Fletcher MVP

Concept

For project Fletcher I would like to spend time trying out different approaches and gaining confidence using different machine learning algorithms - both supervised and unsupervised.

Domain

Natural language processing, formatting the data to be ready for input (tokenization, lemmatization, etc.) supervised learning, unsupervised learning. Ultimately, I would like to be able to predict a review's numeric score from the text in it.

Data

Amazon Fine Food Reviews - dataset consists of reviews of fine foods from Amazon, spanning a period of more than 10 years, including all ~500,000 reviews up to October 2012.

Name	Data Type	Description
Id	Numeric	Unique ID of the review
ProductId	Numeric	The ID of the product being reviewed
UserId	Numeric	Unique identifier of the user writing the review
ProfileName	String	User name
HelpfulnessNumerator	Numeric	Upvoted on a review
HelpfulnessDenominator	Numeric	Downvotes on a review
Score	Numeric	The numeric score the user left for the product being reviewed
Time	DateTime	Date of the review
Summary	String	Title of the review
Text	String	The body of the review

Known unknowns

I have no prior experience with NLP outside of the pair problems and lectures, so I do not know how it will go. I am however very excited to try it out!

I have not yet decided how I'll approach the dataset. It is too big to be used inside a Jupyter notebook so I will either need to subset it or use and AWS EC2 instance.