Natural Language Processing and Python

Dr. Sowmya Vajjala
AbacusNext, Toronto

Big Data Applications track@Mid-west Big Data Summer School
Iowa State University, USA

17 May 2018

About Me

- Data Scientist@AbacusNext, Toronto, Canada (Just started!)
- Faculty at ISU (01/16-05/18)

Today's Agenda

- What is Natural Language Processing and where is it useful?
- NLP in Python - overview
- NLP and Python: Practical examples

Note: Slides and code examples are at:
`https://github.com/nishkalavallabhi/MBDS2018-NLPTutorial`

What is NLP?

1. NLP is a sub-field of Artificial intelligence that is concerned with analyzing, modeling and understanding human language using computational methods.
2. It explores how humans can interact with computers in human languages
3. The eventual goal is to make computers understand (and generate) human languages, and make them communicate with humans like humans

Where is NLP used in real-world?

1. Apple Siri and other such software that can understand and interpret human speech (okay, partially)
2. Google Translate and the likes
3. Search Engines
4. Question Answering (e.g., IBM Watson)
5. News recommendation - related articles features in News websites
6. Sentiment analysis of product reviews on Amazon, for example
7. Spam classification in Gmail, Yahoomail etc
8. Information extraction from text (e.g., identifying calendar entries automatically in gmail)
9. Dialog systems (having interactive conversations with users, to do flight bookings etc)
10. Spelling and grammar checkers

... and many more.

What makes NLP challenging (and useful)?

... to understand that, let us look at some of the tasks involved through some fun video demos of cutting edge technologies.

Where is NLP useful? -1

Google Home demo

older one (2016): `https://www.youtube.com/watch?v=2KpLHdAURGo`

Where is NLP useful? -1

Google Home demo

older one (2016): `https://www.youtube.com/watch?v=2KpLHdAURGo`

latest (Google Duplex):
`https://www.youtube.com/watch?v=d40jgFZ5hXk`

Where is NLP useful? -2

Maluuba reading comprehension demo

Machine reading comprehension is this task where machine reads a text and answers questions about it.
Let us watch this demo video:
`https://www.youtube.com/watch?v=5UXsPtyBlhs`

Another Maluuba demo about reading stories and news articles (watch later):
`https://www.youtube.com/watch?v=QUwsAPO15_U`

Where is NLP useful? -3

from 2011: Watson beats humans in Jeopardy

https://www.youtube.com/watch?v=WFR3lOm_xhE

Okay, this is HCI - where is NLP?

The eventual applications involve spoken interactions with humans, but to develop such applications, that machine should be able to process text and understand it. Without that happening in the background, the foreground will not exist!

Some NLP tasks

Let us take a small text snippet

"'I'm not big, I'm not fancy," she said as she sat in a booth, looking out the window to Lincoln Way. "But I don't mind."

Chinese Homestyle Cooking's owners, Tina and Chung Song, have run the restaurant for almost 20 years from the small building off the corner of Sheldon Avenue and Lincoln Way. But in late October, they'll close their business when the lease on their building runs out.

Tina's start in the restaurant business came in 1982, when she emigrated from Taiwan and began working as a waitress for her sister's restaurants in Des Moines and Ankeny. While in the U.S., she got a call from Chung, who she had grown up with as a child in their home country.

Source: Ames Tribune (`http://goo.gl/zvx9Uw`)

1. When she says "I" in the first sentence, does she mean herself literally?

Let us take a small text snippet

"'I'm not big, I'm not fancy," she said as she sat in a booth, looking out the window to Lincoln Way. "But I don't mind."

Chinese Homestyle Cooking's owners, Tina and Chung Song, have run the restaurant for almost 20 years from the small building off the corner of Sheldon Avenue and Lincoln Way. But in late October, they'll close their business when the lease on their building runs out.

Tina's start in the restaurant business came in 1982, when she emigrated from Taiwan and began working as a waitress for her sister's restaurants in Des Moines and Ankeny. While in the U.S., she got a call from Chung, who she had grown up with as a child in their home country.

Source: Ames Tribune (`http://goo.gl/zvx9Uw`)

1. When she says "I" in the first sentence, does she mean herself literally?
2. What is she referring to? When will we know what is she referring to?

Let us take a small text snippet

"'I'm not big, I'm not fancy," she said as she sat in a booth, looking out the window to Lincoln Way. "But I don't mind."

Chinese Homestyle Cooking's owners, Tina and Chung Song, have run the restaurant for almost 20 years from the small building off the corner of Sheldon Avenue and Lincoln Way. But in late October, they'll close their business when the lease on their building runs out.

Tina's start in the restaurant business came in 1982, when she emigrated from Taiwan and began working as a waitress for her sister's restaurants in Des Moines and Ankeny. While in the U.S., she got a call from Chung, who she had grown up with as a child in their home country.

Source: Ames Tribune (`http://goo.gl/zvx9Uw`)

1. When she says "I" in the first sentence, does she mean herself literally?
2. What is she referring to? When will we know what is she referring to?
3. Who is "She"?

Let us take a small text snippet

"'I'm not big, I'm not fancy," she said as she sat in a booth, looking out the window to Lincoln Way. "But I don't mind."

Chinese Homestyle Cooking's owners, Tina and Chung Song, have run the restaurant for almost 20 years from the small building off the corner of Sheldon Avenue and Lincoln Way. But in late October, they'll close their business when the lease on their building runs out.

Tina's start in the restaurant business came in 1982, when she emigrated from Taiwan and began working as a waitress for her sister's restaurants in Des Moines and Ankeny. While in the U.S., she got a call from Chung, who she had grown up with as a child in their home country.

Source: Ames Tribune (`http://goo.gl/zvx9Uw`)

1. When she says "I" in the first sentence, does she mean herself literally?
2. What is she referring to? When will we know what is she referring to?
3. Who is "She"?
4. What is "home country" in the last sentence?

Let us take a small text snippet

"'I'm not big, I'm not fancy," she said as she sat in a booth, looking out the window to Lincoln Way. "But I don't mind."

Chinese Homestyle Cooking's owners, Tina and Chung Song, have run the restaurant for almost 20 years from the small building off the corner of Sheldon Avenue and Lincoln Way. But in late October, they'll close their business when the lease on their building runs out.

Tina's start in the restaurant business came in 1982, when she emigrated from Taiwan and began working as a waitress for her sister's restaurants in Des Moines and Ankeny. While in the U.S., she got a call from Chung, who she had grown up with as a child in their home country.

Source: Ames Tribune `http://goo.gl/zvx9Uw`

1. What is the main event of this text?

Let us take a small text snippet

"'I'm not big, I'm not fancy," she said as she sat in a booth, looking out the window to Lincoln Way. "But I don't mind."

Chinese Homestyle Cooking's owners, Tina and Chung Song, have run the restaurant for almost 20 years from the small building off the corner of Sheldon Avenue and Lincoln Way. But in late October, they'll close their business when the lease on their building runs out.

Tina's start in the restaurant business came in 1982, when she emigrated from Taiwan and began working as a waitress for her sister's restaurants in Des Moines and Ankeny. While in the U.S., she got a call from Chung, who she had grown up with as a child in their home country.

Source: Ames Tribune `http://goo.gl/zvx9Uw`

1. What is the main event of this text?
2. What is "Chinese Homestyle Cooking" referring to?

Let us take a small text snippet

"'I'm not big, I'm not fancy," she said as she sat in a booth, looking out the window to Lincoln Way. "But I don't mind."

Chinese Homestyle Cooking's owners, Tina and Chung Song, have run the restaurant for almost 20 years from the small building off the corner of Sheldon Avenue and Lincoln Way. But in late October, they'll close their business when the lease on their building runs out.

Tina's start in the restaurant business came in 1982, when she emigrated from Taiwan and began working as a waitress for her sister's restaurants in Des Moines and Ankeny. While in the U.S., she got a call from Chung, who she had grown up with as a child in their home country.

Source: Ames Tribune `http://goo.gl/zvx9Uw`

1. What is the main event of this text?
2. What is "Chinese Homestyle Cooking" referring to?
3. What is the relationship between "Chinese Homestyle cooking" and Tina?

Let us take a small text snippet

"'I'm not big, I'm not fancy," she said as she sat in a booth, looking out the window to Lincoln Way. "But I don't mind."

Chinese Homestyle Cooking's owners, Tina and Chung Song, have run the restaurant for almost 20 years from the small building off the corner of Sheldon Avenue and Lincoln Way. But in late October, they'll close their business when the lease on their building runs out.

Tina's start in the restaurant business came in 1982, when she emigrated from Taiwan and began working as a waitress for her sister's restaurants in Des Moines and Ankeny. While in the U.S., she got a call from Chung, who she had grown up with as a child in their home country.

Source: Ames Tribune `http://goo.gl/zvx9Uw`

1. What is the main event of this text?
2. What is "Chinese Homestyle Cooking" referring to?
3. What is the relationship between "Chinese Homestyle cooking" and Tina?
4. Is Lincoln Way something related to President Lincoln?

Each question I asked is an NLP problem which is not completey solved yet!

Tasks in NLP and the use of Python for NLP

Some libraries I will use today

- For linguistic analysis: NLTK, spacy
- For using machine learning with language data: sklearn, tensorflow, keras
- Text generation example: textrnn library
- Dataset for text classification: a publicly available movie review dataset for sentiment classification.

Code and details at:
`https://github.com/nishkalavallabhi/MBDS2018-NLPTutorial`

Some NLP tasks: Tokenization

- Sentence level tokenization: splitting a text into sentences
- Word level tokenization: splitting a text into tokens (words, punctuations etc)

[Question: What is so particularly challenging about these? Aren't they pretty straight forward?]

What is the big deal about tokenizing?

Issues to consider in a tokenizer

- How many tokens are there in this sentence: "Hmm, I worry uh a lot about next week."

What is the big deal about tokenizing?

Issues to consider in a tokenizer

- How many tokens are there in this sentence: "Hmm, I worry uh a lot about next week."
- Should C.N.N be one token as CNN, or three word tokens?

What is the big deal about tokenizing?

Issues to consider in a tokenizer

- How many tokens are there in this sentence: "Hmm, I worry uh a lot about next week."
- Should C.N.N be one token as CNN, or three word tokens?
- Should phone numbers: 555-333-222 be split into three tokens or one (don't forget it is not written like this all over the world)

What is the big deal about tokenizing?

Issues to consider in a tokenizer

- How many tokens are there in this sentence: "Hmm, I worry uh a lot about next week."
- Should C.N.N be one token as CNN, or three word tokens?
- Should phone numbers: 555-333-222 be split into three tokens or one (don't forget it is not written like this all over the world)
- Mr. Anderson - one token or two?

What is the big deal about tokenizing?

Issues to consider in a tokenizer

- How many tokens are there in this sentence: "Hmm, I worry uh a lot about next week."
- Should C.N.N be one token as CNN, or three word tokens?
- Should phone numbers: 555-333-222 be split into three tokens or one (don't forget it is not written like this all over the world)
- Mr. Anderson - one token or two?
- Doesn't - one or two?

What is the big deal about tokenizing?

Issues to consider in a tokenizer

- How many tokens are there in this sentence: "Hmm, I worry uh a lot about next week."
- Should C.N.N be one token as CNN, or three word tokens?
- Should phone numbers: 555-333-222 be split into three tokens or one (don't forget it is not written like this all over the world)
- Mr. Anderson - one token or two?
- Doesn't - one or two?
- Agent Smith's Matrix - how many tokens are there in Agent Smith's?

What is the big deal about tokenizing?

Issues to consider in a tokenizer

- URLs: Should they considered single token? or split at every underscores, slash etc?
- Chicago-Des Moines flight: If we split this on space, Chicago-Des is one token.
- But splitting on - seperates part-time which is one token.
- Some words are compound words (like with some long German nouns). What will we use to split such words?

What is the big deal about sentence splitting?

Just splitting in full-stop or ? or ! will not do.

- People don't follow conventions or grammar sometimes. Missing capitalization at the start of a sentence, not leaving a space after sentence breaker etc.
- Spoken language, tweets etc - do not follow same conventions as news articles. This diversity may affect the accuracy of our sentence splitting rules.

Tokenization and Sentence Splitting in Python

```
from nltk.tokenize import sent_tokenize,word_tokenize
content = open("text1.txt").read()
sentences = sent_tokenize(content)
for sentence in sentences:
    words_in_this_sentence = word_tokenize(sentence)
    print(sentence)
    print(words_in_this_sentence)
```

Note: My examples are with English texts. But, many of these tools work for a few other languages too.

Some NLP tasks: Pattern Extraction

- Task: Extract the language patterns that exist in textual data (e.g., all dates, all phone numbers, emails, postal addresses etc in text documents).
- Regular expressions are very useful for this.
- More advanced methods (which rely on machine learning) exist to extract unknown patterns from unstructured text documents.

Simple Pattern Extraction Example

```
import re
pattern = re.compile("\d{3}-\d{3}-\d{4}")
string = open("text3.txt").read()
matches = re.findall(pattern,string)
print(matches)
```

-What does this do?

Some NLP tasks: POS Tagging

What is the big deal about automatic tagging?

- Task: Given a sequence of words, return the POS tags for each word.
- An example problem: What is the best tag for a word in a context?
  - I wish to cite this work.
    PRP/I VBP/wish TO/to VB/cite DT/this NN/work ./.
  - He has a wish.
    PRP/He VBZ/has DT/a NN/wish ./.

POS Tagging and Python

```
from nltk.tokenize import sent_tokenize,word_tokenize
from nltk.tag import pos_tag
content = open("text1.txt").read()
sentences = sent_tokenize(content)
for sentence in sentences:
    words_in_this_sentence = word_tokenize(sentence.strip())
    print(sentence.strip())
    print(pos_tag(words_in_this_sentence))
    print()
```

Some NLP tasks: Identifying Named Entities, Noun Chunks etc.

- Task: Identify words that indicate names of persons/organizations etc. Identify groups of noun words that go together in a sentence.
- Use: Information Extraction, Answering questions, search etc.
- Next slide: examples using spaCy Python library.

Noun Chunk Extraction in Python

```
import spacy
from nltk.tokenize import sent_tokenize

nlp = spacy.load('en_core_web_sm')
content = open("text1.txt").read()
sentences = sent_tokenize(content)
for sentence in sentences:
    print(sentence.strip())
    spacified = nlp(sentence.strip())
    for nc in spacified.noun_chunks:
        print(nc)
    print("")
```

Note: You don't have to use nltk here. I am using it just to show you can use parts of different libraries.

Entity Extraction in Python

```
import spacy
from nltk.tokenize import sent_tokenize

nlp = spacy.load('en_core_web_sm')
content = open("text1.txt").read()
sentences = sent_tokenize(content)
for sentence in sentences:
    print(sentence.strip())
    spacified = nlp(sentence.strip())
    for entity in spacified.ents:
        print(entity.text, entity.label_,sep="\t")
    print("")
```

Some Common NLP Tasks: Parsing

- Goal: Build the syntactic structure of a sentence (phrasal structure, or relationship between words such as subject-object etc)
- Use: To do many things e.g., understanding the meaning of a sentence, extract information, answer questions etc.

```
import spacy
from spacy import displacy
sentence = nlp("This is a small sentence to show how a
dependency tree looks like."))
displacy.serve(sentence, style='dep')
```

Some NLP tasks: Language Generation

- Task: Generate text automatically.
- Texts should be grammatically and semantically correct. Should be human like.
- Example uses: Create weather reports, match summaries, reports etc. automatically (without human intervention!)
- There are some software libraries that support the development of NLG systems for some languages currently.

Language Generation in Python

```
from textgenrnn import textgenrnn
textgen = textgenrnn()
textgen.train_from_file('obamaspeechescorpus.txt', num_epochs=1)
textgen.generate_samples()
textgen.generate_to_file('generatedobama.txt', n=5)
generated_texts = textgen.generate(n=5, prefix="America",
temperature=0.2, return_as_list=True)
for text in generated_texts:
  print(text)
```

Source: https://github.com/minimaxir/textgenrnn/blob/master/docs/textgenrnn-demo.ipynb Corpus:

http://www.thegrammarlab.com/?nor-portfolio=corpus-of-presidential-speeches-cops-and-a-clintontrump-corpus

This is not it.

- Obviously, there are more things you can do, and more challenging ones too.
- These libraries that I used provide several powerful functions you can use as starting points.
- Other things you can try:
  - Google Cloud API (I will show one example)
  - Microsoft Azure's Cognitive Services API

Text Classification

Text Classification

- Goal: Learn to categorize text into a set of known categories.
- Example: Email spam classification
- Process: First, show a lot of example texts for each category, and decide on a "feature" representation for text (e.g., representing text as a vector of words in the vocabulary)
- The classification algorithm will then "learn" patterns from the feature vectors, to separate between categories.

Text Classification and Python

- For feature representation: NLTK, Gensim, Spacy etc.
- For classification algorithms: sklearn, keras etc.

Text Classification and Python - Example

- Task: Classification of movie reviews into positive or negative sentiment.
- Data: 1000 examples of positive, and 1000 examples of negative reviews.
- Source of data: `http://www.cs.cornell.edu/people/pabo/movie-review-data/`
- What will I do?: Build a small text classification model, evaluate it, and use it on new reviews.

[Over to code!]

Resources for further study

Online courses etc

1. Coursera courses: Introduction to NLP by Dragomir Radev
2. There are two other courses: one taught by Jurafsky and Manning and another by Michael Collins.
3. I don't think the courses are offered now. You may be able to get lecture videos online somewhere though.
4. Deep Learning for NLP - Stanford course video lectures: `https://goo.gl/hU1YNV`

Text Books

1. Textbook 1: Speech and Language Processing by Jurafsky & Martin (2nd Edition)
   - 2nd Edition is the actual print book, but 3rd Edition draft chapters are already available for free. `https://web.stanford.edu/~jurafsky/slp3/`
2. Textbook 2: Foundations of Statistical Natural Language Processing by Manning and Schütze

General references, resources etc

- NACLO website - good resource for some brainstorming about language processing problems.
  `http://www.nacloweb.org/`
- Access to various publications:
  `http://aclweb.org/anthology/`
- Information about resources for different languages: ACL Wiki
  `http://www.aclweb.org/aclwiki`
- Know about other NLP courses around the world etc:
  ACLWeb again
- Lot of code, datasets and tutorials shared on github
- Brand new results shared on ArXiv pre-print server

General references, resources etc

- NACLO website - good resource for some brainstorming about language processing problems.
  `http://www.nacloweb.org/`
- Access to various publications:
  `http://aclweb.org/anthology/`
- Information about resources for different languages: ACL Wiki
  `http://www.aclweb.org/aclwiki`
- Know about other NLP courses around the world etc:
  ACLWeb again
- Lot of code, datasets and tutorials shared on github
- Brand new results shared on ArXiv pre-print server

Lot of free resources, tutorials shared online. Lot of new jobs being posted. This is a good time to work in NLP!

Thanks!
Questions?

contact: sowmya@iastate.edu

Note: Slides and code examples are at:
`https://github.com/nishkalavallabhi/MBDS2018-NLPTutorial`