



COGS 108 – DATA SCIENCE IN PRACTICE

(DR. BRAD VOYTEK)

ENLIN WEI

SECTIONS A01 & A04

(MONDAYS & FRIDAYS 2PM-3PM)

OH FRIDAYS 12-2PM @ PC ROOM 4

NATURAL LANGUAGE PROCESSING (NLP)

- Helping computers understand and use human language
 - Syntax and usage of words (“tree” is a noun, “move is a verb...but “move” can also be a noun in certain contexts, etc.)
 - Association of certain sentiments with certain words (“dislike” vs. “hate”)
 - Association of certain words with others (“cell” with “biology”, not “rock music”)

NATURAL LANGUAGE PROCESSING (NLP)

- Helping computers understand and use human language

Text

Documents

DETECT LANGUAGE

JAPANESE

CHINESE

ENGLISH

↔

ENGLISH

CHINESE (SIMPLIFIED)

JAPANESE

"Oceanview" dining hall doesn't give good a view of the ocean. I propose calling it "Oceanglimpse".

×

“海景”餐厅无法欣赏到海洋美景。我建议称其为“海洋动物”。

☆

“Hǎijǐng” cāntīng wúfǎ xīnshǎng dào hǎiyáng měijǐng. Wǒ jiànyì chēng qí wèi “hǎiyáng dòngwù”.

99/5000

🔊

📄

✎

🔗

NATURAL LANGUAGE PROCESSING (NLP)

- Helping computers understand and use human language

Text

Documents

DETECT LANGUAGE

JAPANESE

CHINESE

ENGLISH

↔

ENGLISH

CHINESE (SIMPLIFIED)

JAPANESE

"Oceanview" dining hall doesn't give good a view of the ocean. I propose calling it "Oceanglimpse".

×

“海景”餐厅无法欣赏到海洋美景。我建议称其为“海洋动物”。 “ocean fauna” != “oceanglimpse”

☆

“Hǎijǐng” cāntīng wúfǎ xīnshǎng dào hǎiyáng měijǐng. Wǒ jiànyì chēng qí wèi “hǎiyáng dòngwù”.

99/5000

🔊

📄

✎

🔗

WHEN NLP FAILS...

- “Let it Go” translated through 6 or 7 layers of languages and then back into english...

The background is a blue gradient with decorative white circuit-like lines in the corners. These lines consist of straight segments and small circles, resembling a stylized electronic circuit or neural network structure.

NATURAL LANGUAGE TOOLKIT (NLTK)

WHAT IS TOKENIZING?

- <https://www.guru99.com/tokenize-words-sentences-nltk.html>

The background is a blue gradient with decorative white circuit-like lines in the corners. These lines consist of straight segments and small circles, resembling a stylized electronic circuit board.

WHAT IS TF-IDF?



TF-IDF Explained

- So a measure of the relevancy of a word to a document might be:

$$\frac{\textit{Term Frequency}}{\textit{Document Frequency}}$$

Or: Term Frequency * Inverse Document Frequency

That is, take how often the word appears in a document, over how often it just appears everywhere. That gives you a measure of how important and unique this word is for this document

STOPWORDS

- Words that don't mean much if you're trying to consider meaning of a text as a whole; like "I" or "the" or "he"

STOP WORDS

- Words that don't mean much if you're trying to consider meaning of a sentence; like "I" or "the" or "he"

```
(base) C:\Users\enlin>python
Python 3.7.3 (default, Mar 27 2019, 17:13:21) [MSC v.1915 64 bit (AMD64)] :: Anaconda, Inc. on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> from nltk.corpus import stopwords
>>> stopwords
<WordListCorpusReader in '.../corpora/stopwords' (not loaded yet)>
>>> stopwords.words('english')
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]
>>>
```