## Vector Models and Text Preprocessing

Count Vectorizer

# https://www.kaggle.com/shivamkushwaha/bbc-full-text-document-classification

!wget -nc https://lazyprogrammer.me/course\_files/nlp/bbc\_text\_cls.csv

TFIDF Recommender System

# https://www.kaggle.com/tmdb/tmdb-movie-metadata

!wget https://lazyprogrammer.me/course\_files/nlp/tmdb\_5000\_movies.csv

Word Embeddings Demo

# Slower but always guaranteed to work

!wget -nc https://lazyprogrammer.me/course\_files/nlp/GoogleNews-vectors-negative300.bin.gz

# You are better off just downloading this from the source

# https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing

# https://code.google.com/archive/p/word2vec/

# !gdown https://drive.google.com/uc?id=0B7XkCwpI5KDYNlNUTTlSS21pQmM

## Markov Models

Markov Model Classifier / Poetry Generator

!wget -nc https://raw.githubusercontent.com/lazyprogrammer/machine\_learning\_examples/master/hmm\_class/edgar\_allan\_poe.txt

!wget -nc https://raw.githubusercontent.com/lazyprogrammer/machine\_learning\_examples/master/hmm\_class/robert\_frost.txt

## Article Spinner

# https://www.kaggle.com/shivamkushwaha/bbc-full-text-document-classification

!wget -nc https://lazyprogrammer.me/course\_files/nlp/bbc\_text\_cls.csv

## Cipher Decryption

https://lazyprogrammer.me/course\_files/moby\_dick.txt

# is an edit of https://www.gutenberg.org/ebooks/2701

# (I removed the front and back matter)

Test text (note: you can use any text you like):

I then lounged down the street and found,

as I expected, that there was a mews in a lane which runs down

by one wall of the garden. I lent the ostlers a hand in rubbing

down their horses, and received in exchange twopence, a glass of

half-and-half, two fills of shag tobacco, and as much information

as I could desire about Miss Adler, to say nothing of half a dozen

other people in the neighbourhood in whom I was not in the least

interested, but whose biographies I was compelled to listen to.

## Spam Detection

# https://www.kaggle.com/uciml/sms-spam-collection-dataset

!wget https://lazyprogrammer.me/course\_files/spam.csv

## Sentiment Analysis

# https://www.kaggle.com/crowdflower/twitter-airline-sentiment

!wget -nc https://lazyprogrammer.me/course\_files/AirlineTweets.csv

## Text Summarization

# https://www.kaggle.com/shivamkushwaha/bbc-full-text-document-classification

!wget -nc https://lazyprogrammer.me/course\_files/nlp/bbc\_text\_cls.csv

## Topic Modeling

# https://www.kaggle.com/shivamkushwaha/bbc-full-text-document-classification

!wget -nc https://lazyprogrammer.me/course\_files/nlp/bbc\_text\_cls.csv

## Latent Semantic Analysis

!wget -nc https://raw.githubusercontent.com/lazyprogrammer/machine\_learning\_examples/master/nlp\_class/all\_book\_titles.txt

## The Neuron

# https://www.kaggle.com/crowdflower/twitter-airline-sentiment

!wget -nc https://lazyprogrammer.me/course\_files/AirlineTweets.csv

## ANN

TF2 ANN with TFIDF

# https://www.kaggle.com/shivamkushwaha/bbc-full-text-document-classification

!wget -nc https://lazyprogrammer.me/course\_files/nlp/bbc\_text\_cls.csv

## CNN

# https://www.kaggle.com/shivamkushwaha/bbc-full-text-document-classification

!wget -nc https://lazyprogrammer.me/course\_files/nlp/bbc\_text\_cls.csv

## RNN

RNN Text Classification

# https://www.kaggle.com/shivamkushwaha/bbc-full-text-document-classification

!wget -nc https://lazyprogrammer.me/course\_files/nlp/bbc\_text\_cls.csv

NER TF2

# conll 2003

!wget -nc https://lazyprogrammer.me/course\_files/nlp/ner\_train.pkl

!wget -nc https://lazyprogrammer.me/course\_files/nlp/ner\_test.pkl