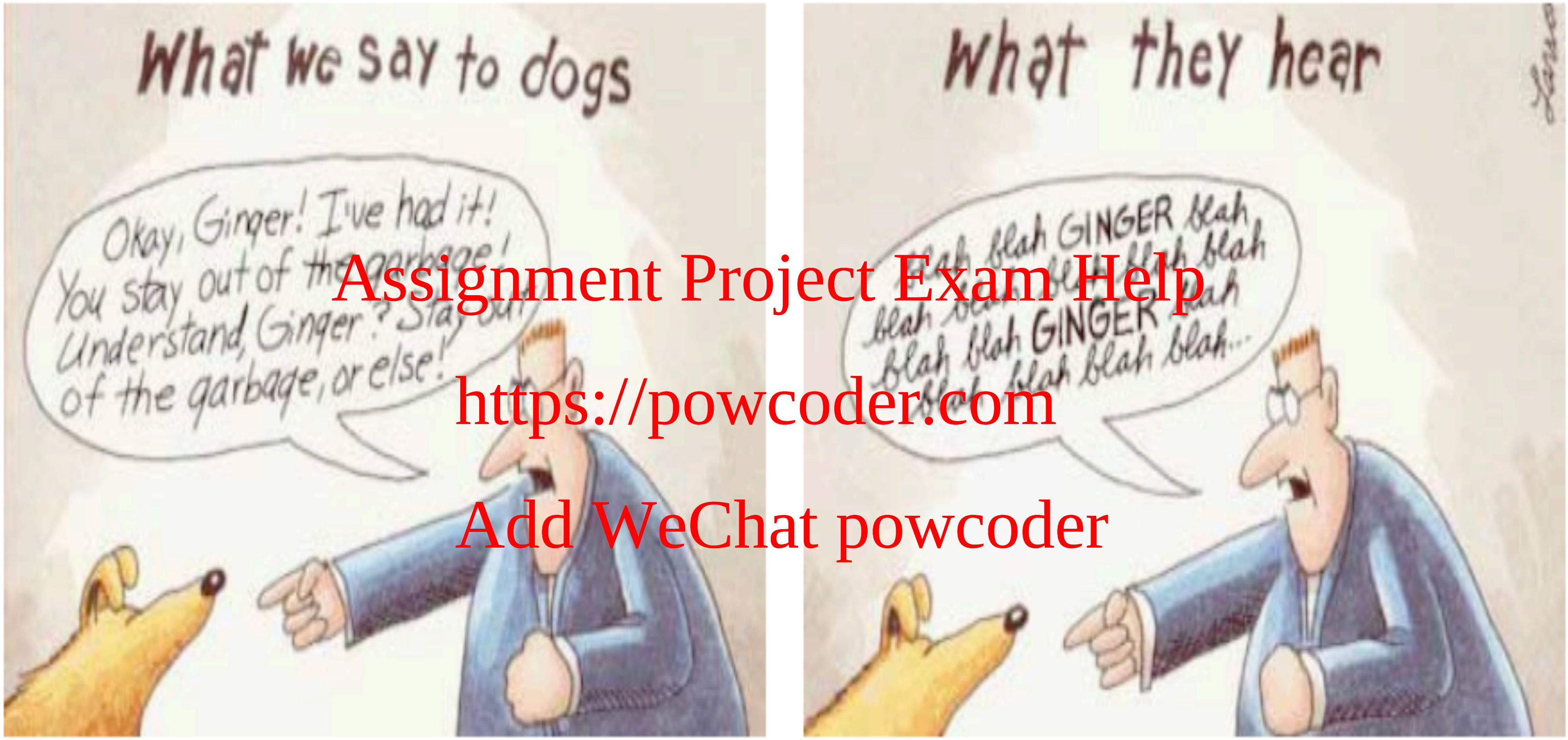
Natural Language Processing Essentials

Lizhen Qu

What is Natural Language Processing?

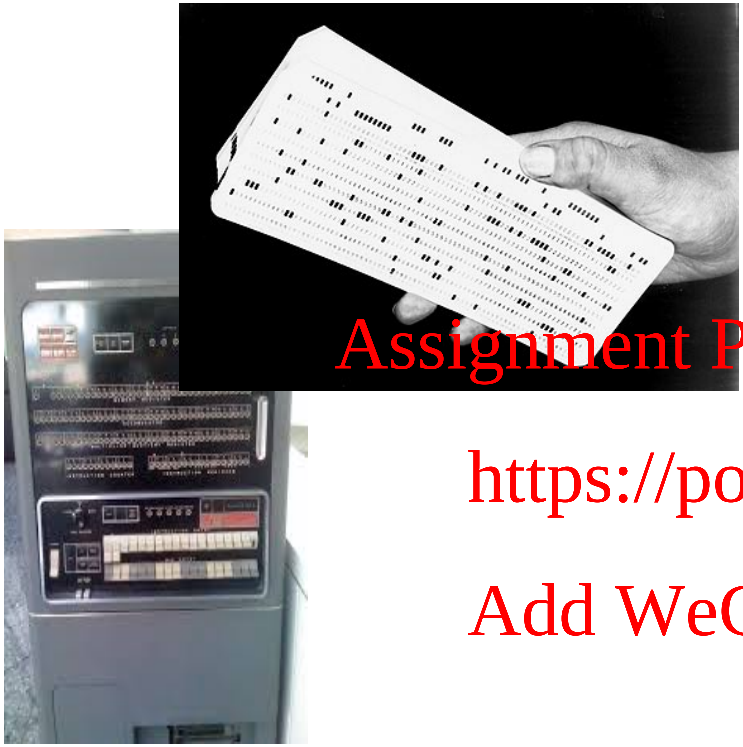

What we say to dogs
Okay, Ginger! I've had it! You stay out of the garbage! Understand, Ginger? Stay out of the garbage, or else!
What they hear
blah blah GINGER blah blah blah blah blah blah blah blah GINGER blah blah blah blah blah...

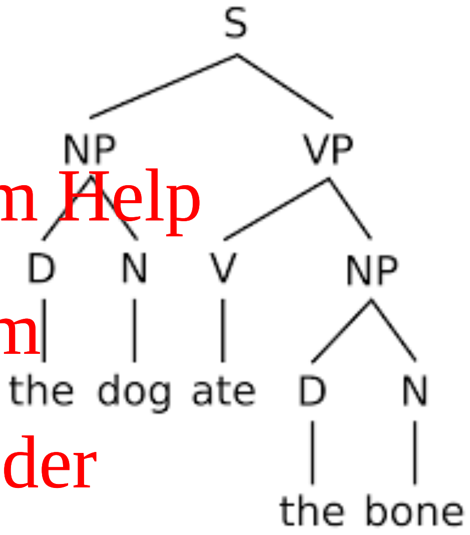
Solving engineering problems that need to analyze or generate natural language text.

Applications of NLP

- Speech recognition.
- Question answering.
- Machine Translation.
- Spelling correction, grammar checking.
- Information extraction.
- Summarization.
- Dialogue systems.

History of Computational Linguistics





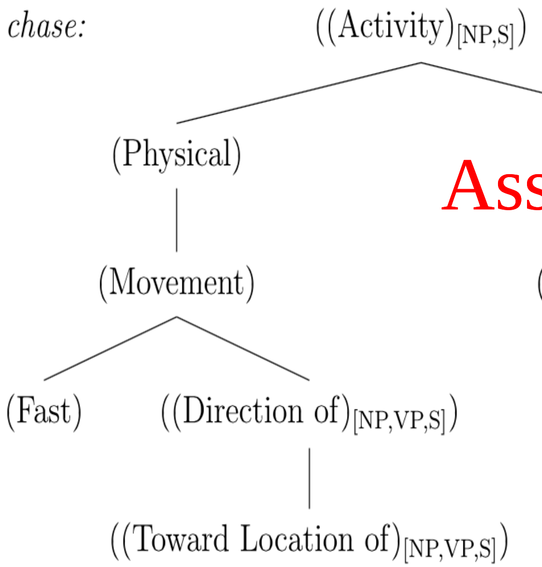
1950s

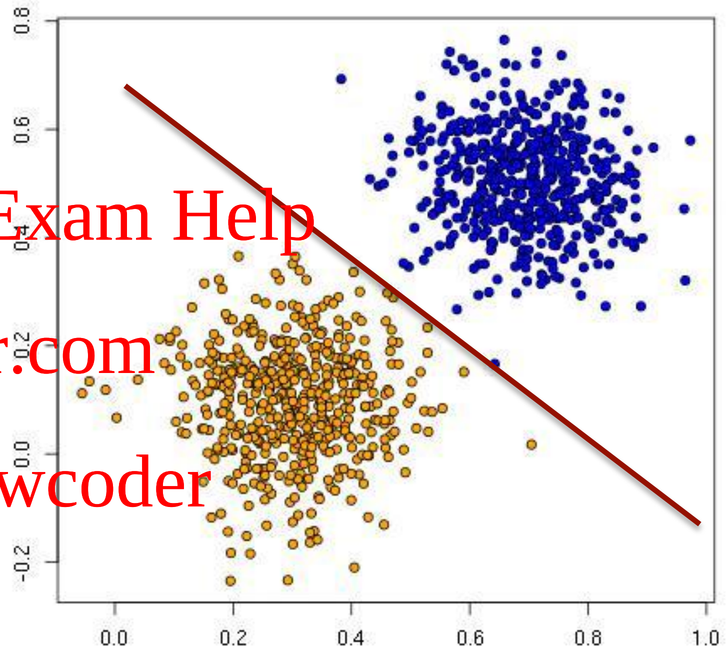
Machine Translation

1960s

Syntax and QA

History of Computational Linguistics





1970s – 1980s

Semantics

1990s – present

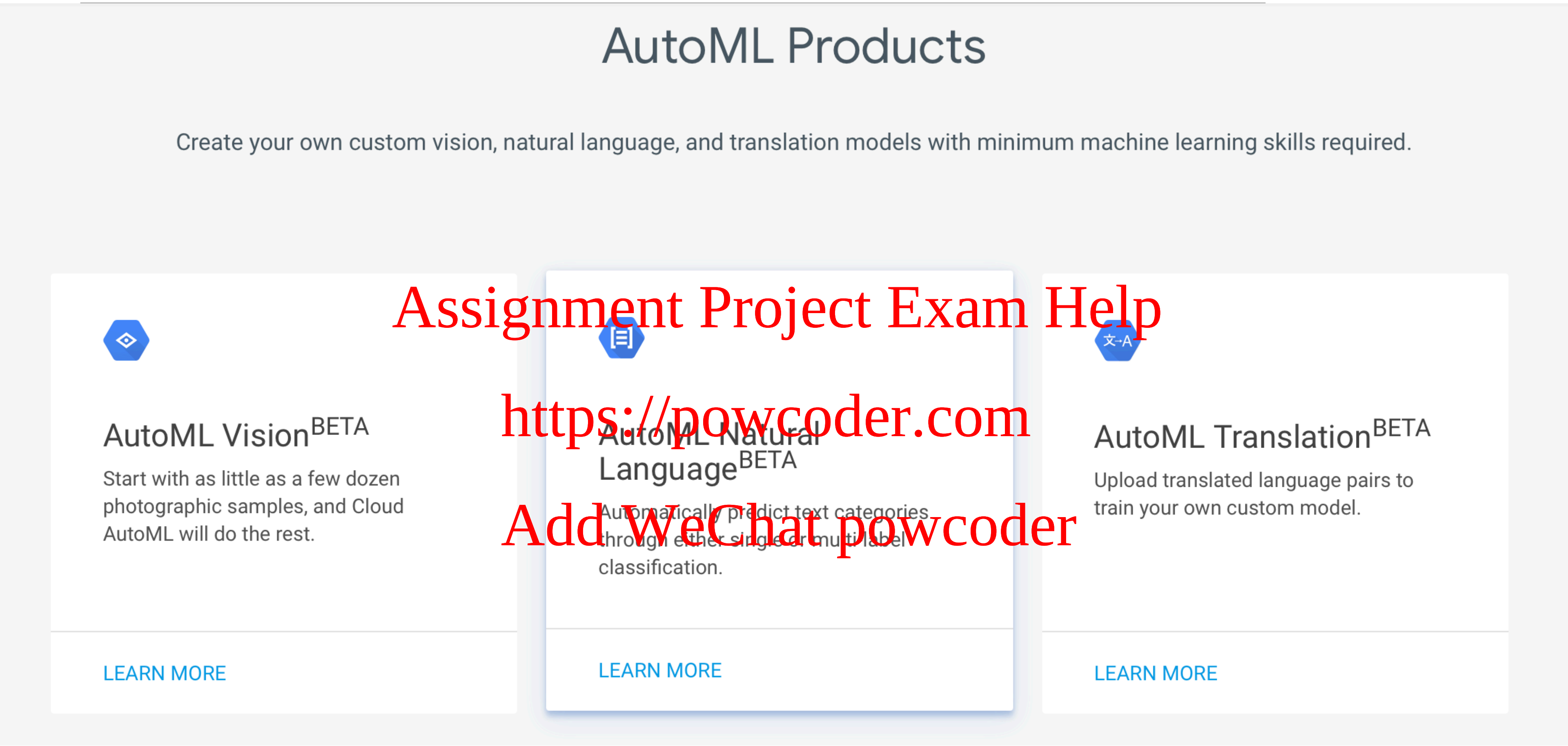
Statistical NLP

Deep Learning

Google Assistant

https://www.youtube.com/watch?v=JvbHu_bVa_g

Google AutoML



AutoML Products

Create your own custom vision, natural language, and translation models with minimum machine learning skills required.

AutoML Vision BETA

Start with as little as a few dozen photographic samples, and Cloud AutoML will do the rest.

AutoML Natural Language BETA

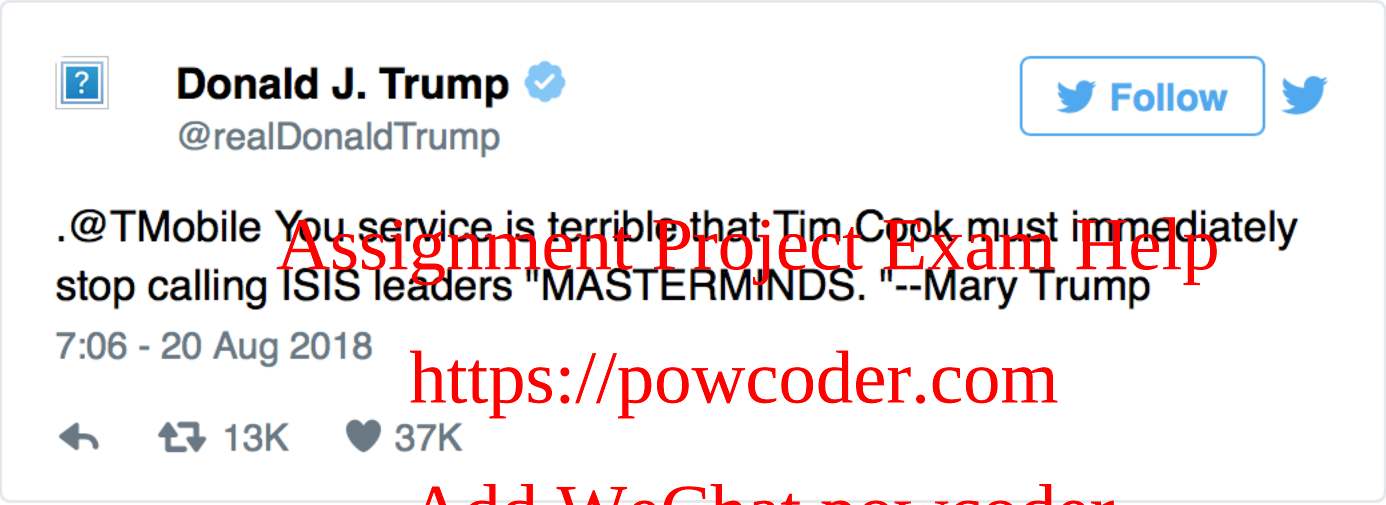
Automatically predict text categories through either single or multi-label classification.

AutoML Translation BETA

Upload translated language pairs to train your own custom model.

https://cloud.google.com/automl/

Trump Speech



Donald J. Trump
@realDonaldTrump

.@TMobile You service is terrible that Tim Cook must immediately stop calling ISIS leaders "MASTERMINDS. "--Mary Trump

7:06 - 20 Aug 2018

https://filiph.github.io/markov/

Obama’s Speech

Good afternoon. God bless you.

The United States will step up to the cost of a new challenges of the American people that will share the fact that we created the problem. They were attacked and so that they have to say that all the task of the final days of war that I will not be able to get this done. The promise of the men and women who were still going to take out the fact that the American people have fought to make sure that they have to be able to protect our part. It was a chance to stand together to completely look for the commitment to borrow from the American people. And the fact is the men and women in uniform and the millions of our country with the law system that we should be a strong stretcks of the forces that we can afford to increase our spirit of the American people and the leadership of our country who are on the Internet of American lives.

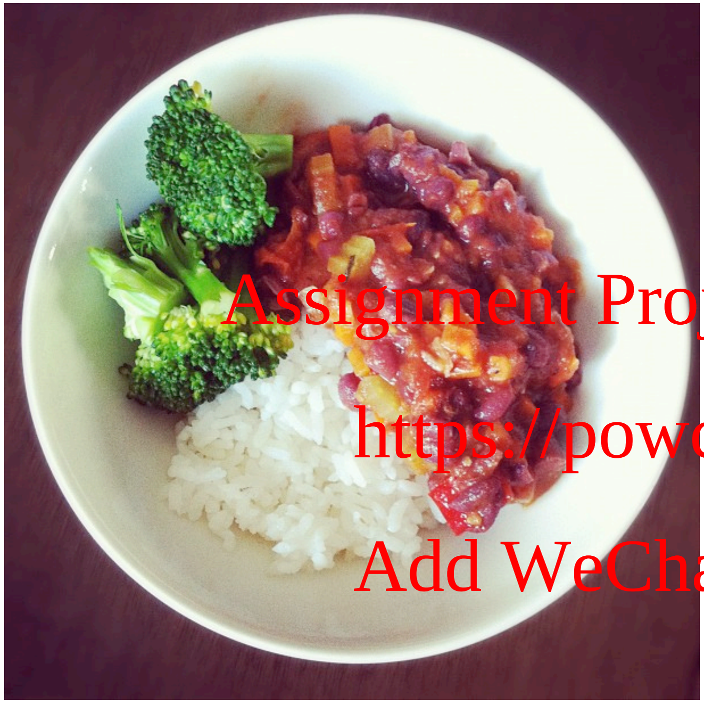
Thank you very much. God bless you, and God bless the United States of America.

Automatically Generate Text

- Machine generated talks.
  - https://www.youtube.com/watch?v=-OodHtJ1saY

- Machine generated text.
  - http://www.cs.toronto.edu/~ilya/fourth.cgi
  - https://medium.com/@samim/obama-rnn-machine-generated-political-speeches-c8abd18a2ea0#.56709ikih
  - https://www.youtube.com/watch?v=EFHyzuqjaok
  - https://filiph.github.io/markov/

Natural Language Grounding



What is the food?
Broccoli, rice, beans etc.
What color is the plate?
White
What is the green on the plate?
Broccoli
Where is the broccoli?
on the plate

Broccoli is on the white plate.
Red sauce is on the plate.
The plate is white.

Screenplay: https://www.youtube.com/watch?v=LY7x2Ihqjmc

Levels of Language

Three years ago, Christina Wu was sitting at the Ivy Café in Canberra studying for her ANU exams. She tweeted about it, wondering if any of her friends were studying nearby.

REQUEST: open the door.
STATEMENT: the door is open.
INFORMATION QUESTION: is the door open?

Levels of Language

- Phonetics and phonology – knowledge about linguistic sounds.
- Morphology – knowledge of the meaningful components of words.
- Syntax – knowledge of the structural relationships between words.
- Semantics – knowledge of meaning.
- Discourse – knowledge about linguistic units larger than a single utterance.
- Pragmatics – knowledge of the relationship of meaning to the goals and intentions of the speaker.

Ambiguity of Natural Language

I made her duck.

i. I cooked waterfowl for her.
ii. I cooked waterfowl belonging to her.
iii. I created the duck she owns.
iv. I caused her to quickly lower her head or body.
v. I waved my magic wand and turned her into a waterfowl.

Overview of the NLP Lectures

- Introduction to natural language processing (NLP).
- Regular expressions, sentence splitting, tokenization, part-of-speech tagging.
- Language models.
- Vector semantics.
- Parsing.
- Compositional semantics.

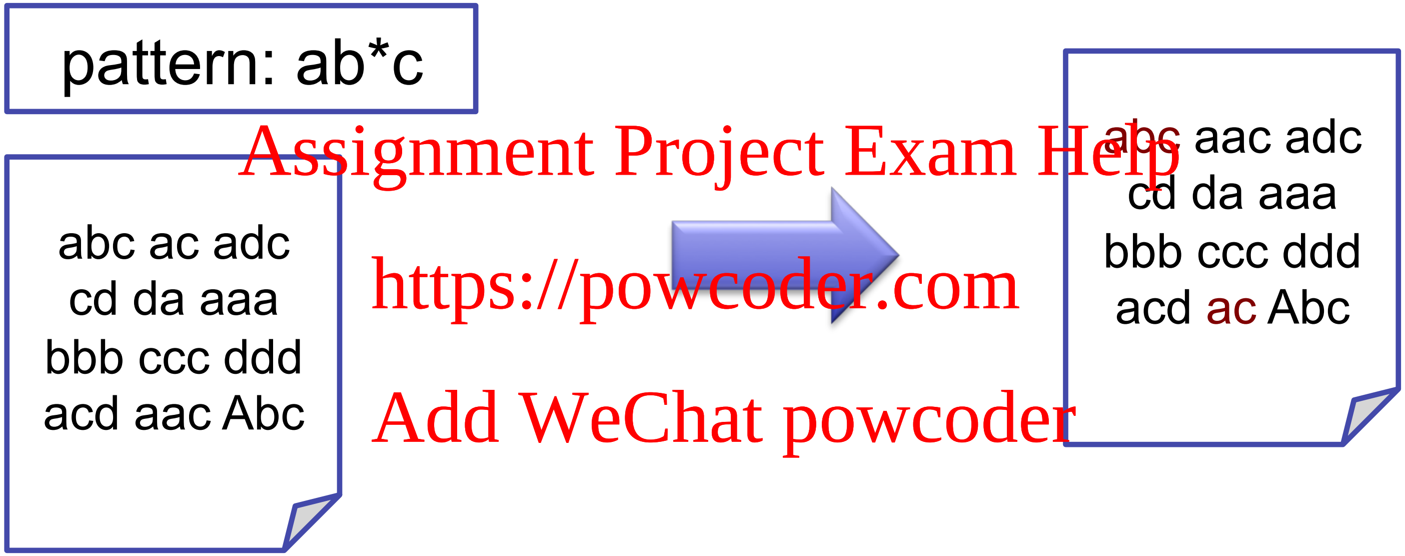
How to Find All Prices?

In January 2014, IBM announced plans to invest more than $1.2bn (£735m) into its data centers and cloud storage business. It plans to build 15 new centers around the world, bringing the total number up to 40 during 2014.[82]

In July 2014, the company revealed it was investing $3 billion over the following five years to create computer functionality to resemble how the human brain thinks. A spokesman said that basic computer architecture had not altered since the 1940s. IBM says its goal is to design a neural chip that mimics the human brain, with 10 billion neurons and 100 trillion synapses, but that uses just 1 kilowatt of power.

Regular Expressions (RE)

- Definition: an regular expression is an algebraic notation for characterizing a set of strings.



pattern: ab*c

abc ac adc
cd da aaa
bbb ccc ddd
acd aac Abc

abc aac adc
cd da aaa
bbb ccc ddd
acd ac Abc

- RE Tools.
  - Perl, Java, Python etc.
  - grep, awk, vi, Emacs, Word etc.

RE Operators

- A single character: a, b, c …
- A sequence of characters: country, fun …
- []: disjunction of characters. E.g. [Cc]ountry .
- [-]: any one character in a range. E.g. [A-Z], [0-9] .
- *: zero or more occurrences. E.g. ab*c .
- +: one or more occurrences. E.g. [0-9]+ .
- X{n}: X occurs exactly n times.
- X{n,m}: X occurs at least n times but no more than m times.
- | : disjunction. E.g. parrot|bird .
- () : capturing group.

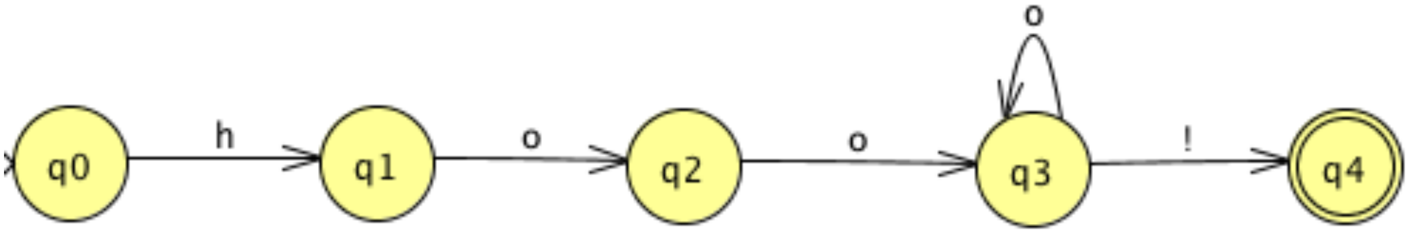
RE Demo

- Demo tool: http://java-regex-tester.appspot.com/

> In January 2014, IBM announced plans to invest more than $1.2bn (£735m) into its data centers and cloud storage business. It plans to build 15 new centers around the world, bringing the total number up to 40 during 2014.[82]
>
> In July 2014, the company revealed it was investing $3 billion over the following five years to create computer functionality to resemble how the human brain thinks. A spokesman said that basic computer architecture had not altered since the 1940s. IBM says its goal is to design a neural chip that mimics the human brain, with 10 billion neurons and 100 trillion synapses, but that uses just 1 kilowatt of power.

- Search pattern?

Alternative Representations

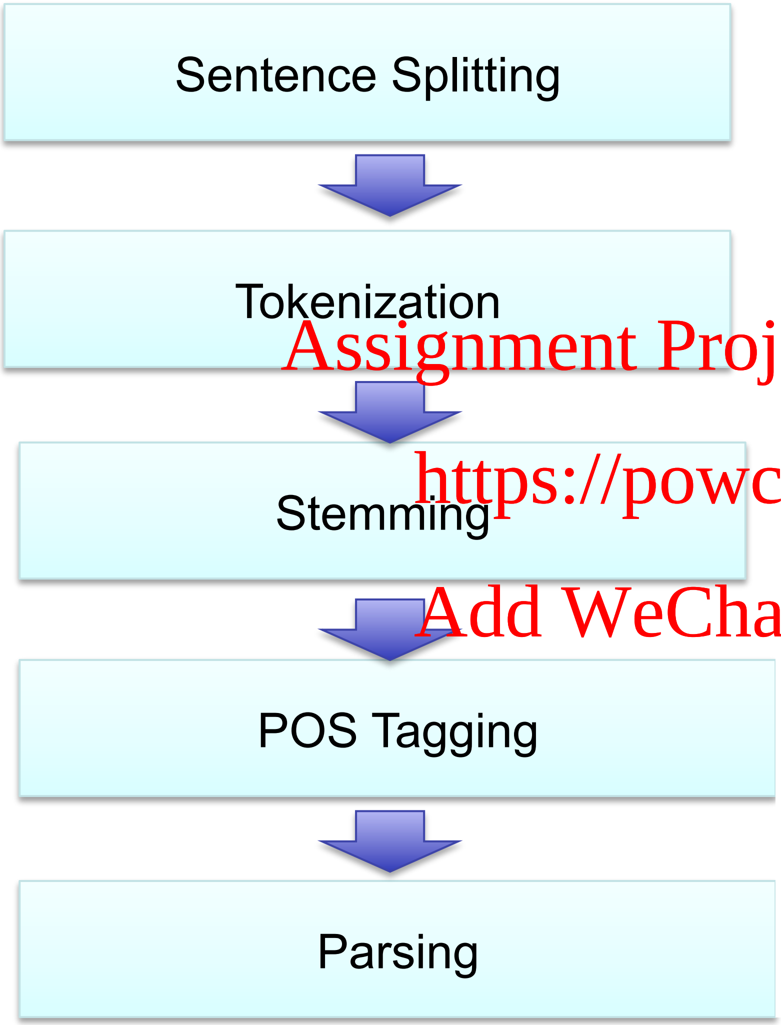


- Regular expressions: hoo+!

- State-transition table:

	Input		
State	h	o	!
q0	q1	∅	∅
q1	∅	q2	∅
q2	∅	q3	∅
q3	∅	q3	q4
q4	∅	∅	∅

A Popular NLP Pipeline



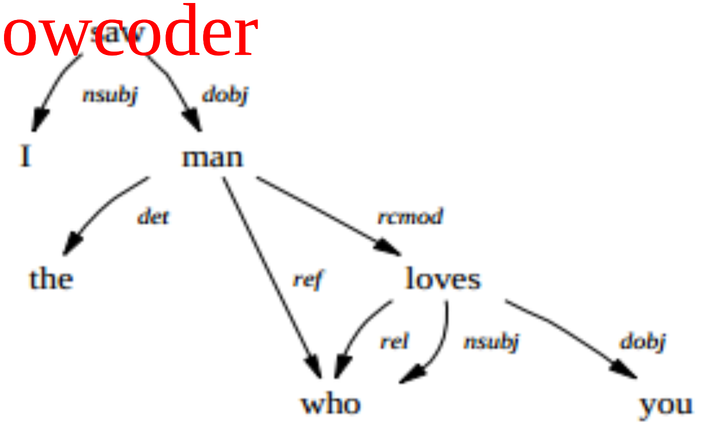
Sentence Splitting

Tokenization

Stemming

POS Tagging

Parsing



Sentence Segmentation

- Problems with sentence splitting.
  - “You reminded me,” she remarked, “of your mother.”

- Sentence boundary detection algorithms.
  - Regular expressions.
  - Rule-based approaches.
  - Machine learning approaches.

- Tools.
  - OpenNLP.
  - Stanford CoreNLP.
  - NLTK.

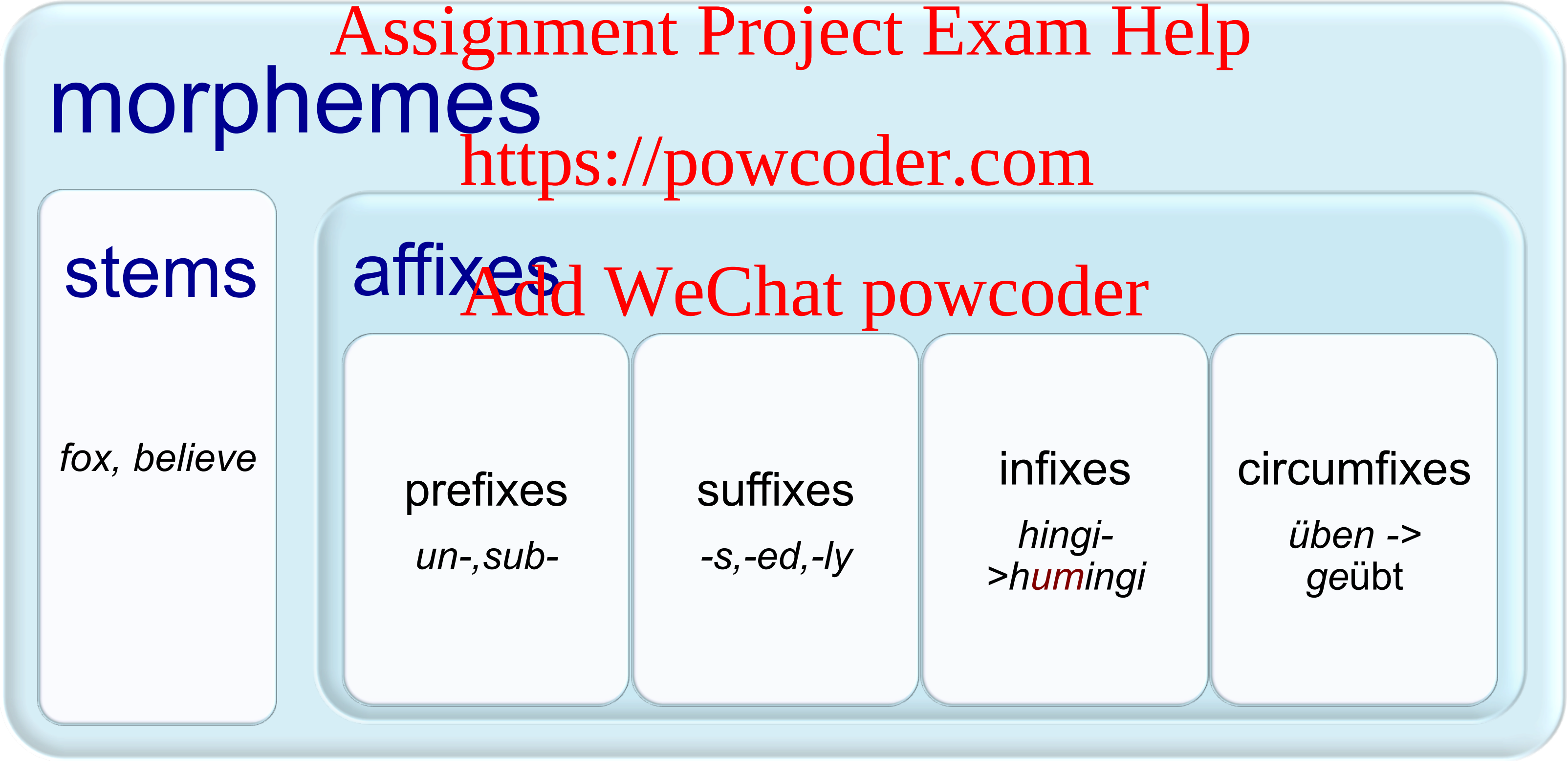
Tokenization

- Divide text into words, numbers, punctuations.

- English
  - Regular expressions?
    E.g. [A-Za-z0-9]+|{\Punct}+
  - Problems.
    - Periods: E.g. Ph.D.., google.com
    - Clitic: E.g. isn't => *is* + *n't (not)*
    - Hyphenation.
      - *co-operate.*
      - *most-visited => most visited.*
  - Tools.
    - OpenNLP.
    - Stanford CoreNLP.
    - NLTK.

Morphology

- Morphology is the study of the way words are built up from smaller meaning-bearing units.
  - Morpheme: minimal meaning-bearing unit in a language.



morphemes

stems

fox, believe

affixes

prefixes

un-,sub-

suffixes

-s,-ed,-ly

infixes

hingi->humingi

circumfixes

üben -> geübt

Combining Morphemes to Create Words

- Inflection.
  - combination of a word stem with a grammatical morpheme.
  - same word class, e.g. *mouse -> mice, chidren's, walks, walked, ate, cut*.
- Derivation.

  - combination of a word stem with a grammatical morpheme.

  - different word class, e.g. appoint -> appointment, *computation -> computational*.

- Compounding.
  - combination of multiple word stems. E.g. *bedroom*.
- Cliticization.
  - combination of a word stem with a clitic. E.g. *I've = I + have*.
  - clitic: a morpheme that acts syntactically like a word, but is reduced in form and attached to another word.

Stemming (Review)

- A crude heuristic process that strips off suffixes.
- Algorithms.
  - Lookup algorithms.
  - Regular expressions.
  - Suffix-stripping algorithms.
    - Porter stemmer.
      http://tartarus.org/martin/PorterStemmer
    - Lovins stemmer.
      http://snowball.tartarus.org/algorithms/lovins/stemmer.html
    - Lancaster stemmer.
      http://www.comp.lancs.ac.uk/computing/research/stemming/
- Tools.
  - http://snowball.tartarus.org
  - NLTK

Porter stemmer.

- Lexicon free stemmer.
- Rewrite rules.
  - ATIONAL → ATE (e.g. *relational, relate*)
  - FUL → ε (e.g. *hopeful, hope*)
  - SSES→ SS (e.g. *caresses, caress*)
- Errors of Commission.
  - *Organization* → *organ*
  - *Policy* → *polici*
- Errors of Omission.
  - *urgency* → *urgenci* (not stemmed to *urgent*)
  - *European*→ *European* (not stemmed to *Europe*)

English Parts of Speech

- Noun (person, place or thing)
  - Singular (NN): dog, fork
  - Plural (NNS): dogs, forks
  - Proper (NNP, NNPS): John, Springfields
  - Personal pronoun (PRP): I, you, he, she, it
  - Wh-pronoun (WP): who, what
- Verb (actions and processes)
  - Base, infinitive (VB): eat
  - Past tense (VBD): ate
  - Gerund (VBG): eating
  - Past participle (VBN): eaten
  - Non 3rd person singular present tense (VBP): eat
  - 3rd person singular present tense: (VBZ): eats
  - Modal (MD): should, can
  - To (TO): to (to eat)

English Parts of Speech (cont.)

- Adjective (modify nouns)
  - Basic (JJ): red, tall
  - Comparative (JJR): redder, taller
  - Superlative (JJS): reddest, tallest
- Adverb (modify verbs)
  - Basic (RB): quickly
  - Comparative (RBR): quicker
  - Superlative (RBS): quickest
- Preposition (IN): on, in, by, to, with
- Determiner:
  - Basic (DT) a, an, the
  - WH-determiner (WDT): which, that
- Coordinating Conjunction (CC): and, but, or…
- Particle (RP): off (took off), up (put up)

Closed vs. Open Class

- ***Closed class*** categories are composed of a small, fixed set of grammatical function words for a given language.
  - Pronouns, Prepositions, Modals, Determiners, Particles, Conjunctions.

- **Open class** categories have large number of words and new ones are easily invented.
  - Nouns (Googler), Verbs (google), Adjectives (geeky), Abverb (chompingly)

Part of Speech Tagging

John	saw	the	saw	and	decided	to	take	it	to	the	table.
NNP	VBD	DT	NN	CC	VBD	TO	VB	PRP	IN	DT	NN

English POS Tagsets

- Original Brown corpus used a large set of 87 POS tags.

- Most common in NLP today is the Penn Treebank set of 45 tags.

- Universal POS tags.
  - http://universaldependencies.org/u/pos/
  - https://arxiv.org/pdf/1104.2086.pdf

Ambiguity in POS Tagging

- "Like" can be a verb or a preposition.
  - I like/VBP candy.
  - Time flies like/IN an arrow.

- "Around" can be a preposition, particle, or adverb.
  - I bought it at the shop around/IN the corner.
  - I never got around/RP to getting a car.
  - A new Prius costs around/RB $25K.

Clues for POS Tagging

- Morphological clues.
  - -ment: government, establishment.
- Syntactic clues.
  - the *near* window.
  - The end is very *near*.
- POS in local context.
  - DET ? NN PUNC
  - DET NN BE ADV ?

POS tagger with LR: http://nlp.stanford.edu/~manning/papers/emnlp2000.pdf
Practice with NLTK: http://www.nltk.org/book/ch05.html

POS Tagger for Download

- Stanford POS tagger.
  - http://nlp.stanford.edu/software/corenlp.shtml
  - http://nlp.stanford.edu/software/tagger.shtml
- Open NLP.

- UIUC POS tagger.
  - http://cogcomp.cs.illinois.edu/demo/pos/results.php

- NLTK.

  - http://text-processing.com/demo/tag
- Twitter POS tagger.
  - http://www.ark.cs.cmu.edu/TweetNLP/
  - http://www.ark.cs.cmu.edu/TweetNLP/cluster_viewer.html

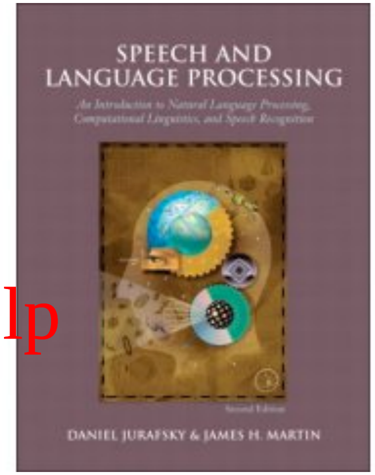
Textbooks

SPEECH and LANGUAGE PROCESSING

An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition

Daniel Jurafsky and James H. Martin

https://web.stanford.edu/~jurafsky/slp3/


SPEECH AND LANGUAGE PROCESSING
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
Second Edition
DANIEL JURAFSKY & JAMES H. MARTIN

Recommend to read at least 4.1 https://lagunita.stanford.edu/c4x/Engineering/CS-224N/asset/slp4.pdf before the next class.