CENTRAL THEMES

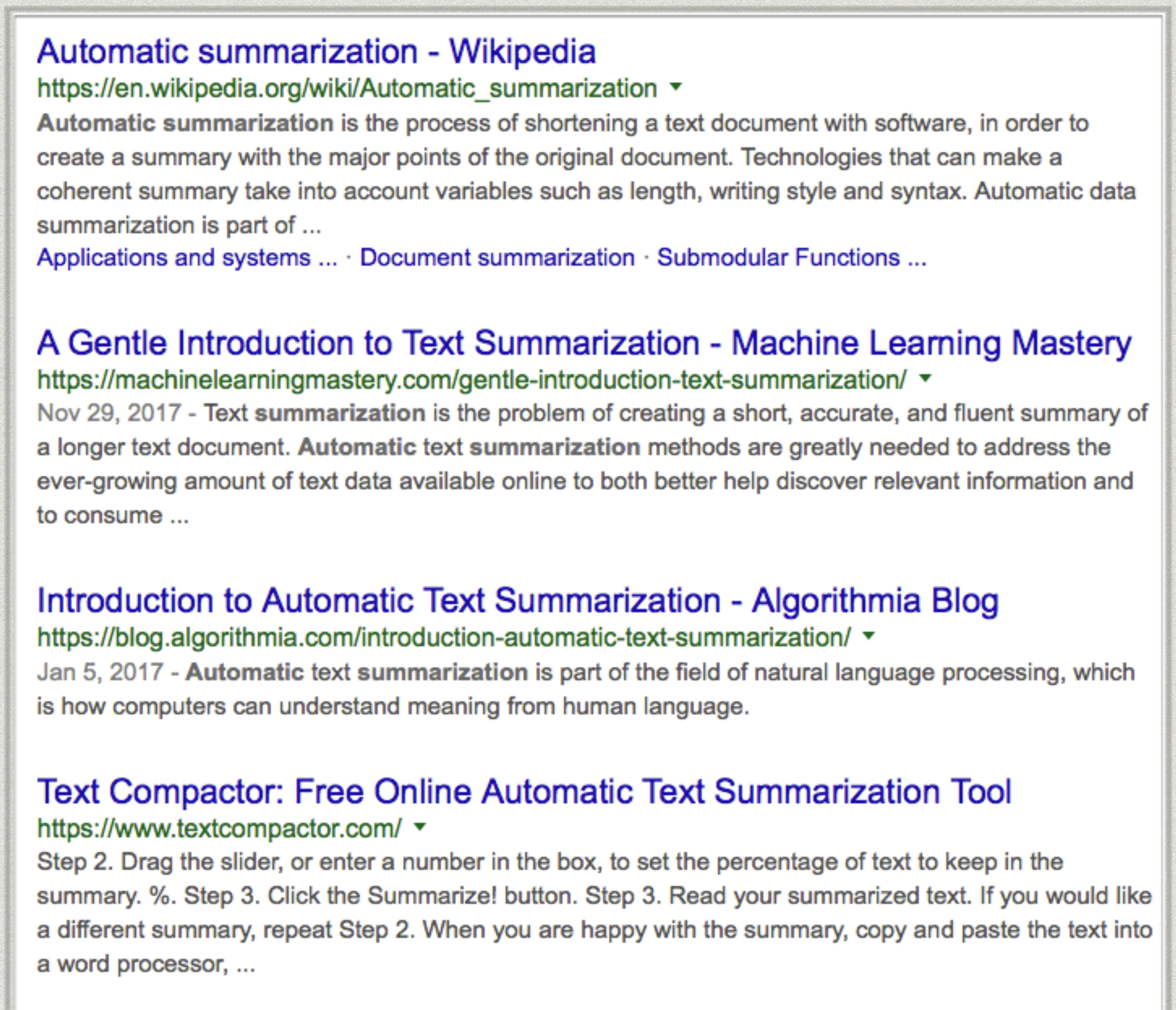
SUMMARIZING TEXT WITH SOCIAL NETWORK ANALYSIS

Shaina Race, PhD

Institute for Advanced Analytics

Automatic Summarization

Keyword-Guided Summarization



Automatic summarization - Wikipedia
https://en.wikipedia.org/wiki/Automatic_summarization
Automatic summarization is the process of shortening a text document with software, in order to create a summary with the major points of the original document. Technologies that can make a coherent summary take into account variables such as length, writing style and syntax. Automatic data summarization is part of ...
Applications and systems ... · Document summarization · Submodular Functions ...

A Gentle Introduction to Text Summarization - Machine Learning Mastery
https://machinelearningmastery.com/gentle-introduction-text-summarization/
Nov 29, 2017 - Text **summarization** is the problem of creating a short, accurate, and fluent summary of a longer text document. **Automatic** text **summarization** methods are greatly needed to address the ever-growing amount of text data available online to both better help discover relevant information and to consume ...

Introduction to Automatic Text Summarization - Algorithmia Blog
https://blog.algorithmia.com/introduction-automatic-text-summarization/
Jan 5, 2017 - **Automatic** text **summarization** is part of the field of natural language processing, which is how computers can understand meaning from human language.

Text Compactor: Free Online Automatic Text Summarization Tool
https://www.textcompactor.com/
Step 2. Drag the slider, or enter a number in the box, to set the percentage of text to keep in the summary. %. Step 3. Click the Summarize! button. Step 3. Read your summarized text. If you would like a different summary, repeat Step 2. When you are happy with the summary, copy and paste the text into a word processor, ...

Editorial Summarization

AI Has A Big (Data) Problem

Mar. 21, 2018 9:41 AM ET | 48 comments | Includes: DDM, DIA, DOG, DXD, EEH, EPS, EQL, FB, FEX, FWDD, HU...

David Trainer Mute

Long/short equity, value, research analyst, Deep Value

MARKETPLACE Value Investing 2.0

Follow

(6,933 followers)

Summary

- We are awash in an ocean of data that grows bigger by the second. And it's a complete and utter mess.
- Poor data quality represents the single largest hurdle for developing useful artificial intelligence.
- It doesn't matter how "smart" machines become if they're fed data that is inaccurate or incomprehensible.
- *This idea was discussed in more depth with members of my private investing community, Value Investing 2.0 .*

The total size of all global data hit 20 zettabytes in 2017. For 99% of people, that number probably means nothing, so picture this: if every 64-gigabyte iPhone were a brick, we could build 80 Great Walls of China with the iPhones needed to store all the world's data.

We are awash in an ocean of data that grows bigger by the second. And it's a complete and utter mess.

90% of web data is unstructured, meaning it's in a format that cannot be easily searched and understood by machines. Poor data quality costs the US economy $3.1 trillion a year, according to IBM Corp. (NYSE:IBM). We have

www.SeekingAlpha.com

POST-PROCESSING OF TOPIC-MINING

- Most topic mining algorithms (LDA, NMF) give a “topic” as a list of words.
- Useful.

interest
rate
rental
home
tax
market
mortgage
income
loan
student
debt

POST-PROCESSING OF TOPIC-MINING

- Most topic mining algorithms (LDA, NMF) give a "topic" as a list of words.
- Useful. But what's more useful?

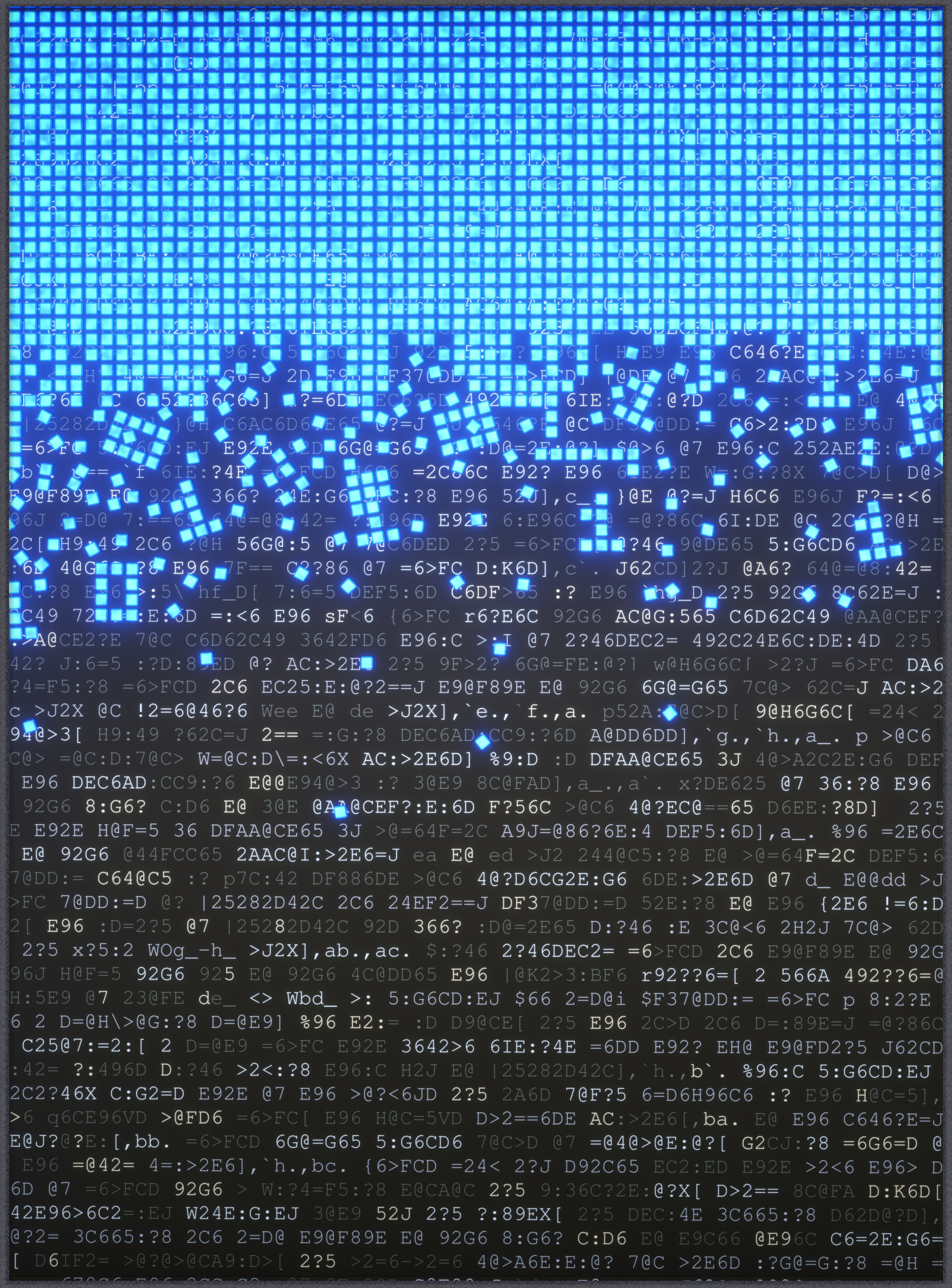
…Some **context** for those words.

interest
rate
rental
home
tax
market
mortgage
income
loan
student
debt

"Further, the rental market will benefit from home prices rising faster than wages and rents, tax legislation reducing the incentives of home ownership, and mounting student loan debt hampering Millennials and delaying first-time home purchases."

TEXT MINING BASICS

TERM-DOCUMENT MATRICES (VECTOR SPACE MODEL)



“Bag of Words”

- Each document considered a collection of unordered words (terms).
- We transform this “bag of words” into a vector of term frequencies.
- Each document then lives in a vector space.
- Its direction in that vector space is what characterizes it semantically.

Term-Document Matrix

(Illustrative Example)

Doc1

My **cat** likes to eat **dog** food. It's insane. He won't eat tuna, but **dog** food? He's all over it.

Doc2

Dog chasing the **cat** around the house. Never gets **tired**. The **cat** is not a **dog** toy! Dumb **dog**.

Doc3

I **injured** my **ankle** playing football yesterday. Bruised and swollen. Maybe **sprained**?

Doc4

So **tired** of being **injured**. My **ankle** just won't get better! I **sprained** it 2 months ago!

$$\mathbf{B} = \begin{array}{lc} & \begin{array}{cccc} doc1 & doc2 & doc3 & doc4 \end{array} \\ \begin{array}{l} \text{“cat”} \\ \text{“dog”} \\ \text{“tired”} \\ \text{“injured”} \\ \text{“ankle”} \\ \text{“sprained”} \end{array} & \begin{pmatrix} 1 & 2 & 0 & 0 \\ 2 & 3 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \end{array}$$

Each document becomes vector in $\mathbb{R}^6$

Term-Document Matrix

- A “document” can be any block of text
    - Articles
    - Paragraphs
    - Sentences
    - Tweets
    - Books
- The user can define/process the text uniquely
    - Terms/Multi-term Phrases
    - Remove **stopwords** (i.e. it, an, the)
    - Stemming (running —> run)

Term-Document Matrix

- Extremely sparse matrix (lots of zeros):
  Most documents do ***not*** contain most terms.
- Inefficient to store so many zeros.
- Sparse matrix format:

$$
\begin{array}{ccc}
\textit{row} & \textit{col} & \textit{value}
\end{array}
$$

$$
\begin{pmatrix}
1 & 1 & 1 \\
1 & 2 & 2 \\
2 & 1 & 2 \\
2 & 2 & 3 \\
3 & 2 & 1 \\
\vdots & \vdots & \vdots
\end{pmatrix}
$$

Term Weighting

- Weight entries in the term-document matrix according to their **global and local prevalence**.
- Downgrade words for frequent appearance in the corpus.
- Most common term weighting scheme: **TF-IDF**
    - Term Frequency - Inverse Document Frequency
    - Scale each row by the log of inverse of proportion of documents containing that term

$$\mathrm{idf}(term) = \log\left(\frac{\#\,\text{documents}}{\#\,\text{documents containing term t}}\right)$$

Measuring Similarity Between Documents

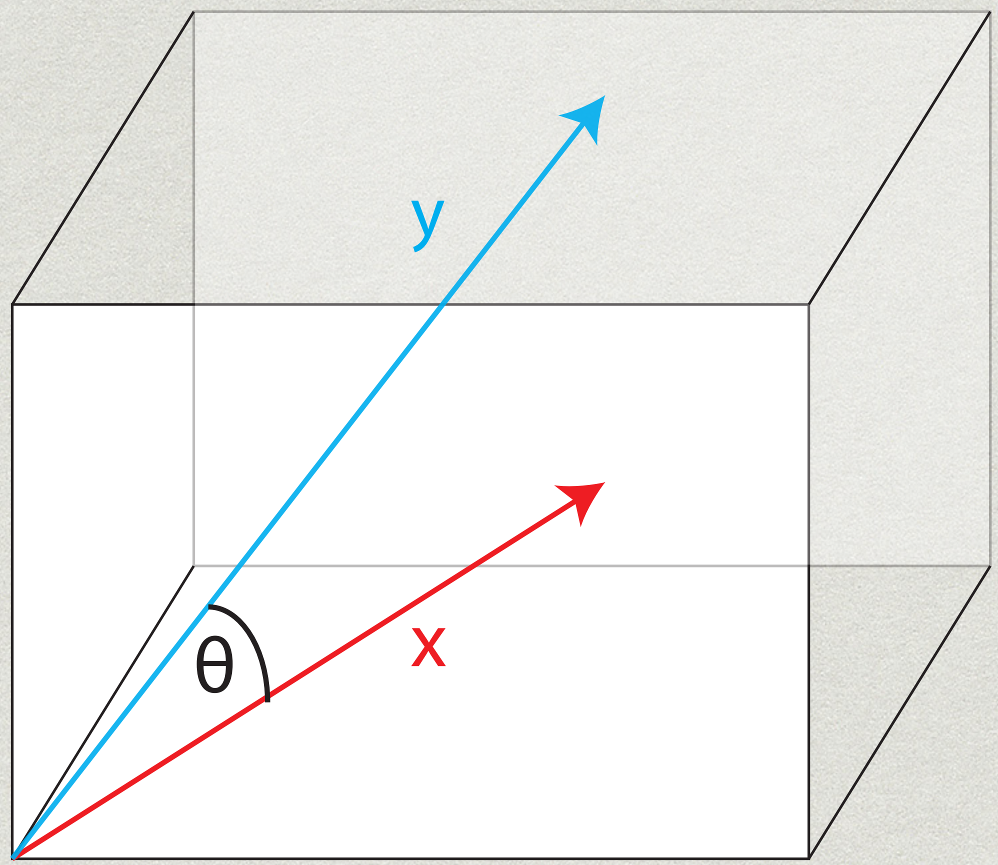
- A document's **direction** in space is what characterizes it semantically.
- The length (magnitude/norm) of the vector depends more on length of document than content
- Look at the **angle between document vectors to characterize similarity** rather than norm distance

Cosine Similarity

- For text vectors, $0 \leq \cos(\theta) \leq 1$

  ($\cos(\theta) \nless 0$ b/c $\mathbf{x},\mathbf{y} \geq 0$)

- $\cos(\theta) = 0$ means no terms shared
- $\cos(\theta) = 1$ means same terms in same proportions



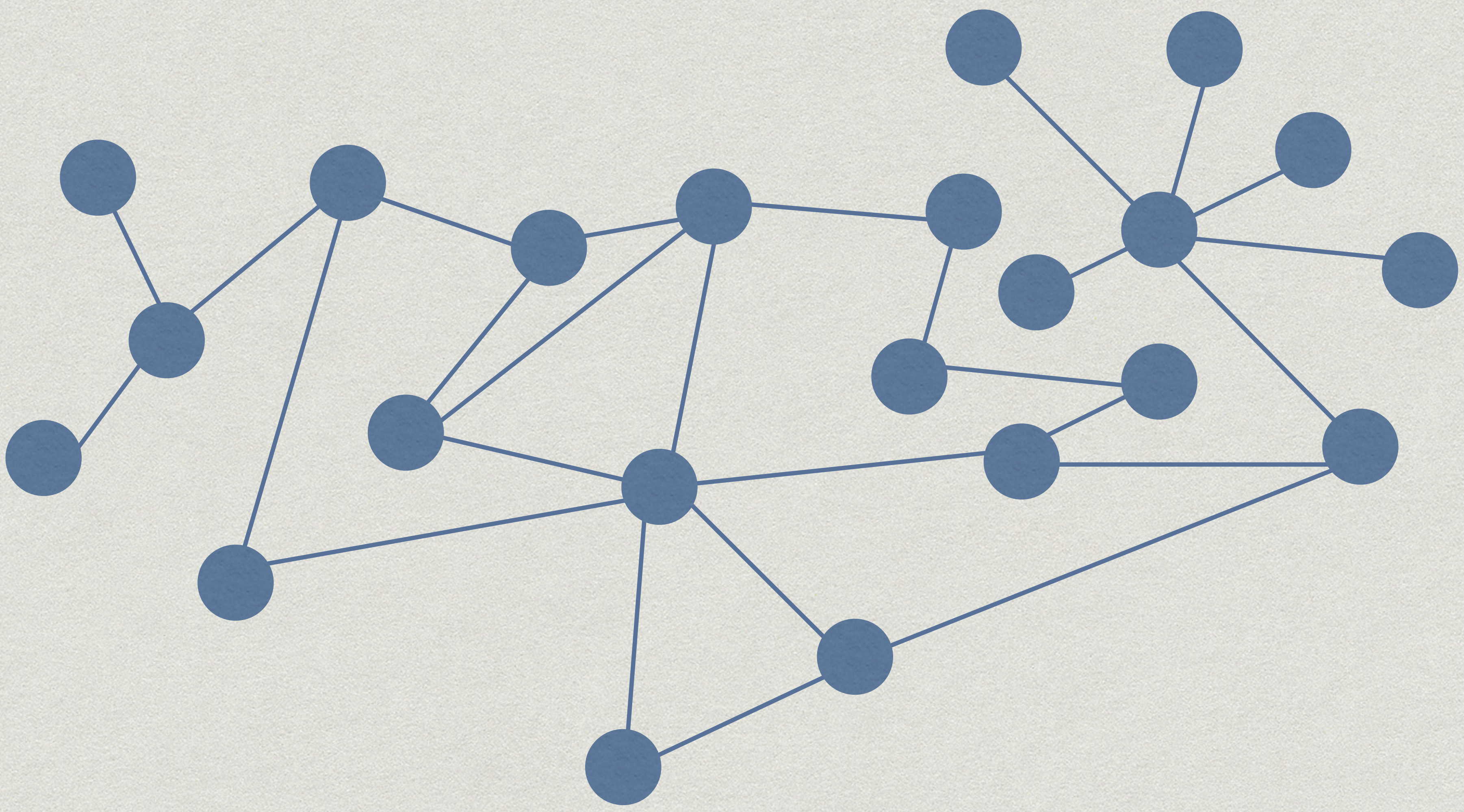
$$\cos(\theta) = \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|}$$

SOCIAL NETWORK ANALYSIS BASICS

GRAPHS
DEGREE
CENTRALITY

NETWORK/GRAPH

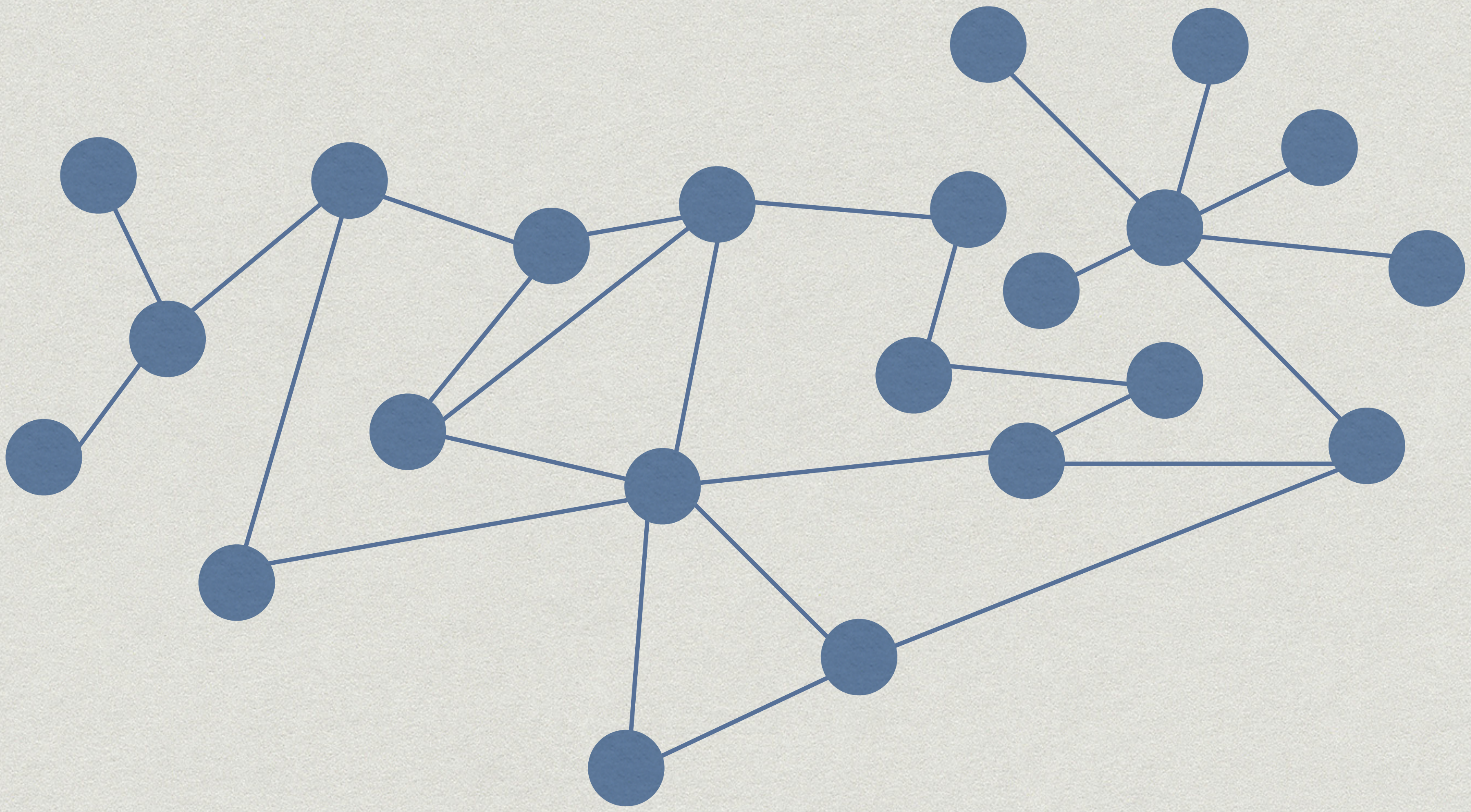
Collection of nodes connected by edges



(Binary) Adjacency Matrix: $$\mathbf{A}_{ij} = \begin{cases} 1 & \text{if node } i \text{ connected to node } j \\ 0 & \text{otherwise} \end{cases}$$

Network/Graph

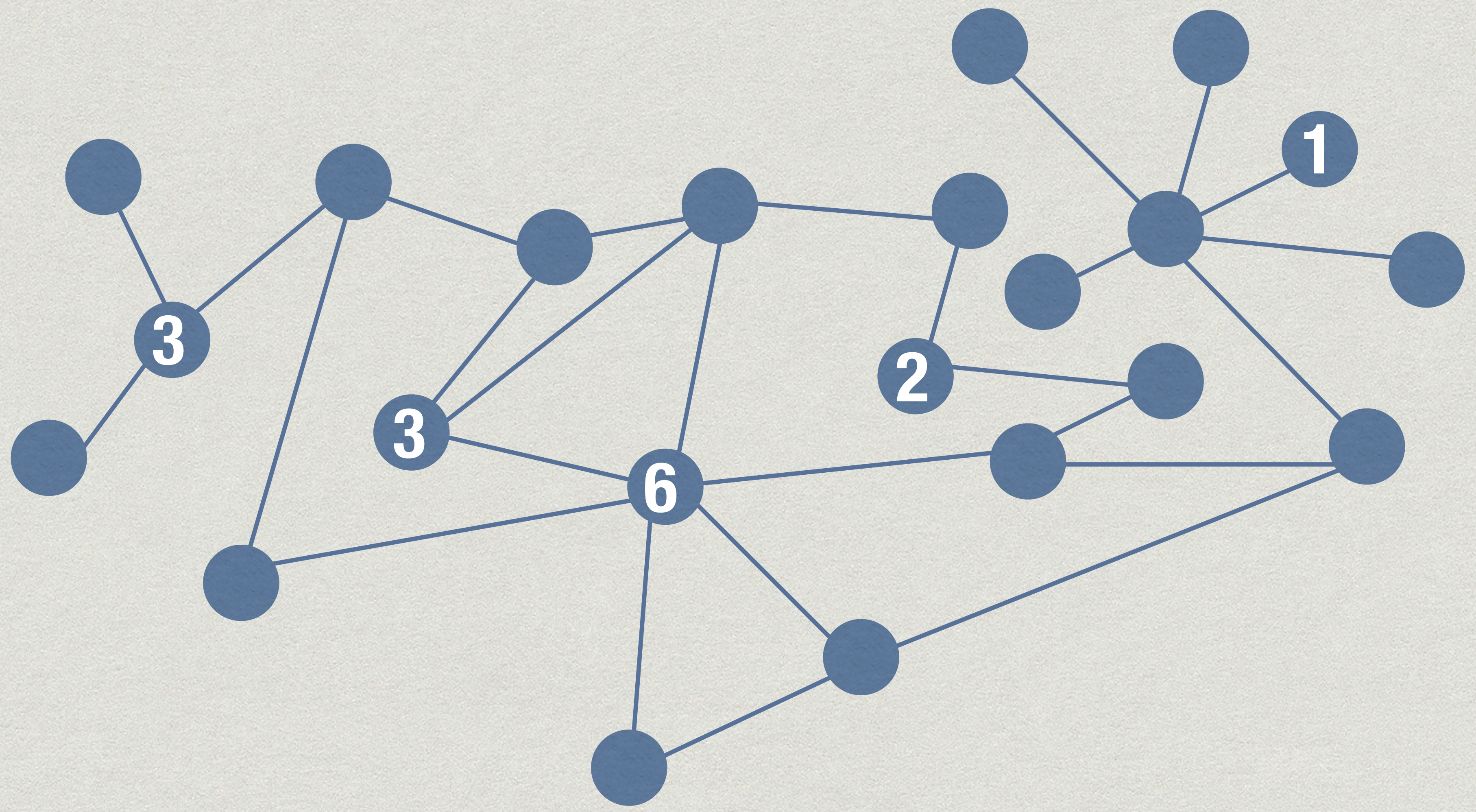
Collection of nodes connected by edges



(Weighted) Adjacency Matrix: $\mathbf{A}_{ij} = \begin{cases} w_{ij} & \text{if node } i \text{ connected to node } j \\ 0 & \text{otherwise} \end{cases}$

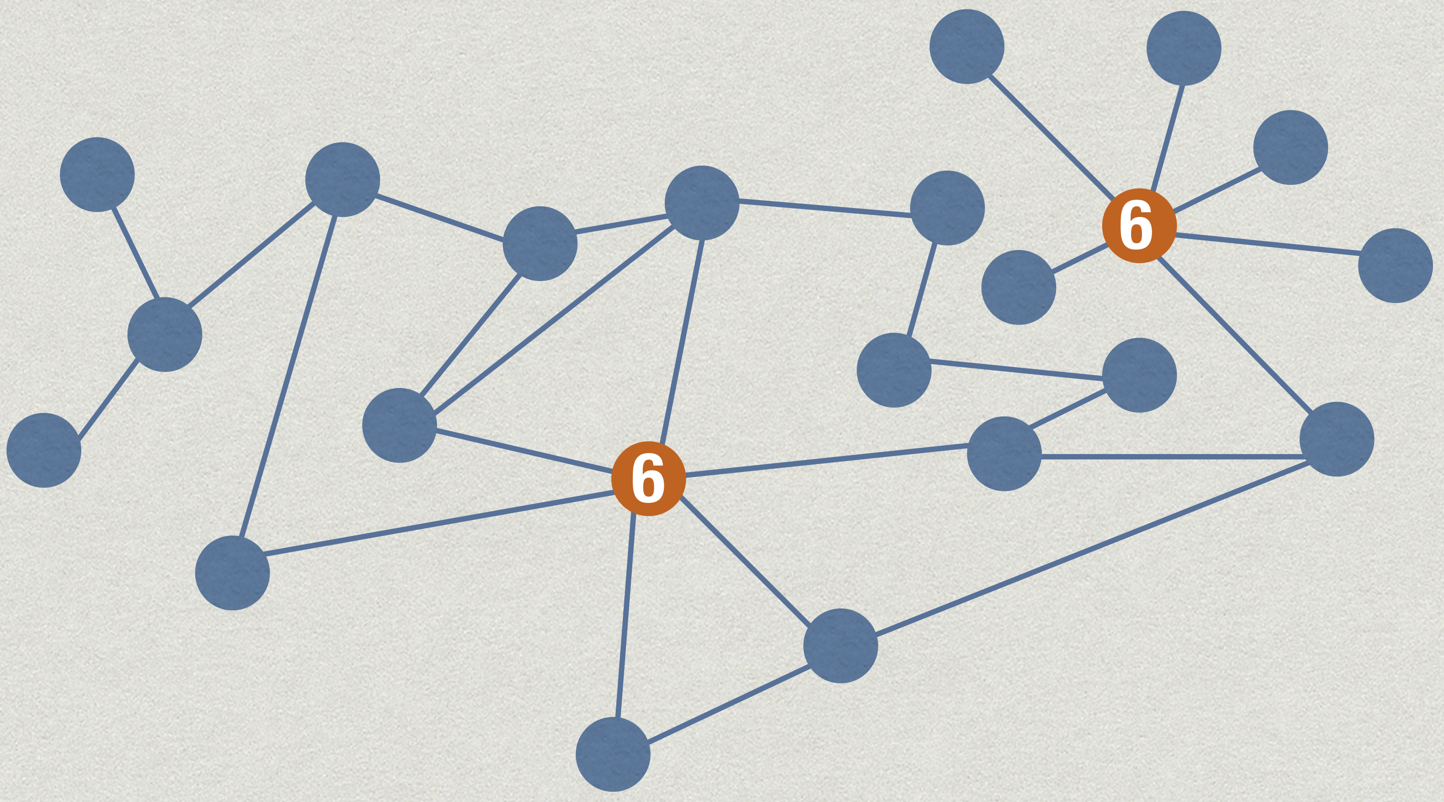
Vertex Degree

The **degree** of a vertex is the number of edges adjacent to that vertex.



CENTRALITY
MEASURING INFLUENCE

- Degree is the simplest measure of **centrality**
- Nodes with more connections have more influence
- Problem: Local Measure. ***Who*** are connections?



Grand Idea:

- Let each node's centrality be the sum of its neighbors' centralities.
- Let c_i be the centrality of node i. Let N_i be the set of neighbors for node i.

$$c_i = \sum_{j \in \{N_i\}} c_j$$

- In matrix form, we'd write $\mathbf{c} = \mathbf{A}\mathbf{c}$ where $\mathbf{A}$ is the adjacency matrix!

EIGENVECTOR CENTRALITY

$$\mathbf{Ac} = \mathbf{c}$$

- This equation has no solutions unless 1 is an eigenvalue of $\mathbf{A}$.

- Refine our grand idea:

Perron-Frobenius
$\Rightarrow \lambda_1$

- Let each node's centrality be positive and <**PROPORTIONAL TO**> the sum of its neighbors' centralities.

$$\mathbf{Ac} = \lambda\mathbf{c}$$

CENTRAL SENTENCES

USING CENTRALITY TO FIND KEY SENTENCES IN A DOCUMENT

Network of Sentences

- Parse a document into sentences, creating a vector for each sentence.
- Compute the cosine similarity between each sentence
- Connect two sentences if their cosine similarity is greater than some threshold
- Use the cosine similarity as the weight of the edge between the sentences.

Sample Data

- Articles from investment research website www.SeekingAlpha.com
- Each has an editorial summary at the top of the article.
- Two Sample Articles:
  - Decline In Housing Affordability To Benefit REITs
  - AI has a Big (Data) Problem

With home values interest rates expected
home affordability index will contin
mortgages will drive housing market
versely, rising interest rate environment reflection strong economic
bring increased wages increased rents, ultimately benefitting REITs
The expected rise home
will driven low inventory
A strong economy
r pace home prices rents growing proportionally
property affordable attractive option buying benefitting residential REITs
" The rise student loan debt negatively impact
save payment ability obtain financing
owing costs will reduce home af
driving consumers towards rental market
trillion stude
bt, double
ayment"
Rising costs, combined tax legislation mounting studen
debt, will continue negatively impact hom
egislation reducing incent
home ownership, mounting student loan debt hampering Millennials delaying first-time home purchases
nued decline
me affordability will push - home owners rental market
The decrease home affordability wil
nefit rental market residential REITs
Because positive factors rental marke
, towards lower
-week range $
Perhaps greatest factor wi
The decline driven growth single family home prices
acing income growth coupled increased mortgage rates
near botton
week range $
The decline
iShares Residential
" The lower tax liability dividend
ake REITs attractive investors
pgrade expensive homes may dissuade
main existing homes, contributing lack inventory less expensive homes
The increased tax liability homeowners will make h
ownership less a
e growth line rent g
median rental prices grown roughly % since
REZ
e cost undergraduate education inc
ed percent since , incomes increased percent
As //, RE
are price $

Compute Sentence Centrality

RESULTS

Most Central Sentences, in order:

Further, the **rental market will benefit from home prices rising faster than wages and rents, tax legislation reducing the incentives of home ownership, and mounting student loan debt hampering Millennials and delaying first-time home purchases.**

The **decrease in home affordability will benefit the rental market and residential REITs.**

Actual Editorial Summary

Richard J Dingraudo Mute

Long/short equity, macro, homebuilders, value

Follow

(14 followers)

Summary

- Further increases in home values and interest rates are expected in 2018.
- Rising education costs have resulted in mounting student loan debt hindering potential first-time homebuyer's ability to save.
- The 2017 tax legislation reduces incentives of home ownership which may push more individuals into the rental market.
- Given REITs are considered pass through entities, the new tax legislation will reduce the tax liability on dividends paid to REIT investors.
- While a rising interest rate environment negatively impactsREITs by increasing borrowing costs, strong economic conditions, coupled withthe factors above, will benefit the rental market and increase residential REITvalues.

Results

Most Central Sentences, in order:

AI and machine learning may be able to replace those 9% of data scientists who are mining data for patterns, but it **will still need** the 80% **working on collecting, cleaning and organizing data.**

Structuring Data: More About Team Than Technology
As long as financial data remains unstructured, existing machine learning tools cannot process it effectively.

Structuring Financial Data: Not as Easy as Most Think
In theory, **financial data in filings would be more structured and standardized, or we could make it that way easily.**

Actual Editorial Summary

David Trainer Mute

Long/short equity, value, research analyst, Deep Value

MARKETPLACE Value Investing 2.0

Follow

(6,990 followers)

Summary

- We are awash in an ocean of data that grows bigger by the second. And it's a complete and utter mess.
- Poor data quality represents the single largest hurdle for developing useful artificial intelligence.
- It doesn't matter how "smart" machines become if they're fed data that is inaccurate or incomprehensible.

Method Summary

- Create a sentence similarity matrix using cosine similarity.
- Change to zero any values below some threshold. The result is your adjacency matrix.
- Compute dominant eigenvector of that matrix.
- Rank sentences according to entries in dominant eigenvector.

Adaptation for Topic Models

- Use collection of documents with highest association with a given topic to form corpus
  - Chose either one document as a summary document

    OR

  - Chose several documents and one sentence from each of those documents to serve as a summary.

Other Adaptations

What does this document/group of documents say about <keyword>?

- Weight topic keywords higher in the term document matrix

R Code for Today's Examples

http://www4.ncsu.edu/~slrace/textMiningviaSNA.R

http://www4.ncsu.edu/~slrace/SAarticle.txt

http://www4.ncsu.edu/~slrace/SAarticle2.txt