UNIVERSITY OF
SAN FRANCISCO

Master of Science
in Analytics

Overview

Natural Language Processing

Twitten By…?

[redacted]

Big announcement by Ford today. Major investment to be made in three Michigan plants. Car companies coming back to U.S. JOBS! JOBS! JOBS!

Donald J. Trump
@realDonaldTrump

Big announcement by Ford today. Major investment to be made in three Michigan plants. Car companies coming back to U.S. JOBS! JOBS! JOBS!

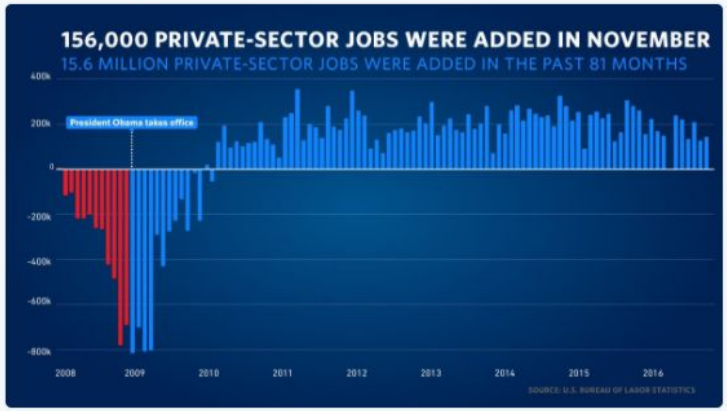
Mar. 28, 2017, 6:36 a.m.

[redacted]

Facing the worst financial crisis in 80 years, you delivered the longest streak of job growth in our history.

President Obama @POTUS44 · Jan 1

Facing the worst financial crisis in 80 years, you delivered the longest streak of job growth in our history.



1.3K 27K 81K

Why study NLP for Data Science?

- Sutton's Law: because that's where the money is
- Because that's what we were promised
- Because not all data is structured
  - It's easier to draw conclusions from SQL
  - But humans do not live in structures
- What is "NLP for Data Science?"
  - Sentiment Analysis ("Voice of the customer")
  - Topic Modelling
  - Information Retrieval / Extraction

Metropolis poster available at https://en.wikipedia.org/w/index.php?curid=8913129
Sonny's head by Shao19 - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=7395491
HAL 9000 by Cryteria - Own work, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=11651154
Data image available at https://en.wikipedia.org/w/index.php?curid=12543502

What is NLP?

- The ability of a computer program to process human language
  - ***Computer program*** — early versions were FSAs/FSTs
  - ***Process*** — natural language understanding vs. natural language generation
  - ***Human language*** — artificial languages... (written) text... speech
- @emilymbender was asked (roughly): what should NLP people know?

  - Language has structure beyond linear order of words.
  - That structure is useful to leverage if we're interested in extracting meaning.
  - [...]
  - The structure of any given language is fairly consistent across genres [...].
  - But languages vary in the structures that they use [...].

Why is NLP hard?

- Ambiguity
  - Lexical — what's the sense of a word?

    I made her duck.
  - Attachment — who did what to whom with what?

    He saw the woman with the telescope.
  - Structural

    The horse raced past the barn fell.
- Multi-word expressions, metaphor, simile, poetry, novelty, etc.

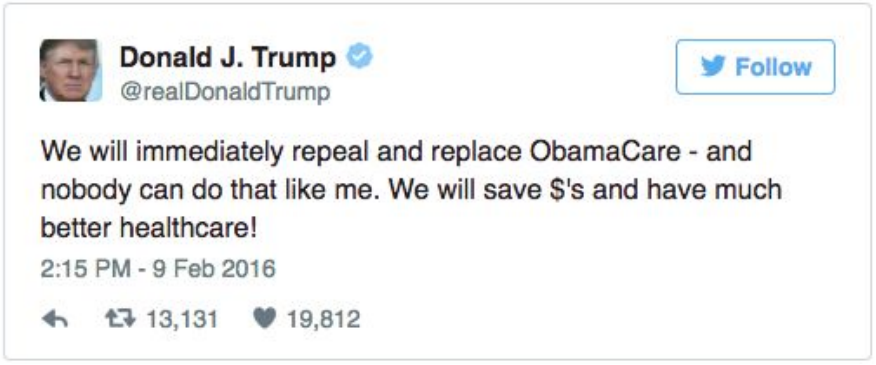
  "Water from the ocean could sail a boat..."
- Pragmatics, discourse

  Ann: Do you like Italian food?

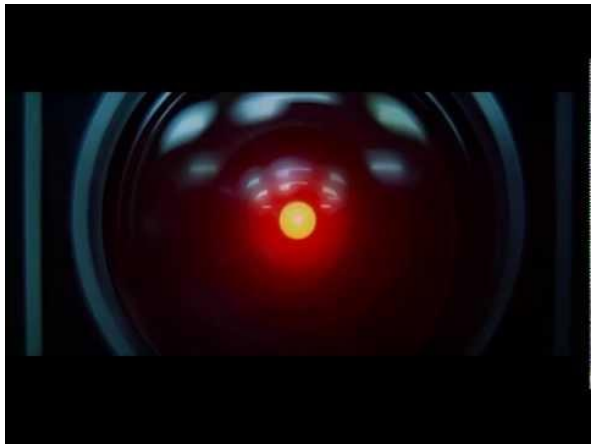
  Mohammad: I like Thai.

Sometimes, we are the problem

- Newspaper headlines:
  - Farmer bill dies in house
  - Iraqi head seeks arms
  - Red tape holds up new bridge
- Translations:
  - All your base are belong to us
  - The spirit is willing, but the flesh is weak >> The vodka is good, but the meat is rotten
- Lies, damned lies ~~and statistics~~ poets, priests & politicians
  - *It depends upon what the meaning of the word "is" is* — Bill Clinton, 1998
  - *Let's fund our [National Health Service] instead* — Nigel Farrage, 2016
  - *I never said repeal it and replace it within 64 days* — Donald Trump



What we do without thinking



Dave: Open the pod bay doors, HAL.

HAL needs to understand that Dave is issuing a command… and what the command is.

HAL 9000: I'm sorry, Dave. I'm afraid I can't do that.

Contraction for "I am."

Politeness

Knows who is speaking.

Reference to the command.

Contraction for "cannot."

Can Machines Think?

- The Turing Test (1950)
  - Simplifies the question: can you distinguish between a person and a machine?
  - Very few systems have passed the Turing Test
- ELIZA (1966)
  - Rogerian psychotherapist — simple rules about reflecting a conversation
  - Passed the Turing test

    User: You are like my father in some ways.
    ELIZA: What resemblance do you see
    User: You are not very aggressive, but I think you don't want me to notice that.
    ELIZA: What makes you think I am not aggressive

This Course

- What we'll study:
  - Human language (examples, etc.)
  - Linguistic and psychological theories
  - Algorithms
  - Applications (systems)
- More specifically:
  - Regular Expressions, Morphology …
  - Syntax —> Semantics
  - Stuff useful for data science: Sentiment Analysis, Topic Modelling, Information Retrieval
  - Applications (NLTK, Mallet & gensim) for building NLU systems
- Schedule — online as a Google Spreadsheet massively subject to change
- Communication — #nlp_2017