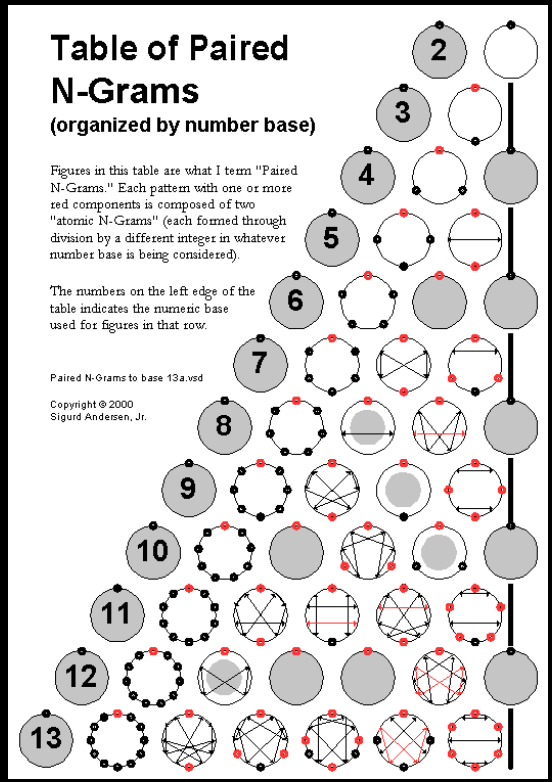
Quant Quest

Deep Patel

What are we analyzing?

General Approach- Other Approaches

- Other approaches
  - N- Grams
  - Rank System
  - LDA (Latent Dirichlet Allocation)



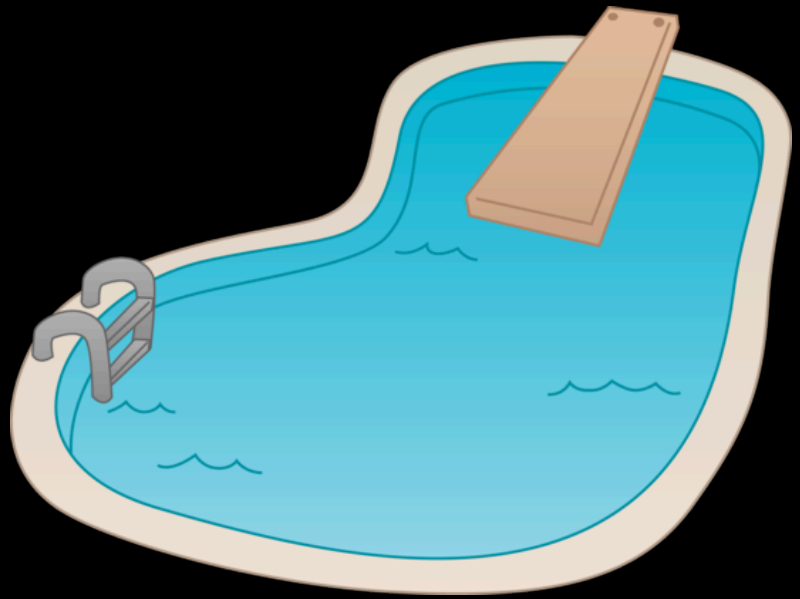
General Approach- Why Latent Semantic Indexing?

- We chose LSI because of it’s
  - Efficiency
  - Analytical value
  - Practicality

General Approach- What is LSI?

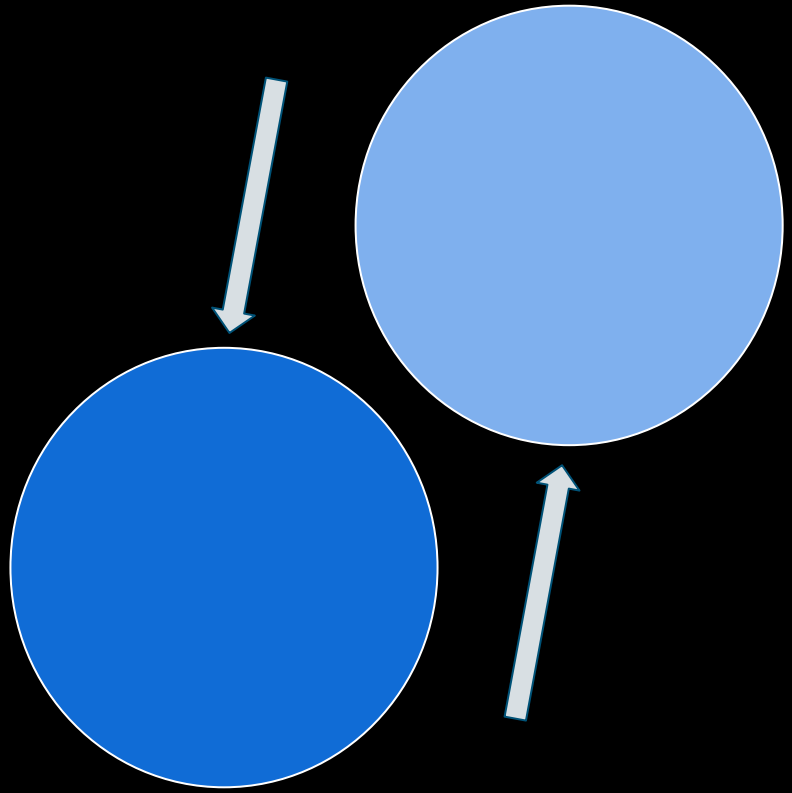
Latent semantic indexing is an indexing and retrieval method that identifies patterns and relationships between the words in a piece of text.

General Approach- LSI Process



General Approach- LSI Process

General Approach- LSI Process



General Approach - Code Weaknesses

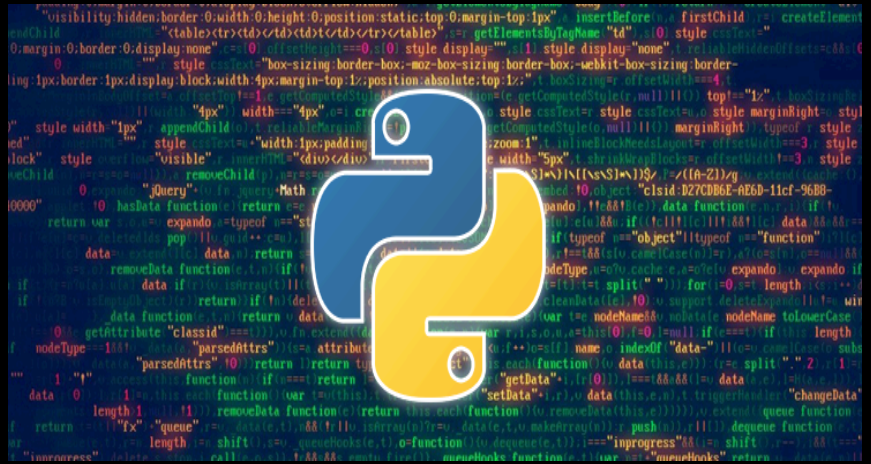
- Our code would not be able to
  - Link companies without Wikipedia pages
  - Relate information that’s not there

General Approach - Time Constraints

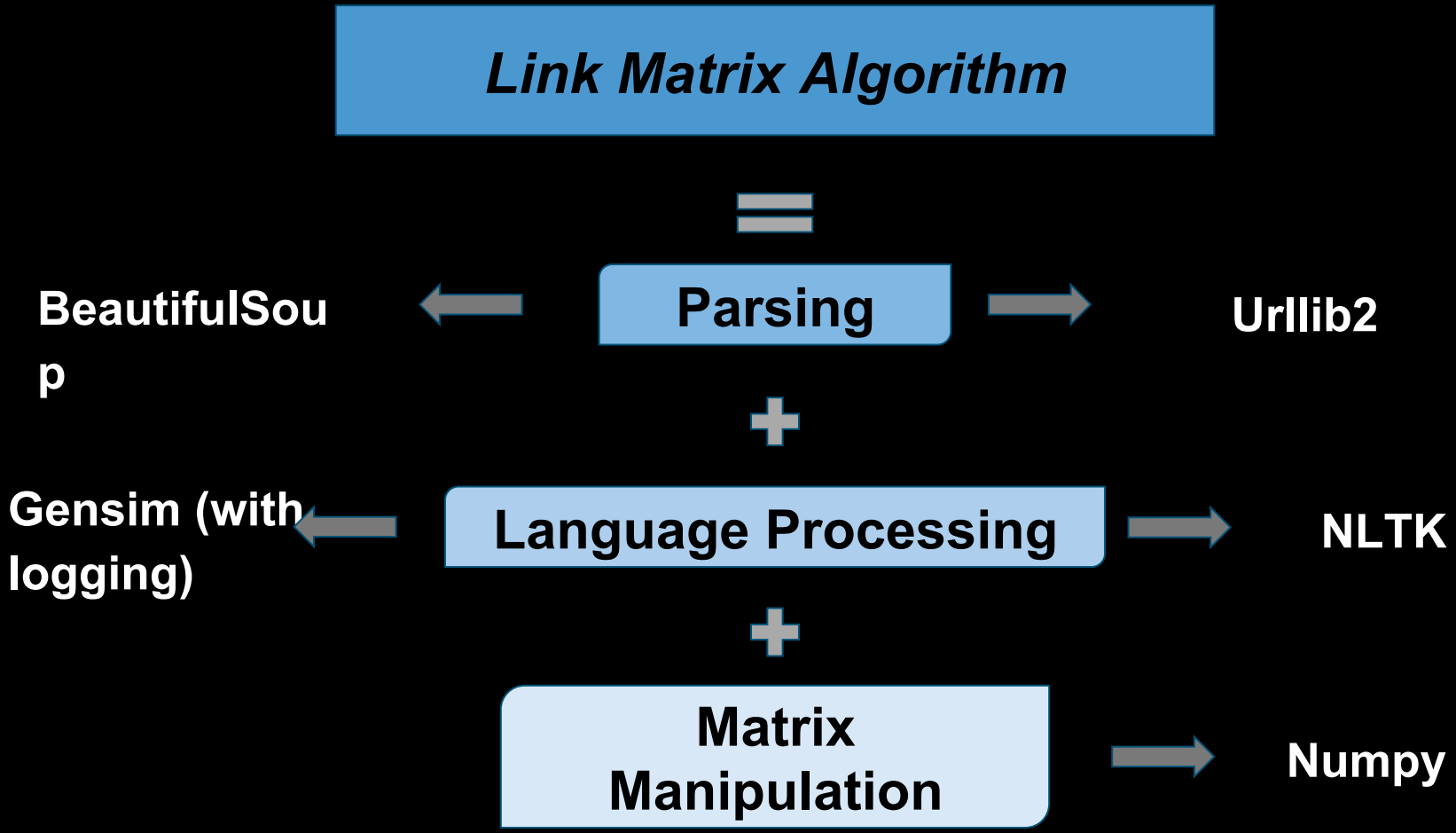
- If we had more time we could have
  - Used more pages, more information
  - Created a larger corpus
  - Used more generators

Technical Approach

Language: Python



- Brevity & Code Readability
- Useful modules
- Multifunctional
- Faster than most Dynamic Languages


Link Matrix Algorithm
=
BeautifulSoup
Parsing
Urllib2
+
Gensim (with logging)
Language Processing
NLTK
+
Matrix Manipulation
Numpy

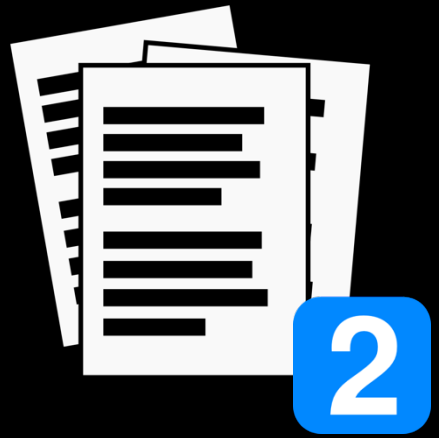
Technical Approach: Overcoming Obstacles

➔ Some companies have multiple stock classes
- Solution: remove duplicates using

➔ Companies without Wikipedia page
- Solution: use sector & sub-industry

So what does the program actually do?

Wikipedia Documents

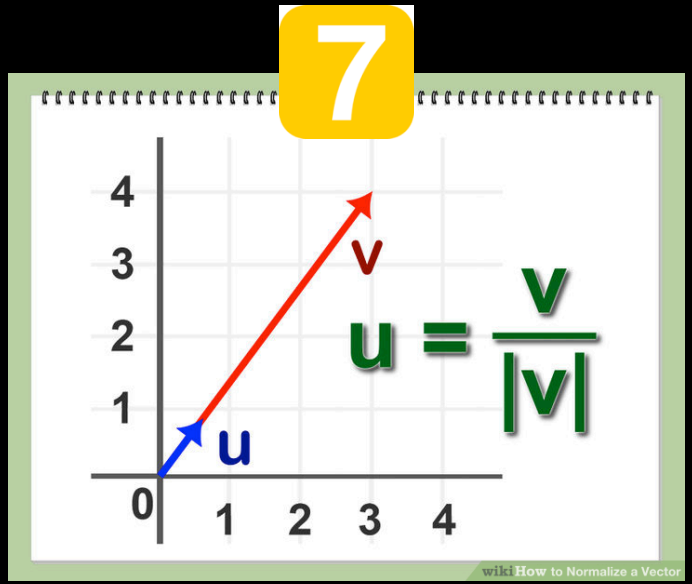


Stop Words

chocolate	vanilla	chip	butter	
1896	1193	705	659	
mint	peanut	with	coffee	
493	385	352	311	
strawberry	pecan	cookie	dough	
305	262	259	238	217
choc	cookies	cherry	swirl	raspberry
206	194	190	171	157
flavors	french	black	caramel	like
146	131	127	123	119
ice	the	rocky	peach	anything
114	112	103	101	100
road	bean	tracks	cup	almond
100	93	85	77	72
something	moose	neopolitan	pistachio	brownie
72	71	68	67	64
marshmallow	flavor	walnut	nuts	fruit
64	63	62	60	56
maple	any	dark	orange	banana
55	53	51	51	50
mocha	ripple	that	chunks	etc
50	50	50	49	49
cake	all	would	coconut	carmel
41	39	39	38	37
cheesecake	some	good	chips	for
36	36	34	33	32

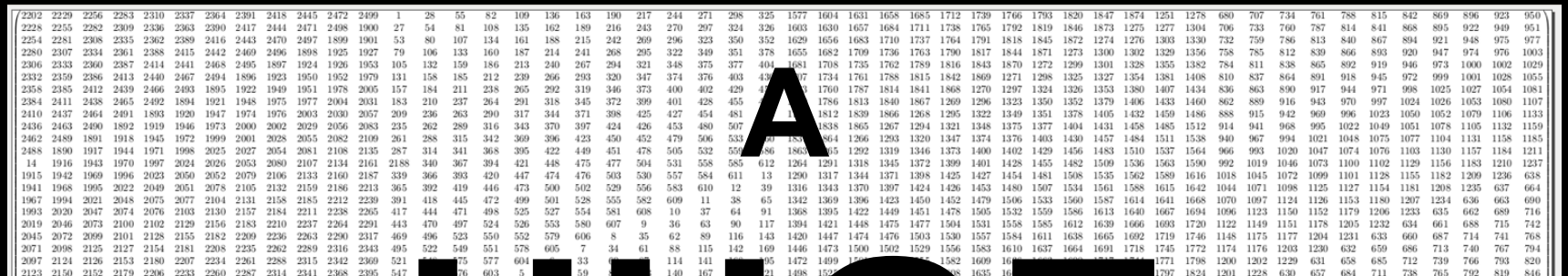
Corpus

Of Links
Using LSI Model
& Similarity Queries



NORMALIZE

FINAL OUTPUT



What did we accomplish?

460
SECONDS

ACCURACY

TOP 10 CORRELATIONS

Marriott Int'l	Wyndham Worldwide	0.00941
Marriott Int'l	Host Hotels & Resorts	0.00920
Marriott Int'l	Starwood Hotels & Resorts	0.00945
Wyndham Worldwide	Marriott Int'l	0.00942
Tiffany & Co.	Time Warner Co.	0.00951
Marathon Oil Corporation	ConocoPhillips	0.00913
Starwood Hotels & Resorts	Marriott Int'l	0.00907
Wyndham Worldwide	Host Hotels & Resorts	0.00906
Starwood Hotels & Resorts	Wyndham Worldwide	0.00912
NASDAQ OMX GROUP	CME Group Inc.	0.00891

EXAMPLES

Case 1: Marathon Oil Corporation and ConocoPhillips

Case 2: Tiffany & Co. and Time Warner Inc.

A WHOLE NEW DEFINITION OF “DIVERSIFICATION”

Thank You!

Questions?