# Natural Language Processing

## 15.S60 - Computing in Optimization and Statistics

Colin Pawlowski

MIT Operations Research Center

January 12, 2017

# Introduction to Text Analytics

The Big Question: **How do computers understand text?**

- So far, we have assumed that the inputs to our machine learning models are all numeric.
- To handle categorical variables for Lasso, we converted our data to an array of binary variables.
- Is there a way to process text directly?

Our goal is to translate *human language* into *programming language*.

# Bag-of-Words

The simplest method for processing text is to use **Bag-of-Words**.

- **Idea:** Ignore the order of words in each sentence and the word meanings, and just count the frequency of each word in the document.
- **Example:** "Twelve astronauts have walked on the moon, and over five hundred people have been in outer space. Currently, two astronauts from the USA are aboard the International Space Station."

Table: Example Bag-of-Words

| **Word** | aboard | and | are | astronauts | ... | walk |
|---|---|---|---|---|---|---|
| **Count** | 1 | 1 | 1 | 2 | ... | 1 |

- For each document, we obtain a vector of word counts.

# Bag-of-Words

Table: Example Bag-of-Words

| **Word** | aboard | and | are | astronauts | ... | walk |
|---|---|---|---|---|---|---|
| **Count** | 1 | 1 | 1 | 2 | ... | 1 |

- In addition, before running Bag-of-Words, we typically do some text pre-processing, including:
  - Converting all text to lower case
  - Removing all punctuation
  - Stemming the document (i.e. "walked" $\rightarrow$ "walk")
- Assuming that there are $N$ words in the dictionary, the final output of Bag-of-Words will be integer feature vectors in $\mathbb{R}^N$.
- We will use Bag-of-Words in R to build machine learning models using raw text data of Airbnb reviews from `reviews.csv`.

# Demo of IBM Watson Natural Language Classifier


`https://natural-language-classifier-demo.mybluemix.net/`