Regular Expressions

25 January 2022
Worksheet due now

(Optional) Assignment 0 due 10 p.m. on Wednesday.
Background
Our textbook is called *Speech and Language Processing*, but this class will be focused on text.

Why might we do that?
cor-pus |ˈkɔrpəs|

noun (plural corpora |ˈkɔrpəra| or corpuses)

1 a collection of written texts, especially the entire works of a particular author or a body of writing on a particular subject: the Darwinian corpus.
   a collection of written or spoken material in machine-readable form, assembled for the purpose of studying linguistic structures, frequencies, etc.

2 Anatomy the main body or mass of a structure.
   the central part of the stomach, between the fundus and the antrum.

ORIGIN
late Middle English (denoting a human or animal body): from Latin, literally ‘body’. corpus (sense 1) dates from the early 18th century.
What *isn’t* a text corpus?
Fundamental idea: Search
LITERARY CHARACTER APB

$400:

His victims include Charity Burbage, Mad Eye Moody & Severus Snape; he’d be easier to catch if you’d just name him!
“And anytime you feel the pain, hey” this guy “refrain, don’t carry the world upon your shoulders”
OLYMPIC ODDITIES

$800:
In the 2004 opening ceremonies, a sole member of this team opened the Parade of Nations; the rest of his team closed it.
Some of these questions are really hard for people because you need to know so much trivia, and our meat minds are bad at that.

But computers can store a lot; Watson had 21.6 TB of storage, back in 2011.
LINUS’S LAW: “Given enough eyeballs, all bugs are shallow.”
“Given enough text, all questions are easy.”
Searching for information
“[Regular expressions] are particularly useful for searching in texts, when we have a pattern to search for and a corpus of texts to search through. A regular expression search function will search through the corpus, returning all texts that match the pattern.”

Jurafsky & Martin, § 2.1
Consider learning when people were born and when they died.
Consider learning when people were born and when they died.

What do you search for?
Let’s try some of these out!

We’ll look at an old snapshot of English Wikipedia, and instead of programming, we’ll use a standard Unix tool: **grep**.
“… the Unix command-line tool grep takes a regular expression and returns every line of the input document that matches the expression.”

Jurafsky & Martin, § 2.1
grep lets you search files for text

$ grep bananas foo.txt

Here are some of my favourite grep command line arguments!

-E aka egrep
- Recursive! Search all the files in a directory.

-v invert match: find all lines that don't match
- only print the matching part of the line (not the whole line)

-i case insensitive

-a Show context for your search:
  $ grep -A 3 foo will show 3 lines of context after a match
- Search binaries: treat binary data like it's text instead of ignoring it!

-A, -B

-F aka fgrep
- don't treat the match string as a regex
  eg $ grep -F ...

grep alternatives
  awk, ag, ripgrep (better for searching code!)
born in \[0-9]\{4\}

Don’t want to match places or other descriptions, e.g.,

born in New York
born in poverty

died in \[0-9]\{4\}
(died|passed away) in [0–9]{4}
born in ([0-9]{4}|[0-9]+ (AD|BC))
born in ([0-9]{4}|[0-9]+ (AD|BC|CE|BCE))
born on the 8th of May, 1885
Another pattern:

```
born on .+ [0-9]{4}
```
David Bowie

From Wikipedia, the free encyclopedia

For other uses, see David Bowie (disambiguation).

David Robert Jones OAM (8 January 1947 – 10 January 2016), known professionally as David Bowie (/ˈboʊi/ BOH-ee),[1] was an English singer-songwriter and actor. A leading figure in the music industry, Bowie is regarded as one of the most influential musicians of the 20th century. He was acclaimed by critics and musicians, particularly for his innovative work during the 1970s. His career was marked by reinvention and visual presentation, and his music and stagecraft had a significant impact on popular music.

Bowie developed an interest in music as a child. He studied art, music and design before embarking on a professional career as a musician in 1963. "Space Oddity", released in 1969, was his first top-five entry on the UK Singles Chart. After a period of experimentation, he re-emerged in 1972 during the glam rock era with his flamboyant and androgynous alter ego Ziggy Stardust. The character was spearheaded by the success of Bowie's single "Starman" and album The Rise and Fall of Ziggy Stardust and the Spiders from Mars, which won him widespread popularity. In 1975, Bowie's style shifted towards a sound he characterised as "plastic soul", initially alienating
We could keep going!
These kind of searches let us learn lots of information that’s stated in text.

Which companies bought which other companies.
What state is a town in?
Which musicians made which albums?
Adding in syntactic relations, these form the basis for a powerful basic technique called *Hearst patterns*:

“\(NP_y \text{ such as } NP_x\)”

\(\rightarrow x \text{ is a } y\)

cartoon characters such as Snoopy

\(\rightarrow \text{Snoopy is a cartoon character}\)
When we use these patterns to look for information, how do we know how we’re doing?
Eliza and chatbots
There’s a long history of making computers we can talk to, some serious, some not so serious.
Men are all alike.
IN WHAT WAY
They're always bugging us about something or other.
CAN YOU THINK OF A SPECIFIC EXAMPLE
Well, my boyfriend made me come here.
YOUR BOYFRIEND MADE YOU COME HERE
He says I'm depressed much of the time.
I AM SORRY TO HEAR YOU ARE DEPRESSED
It's true, I am unhappy.
DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY
I need some help, that much seems certain.
WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP
Perhaps I could learn to get along with my mother.
TELL ME MORE ABOUT YOUR FAMILY
My mother takes care of me.
WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU
My father.
YOUR FATHER
You are like my father in some ways.
WHAT RESEMBLANCE DO YOU SEE
You are not very aggressive but I think you don't want me to notice that.
WHAT MAKES YOU THINK I AM NOT VERY AGGRESSIVE
You don't argue with me.
WHY DO YOU THINK I DON'T ARGUE WITH YOU
You are afraid of me.
DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU
My father is afraid of everybody.
WHAT ELSE COMES TO MIND WHEN YOU THINK OF YOUR FATHER
Bullies.
DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOYFRIEND MADE YOU COME HERE

“ELIZA is a surprisingly simple program that uses pattern matching to recognize phrases like ‘I need X’ and translate them into suitable outputs like ‘What would it mean to you if you got X?’ This simple technique succeeds in this domain because ELIZA doesn’t actually need to know anything to mimic a Rogerian psychotherapist.”

Jurafsky & Martin, § 2
Just as we can match parts of a sentence when we’re doing information extraction, we can match parts of a user’s input to our chatbot.
A formal notation in which to represent the decomposition template is:

(0 YOU 0 ME)

and the reassembly rule

(WHAT MAKES YOU THINK I 3 YOU).

The “0” in the decomposition rule stands for “an indefinite number of words” (analogous to the indefinite dollar sign of Comit) [6] while the “3” in the reassembly rule indicates that the third component of the subject decomposition is to be inserted in its place.

A formal notation in which to represent the decomposition template is:

(0 YOU 0 ME)

and the reassembly rule

(WHAT MAKES YOU THINK I 3 YOU).

```python
import re

patient = "I hate that you judge me so much."

m = re.match("(.*) (you) (.*) (me)" , patient)

print("What makes you think I " + m.group(3) + " you?")
```
How do we evaluate a chatbot?
What does it mean for a computer to understand language?
MIND
A QUARTERLY REVIEW OF
PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing
“I shall replace the question ['can machines think?'] by another, which is closely related to it and is expressed in relatively unambiguous words.

“The new form of the problem can be described in terms of a game which we call the ‘imitation game’…”
“How long is your hair?”
“This is so 1950 it hurts.”

“How long is your hair?”
Big brain time
Big brain time

“How long is your hair?”
Big brain time

“How long is your hair?”

“It’s a cute pixie cut lol”
Turing suggested talking over teletypes, i.e., texting with the computer.

But we often prefer speech – and high-quality speech recognition and synthesis turns out to be one of the easier parts of making an intelligent machine!
Q: Please write me a sonnet on the subject of the Forth Bridge.
A: Count me out on this one. I never could write poetry.
Q: Add 34957 to 70764
A: (Pause about 30 seconds and then give as answer) 105621.
Q: Do you play chess?
A: Yes.

Q: I have K at my K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What do you play?
A: (After a pause of 15 seconds) R-R8 mate.
Does this match what you think it means to understand language?
Acknowledgments

This class incorporates material from:

Jurafsky & Martin, *Speech and Language Processing*, 3rd ed. draft