Natural Language Processing

Data Skills for Empirical Research

Winter, 2020

NORTHWESTERN UNIVERSITY

Kellogg School of Management

Dealing with Unstructured Data

Vast quantities of information are encoded as unstructured data, in the form of natural language text.

But it can be hard to make this information available for computational analysis at scale.

Information Extraction

As we learned in the regular expression workshop, regex are great tackling lots of information extraction problems.

Information Extraction

But sometimes regex are inadequate for the job.

Information Extraction

The internet company spent $1 billion on a large office park near its headquarters in Mountain View, California, according to the Mercury News, and has now spent at least $2.8 billion on properties in Mountain View, Sunnyvale and San Jose over the last two years.
In this case, Google is purchasing property that it's already been leasing. The company is the main tenant of the 12 buildings that comprise the 51.8-acre Shoreline Technology Park. Google declined to comment on its purchase.

Earlier this month, Google agreed to pay an additional $110 million for 10.5 acres for a new campus in downtown San Jose, with the possibility of buying about 11 more acres. The city will vote on the plans in early December.
It's also been a big year for Google property purchases outside of Silicon Valley.

In the first quarter, the company spent $2.4 billion to buy New York City's Chelsea Market. Chief Financial Officer Ruth Porat said that the company favors "owning rather than leasing real estate when we see good opportunities."

As for leases, Google just signed on for a massive new space in downtown San Francisco.
A little less than a decade later, dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation.

Information Extraction

The internet company spent $1 billion on a large office park near its headquarters in Mountain View, California, according to the Mercury News ORG , and has now spent at least $2.8 billion on properties in Mountain View, Sunnyvale and San Jose over the last two years.
In this case, Google ORG is purchasing property that it's already been leasing. The company is the main tenant of the 12 buildings that comprise the 51.8-acre Shoreline Technology Park. Google ORG declined to comment on its purchase.

Earlier this month, Google ORG agreed to pay an additional $110 million for 10.5 acres for a new campus in downtown San Jose, with the possibility of buying about 11 more acres. The city will vote on the plans in early December.
It's also been a big year for Google ORG property purchases outside of Silicon Valley.

In the first quarter, the company spent $2.4 billion to buy New York City's Chelsea Market. Chief Financial Officer Ruth Porat said that the company favors "owning rather than leasing real estate when we see good opportunities."

As for leases, Google ORG just signed on for a massive new space in downtown San Francisco.
A little less than a decade later, dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation.

Information Extraction

The internet company spent $1 billion on a large office park near its headquarters in Mountain View, California, according to the Mercury News ORG , and has now spent at least $2.8 billion on properties in Mountain View, Sunnyvale and San Jose over the last two years.
In this case, Google ORG is purchasing property that it's already been leasing. The company is the main tenant of the 12 buildings that comprise the 51.8-acre Shoreline Technology Park. Google ORG declined to comment on its purchase.

Earlier this month, Google ORG agreed to pay an additional $110 million for 10.5 acres for a new campus in downtown San Jose, with the possibility of buying about 11 more acres. The city will vote on the plans in early December.
It's also been a big year for Google ORG property purchases outside of Silicon Valley.

In the first quarter, the company spent $2.4 billion to buy New York City's Chelsea Market PERSON . Chief Financial Officer Ruth Porat PERSON said that the company favors "owning rather than leasing real estate when we see good opportunities."

As for leases, Google ORG just signed on for a massive new space in downtown San Francisco.
A little less than a decade later, dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation.

Information Extraction

The internet company spent $1 billion on a large office park near its headquarters in Mountain View GPE , California GPE , according to the Mercury News ORG , and has now spent at least $2.8 billion on properties in Mountain View GPE , Sunnyvale GPE and San Jose GPE over the last two years.
In this case, Google ORG is purchasing property that it's already been leasing. The company is the main tenant of the 12 buildings that comprise the 51.8-acre Shoreline Technology Park LOC . Google ORG declined to comment on its purchase.

Earlier this month, Google ORG agreed to pay an additional $110 million for 10.5 acres for a new campus in downtown San Jose GPE , with the possibility of buying about 11 more acres. The city will vote on the plans in early December.
It's also been a big year for Google ORG property purchases outside of Silicon Valley LOC .

In the first quarter, the company spent $2.4 billion to buy New York City's GPE Chelsea Market PERSON . Chief Financial Officer Ruth Porat PERSON said that the company favors "owning rather than leasing real estate when we see good opportunities."

As for leases, Google ORG just signed on for a massive new space in downtown San Francisco GPE .
A little less than a decade later, dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation.

Information Extraction

The internet company spent $1 billion on a large office park near its headquarters in Mountain View GPE , California GPE , according to the Mercury News ORG , and has now spent at least $2.8 billion on properties in Mountain View GPE , Sunnyvale GPE and San Jose GPE over the last two years DATE .
In this case, Google ORG is purchasing property that it's already been leasing. The company is the main tenant of the 12 buildings that comprise the 51.8-acre Shoreline Technology Park LOC . Google ORG declined to comment on its purchase.

Earlier this month DATE , Google ORG agreed to pay an additional $110 million for 10.5 acres for a new campus in downtown San Jose GPE , with the possibility of buying about 11 more acres. The city will vote on the plans in early December DATE .
It's also been a big year DATE for Google ORG property purchases outside of Silicon Valley LOC .

In the first quarter DATE , the company spent $2.4 billion to buy New York City's GPE Chelsea Market PERSON . Chief Financial Officer Ruth Porat PERSON said that the company favors "owning rather than leasing real estate when we see good opportunities."

As for leases, Google ORG just signed on for a massive new space in downtown San Francisco GPE .
A little less than a decade later DATE , dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation.

Information Extraction

The internet company spent $1 billion MONEY on a large office park near its headquarters in Mountain View GPE , California GPE , according to the Mercury News ORG , and has now spent at least $2.8 billion MONEY on properties in Mountain View GPE , Sunnyvale GPE and San Jose GPE over the last two years DATE .
In this case, Google ORG is purchasing property that it's already been leasing. The company is the main tenant of the 12 CARDINAL buildings that comprise the 51.8-acre Shoreline Technology Park LOC . Google ORG declined to comment on its purchase.

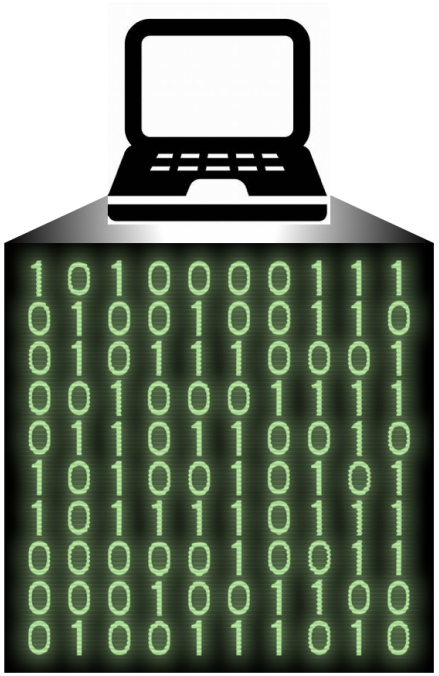
Earlier this month DATE , Google ORG agreed to pay an additional $110 million MONEY for 10.5 acres QUANTITY for a new campus in downtown San Jose GPE , with the possibility of buying about 11 more acres QUANTITY . The city will vote on the plans in early December DATE .
It's also been a big year DATE for Google ORG property purchases outside of Silicon Valley LOC .

In the first quarter DATE , the company spent $2.4 billion MONEY to buy New York City's GPE Chelsea Market PERSON . Chief Financial Officer Ruth Porat PERSON said that the company favors "owning rather than leasing real estate when we see good opportunities."

As for leases, Google ORG just signed on for a massive new space in downtown San Francisco GPE .
A little less than a decade later DATE , dozens CARDINAL of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation

NLP to the Rescue

A field concerned with the ability of a computer to understand, analyze, manipulate, and potentially generate human language



NLP to the Rescue

Autocomplete

natural language processing is

natural language processing issues
natural language processing is **fun**
natural language processing is
natural language processing is **divided into**
natural language processing is **machine learning**
natural language processing israel
natural language processing is **used for what**
natural language processing is **a field of quizlet**

Spam Filter

Autocorrect

NLP Applications

Supervised Methods

- Information Extraction
  - Named Entities (people, places, organizations)
  - Relations (location, times, parts)
  - Temporal Expressions (Dec. 25th, 10 years ago, next week)
  - Events (verb semantics)
- Entity Linking (map mentions onto normalized data)
- Text Classification (e.g., sentiment analysis)

NLP Applications

Unsupervised Methods

- Text Feature Analysis (word counts, average word length, complexity)
- Topic Modeling (discovering latent structure in a collection of texts)
- Language Models (probability of a sequence)

NLP Tasks

Lower-Level Tasks

- Sentence splitting
- Tokenization (`The quick red fox` → `{The, quick, red, fox}`)
- Lemmatizing / stemming (`{bought, buys, buying}` → `buy`)
- Removing "stop" words, punctuation (`a, an, the, for, to, ?, !`)
- Vectorizing documents
- Word embeddings

NLP Tasks

Higher-Level Tasks

- Part of Speech Tagging
- Word Sense Disambiguation
- Coreference Resolution
- Semantic Role Labeling
- Syntactic Parsing
- Dialogue Modeling

Resources

- Introductory textbook: https://web.stanford.edu/~jurafsky/slp3
- Python resources:
  - spaCy: https://spacy.io/
  - scikit-learn: https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html
- R resources:
  - https://cran.r-project.org/web/views/NaturalLanguageProcessing.html
  - tidytext: https://cran.r-project.org/web/packages/tidytext/index.html
  - tm: https://cran.r-project.org/web/packages/tm/index.html