Word Embeddings

Michael Szczepaniak

A bit about me...

- 6 years as Sr. Data Scientist
  - operations research for defense contractor
  - top secret clearance
- 7 years as software engineer
  - couple of small companies
- Last two degrees
  - MS Data Science, RIT (2024)
    - Capstone: LLM assisted model development [1, 2]
  - BS Computer Science, CSU

RIT | Rochester Institute of Technology

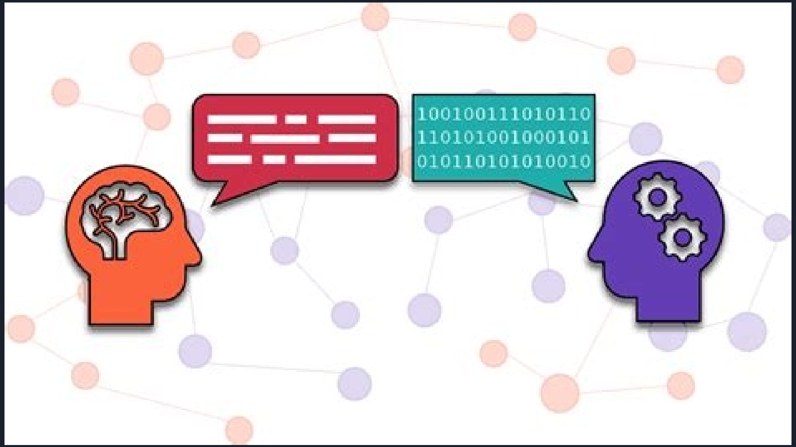
COLORADO STATE UNIVERSITY

What we'll cover

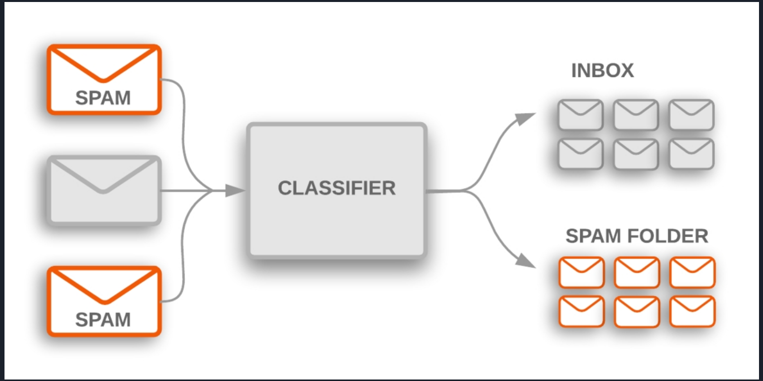
- A few important NLP tasks
- Getting ML models to understand text
  - text processing
  - one hot encoding vs. word embeddings
- Coding examples
  - word similarity
  - relationships between words
  - logistic regression
- Summary

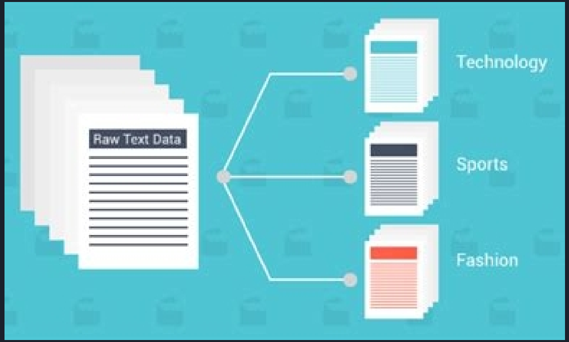
A few important NLP tasks

- Text classification
- Part of speech (PoS) tagging
- Named entity recognition (NER)



Text classification



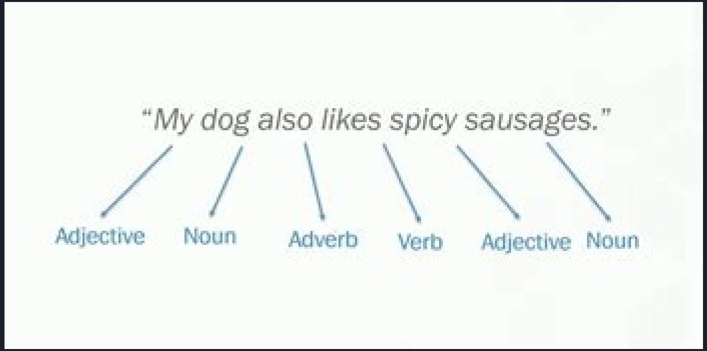


- What bucket do we put a document into?
    - binary: spam or not spam
    - multi-class: tech, sports or fashion
- Lots of models
    - Logistic regression
        - Why try this first?
    - Tree-base methods
    - Neural networks

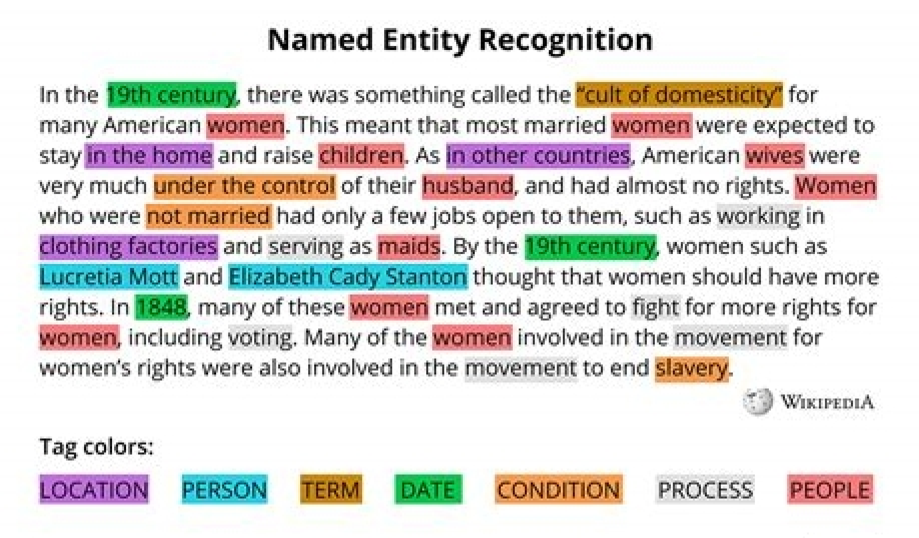
Part of Speach (PoS) tagging

- Identify what part of speech each word is in a sentence or document.
- First step in word sense disambiguation
  - WSD arises from words having more than one meaning:

I need to get to the bank to deposit my canoe.



Named Entity Recognition (NER)



Named Entity Recognition

In the 19th century, there was something called the "cult of domesticity" for many American women. This meant that most married women were expected to stay in the home and raise children. As in other countries, American wives were very much under the control of their husband, and had almost no rights. Women who were not married had only a few jobs open to them, such as working in clothing factories and serving as maids. By the 19th century, women such as Lucretia Mott and Elizabeth Cady Stanton thought that women should have more rights. In 1848, many of these women met and agreed to fight for more rights for women, including voting. Many of the women involved in the movement for women's rights were also involved in the movement to end slavery.

WIKIPEDIA

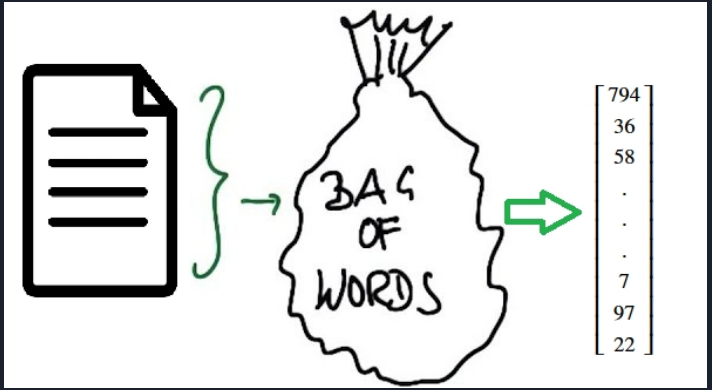
Tag colors:

LOCATION PERSON TERM DATE CONDITION PROCESS PEOPLE

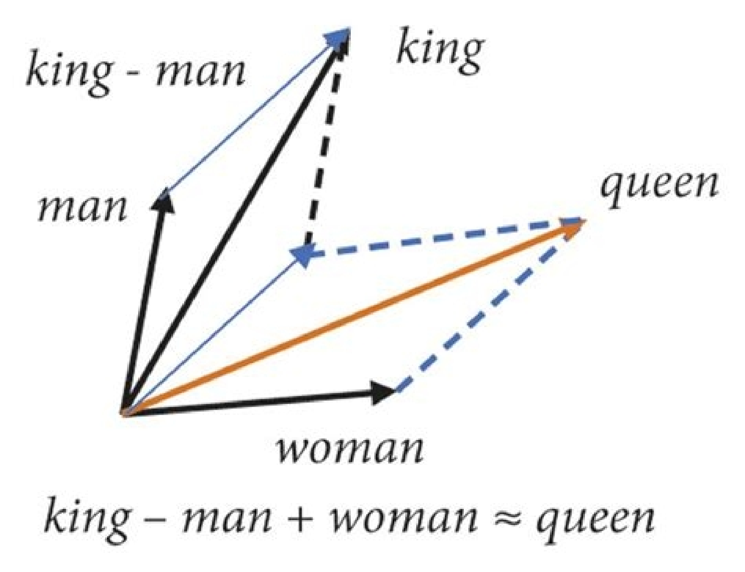
- Categorize specific types of words
  - person, organization, place/location,...
- Use cases
  - id customer names in customer service transcript
  - determine location of a user in a social media post
  - content discovery

Getting ML models to understand text

- Why we need encoding
  - models built for numeric inputs
- Simplest way to convert text to numbers
  - treat each word as a categorical variable: one hot encoding (vector of integers)
- Pros of one hot encoding
  - simple word encoding
  - simple document encoding
- Cons of one hot encoding
  - words have no relationship to each other
  - word vector size = | vocabulary | (27 - 57k words [5])



Getting ML models to understand text (cont.)



- Word embeddings capture relationships
  - vectors of real number (not integers)
- Number of dimensions: meta-parameter
  - performance vs. resolution trade-off
- Pros of word embeddings
  - captures relationship between words
  - higher information density
- Cons of word embeddings
  - must choose: canned vs. custom
  - must choose algorithm (e.g. GloVe, Word2vec, et. al.)

Coding examples

This section done in a python jupyter notebook

Summary

❑ A few of the important NLP tasks

- text classification
- PoS tagging
- NER
- WSD

❑ NLP ML model need numeric input

- one hot encoding (simplest)
- word embeddings (pre-trained /canned or custom)
- pros and cons: one hot encoding vs. word embeddings

Summary (cont.)

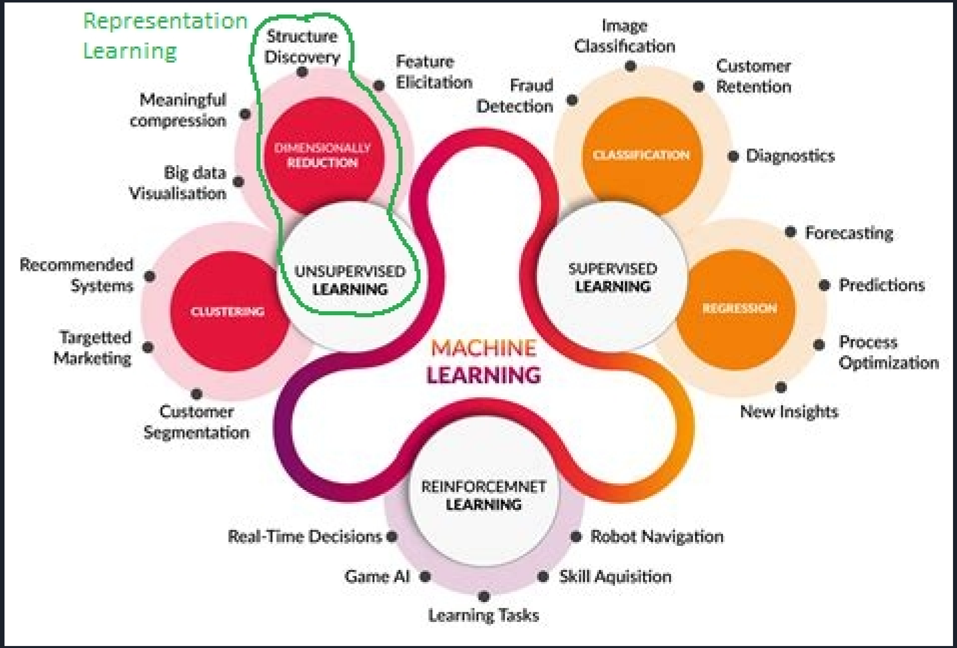
- ❑ Coding examples: word similarity & relationships
  - words with similar meaning closest together in embedding space
  - ([king] - [man]) more similar to ([queen] - [woman]) than ([cat] - [dog])

- ❑ Coding example: logistic regression
  - Doing logistic regression as a first model is a good idea
  - text processing
  - word encoding → document encoding
  - information density: word embeddings > one hot encoding

- ❑ Custom or canned embeddings?
  - Cost versus performance trade-off

Appendices

Appendix A - What are word embeddings?



- Mapping of words to vectors
- Results from representation learning
  - Unsupervised learning
- More than one way to create
  - GloVe (Stanford)
  - Word2Vec (Google)
  - Canned vs. Custom
  - Custom: must select algorithm

Appendix B - Text processing example

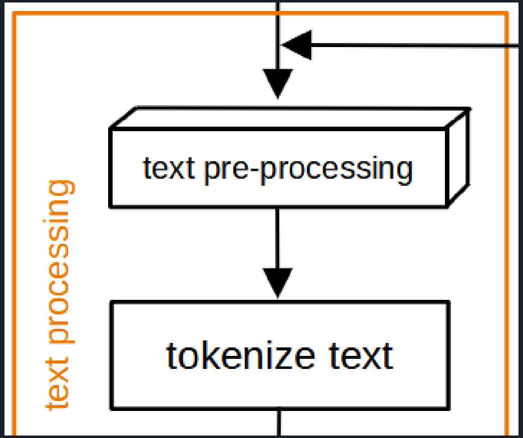
I can't bloody wait!! Sony Sets a Date For Stephen King⑨Ûªs ⑨Û÷The Dark Tower⑨Ûª
#stephenking #thedarktower http://t.co/J9LPdRXCDE @bdisgusting

➔ normalize URLs

I can't bloody wait!! Sony Sets a Date For Stephen King⑨Ûªs ⑨Û÷The Dark Tower⑨Ûª
#stephenking #thedarktower <url> @bdisgusting

➔ normalize twitter special chars

I can't bloody wait!! Sony Sets a Date For Stephen King⑨Ûªs ⑨Û÷The Dark Tower⑨Ûª
<hashtag>stephenking <hashtag>thedarktower <url> <user>bdisgusting



➔ expand contractions

I can not bloody wait!! Sony Sets a Date For Stephen King⑨Ûªs ⑨Û÷The Dark Tower⑨Ûª
<hashtag>stephenking <hashtag>thedarktower <url> <user>bdisgusting

➔ remove stop words

~~I can~~ not bloody wait!! Sony Sets ~~a~~ Date ~~For~~ Stephen King⑨Ûªs ⑨Û÷The Dark Tower⑨Ûª
<hashtag>stephenking <hashtag>thedarktower <url> <user>bdisgusting

➔ remove punc, lemmatize, lower case, remove singletons and OOV words

not bloody wait!! Sony Sets Date Stephen King⑨Ûªs ⑨Û÷The Dark Tower⑨Ûª
<hashtag>stephenking <hashtag>thedarktower <url> <user>bdisgusting

not bloody wait sony set date stephen dark <hashtag> <hashtag> <url> <user>

➔ tokenize

not | bloody | wait | sony | set | date | stephen | dark | <hashtag> | <hashtag> | <url> | <user>

References

1. Large Language Model Assisted Model Development - Final paper
   https://github.com/MichaelSzczepaniak/llmamd/blob/main/docs/Final_paper.pdf

2. Large Language Model Assisted Model Development project code, data, notebooks
   https://github.com/MichaelSzczepaniak/llmamd

3. Word Embeddings notebook and supporting materials
   https://github.com/MichaelSzczepaniak/WordEmbeddings/

4. What is Named Entity Recognition (NER): Benefits, Use Cases
   https://www.expressanalytics.com/blog/what-is-named-entity-recognition-ner-benefits-use-cases-algorithms/

5. How many words do we know?
   https://doi.org/10.3389/fpsyg.2016.01116