Financial Sentiment Analysis

Jack Boydell | Andrew Raymond

MATH 280

Introduction to Dataset, Goals of Computation, Connections to In-Class Material

Dataset: Financial Sentiment Analysis dataset on Kaggle with data pulled from FiQA and Financial PhraseBank

Snapshot of the dataset

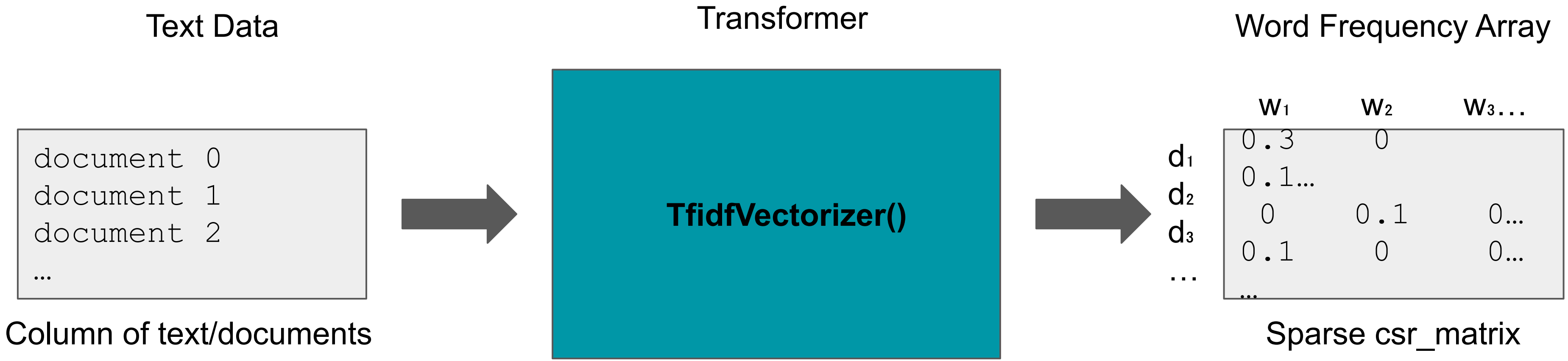
	Sentence	Sentiment
0	The GeoSolutions technology will leverage Bene...	positive
1	$ESI on lows, down$1.50 to $2.50 BK a real po...	negative
2	For the last quarter of 2010 , Componenta 's n...	positive

Goals: explore a Natural Language Processing application of interpretable singular value decomposition (topic extraction, word frequency), incorporate Machine Learning abilities, illustrate connections to in-class material

Connections: Singular Value Decomposition, Matrix Factorization (Multiplication), Kullback-Leibler Loss Function, Conditional Probability

Word Frequency Array with TfidfVectorizer()

The TfidfVectorizer() from the scikit-learn package in Python turns text data into a word frequency array with different words as the columns and each document/fragment of text as the rows. This allows us to work with text data in a numerical format.



Note: sklearn.feature_extraction TfidfVectorizer returns a sparse matrix, a matrix of mostly zero entries (this makes sense)… this is important to remember for the next slide!

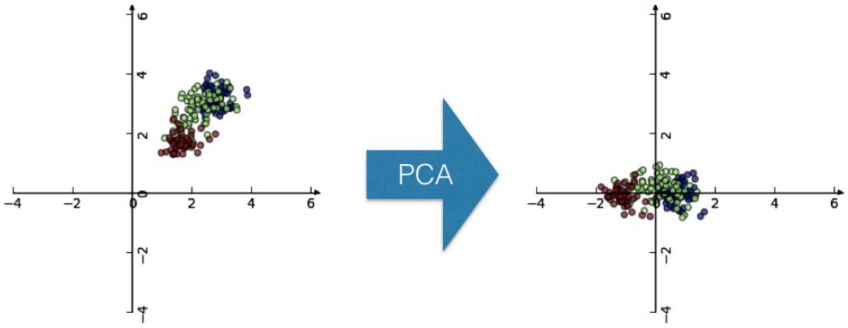
Singular Value Decomposition: PCA vs. Truncated SVD

Since the output of our Tfidf Vectorizer is a sparse matrix (a matrix of mostly zero entries), PCA is not an effective method for dimension reduction. The use of Truncated SVD allows us to find the number of principal components to include in our next step.

Recall the singular value decomposition application we learned about in class!

Traditional PCA

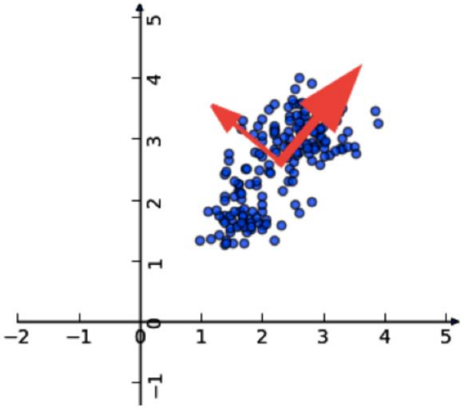
Centers the data with mean of 0 before performing SVD ✅



Images courtesy of Datacamp

Truncated SVD

Does not center the data before performing SVD ❌



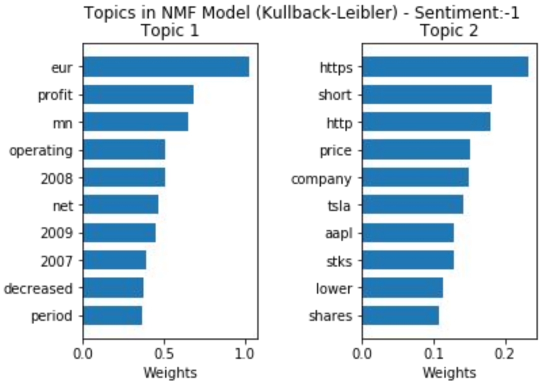
Non-negative Matrix Factorization (NMF)

Non-negative Matrix Factorization (NMF) is a dimension reduction technique that performs topic extraction when applied to text data in the form of a word frequency array. NMF finds two matrices (W, H) whose product approximates the non-negative matrix X (word frequency array)... matrix multiplication! We used the Kullback-Leibler loss function to be the minimized divergence between X and the product WH.

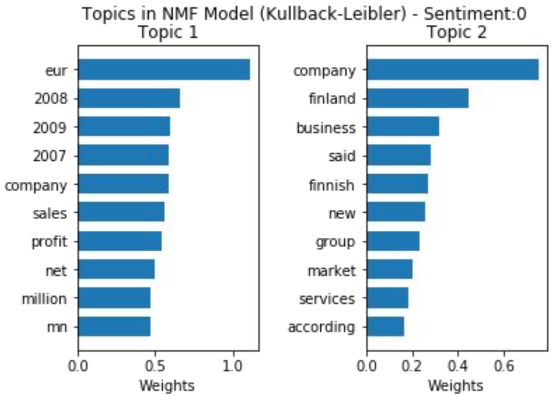
Recall the cross entropy loss function (Kullback-Leibler divergence) application we learned about in class!

NMF expresses documents as combinations of topics, represented in its model components. Samples can be reconstructed as (feature values x NMF components).

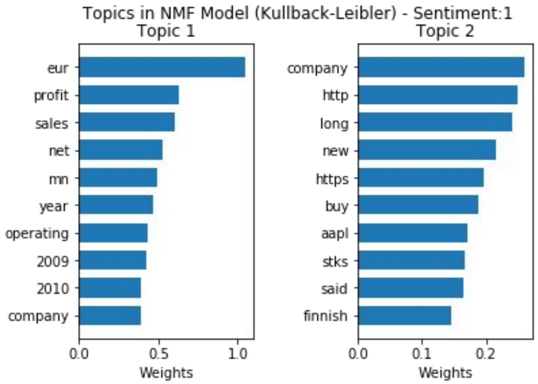
Negative



Neutral

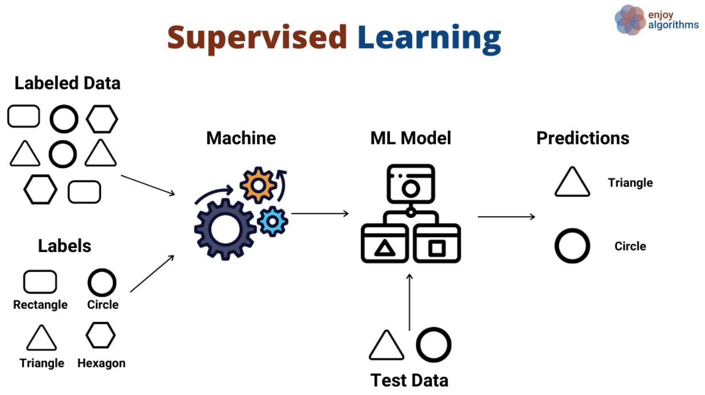


Positive



Further Analysis

- Using the word frequency array from NMF we can:
  - Creating predictive machine learning algorithms
  - Create features or predictors for classification models or neural networks
- However there are limitations
  - In this instance we attempted to create a basic classification model using the most occurring topics



Conditional Probability

- Probability the sentiment (positive, neutral, negative) given the word “profit” occurs in the sentence

Ex. for word: “Profit”	601/5842	Conditional Probability
Positive	226/601	~37%
Neutral	215/601	~35%
Negative	160/601	~26%

- “Profit” as a predictor would not be useful!
- This trend continued with the most frequent occurring words from the frequency array
- Given the limitations that the dataset provides training an effective classification model would require “slicker” features (predictors) and more