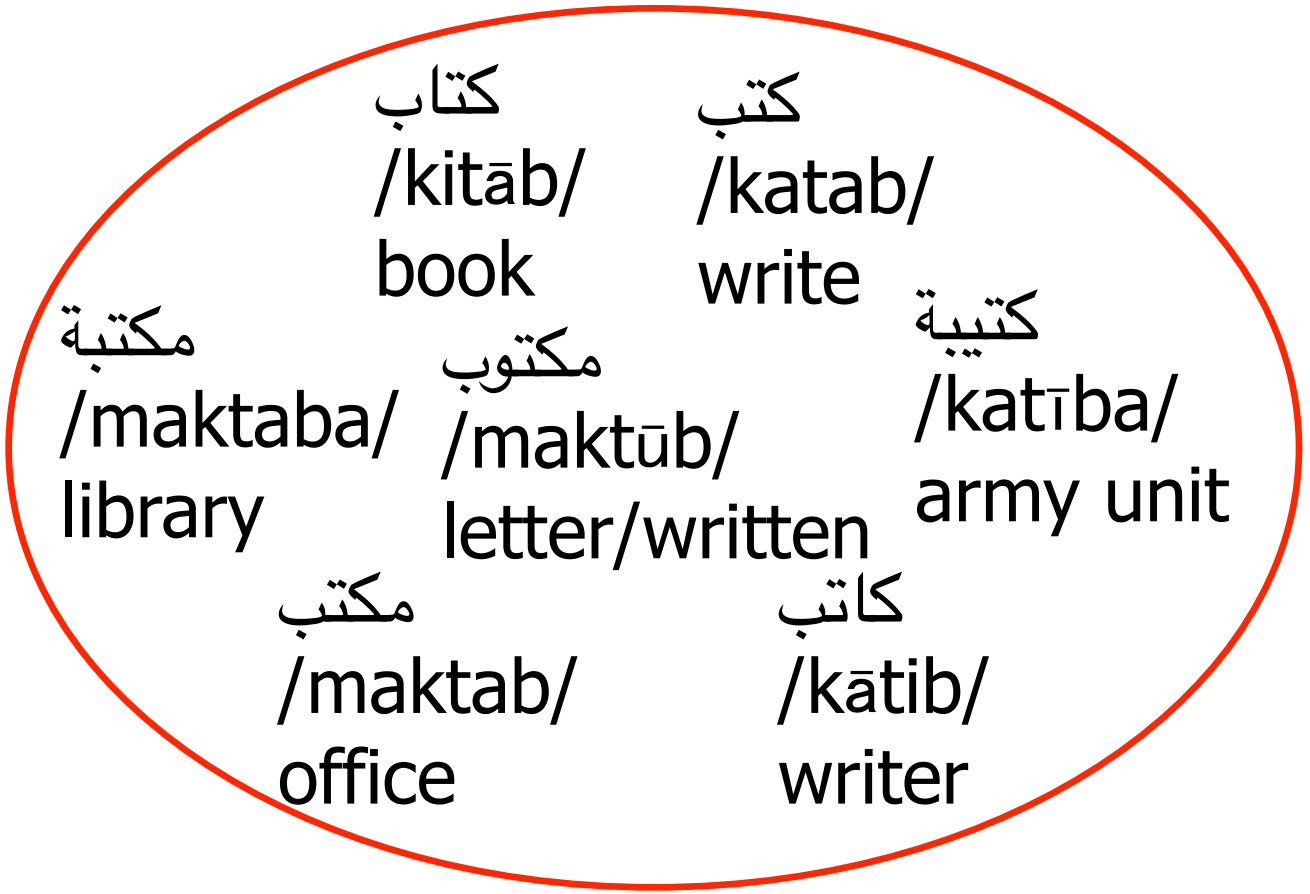
-al	computation	computational
-able	embrace	embraceable
-less	clue	clueless

Example: *Compute*

- Many paths are possible…
- Start with compute
  - Computer -> computerize -> computerization
  - Computer -> computerize -> computerizable
- But not all paths/operations are equally good (allowable?)
  - Clue
    - Clue -> clueless
    - Clue -> *clueable

Derivational Templatic Morphology
Arabic Root Meaning

writing-related = KTB = ك ت ب



Morphological Function
Derivation and Inflection

	Inflectional Operations	Derivational Operations
Part-of-Speech	Do not change POS (same lexeme)	**Often** change POS (different lexeme)
Meaning	Syntactically conditioned information, e.g., gender, number, case	Lexical meaning
Regularity	More regular	More idiomatic
Obligatoriness	Obligatory	Optional

Do you know any Swahili?

Swahili Morphology

(Eastern Congo Dialect)

1. Ninasema.	'I speak.'	6. Niliona.	'I saw.'
2. Wunasema.	'You speak.'	7. Ninawaona.	'I see them.'
3. Anasema.	'She speaks.'	8. Niliwuona.	'I saw you.'
4. Wanasema.	'They speak.'	9. Ananiona.	'She sees me.'
5. Ninaona.	'I see.'	10. Wutakaniona.	'You will see me.'

- What is the grammar of the verb morphology in this dialect of Swahili?
- How do we say?

_____	_____	_____
'She saw them.'	'I will see you'	'She saw me.'

Swahili Morphology

(Eastern Congo Dialect)

- What is the grammar of the verb morphology in this dialect of Swahili?

  Subject→Tense→Object→Verb

  Subject = {Ni/I, Wu/You, A/She, Wa/They}

  Object = {Ni/I, Wu/You, A/She, Wa/They}

  Tense = {na/present, li/past, taka/future}

  Verb = {sema/speak, ona/see}

- How do we say?

Aliwaona	Nitakawuona	Aliniona
'She saw them.'	'I will see you'	'She saw me.'

Next Time

- Read J+M Chap 3 (3.2 up to 3.8)
- Assignment #1 due Feb 18 midnight