Corporate Annual Report Textual Analysis

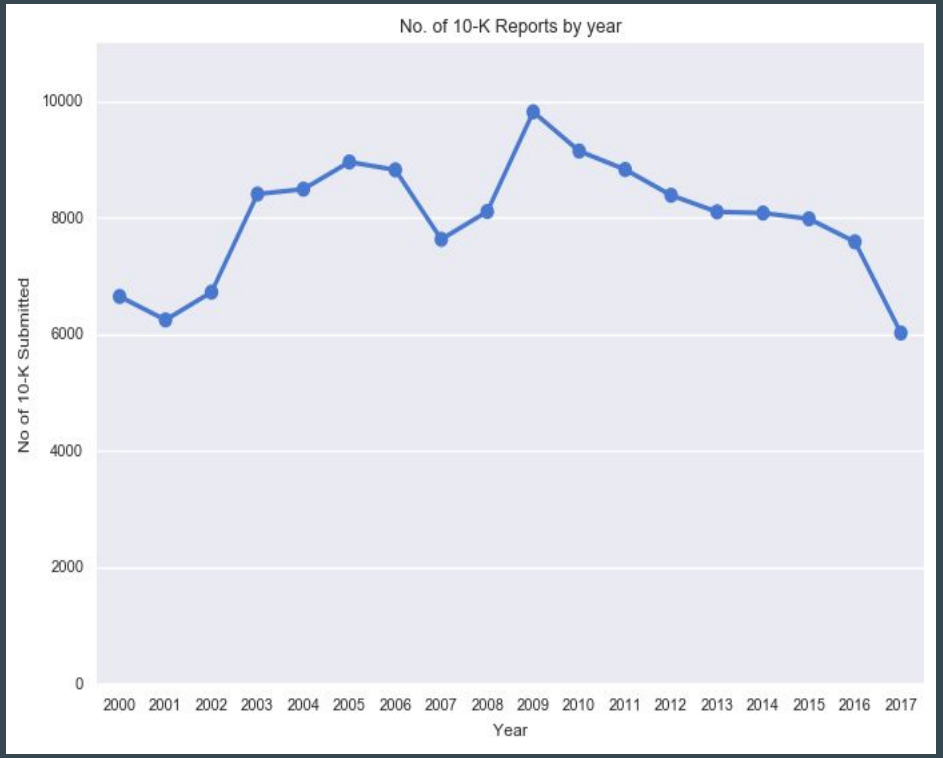
By Tribots Incorporated

Overview

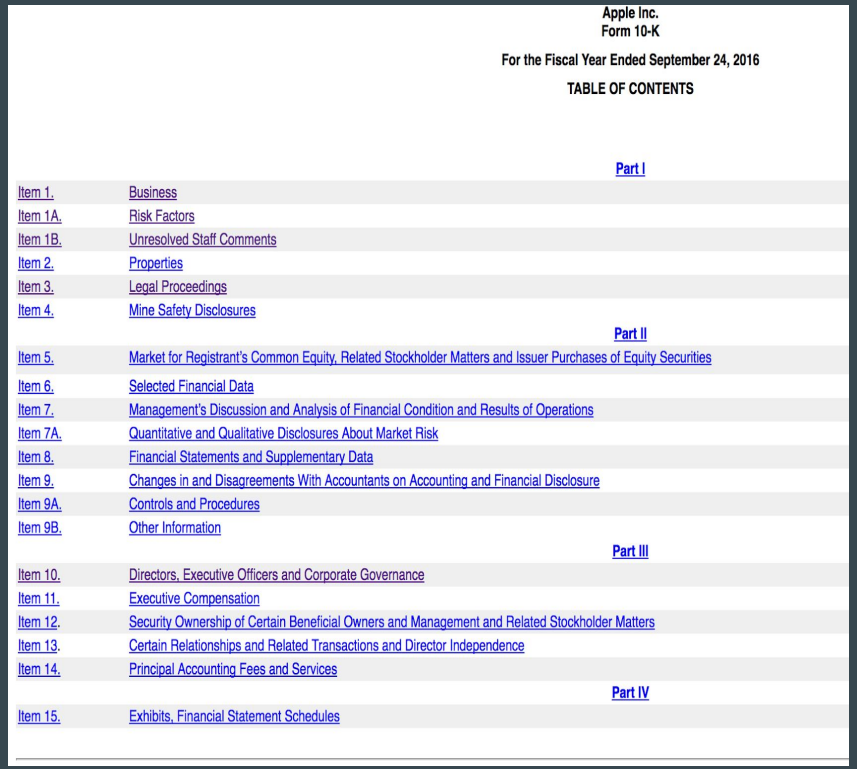
1. Corporate Annual Reports
2. Research Question
3. Data Acquisition
4. Analysis and Application
5. Future Direction

Corporate Annual Reports (10-K)

1. Mandated by the U.S. Securities and Exchange Commission (SEC) for publicly traded companies
    a. Quarter filling (10-Qs)
    b. Annual filing (10-Ks)
        i. 60-90 days after the end of FY
2. Basis of intelligent investment
3. “Other guys read Playboy, I read annual reports." -Warren Buffett



Research Question



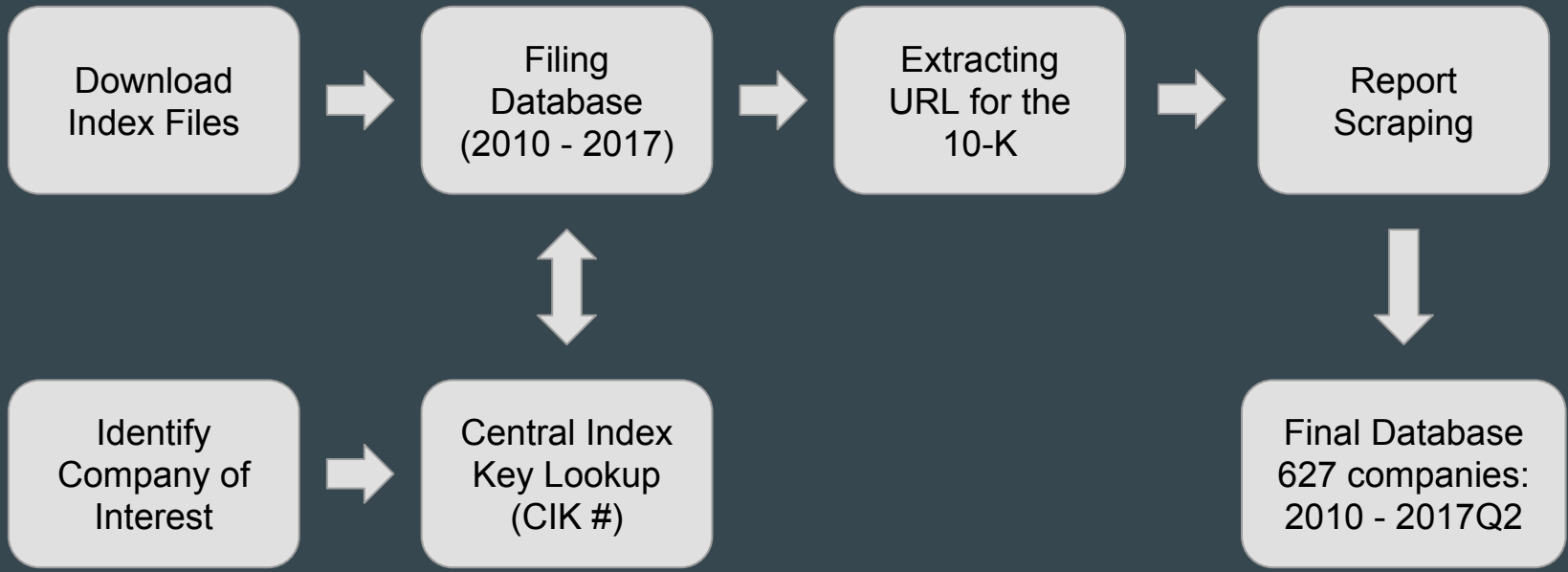
Apple Inc.
Form 10-K

For the Fiscal Year Ended September 24, 2016

TABLE OF CONTENTS

- “It’s really boring to read 10-Ks” - Daniel Rim
- The job of a financial analyst is to read carefully each 10-Ks or other corporate filings to do due diligence
- If there is a way to process and analyze 10-K efficiently using NLP to help guide investment decision?
- Classification system that helps us to identify declining companies based on their stock prices

Web Scraping - SEC.gov



Finding Numeric Representation of Texts

Bag of Words:

1. Count of words (frequency)
2. Text Cleaning:
    a. Remove punctuations, numbers, spaces, and newlines
    b. Remove stop words: it's, to, really
3. Lemmatization and Stemming
    a. Boring => bored
    b. Read
4. Count the times that each vocab occurs
5. Term Frequency/Inverse Document Frequency (TF*IDF)
    a. Weights
    b. Reduce the importance of frequent words

IDF = log(total no. of document/# of the document with the word)

Example

1. "It's really boring to read 10-Ks!" (Daniel Rim)
2. ["it's", "really","boring","to","read","10-ks","!"]
3. ["read", "bore"]

Intro to Word to Vec

Word2Vec is a method that would try to represent the document into a vector form

General idea is that for the vocabularies used in the document, one would come up with a probability distribution of words that would be used in a nearby expression

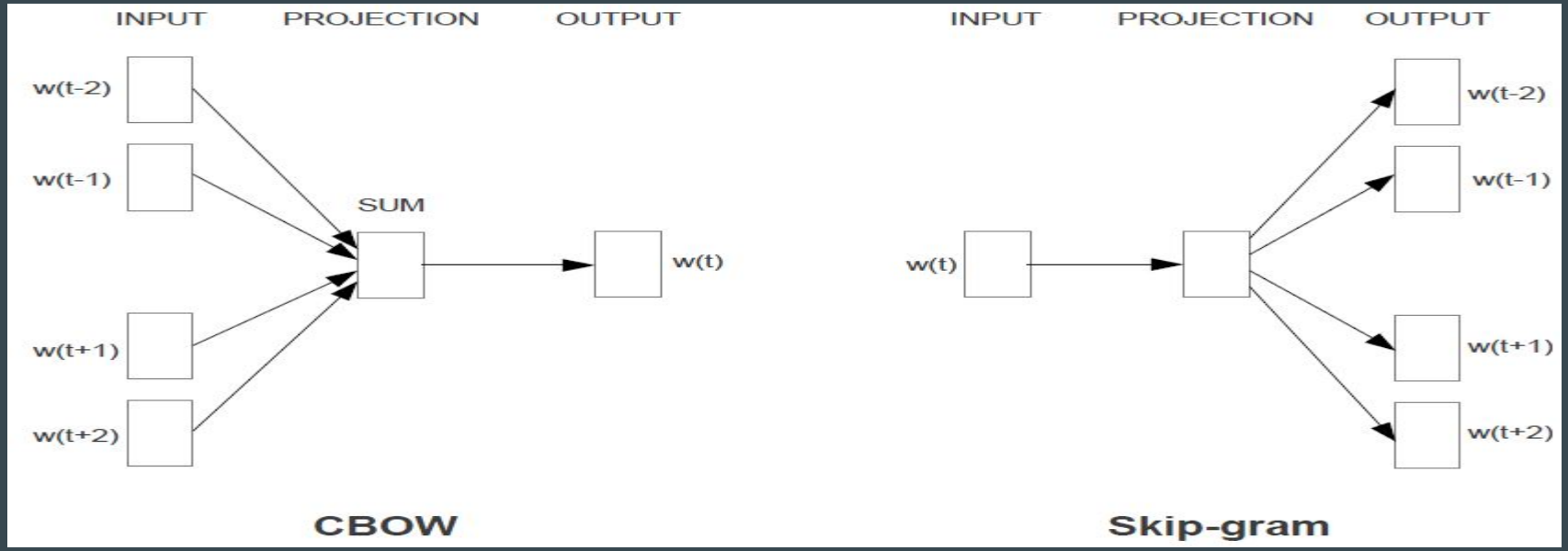
2 Methods for doing this is continuous-bag-of-words and skip-gram

Continuous-bag-of-words tries to produce probability distribution of a word given a list of words

Skip-gram tries to provide probability distribution of words that show up in similar content given a vocabulary

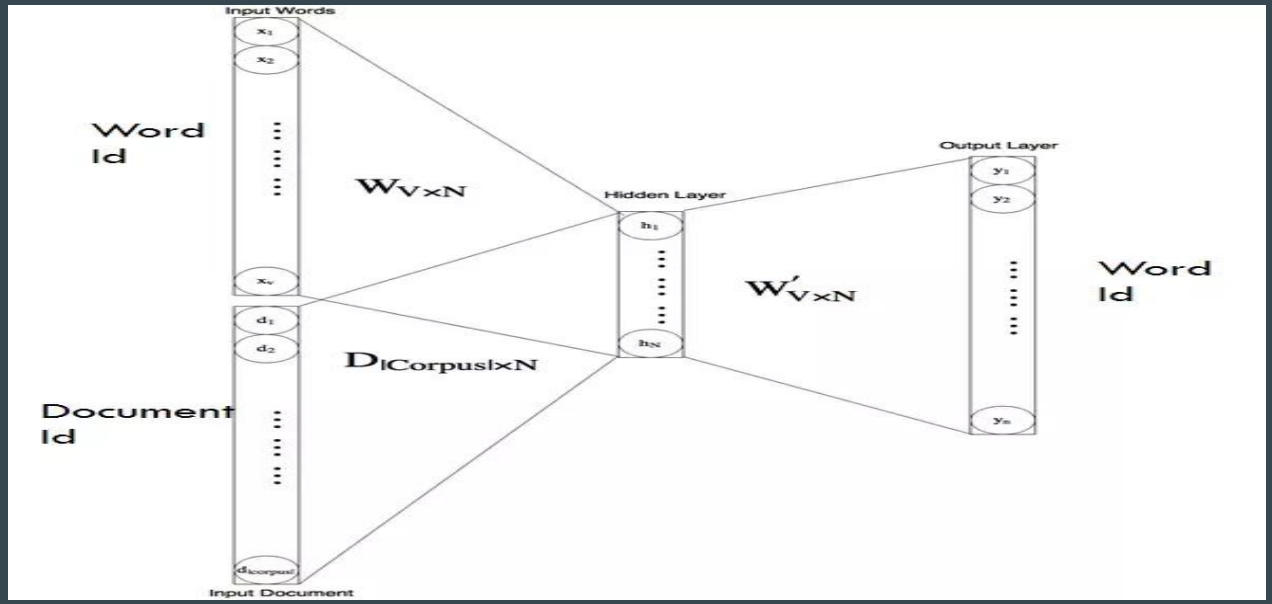
CBOW and Skip-Gram

Picture of difference between CBOW and Skip-Gram



Introduction to Doc to Vec

Doc2Vec tries to incorporate document characteristic when computing word2vec



Authorship attribution

We have also tried to apply authorship attribution analysis to the reports

It seems that doc2vec analysis is not quite effective at distinguishing authors unless the writings are drastically different

One methodology is to use key features that would give characteristics about the authors and use k-means clustering

Authorship attribution

Lexical Features:

- average number of words per sentence
- sentence length variation(standard deviation of words per sentence)
- Lexical diversity(number of unique words/words used in document)

Results on known different authors

Misclassification Rate	A Tale of Two Cities vs. The Great Gatsby	The Great Gatsby vs Sherlock Holmes	A Tale of Two Cities vs. Shakespeare	Amazon 10K vs. Altria 10K	Roms/Gal/1Cor vs Eph/Col	Roms/Gal/1Cor vs 1,2Tim/Titus
Lexical	14%	0%	46%	0%	0%	0%
Syntax	14%	25%	46%	0%	0%	17%
BOW	46%	38%	46%	0%	0%	0%

Authorship attribution

Syntactical:

Frequencies for common Parts of Speech types(singular/plural noun, proper noun, determiner, preposition/conjunction,adjective)

Bag of Words:

Count the most common words in the documents and apply clustering

Recurrent Neural Networks

Neural Network has been around since 1940s but were not so useful until about 5 years ago

Whereas regular Neural Networks does not have a sense of time, Recurrent Neural Networks try to capture time element while working with neural networks by using previous output as another input

Recurrent Neural Networks by itself is not as effective so people have incorporated LSTM to forget or remember previous outputs

Generating Corporate Report using RNN/LSTM

10K reports were too large for us to train using RNN

Following the example of writing a similar story in Aesop's fable, we have written summary using company profiles in CNBC

"the bell outwit met . nobody will all mouse got up and said that is all very well , but he thought would meet the case . you will all agree , said attached chief'"(Aesop's Fable LSTM RNN output)

"other service offerings for play and retailers through the consumer and sale beverages. Hasbro, The Investment segment focuses sales and facilitates that sectors. Wholesale in food countries in China its in fundamental storage, equity, and"(somewhat makes sense)

Generating Other Texts using RNN/LSTM

People have tried to use RNN/LSTM in texts more interesting than corporate annual reports

http://www.thedailybeast.com/meet-the-robot-writing-friends-sequels

> “Chandler: (in a muffin) (Runs to the girls to cry) Can I get some presents.”

Impressive but still have long way to go

Other Examples

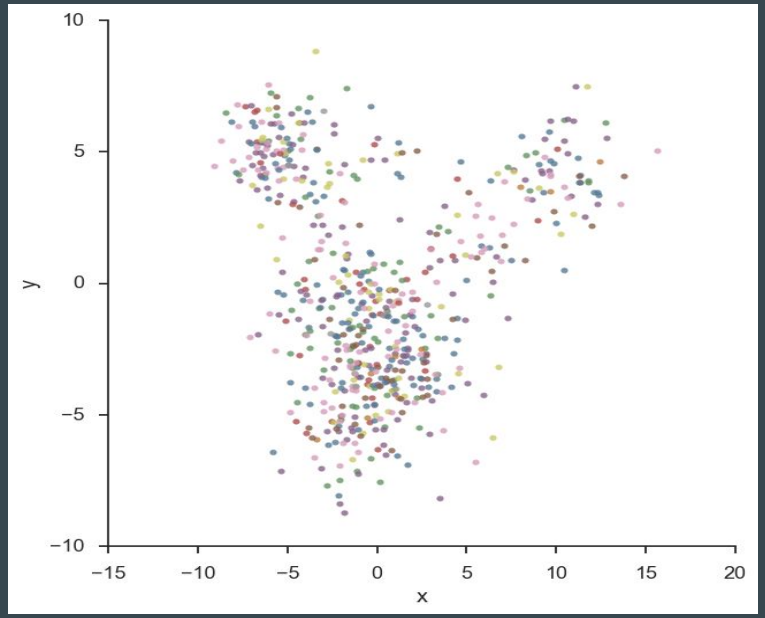
patrician patrician patrician uninterrupted, uninterrupted, slaves; in the Middle Ages, feudal lords, vassals, guild-masters, journeymen, apprentices, serfs; in almost all of these classes, again, subordinate gradations. The modern bourgeois society that has sprouted of oppression, new forms of struggle in place of the old ones.(Marx)

sun is new ever returning on its course. All streams flow into the sea, yet the sun. Is there anything of which one can say, “Look! This is something new”? It was here already, our time. No one remembers the former generations, and even those yet to come will not be remembered by those who follow them more than the place the streams come from, there they return again. (Ecclesiastes)

Visualizing Doc2Vec - Visualize

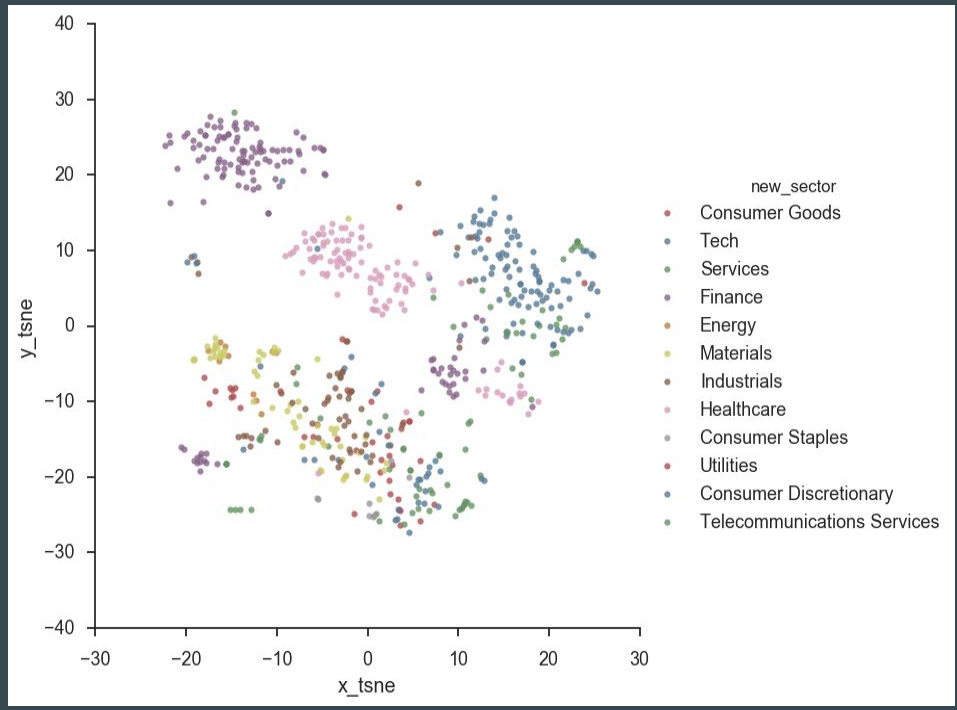
PCA

- 2 PCs
- Global structure - trying to maximize variance and preserve dissimilarity



t-SNE: t-Distributed Stochastic Neighbor Embedding

- Preserving local similarity to identify patterns



Similarity between Companies Using Doc2Vec

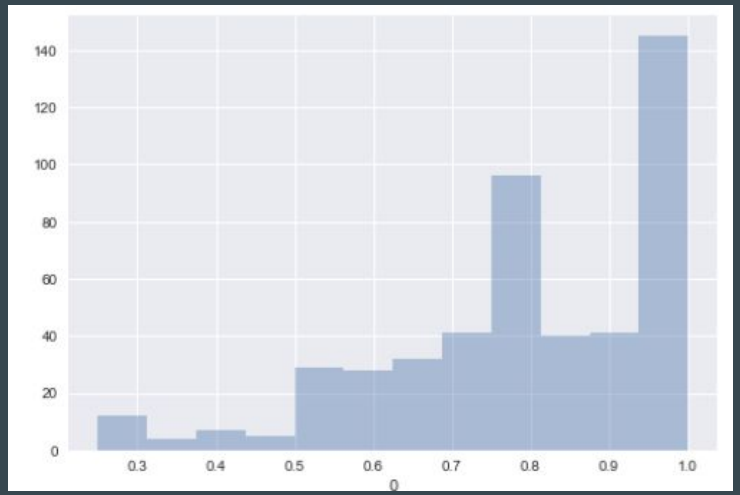
First, we compute the Similarity between Netflix and other companies.

We could see that most of the companies that are in the technology and media industry.

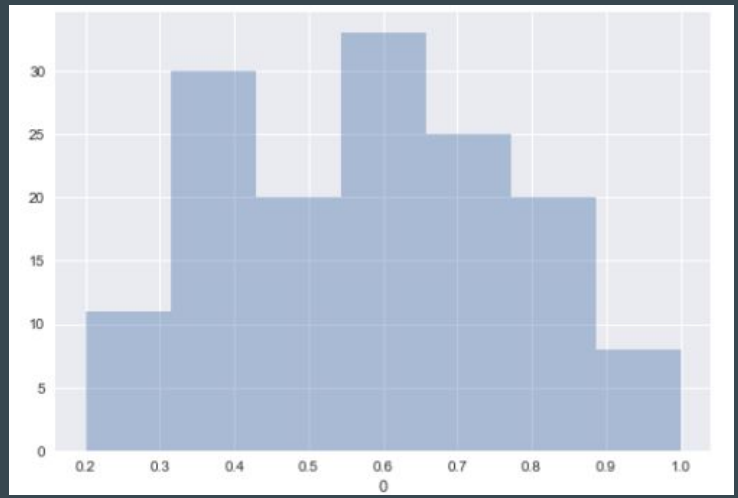
Company name	Similarity
ELECTRONIC ARTS INC.	0.7765984535217285
YAHOO INC	0.7334882020950317
WALT DISNEY CO	0.6724297404289246
TIME WARNER INC	0.633676290512085
MICROSOFT CORP	0.6339473724365234

Winners in similar classification by Doc2Vec

Distribution of number of winners in similar group in winner group



Distribution of number of winners in similar group in loser group



Good and Bad Companies Classification

If the company's stock price has decreased by 50% we have considered them "bad"

If the company's stock price has increased by more than 100% we have considered them "good"

SVM for Doc2Vec

Confusion Matrix	Prediction Losers	Prediction Winners
Actual Losers	17	17
Actual Winners	27	128

After running SVM, we get 76.7% accuracy rate, 82.5% sensitivity,

50 %specificity,

It seems that SVM provides a reasonable result in classifying the test set

Of course one can track the performance of this method by tracking the winner performance

Logistic Classification For Doc2Vec

Confusion Matrix	Prediction Losers	Prediction Winners
Actual Losers	24	20
Actual Winners	36	109

70.4% Accuracy Rate, 76.7% accuracy rate, 75.1% sensitivity and 54.5 %specificity for Logistic Classification on the test set

Slightly lower than than the accuracy rate with SVM

Random Forest For Doc2Vec

Confusion Matrix	Prediction Losers	Prediction Winners
Actual Losers	11	33
Actual Winners	5	130

Random Forest Accuracy result comes out to be 79.9% accuracy, 96.2% specificity, 25% sensitivity.

These accuracy rates are quite high but of course to use them for future investment would incur risk as future stock price movement can be drastically different from past 4 years

Future Direction

- More extensive research:
  - Examine 10-K by section (i.e., risks, business overview) vs, entire document
  - Incorporate other quantitative measures (i.e., valuation models)
  - Incorporate 10-Qs that might better reflect
  - Incorporate historical data
- Doc2vec can be improved by using recurrent neural networks?

Thank you!

Any Questions?