Before We Begin

About Me

- University of Chicago, Mathematics
- Narrative Science, Software Engineer - Test

Citations

- Big thanks to my colleagues at Narrative Science!
- Special thanks to Zachary Ernst, Trunk Club data engineer.
- Thanks to the NLTK project: Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. O'Reilly Media Inc.

Pluralizing Noun Phrases

Getting Algorithms to Learn English Language Quirks

Let's Do An Experiment

What is the plural form of...

Field Engineer, Telecommunications

Vice President of Sales

Press Secretary

Able-bodied seaman

Airman First Class

Midwife

A Naive Algorithm Would Say...

Field Engineer, **Telecommunicationses**

Vice President of **Saleses**

Press **Secretarys**

Able-bodied **seamans**

Airman First **Classes**

Midwifes

The Puzzle

.NET Software Developer / Programmer
3d Artist
911 Dispatcher
A-Operator
ABAP Developer
ASIC Design Engineer
ASIC Engineer
ASP.NET Developer
Able Bodied Seaman
Abstractor
Academic Advisor
Academic Advisor (College/University)
Academic Affairs Dean
Academic Counselor
Academic Dean
Academic Director
Account Analyst
Account Clerk
Account Coordinator
Account Development Manager

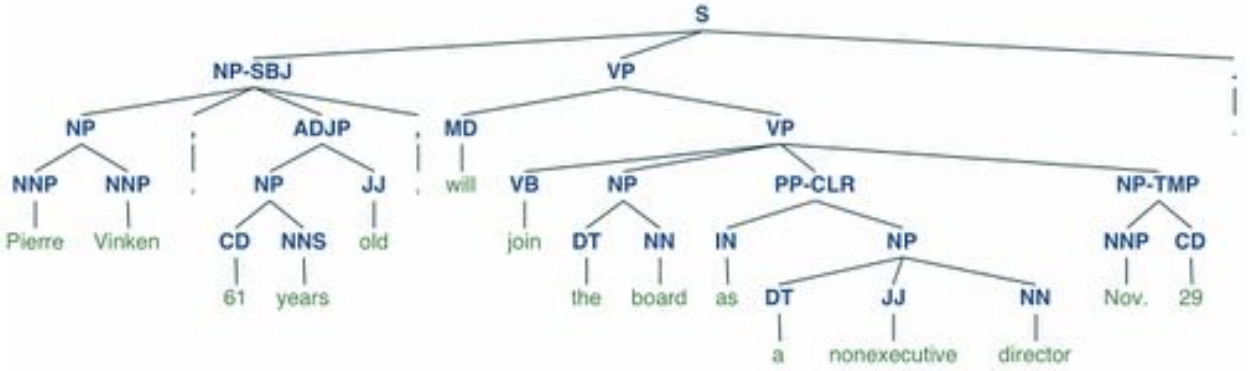
Director of Property Management
Director of Provider Relations
Director of Public Relations (PR)
Director of Public Works, City
Director of Purchasing
Director of Reimbursement
Director of Revenue Cycle Management
Director of Revenue Management
Director of Sales Operations
Director of Sales and Marketing
Director of Sales, Hotels/Hospitality
Director of Special Education
Director of Strategic Alliances
Director of Strategic Planning
Director of Strategy
Director of Surgical Services
Director of Transportation
Director of Vendor Management
Director of Web Content Management
Director of Youth and College Ministries

Tools We Used

Python's Natural Language Toolkit (NLTK): http://www.nltk.org/

Display a parse tree:

```
>>> from nltk.corpus import treebank
>>> t = treebank.parsed_sents('wsj_0001.mrg')[0]
>>> t.draw()
```



Tools We Used

Python's Inflection library: https://inflection.readthedocs.io/en/latest/

inflection.pluralize(*word*) [source]

Return the plural form of a word.

Examples:

```
>>> pluralize("post")
"posts"
>>> pluralize("octopus")
"octopi"
>>> pluralize("sheep")
"sheep"
>>> pluralize("CamelOctopus")
"CamelOctopi"
```

inflection.ordinalize(*number*) [source]

Turn a number into an ordinal string used to denote the position in an ordered sequence such as 1st, 2nd, 3rd, 4th.

Examples:

```
>>> ordinalize(1)
"1st"
>>> ordinalize(2)
"2nd"
>>> ordinalize(1002)
"1002nd"
>>> ordinalize(1003)
"1003rd"
>>> ordinalize(-11)
"-11th"
>>> ordinalize(-1021)
"-1021st"
```

Step 0: Identify the Cases

- Modifier + Noun: `Deputy Sheriff`
- Special Characters:
  - `Closing Agent, Title`
  - `Coding/Insurance Specialist`
  - `Crime Scene Investigator (CSI)`
- Noun + Modifier: `Attorney General`

Step 1: Clean Up

- “Coding/Insurance Specialist” → “Coding and Insurance Specialist”

- “Crime Scene Investigator (CSI)” → “Crime Scene Investigator”

Step 2: Tokenize the Strings

```
>>> from ntlk.tokenize import word_tokenize

>>> word_tokenize(Head of Medical Records)

[“head”, “of”, “medical”, “records”]
```

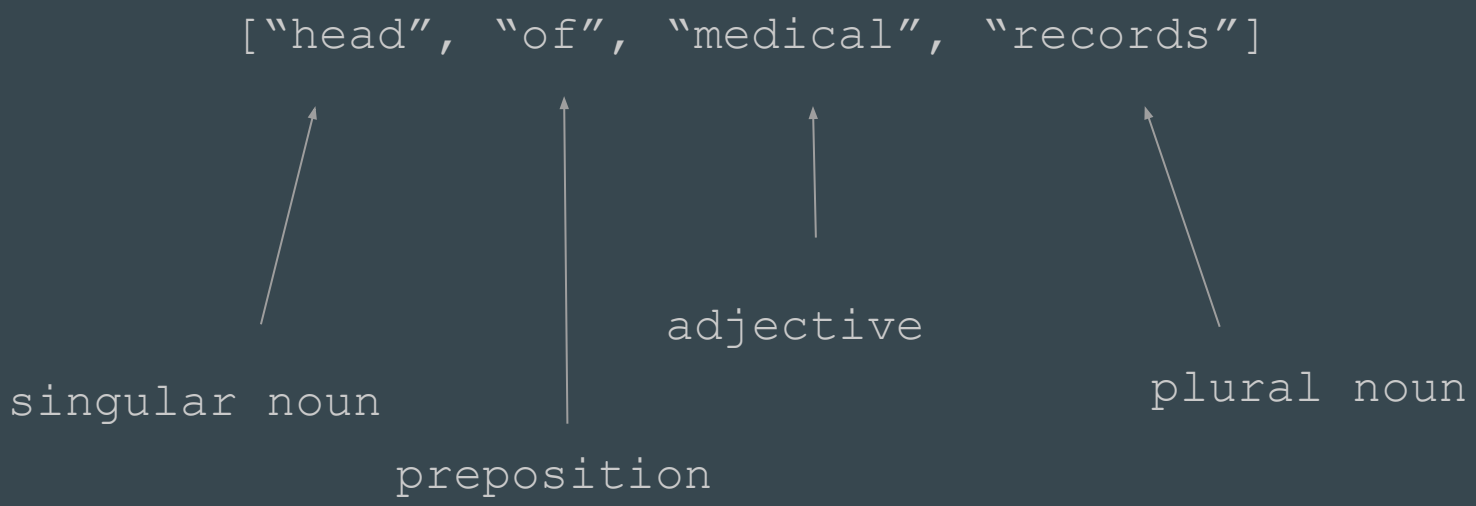
Step 3: Tag Parts of Speech

```
>>> from nltk.tag import pos_tag

>>> pos_tag([“head”, “of”, “medical”, “records”])

[('head', 'NN'), ('of', 'IN'), ('medical', 'JJ'),
('records', 'NNS')]
```

Step 3: Tag Parts of Speech



Step 4: Manipulate The Parts of Speech

- If there's only one noun in the phrase, pluralize it:

  `"Deputy" + "Sheriffs"`

- If the phrase has multiple nouns, pluralize the first one that immediately precedes a preposition:

  `"Heads" + "of" + "Medical" + "Records"`

Test Time!

Regular English Plurals :)

“machine learner” → “machine learners”

“charismatic politician” → “charismatic politicians”

“gold medalist” → “gold medalists”

“non-binary identity” → “non-binary identities”

Irregular English Plurals :)

“groomsman” → “groomsmen”

“The Good Wife” → “the good wives”

“strong ox → strong oxen”

Prepositional Phrases :)

“truth in wine” → “truths in wine”

“doughnut with coffee” → “doughnuts with coffee”

“woman of steel” → “women of steel”

“young person in love” → “young people in love”

Non-English Phrases :(

“polyhedron” → “polyhedrons”

“schema” → “schemas”

“illustrious alumna” → “illustrious alumnas”

“samurai in training” → “samurais in training”

What Did We Learn?

- Data is dirty...like, really dirty.
- Our algorithm seemed to struggle with borrowed (non-English) phrases...but maybe those forms were fine?
- Natural language processing and generation are hard! But there are extremely powerful tools for tackling the job.

Thank You For Listening!

NarrativeScience

TELL THE STORIES HIDDEN IN YOUR DATA™

www.narrativescience.com

Other Resources

- Princeton University's WordNET: https://wordnet.princeton.edu/
- Carnegie Mellon University's pronunciation dictionary: http://www.speech.cs.cmu.edu/comp.speech/Section1/Lexical/cmu-dict.html
- Brown's American English Corpus: http://www.hit.uib.no/icame/brown/bcm.html
- The Penn Treebank Project: https://www.cis.upenn.edu/~treebank/

Please write to me! daphne.s.kao@gmail.com