BIG DATA ANALYTICS: NATURAL LANGUAGE PROCESSING (NLP) TECHNIQUES

DR. TAMMY ROBERTSON
AGENDA

• THE APPLICATION OF NLP
• THE ROLE OF MACHINE LEARNING
• NLP TECHNIQUES AND METHODS
• STEPS FOR EXTRACTING INFORMATION USING NLP
BIG DATA ANALYTICS

BIG DATA & DATA ANALYTICS SERVICES

- Open-source Big Data Analytics Solutions
- Databases, Data Warehousing and ETL
- Investigative, Targeted, and Predictive Analytics
- Data Science / Analytics SMEs in various tools
- Statistical Reporting and Visualization
- Business Intelligence Dashboards
BIG DATA ANALYTICS

- Big Data Analytics Renaissance
  - Smart Algorithms
  - Greater Computing Power
  - Open Source Technologies
ARTIFICIAL INTELLIGENCE (AI)

- Definition is a moving target
  - *Intelligent Machines*
  - *Capable of Learning*
- Use Cases
  - *Image Recognition*
  - *Self-Driving Cars*
  - *Robotic Prosthetics*
  - *Personalized Bots*
  - *Cancer Detection*
MACHINE LEARNING

- Supervised Learning
  - Regression
  - Neural Networks
  - Bayesian
- Deep Learning
  - Decision Trees
- Unsupervised Learning
  - Clustering
  - Association Rules
NATURAL PROCESSING LANGUAGE

• Subdiscipline
  • Computer Science
  • Computational Linguistics
  • Artificial Intelligence
• Human Language Interpreter
NATURAL LANGUAGE PROCESSING
1. Install Python (Latest Version)
2. Install NLTK (shell)
   pip install nltk
1. Download Corpora (Python shell)
   import nltk
   nltk.download()
HOW IT WORKS

• Python

• NLTK
  • Documentation: https://www.nltk.org/
  • NLTK HOWTOs: http://www.nltk.org/howto/
HOW IT WORKS

• NLTK Data Folder
  • Libraries
  • Corpora
  • Trained Models
  • Datasets
HOW IT WORKS

Tokenization
Breaking down text into words and sentences

Stopped Words
Filtering common words

Stemming
Removing ends of words

Parts-of-Speech
Identifying part-of-speech

Chunking
Identifying common groups of words

Word Sense Disambiguation (WSD)
Identifying the context in which the word occurs
Some Text: Hello Dr. Robertson, how are you doing today? The weather is great, and NLT is awesome. Have a great Symposium! You shouldn't use contractions in academic writing.

Sentence Tokenizer: ['Hello Dr. Robertson, how are you doing today?' │ 'The weather is great, and NLT is awesome.' │ 'Have a great Symposium!' │ "You shouldn't use contractions in academic writing."]

Word Tokenizer: ['Hello' │ 'Dr.' │ 'Robertson' │ ',' │ 'how' │ 'are' │ 'you' │ 'doing' │ 'today' │ '?' │ 'The' │ 'weather' │ 'is' │ 'great' │ ',' │ 'and' │ 'NLT' │ 'is' │ 'awesome' │ '.' │ 'Have' │ 'a' │ 'great' │ 'Symposium' │ '!' │ 'You' │ 'should' │ "n't" │ 'use' │ 'contractions' │ 'in' │ 'academic' │ 'writing' │ '.']
STOPPED WORDS

• Some Text: Hello Dr. Robertson, how are you doing today? The weather is great, and NLT is awesome. Have a great Symposium! You shouldn't use contractions in academic writing.

• Word Tokenizer: ['Hello' 'Dr.' 'Robertson' ',' 'how' 'are' 'you' 'doing' 'today' '?' 'The' 'weather' 'is' 'great' ',' 'and' 'NLT' 'is' 'awesome' '.' 'Have' 'a' 'great' 'Symposium' '!' 'You' 'should' 'n't' 'use' 'contractions' 'in' 'academic' 'writing' '.']

• Stopped Words: ['Hello' 'Dr.' 'Robertson' ',' 'today' '?' 'The' 'weather' 'great' ',' 'NLT' 'awesome' '.' 'Have' 'great' 'Symposium' '!' 'You' 'n't' 'use' 'contractions' 'academic' 'writing' '.']
STOPPED WORDS LIST

{'at', 'have', 'whom', 'doing', 'will', 'as', 'didn', "aren't", 't', 'further', 'the', 'a', 'was', 'so', 'mustn', "mustn't", "isn't", 'hadn', 'on', "shouldn't", 'weren', "weren't", 'why', 'me', 'ours', 'other', 'again', 'hers', 'wasn', 's', 'myself', 'only', 'ain', 'not', 'should', 'here', 'needn', "needn't", "it's", 'is', 'but', 'in', 'once', 'where', "hasn't", 'how', "haven't", 'himself', "mightn't", 'he', "should've", 'wouldn', 'you'd', 'they', 'because', 'been', 'couldn', 'my', 'i', 'and', 'while', 'for', 'having', 'won't', 'being', 'their', 'you're', 'when', 'that', 'before', 'd', "shan't", 'these', 'we', 'she', 'ma', 'you', 'haven', "she's", 'did', 'off', 'now', 'doesn', 'both', 'be', 'yours', "wouldn't", 'your', 'between', 'which', 'its', 'y', 'each', 'what', 'those', 'up', 'll', 'above', 'own', 'shouldn', 'about', 'just', 'out', 'mightn', 'had', 'who', 'this', 'then', 'below', 'all', 'through', 'aren', 'isn', 'am', 'most', 'it', 'them', 'can', 'an', 'themselves', 'very', 'herself', 'has', 'of', 'any', 'shan', 'under', 'theirs', 'such', "doesn't", 'ourselves', 'than', 'too', 're', 'won', 'him', 'itself', 'her', "wasn't", 'few', 'does', 'his', "that'll", 'with', "you'll", 'no', 'against', 'or', 'some', 'm', 'do', 'nor', 'over', 'same', 'yourself', 'by', 'during', 'our', 'were', 've', 'are', 'if', 'after', 'to', "don't", 'more', 'yourselves', "you've", 'down', 'from', 'don', "didn't", "hadn't", 'there', 'o', 'into', "couldn't", 'until', 'hasn'}
STEMMING

- Close
- Closer
- Closing
- Closed
- Closely
STEMMING

• Some Words: close, closer, closing, closed, closely

Stemmed Words: close
closer
close
close

Some Text: While watching closely, look closer. Is the store close to closing or is it closed?

Stemmed Sentence:
while the close
watch store close
, to close
look or
closer is
. it
LEMMATIZING

- students
- cacti
- geese
- mouse
- closely
- closer
- closed
- better (as adjective)
- best (as adjective)
- run
- run (as verb)

- student
- cactus
- goose
- mouse
- closely
- closer
- closest
- closed
- good
- best
- run
- run
### PARTS-OF-SPEECH

<table>
<thead>
<tr>
<th>POS</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>Coordinating conjunction</td>
<td>there is ... think of it like &quot;there exists&quot;</td>
</tr>
<tr>
<td>CD</td>
<td>Cardinal digit</td>
<td>1</td>
</tr>
<tr>
<td>DT</td>
<td>Determiner</td>
<td>the</td>
</tr>
<tr>
<td>EX</td>
<td>Existential there</td>
<td>there is ... think of it like &quot;there exists&quot;</td>
</tr>
<tr>
<td>FW</td>
<td>Foreign word</td>
<td>am</td>
</tr>
<tr>
<td>IN</td>
<td>Preposition/subordinating conjunction</td>
<td>to</td>
</tr>
<tr>
<td>JJ</td>
<td>Adjective</td>
<td>big</td>
</tr>
<tr>
<td>JJR</td>
<td>Adjective, comparative</td>
<td>bigger</td>
</tr>
<tr>
<td>JJS</td>
<td>Adjective, superlative</td>
<td>biggest</td>
</tr>
<tr>
<td>LS</td>
<td>List marker</td>
<td>1)</td>
</tr>
<tr>
<td>MD</td>
<td>Modal</td>
<td>could, will</td>
</tr>
<tr>
<td>NN</td>
<td>Noun, singular ‘desk’</td>
<td>desk</td>
</tr>
<tr>
<td>NNS</td>
<td>Noun plural</td>
<td>desks</td>
</tr>
<tr>
<td>NNP</td>
<td>Proper noun, singular</td>
<td>Harrison</td>
</tr>
<tr>
<td>NNP5</td>
<td>Proper noun, plural</td>
<td>Americans</td>
</tr>
<tr>
<td>PDT</td>
<td>Predeterminer</td>
<td>all the kids</td>
</tr>
<tr>
<td>POS</td>
<td>Possessive ending</td>
<td>parent's</td>
</tr>
<tr>
<td>PRP</td>
<td>Personal pronoun</td>
<td>I, he, she</td>
</tr>
<tr>
<td>PRPS</td>
<td>Possessive pronoun</td>
<td>my, his, hers</td>
</tr>
<tr>
<td>RB</td>
<td>Adverb</td>
<td>very, silently,</td>
</tr>
<tr>
<td>RBR</td>
<td>Adverb, comparative</td>
<td>better</td>
</tr>
<tr>
<td>RBS</td>
<td>Adverb, superlative</td>
<td>best</td>
</tr>
<tr>
<td>RP</td>
<td>Particle</td>
<td>give up</td>
</tr>
<tr>
<td>TO</td>
<td>To</td>
<td>go 'to' the store.</td>
</tr>
<tr>
<td>UH</td>
<td>Interjection</td>
<td>errrrrrr</td>
</tr>
<tr>
<td>VB</td>
<td>Verb, base form</td>
<td>take</td>
</tr>
<tr>
<td>VBD</td>
<td>Verb, past tense</td>
<td>took</td>
</tr>
<tr>
<td>VBG</td>
<td>Verb, gerund/present participle</td>
<td>taking</td>
</tr>
<tr>
<td>VBN</td>
<td>Verb, past participle</td>
<td>taken</td>
</tr>
<tr>
<td>VBP</td>
<td>Verb, sing. present, non-3d</td>
<td>take</td>
</tr>
<tr>
<td>VBP</td>
<td>Verb, sing. present, non-3d</td>
<td>take</td>
</tr>
<tr>
<td>VBZ</td>
<td>Verb, 3rd person sing. present</td>
<td>takes</td>
</tr>
<tr>
<td>WDT</td>
<td>Wh-determiner</td>
<td>which</td>
</tr>
<tr>
<td>WP</td>
<td>Wh-pronoun</td>
<td>who, what</td>
</tr>
<tr>
<td>WPS</td>
<td>Possessive wh-pronoun</td>
<td>whose</td>
</tr>
<tr>
<td>WRB</td>
<td>Wh-adverb</td>
<td>where, when</td>
</tr>
</tbody>
</table>
This solemn occasion marks the 196th time that a President of the United States has reported on the State of the Union since George Washington first did so in 1790.
REGULAR EXPRESSIONS

Identifiers
- \d = any number
- \D = anything but a number
- \s = space
- \S = anything but a space
- \w = any letter
- \W = anything but a letter
- . = any character, except for a new line
- \ = space around whole words
- \ = period. must use backslash, because . normally means any character.

Modifiers
- {1,3} = for digits, u expect 1-3 counts of digits, or "places"
- + = match 1 or more
- ? = match 0 or 1 repetitions.
- * = match 0 or MORE repetitions
- $ = matches at the end of string
- ^ = matches start of a string
- | = matches either/or. Example x|y = will match either x or y
- [] = range, or "variance"
- {x} = expect to see this amount of the preceding code.
- {x,y} = expect to see this x-y amounts of the preceding code.
- \r = carriage return

Characters to REMEMBER TO ESCAPE IF USED!
- . + * ? [ ] $ ^ ( ) { } |

White Space Charts
- \n = new line
- \s = space
- \t = tab
- \e = escape
- \f = form feed

Brackets
- [] = quant[ia]tative = will find either quantitative, or quantatative.
- [a-zA-Z] = return any lowercase letter a-z
- [1-5a-qA-Z] = return all numbers 1-5, lowercase letters a-q and uppercase A-Z
CHUNKING

```python
>>> import nltk
>>> nltk.download()
```

```bash
<table>
<thead>
<tr>
<th>File</th>
<th>Size</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChunkingSubtree.py</td>
<td>441</td>
<td>.py</td>
</tr>
<tr>
<td>Chunking.py</td>
<td>441</td>
<td>.py</td>
</tr>
<tr>
<td>symposium.py</td>
<td>122</td>
<td>.py</td>
</tr>
<tr>
<td>symposium.exe</td>
<td>122</td>
<td>.exe</td>
</tr>
</tbody>
</table>
```

```
>>> from Chunking import Chunk
>>> Chunk
```

```
>>> print(Chunk)
```
CHUNKING
CHINKING
NAMED ENTITY RECOGNITION
Program

- plan.n.01
- plan
- a series of steps to be carried out or goals to be accomplished
- ['they drew up a six-step plan', 'they discussed plans for a new bond issue']


