Introduction to Data Science

Text Data and Analysis

LNMIIT

Different Levels

- Phonetics
- Morphology
- Syntax
- Semantics

- Pragmatics
- Discourse

LNMIIT

Progress

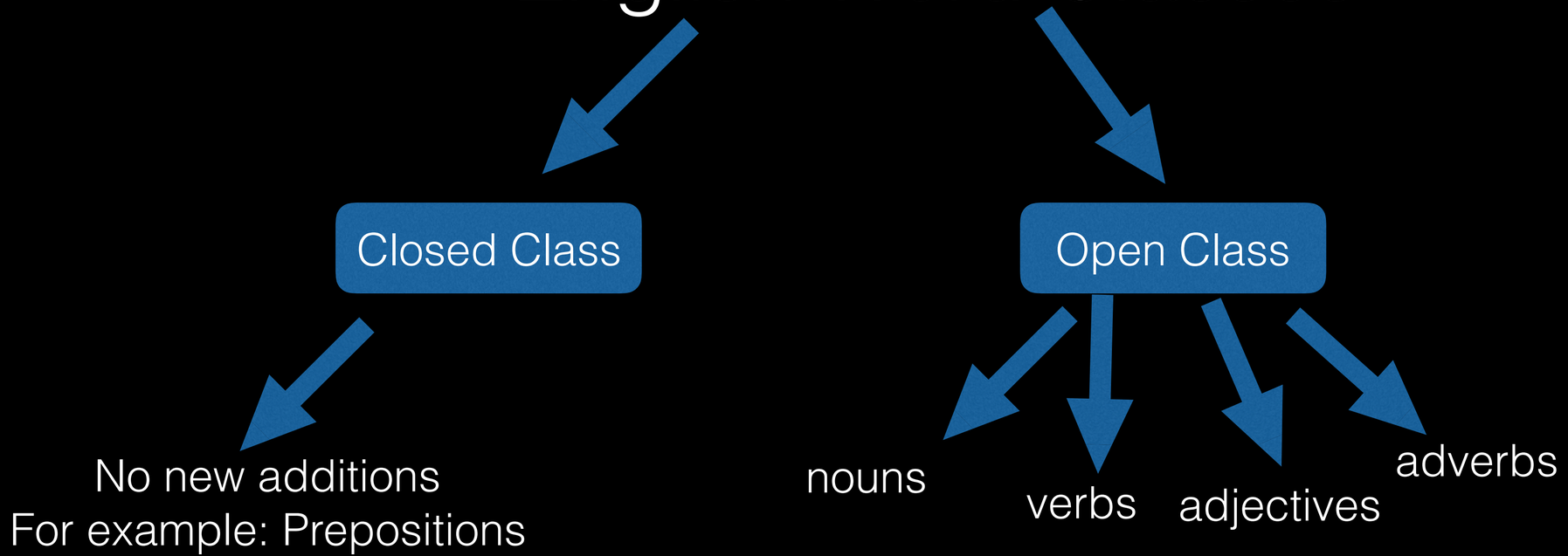
- Almost Done:
  - Spam versus Ham - 99% accuracy
  - PoS - 97%
  - NER - 97%

- Good Progress:
  - Sentiment Analysis
  - Wordsense disambiguation
  - Parsing
  - Machine Translation
  - Information Extraction

- Hard problems:
  - Question Answering systems
  - Paraphrase
  - Summarisation
  - Dialogue

LNMIIT

English Word Classes



LNMIIT

Part-of-Speech Tagging

- Process of assigning a part of speech pr other syntactic class marker to each word in the corpus
- Tokenisation is required, in general, before PoS Tagging
- Input: A string of words and a special tagset
- Output: A single best tag for each word

LNMIIT

Penn Treebank POS Tagset

Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	*and, but, or*	SYM	symbol	+,%, &
CD	cardinal number	*one, two, three*	TO	"to"	*to*
DT	determiner	*a, the*	UH	interjection	*ah, oops*
EX	existential 'there'	*there*	VB	verb, base form	*eat*
FW	foreign word	*mea culpa*	VBD	verb, past tense	*ate*
IN	preposition/sub-conj	*of, in, by*	VBG	verb, gerund	*eating*
JJ	adjective	*yellow*	VBN	verb, past participle	*eaten*
JJR	adj., comparative	*bigger*	VBP	verb, non-3sg pres	*eat*
JJS	adj., superlative	*wildest*	VBZ	verb, 3sg pres	*eats*
LS	list item marker	*1, 2, One*	WDT	wh-determiner	*which, that*
MD	modal	*can, should*	WP	wh-pronoun	*what, who*
NN	noun, sing. or mass	*llama*	WP$	possessive wh-	*whose*
NNS	noun, plural	*llamas*	WRB	wh-adverb	*how, where*
NNP	proper noun, singular	*IBM*	$	dollar sign	*$*
NNPS	proper noun, plural	*Carolinas*	#	pound sign	#
PDT	predeterminer	*all, both*	"	left quote	' or "
POS	possessive ending	*'s*	"	right quote	' or "
PRP	personal pronoun	*I, you, he*	(	left parenthesis	[, (, {, <
PRP$	possessive pronoun	*your, one's*	)	right parenthesis	],), }, >
RB	adverb	*quickly, never*	,	comma	,
RBR	adverb, comparative	*faster*	.	sentence-final punc	. ! ?
RBS	adverb, superlative	*fastest*	:	mid-sentence punc	: ; ... – -
RP	particle	*up, off*			

LNMIIT

PoS Tagging

- Example of output:
  - *Book/VB that/DT flight/NN*
    - *Book* is ambiguous — it may be NN or VB
  - *Does/VBZ that/DT flight/NN serve/VB dinner/NN ?/.*
    - *that can be a determiner or complementiser*
      - *Does that flight serve dinner*
      - *I though that it will be rain today*

Part of PoS tagging is to disambiguate

LNMIIT

PoS Tagging

- How hard is tagging problem?
  - Good news: Most words in English are unambiguous — that is, they have only one tag
  - Bad news: many of the common English words are ambiguous!
    - can — auxiliary (‘to be able’), a noun or a verb

LNMIIT

PoS Tagging

- DeRose (1988) reports that
  - 11.5% words in *Brown corpus* are ambiguous
  - 40% of *Brown tokens* are ambiguous
  - Fortunately many of the 40% ambiguous tokens are easy to disambiguate
  - All choices are not equally likely — *a* will be mostly a determiner than being part of an acronym or an initial

LNMIIT

PoS Tagging

- Usage of the word *back*
  - *The back door* = JJ
  - *On my back* = NN
  - *Win the voters back* = RB
  - *Promised to back the bill* = VB

LNMIIT

PoS Tagging Algorithmic Approches

- Rule-based tagger — EngCG
    - it is based on the Constraint Grammar architecture of Karlsson et al (1995)
- HMM PoS Tagging

LNMIIT

Word Sense

- **Word Senses:** Meaning of a lemma varies with respect to the context. For example:
  - *Instead, a bank can hold the investments in a custodial account in the client's name — sense 1*
  - But as agriculture burgeons on the east bank, the river will shrink even more — sense 2
  - *While some banks give blood only to the needy as a service, others may do it as a business — sense 3*
  - *The bank is on the corner of Nassau and Witherspoon — sense 4*

Sense1 and Sense 2: Hononyms, Homonymy

Sense1 and Sense 3: Polysemy

Sense 4: Metonymy
Example: I really love Jufrasky

LNMIIT

Word Sense

- More example on the verb *Serve*:
  - *They rarely serve red meat, preferring to prepare seafood, poultry or game birds.*
  - *He served as U.S. ambassador to Norway in 1976 and 1977.*
  - *He might have served his time, come out and led an upstanding life.*
- Three senses of *Serve*

LNMIIT

Word Sense

- For determining if two senses are distinct is to conjoin two uses of a word in a single sentence; this kind of conjunction of antagonistic readings is called ***zeugma***. Consider the following examples:
  - Which of those flights serve breakfast?
  - Does Midwest Express serve Philadelphia?
  - Does Midwest Express serve breakfast and Philadelphia?
  - Does Midwest Express serve breakfast and lunch?

right	*adj*. located nearer the right hand esp. being on the right when facing the same direction as the observer.
left	*adj*. located nearer to this side of the body than the right.
red	*n*. the color of blood or a ruby.
blood	*n*. the red liquid that circulates in the heart, arteries and veins of animals.

LNMIIT

Word Sense

American Heritage Dictionary

right	*adj*. located nearer the right hand esp. being on the right when facing the same direction as the observer.
left	*adj*. located nearer to this side of the body than the right.
red	*n*. the color of blood or a ruby.
blood	*n*. the red liquid that circulates in the heart, arteries and veins of animals.

Definitions are circular in nature!

LNMIIT

Word Sense

- Word Sense relations are embodied in on-line databases like **WordNet**

LNMIIT

Relations Between Senses

- Synonymy:
  - meaning of two senses of two different words (lemmas) are identical or nearly identical we call them as Synonyms
    - Example: couch/sofa, vomit/throw up, car/automobile
  - More formal definition: Two words are synonymous if they are substitutable one for the other in any sentence without changing the truth conditions of the sentence.

LNMIIT

Relations Between Senses

- Two words may be synonymous but still they may not have an identical meaning:
  - I came home by automobile / car
  - I am thirsty give me H2O / water
- In practice the word synonym is therefore commonly used to describe a relationship of approximate or rough synonymy

LNMIIT

Relations Between Senses

- Consider the following ATIS sentences, since we could swap big and large in either sentence and retain the same meaning:
  - How big is that plane?
  - Would I be flying on a large or small plane?
- But consider the following WSJ sentence where we cannot substitute large for big:
  - Miss Nelson, for instance, became a kind of big sister to Benjamin
  - Miss Nelson, for instance, became a kind of large sister to Benjamin

LNMIIT

Relations Between Senses

- Antonyms are words with opposite meanings
    - long / short, big / little, fast / slow, cold / hot, dark / light, rise / fall, up / down, in / out
- Two senses can be antonyms if they define a binary opposition, or are at opposite ends of some scale. This is the case for long/short, fast/slow, or big/little, which are at opposite ends of the length or size scale.
    - Another groups of antonyms is reversives which describe some sort of change or movement in opposite directions such as rise/fall or up/down.
- From one perspective, antonyms have very different meanings, since they are opposite.
- From another perspective, they have very similar meanings, since they share almost all aspects of their meaning except their position on a scale, or their direction

LNMIIT

Relations Between Senses

- Hyponym: If the first sense is more specific than the second sense
    - For example: car is a hyponym of vehicle; dog is a hyponym of animal, and mango is a hyponym of fruit
- Hypernym: We say that vehicle is a hypernym of car, and animal is a hypernym of dog. The word superordinate is often used instead of hypernym
- Class denoted by the superordinate extensionally includes the class denoted by the hyponym
- Hypernymy can also be defined in terms of entailment. Under this definition, a sense A is a hyponym of a sense B if everything that is A is also B

LNMIIT

Relations Between Senses

- The term ontology usually refers to a set of distinct objects resulting from an analysis of a domain, or microworld.
- A taxonomy is a particular arrangement of the elements of an ontology into a tree-like class inclusion structure.

LNMIIT

Relations Between Senses

- Meronymy: the part-whole relation
  - For Example: A leg is part of a chair; a wheel is part of a car
  - We say that wheel is a meronym of car, and car is a holonym of wheel
- Semantic field is a model of a more integrated, or holistic, relationship among entire sets of words from a single domain.
  - Consider the following set of words: reservation, flight, travel, buy, price, cost, fare, rates, meal, plane
  - FrameNet project (Baker et al., 1998),

LNMIIT

WordNet

- The most commonly used resource for English sense relations is the WordNet lexical database (Fellbaum, 1998)
- WordNet consists of three separate databases, one each for nouns and verbs, and a third for adjectives and adverbs
- In WordNet closed class words are not included
- Each database consists of a set of lemmas, each one annotated with a set of senses
- The WordNet 3.0 release has 117,097 nouns, 11,488 verbs, 22,141 adjectives, and 4,601 adverbs
- The average noun has 1.23 senses, and the average verb has 2.16 senses
- WordNet can be accessed via the web or downloaded and accessed locally.

LNMIIT

WordNet

The noun "bass" has 8 senses in WordNet.
1. bass[1] - (the lowest part of the musical range)
2. bass[2], bass part[1] - (the lowest part in polyphonic music)
3. bass[3], basso[1] - (an adult male singer with the lowest voice)
4. sea bass[1], bass[4] - (the lean flesh of a saltwater fish of the family Serranidae)
5. freshwater bass[1], bass[5] - (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
6. bass[6], bass voice[1], basso[2] - (the lowest adult male singing voice)
7. bass[7] - (the member with the lowest range of a family of musical instruments)
8. bass[8] - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

The adjective "bass" has 1 sense in WordNet.
1. bass[1], deep[6] - (having or denoting a low vocal or instrumental range)
"a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"

- There are eight senses for the noun and one for the adjective, each of which has a gloss (a dictionary-style definition), a list of synonyms for the sense (called a synset), and sometimes also usage examples (shown for the adjective sense)
- WordNet does not have pronunciation of the word like a dictionary

LNMIIT

WordNet

Relation	Also called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	*breakfast*^{1} → *meal*^{1}
Hyponym	Subordinate	From concepts to subtypes	*meal*^{1} → *lunch*^{1}
Member Meronym	Has-Member	From groups to their members	*faculty*^{2} → *professor*^{1}
Has-Instance		From concepts to instances of the concept	*composer*^{1} → *Bach*^{1}
Instance		From instances to their concepts	*Austen*^{1} → *author*^{1}
Member Holonym	Member-Of	From members to their groups	*copilot*^{1} → *crew*^{1}
Part Meronym	Has-Part	From wholes to parts	*table*^{2} → *leg*^{3}
Part Holonym	Part-Of	From parts to wholes	*course*^{7} → *meal*^{1}
Antonym		Opposites	*leader*^{1} → *follower*^{1}

Noun Relations

LNMIIT

WordNet

Relation	Definition	Example
Hypernym	From events to superordinate events	$fly^9 \rightarrow travel^5$
Troponym	From a verb (event) to a specific manner elaboration of that verb	$walk^1 \rightarrow stroll^1$
Entails	From verbs (events) to the verbs (events) they entail	$snore^1 \rightarrow sleep^1$
Antonym	Opposites	$increase^1 \Longleftrightarrow decrease^1$

Verb Relations

LNMIIT

Word Sense Disambiguation

- Two variants of the generic WSD task:
  - In the lexical sample task: A small pre-selected set of target words is chosen, along with an inventory of senses for each word from some lexicon. For each word, a number of corpus instances (context sentences) can be selected and hand-labeled with the correct sense of the target word in each. Classifier systems can then be trained using these labeled examples. Unlabeled target words in context can then be labeled using such a trained classifier.
  - Early work in word sense disambiguation focused solely on lexical sample tasks of this sort, building word-specific algorithms for disambiguating single words like line, interest, or plant.
  - In the all-words task: Systems are given entire texts and a lexicon with an inventory of senses for each entry, and are required to disambiguate every content word in the text. The all-words task is very similar to part-of-speech tagging, except with a much larger set of tags, since each lemma has its own set.