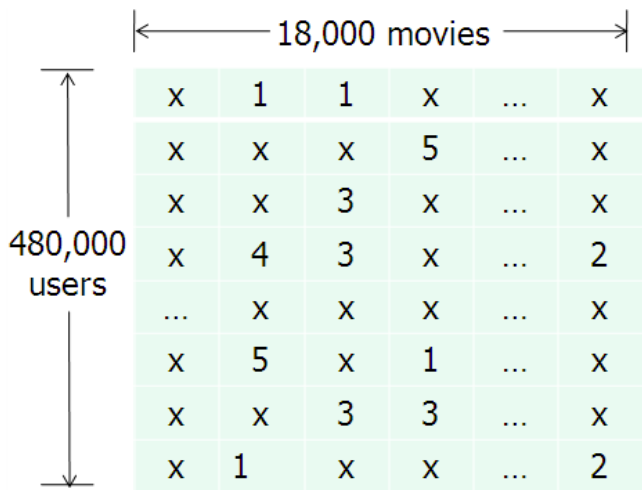
GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

TEXT PROCESSING / NATURAL LANGUAGE PROCESSING

LAST TIME:

I. IMPLICIT VS EXPLICIT DATA
II. CONTENT-BASED FILTERING
III. COLLABORATIVE FILTERING



QUESTIONS?

WHAT WAS THE MOST INTERESTING THING YOU LEARNT?

WHAT WAS THE HARDEST TO GRASP?

I. WHAT IS NATURAL LANGUAGE PROCESSING

II. NLP APPLICATIONS

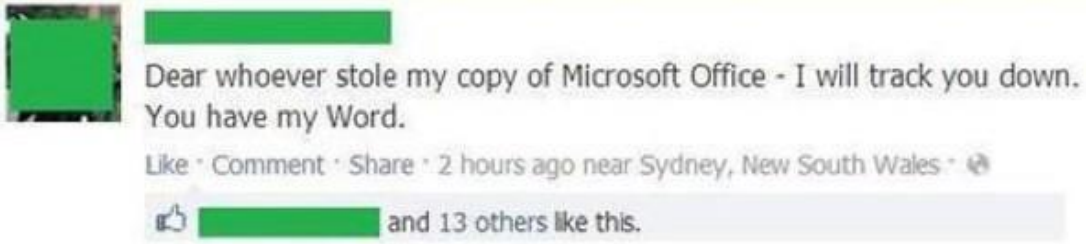
III. BASIC NLP PRACTICE

I. WHAT IS NATURAL LANGUAGE PROCESSING

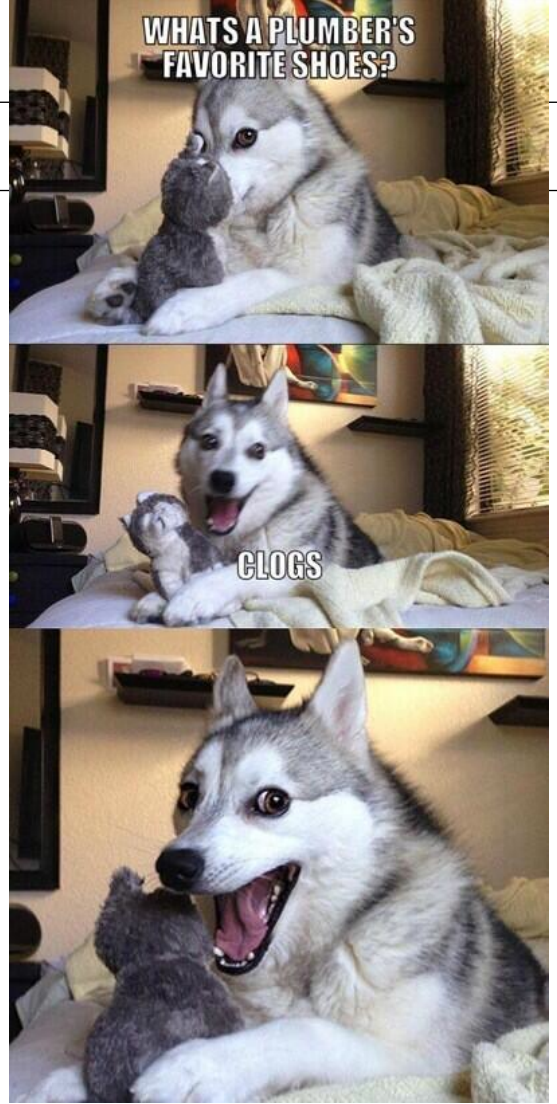
The interface between human and computer language

Consider lexical ambiguity resolution:

The selection of one of multiple possible meanings of a phrase.



Source: @WowSoPunny



Humans are great at this.

Computers are not.

How many times have you yelled after python throws an error,
“You know what I meant!”?

How do we teach computers to understand human language?

How do we discern the meaning of **sea** in the following sentences?

"The driftwood floated in from the **sea**."

"My cousin is dealing with a **sea** of troubles."

Large body of water:
"The driftwood floated in from *the* ***sea***."

Figurative large quantity:
"My cousin is dealing with a ***sea*** *of* troubles."

You can make rules-based models.

But these are fragile.

Language isn’t static.

Semantic models perform very well, but are slow.

Statistical models with the right features can carry us really far.

II. NLP APPLICATIONS

Take two minutes and jot down any common, real-world examples of NLP that you can think of.

Take two minutes and jot down any common, real-world examples of NLP that you can think of.

Find your 2 nearest neighbors and see if you can come up with more.

Take two minutes and jot down any common, real-world examples of NLP that you can think of.

Find your 2 nearest neighbors and see if you can come up with more.

What did you come up with?

Let’s run down some common, well-known tasks.

Speech Recognition

“Hello, HAL. Do you read me? Do you read me, HAL?”
“Affirmative, Dave. I read you.”

The best speech recognition software incorporate language models along with the audio signal.

Machine Translation

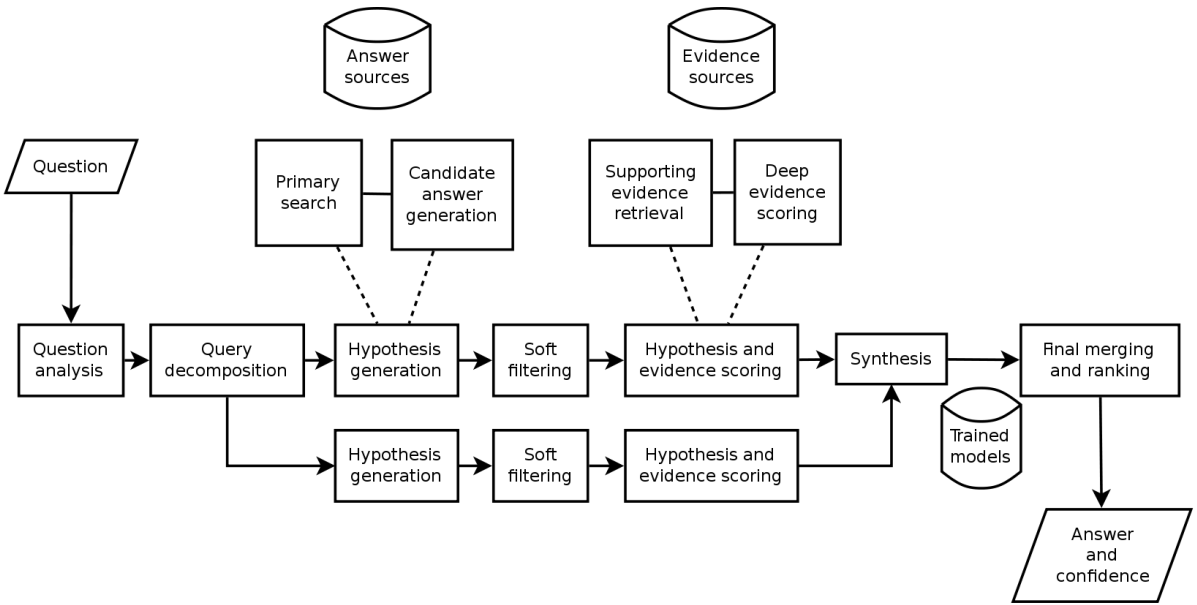
Google Translate.

Are there even any competitors?

They’re able to incorporate trillions of words into their language models and elicit user feedback on results.

Question Answering

IBM Watson, Wolfram Alpha

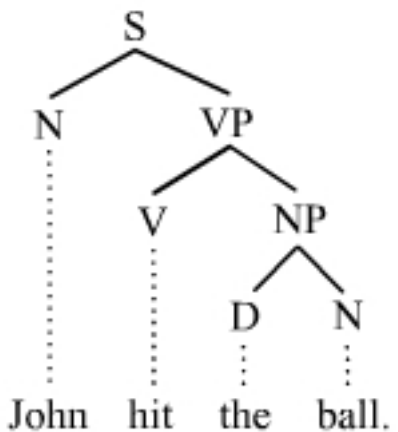


Most applications are less visible.

Part of Speech Tagging/Parse trees

Aids in many other NLP tasks

e.g. *Named Entity Recognition*



Constituency-based parse tree

Topic Modeling

Finding latent groupings of documents based on the words therein.

Different topics generate words at particular frequencies, so you can work backwards from the words in a document to the topics.

Useful for news aggregators, or segmenting a proprietary corpus.

Sentiment Analysis

Determining the emotional content of a document.

Most often applied to tweets with marketing implications.

Many approaches proposed.

At this point, you may be imagining some uses of NLP in your own work.

III. BASIC NLP PRACTICE

First order of business:

Split text into sentences or words.

If you want to parse sentences, it helps to have sentences.

Relatively easy for English.

Sentences end with periods, words are separated by spaces.

There are some oddballs, though.

“Dr. Oz promised me magic weight-loss pills; they didn’t work.”

“omg the food was so gross the portions were tiny ill never go back”

We went over easier examples, but you can imagine difficulties in other languages.

Luckily, statistical models can tolerate some level of messy data.

Second order of business:

normalize word forms

LinkedIn sees 6,000+ variations on the job title, “Software Engineer”

They see 8,000+ variations on the company, “IBM”

They have to recognize all of these and understand they are the same.

On a smaller scale, it is often useful to strip away conjugations and other modifiers.

This is called stemming.

science, scientist => scien

swim, swimming, swimmer => swim

The resulting text is unreadable, but retains semantic content.

The classic, standard English stemmer is the Porter stemmer.

Stemming is very useful to reduce feature set size.

It can sometimes degrade performance.

(Why?)

"Certain things have come to light. And, you know, has it ever occurred to you, that, instead of, uh, you know, running around, uh, uh, blaming me, you know, given the nature of all this new sh*t, you know, I-I-I-I... this could be a-a-a-a lot more, uh, uh, uh, uh, uh, uh, complex, I mean, it's not just, it might not be just such a simple... uh, you know?"

-The Dude, The Big Lebowski (1998)

Some words are so very common that they provide no information to a statistical language model.

We should remove these *stop words*.

Note: different languages have different stop words, and they may have meaning in other languages.

Aside from looking up a list, how can you find stop words?

Term frequency

$N_{term}/N_{terms\ in\ document}$

Document frequency

$N_{documents\ containing\ term}/N_{documents}$

Stop words will have a high document frequency

What about highly discriminative words?

tf-idf

term frequency-inverse document frequency:

$$(N_{term}/N_{terms\ in\ document}) * \log(N_{documents}/N_{documents\ containing\ term})$$

Largest for words that occur more frequently in a document, but occur in fewer documents overall.

Stop word removal and tf-idf weighting are reliable ways to improve many natural language models.

SENTIMENT CLASSIFICATIONS

Patrick Stewart ✓
@SirPatStew

Follow

Tomorrow @sunnyozell and her brilliant band are playing a free show at my Old Vic Theatre pub from 50 years ago! @olddukebristol

Reply Retweet Favorite ••• More

RETWEETS	FAVORITES
52	163

11:50 AM - 25 Jul 2014

Linus Torvalds
@Linus__Torvalds

Follow

Microsoft isn't evil, they just make really crappy operating systems.

Reply Retweet Favorite ••• More

RETWEETS	FAVORITES
41	18

11:05 AM - 29 Jan 2013

Marketers want to know whether their audience's engagement with their brand is positive or negative in nature.

Some say the stock market fluctuates with the mood on Twitter.

Finding training data is really hard.

Humans are good at handling ambiguity, but that's not the same as being accurate.

Miscommunication happens all the time

Natural Language Processing comprises many, very hard problems.

Most are outside the scope of the course, but hopefully you have an idea of what questions to ask if you encounter an NLP problem.

Any more ideas for uses of NLP at your work?