COGS 108 – DATA SCIENCE IN PRACTICE

(DR. BRAD VOYTEK)

ENLIN WEI

SECTIONS A01 & A04

(MONDAYS & FRIDAYS 2PM-3PM)

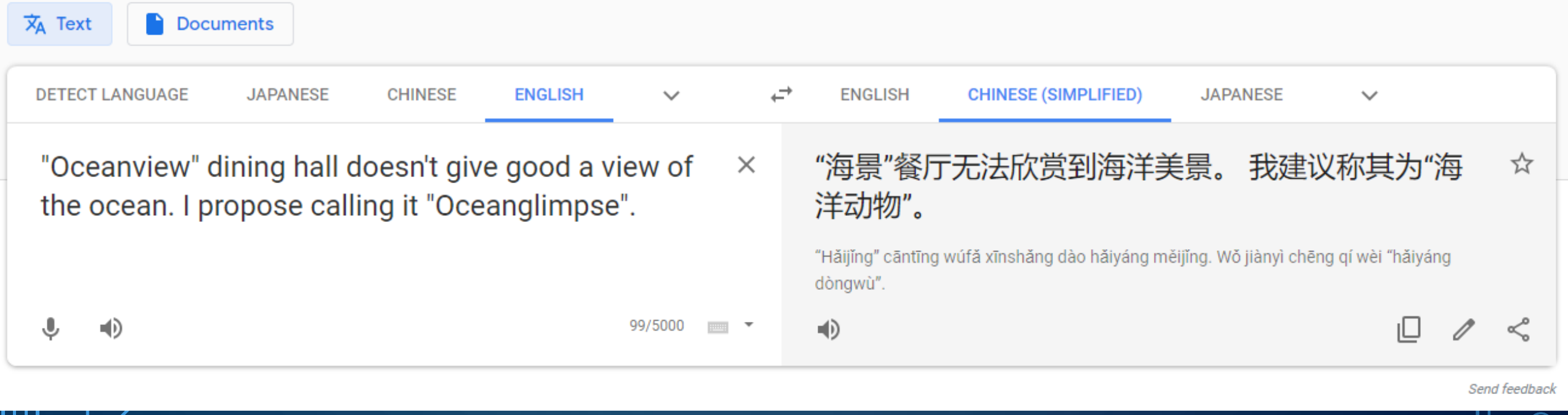
OH FRIDAYS 12-2PM @ PC ROOM 4

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”)

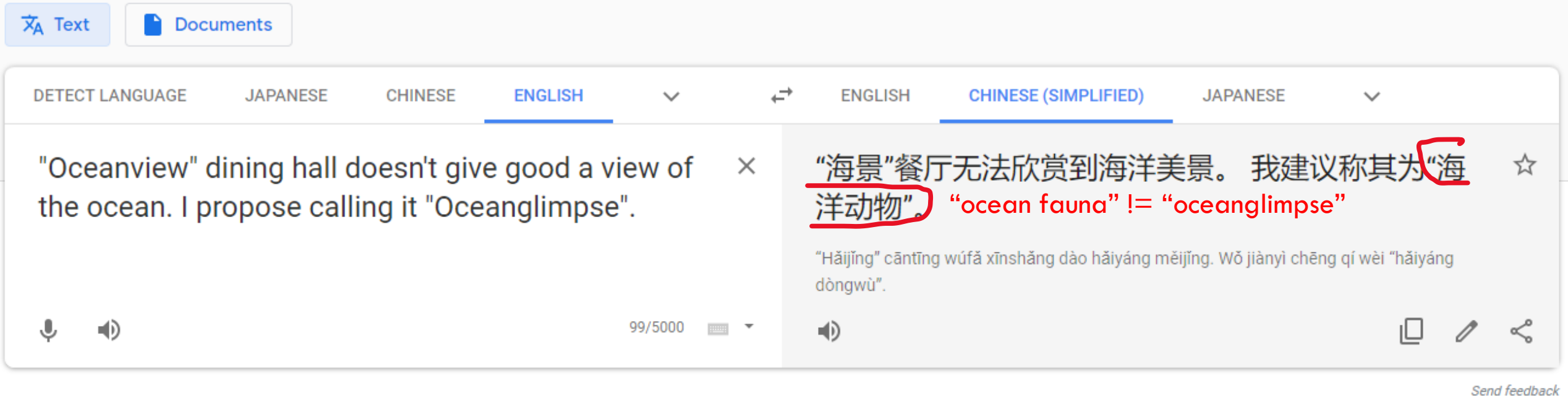
NATURAL LANGUAGE PROCESSING (NLP)

- Helping computers understand and use human language



NATURAL LANGUAGE PROCESSING (NLP)

- Helping computers understand and use human language



WHEN NLP FAILS…

- "Let it Go" translated through 6 or 7 layers of languages and then back into english…

NATURAL LANGUAGE TOOLKIT (NLTK)

WHAT IS TOKENIZING?

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

WHAT IS TF-IDF?

TF-IDF Explained

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

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

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', 'you
rs', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'i
ts', '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', 'doe
s', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'fo
r', '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", 'doe
sn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mus
tn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won
', "won't", 'wouldn', "wouldn't"]
>>>
```