# Topic 2
# Getting data

Web scraping and BeautifulSoup

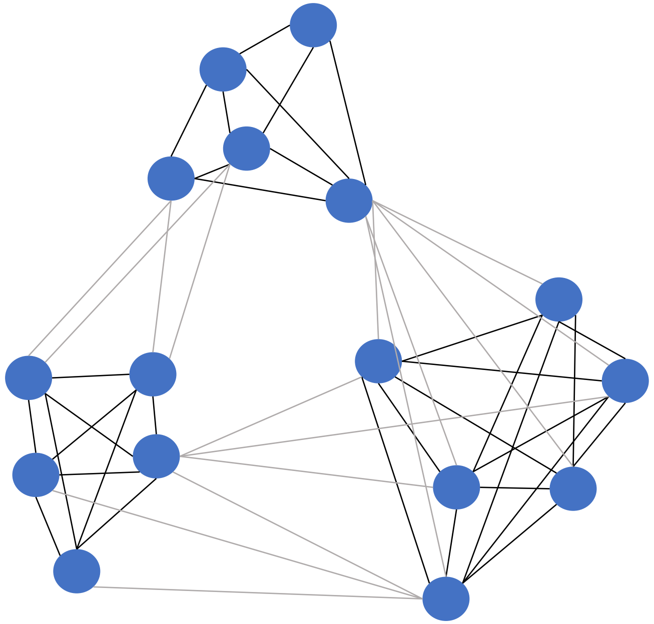
# Examples of Datasets

- How/Where do we find a dataset?
- Do we have to use just text when applying NLP methods?

# Public Datasets

- NLTK/SpaCy corpuses
- Kaggle (check licensing)
- Google datasets search engine
- Compiled lists of available datasets

# Networks



g1, g5, g4, g1, g7, …

g1, g4, g7, g5, g7, …

g2, g3, g4, g1, g7, …

g2, g3, g4, g1, g5, …

g3, g2, g4, g1, g7, …

.
.
.

The bigger the weight on an edge, the more likely a walk will take you through that edge

# We can extract dataset ourselves

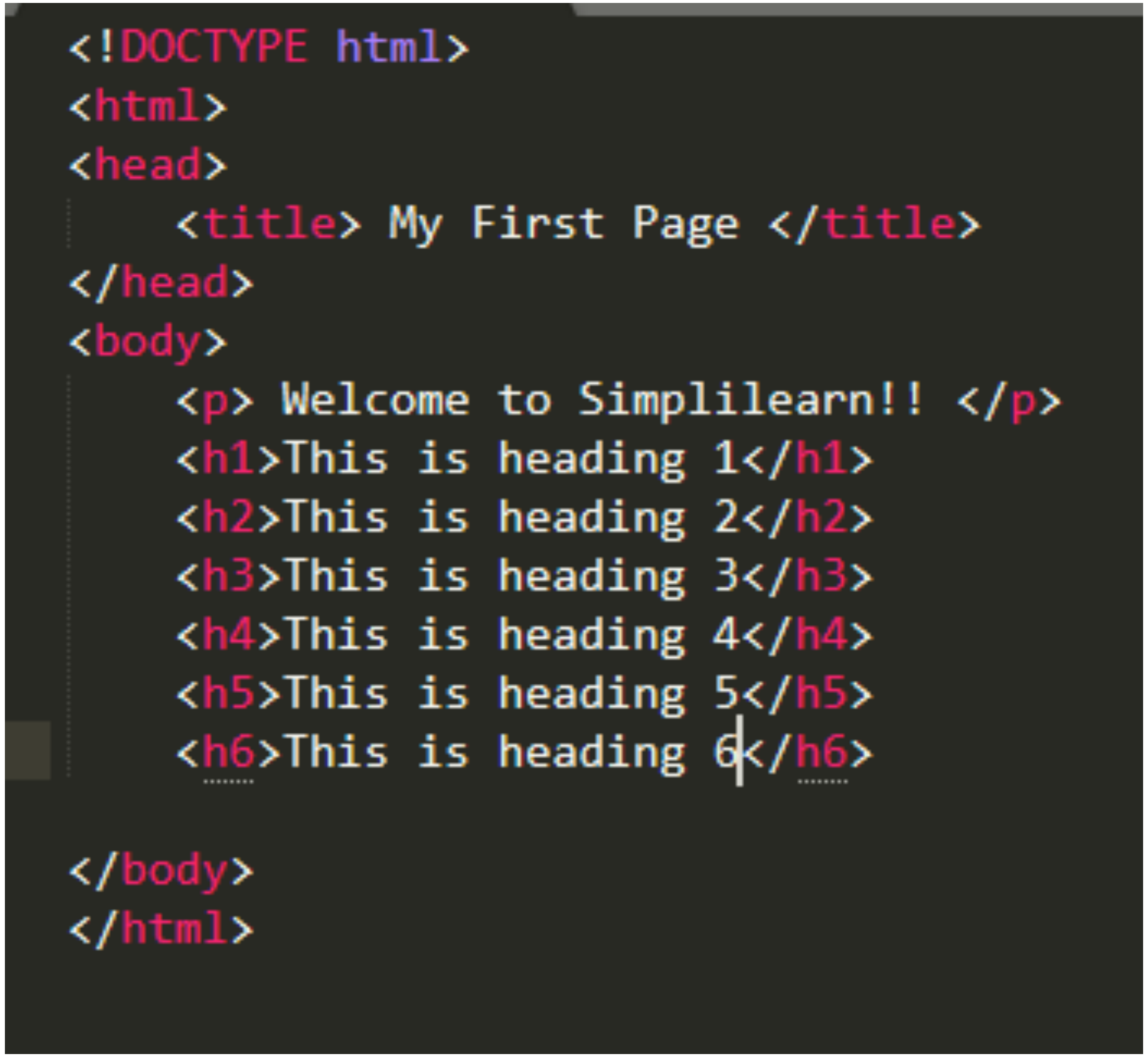
- BeautifulSoup documentation
  - Python library for pulling data out of HTML and XML
  - Making a soup

# Requests library

- Standard for making HTTP requests in Python
- Status code informs you of the status of the request
  - 200 means your request was successful

# Dealing with HTML

- Extracting information from between tags in BeautifulSoup with .text
- HTML parsers
- Regex



```html
<!DOCTYPE html>
<html>
<head>
    <title> My First Page </title>
</head>
<body>
    <p> Welcome to Simplilearn!! </p>
    <h1>This is heading 1</h1>
    <h2>This is heading 2</h2>
    <h3>This is heading 3</h3>
    <h4>This is heading 4</h4>
    <h5>This is heading 5</h5>
    <h6>This is heading 6</h6>

</body>
</html>
```

# IMSDB Example

- Open up the Jupyter Notebook for today and take a look at the IMSDB script extraction example

# Activity

- Choose a website that you would like to extract text from
  - Check the HTML to find the tags or class you wish to find
- Extract content using BeautifulSoup

# Web scraping Tips

- Sleep
- Randomize
- Make sure you are allowed to scrape the site!
- Develop using a local snapshot of the HTML
- Does size/value of data justify maintenance cost?
- Try to minimize dependence on markup details that seem most likely to change

# Augmenting Data

- Sometime, the data available isn’t enough for us to get anything useful and can take a lot of time, expertise and money that isn’t available

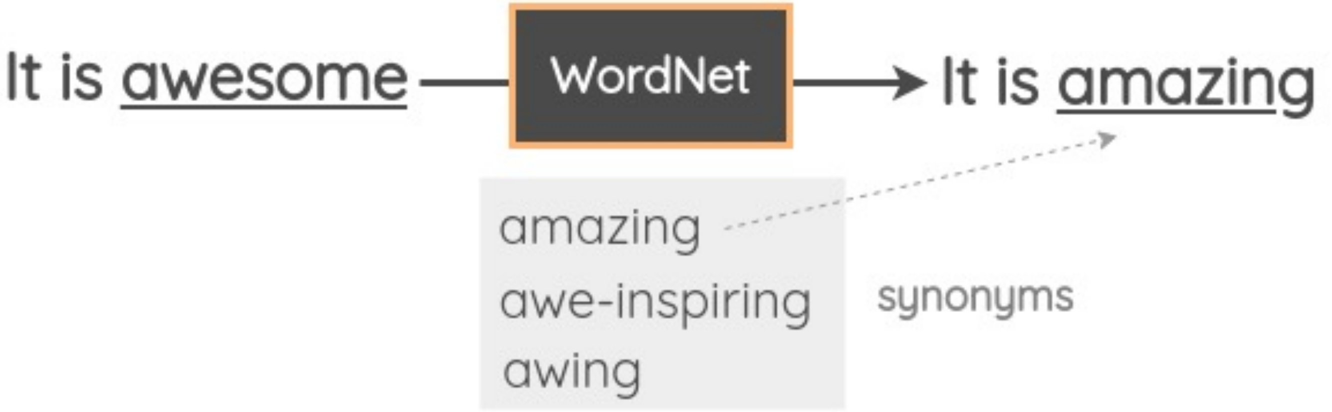
- **Augmentation: In the context of NLP, augmentation is the process of taking a small dataset and applying tricks/hacks in order to generate more data**

# Techniques for Augmenting Text

- **Synonym replacement**
- **TF-IDF-based word replacement**
- **Replacing entities**
- Bigram flipping
- Adding noise
- Back translation

More advanced:

- Snorkel
- EDA
- Active learning



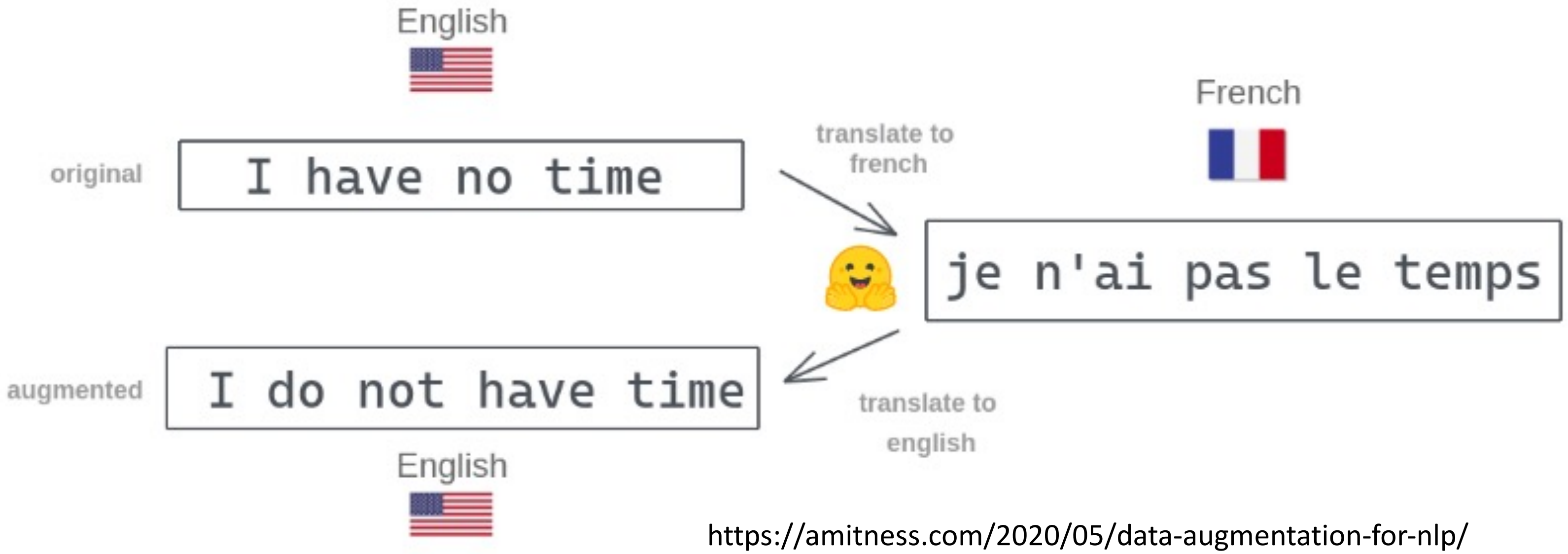
https://amitness.com/2020/05/data-augmentation-for-nlp/

# Techniques for Augmenting Text

- Synonym replacement
- TF-IDF-based word replacement
- Replacing entities
- **Bigram flipping**
- **Adding noise**
- **Back translation**

More advanced:

- Snorkel
- EDA
- Active learning



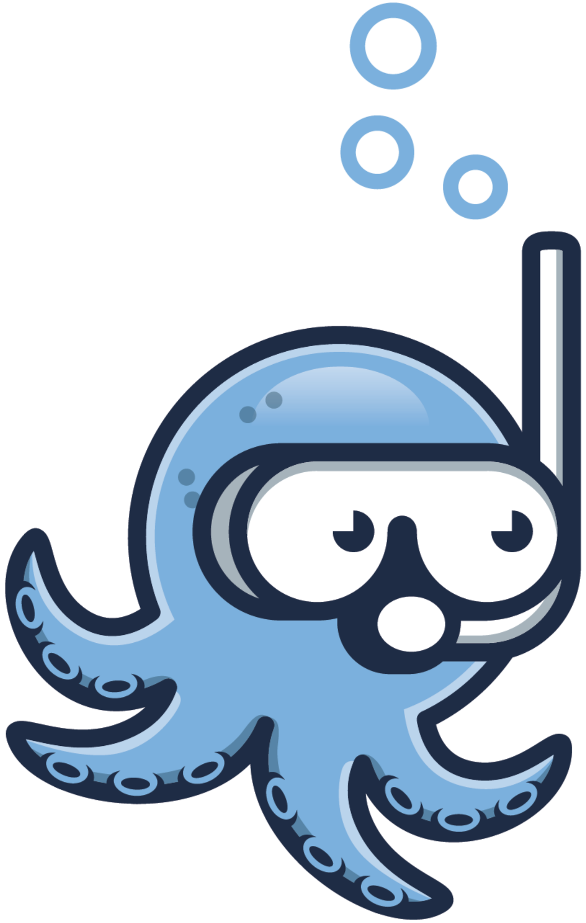
https://amitness.com/2020/05/data-augmentation-for-nlp/

# Techniques for Augmenting Text

- Synonym replacement
- TF-IDF-based word replacement
- Bigram flipping
- Replacing entities
- Adding noise
- Back translation

**More advanced:**

- **Snorkel**
- **EDA**
- **Active learning**



# Next time:

- Text representation